UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

BIOLOGÍA EVOLUTIVA

RECONSTRUCCIÓN DE LOS GENOMAS ANCESTRALES DE LOS LINAJES PRINCIPALES DE LOS DOMINIOS ARCHAEA Y BACTERIA

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRA EN CIENCIAS BIOLÓGICAS

PRESENTA:

CORAL CRUZ GONZÁLEZ LUNA

TUTOR PRINCIPAL DE TESIS:

DR. ARTURO CARLOS II BECERRA BRACHO
FACULTAD DE CIENCIAS, UNAM

COMITÉ TUTOR:

DRA. LUISA ISAURA FALCÓN ÁLVAREZ
INSTITUTO DE ECOLOGÍA, UNAM
DR. LUIS DAVID ALCARAZ PERAZA
FACULTAD DE CIENCIAS, UNAM

CIUDAD UNIVERSITARIA, CIUDAD DE MÉXICO, ABRIL 2018

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

BIOLOGÍA EVOLUTIVA

RECONSTRUCCIÓN DE LOS GENOMAS ANCESTRALES DE LOS LINAJES PRINCIPALES DE LOS DOMINIOS ARCHAEA Y BACTERIA

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRA EN CIENCIAS BIOLÓGICAS

PRESENTA:

CORAL CRUZ GONZÁLEZ LUNA

TUTOR PRINCIPAL DE TESIS:

DR. ARTURO CARLOS II BECERRA BRACHO
FACULTAD DE CIENCIAS, UNAM

COMITÉ TUTOR:

DRA. LUISA ISAURA FALCÓN ÁLVAREZ
INSTITUTO DE ECOLOGÍA, UNAM
DR. LUIS DAVID ALCARAZ PERAZA
FACULTAD DE CIENCIAS, UNAM

CIUDAD UNIVERSITARIA, CIUDAD DE MÉXICO, ABRIL 2018

UNAM
POSGRADO
Ciencias Biológicas

COORDINACIÓN

POSGRADO EN CIENCIAS BIOLÓGICAS
FACULTAD DE CIENCIAS
DIVISIÓN ACADÉMICA DE INVESTIGACIÓN Y POSGRADO

OFICIO FCIE/ DAIP /358/2018

ASUNTO: Oficio de Jurado

Lic. Ivonne Ramírez Wence
Directora General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día **5 de marzo de 2018** se aprobó el siguiente jurado para el examen de grado de **MAESTRA EN CIENCIAS BIOLÓGICAS** en el campo de conocimiento de **Biología Evolutiva** del (la) alumno(a) **CRUZ GONZÁLEZ LUNA CORAL** con número de cuenta **407088216** con la tesis titulada **"Reconstrucción de los genomas ancestrales de los linajes principales de los dominios Archaea y Bacteria"**, realizada bajo la dirección del (la) **DR. ARTURO CARLOS II BECERRA BRACHO:**

Presidente:	**DR. ANTONIO EUSEBIO LAZCANO-ARAUJO REYES**
Vocal:	**DRA. ALICIA NEGRÓN MENDOZA**
Secretario:	**DR. LUIS DAVID ALCARAZ PERAZA**
Suplente:	**DR. RAFAEL CAMACHO CARRANZA**
Suplente:	**DRA MARÍA COLÍN GARCÍA**

Sin otro particular, me es grato enviarle un cordial saludo.

A T E N T A M E N T E
"POR MI RAZA HABLARA EL ESPÍRITU"
Ciudad Universitaria, Cd. Mx., a 9 de abril de 2018

POSGRADO EN CIENCIAS BIOLOGICAS
COORDINACIÓN

DR. ADOLFO GERARDO NAVARRO SIGÜENZA
COORDINADOR DEL PROGRAMA

AGNS/VMVA/ASR/mnm

Unidad de Posgrado • Coordinación del Posgrado en Ciencias Biológicas Edificio D, 1er. Piso, Circuito de Posgrados Cd. Universitaria
Delegación Coyoacán C.P. 04510 Cd. Mx. Tel. 5623 7002 http://pcbiol.posgrado.unam.mx

Agradecimientos

Al Posgrado en Ciencias Biológicas de la Universidad Nacional Autónoma de México, y a la UNAM misma, por brindarme una educación de la más alta calidad.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT), por el apoyo económico que hizo posible la realización de mi proyecto de maestría (CVU: 697387 / Número de becario: 583827). A los apoyos Beca Mixta (CONACyT) y PAEP (PCB, UNAM) que permitieron mi participación en una estancia de investigación (Universidad de Connecticut, agosto-septiembre 2016), y al apoyo PAPIIT-UNAM (IN223916) gracias al cual tuve la oportunidad de asistir a un congreso internacional (ISSOL, julio 2017).

A mi tutor principal de tesis, mi gran maestro el Dr. Arturo Carlos II Becerra Bracho, por la confianza que me ha tenido, por el apoyo, la paciencia, la disposición y la motivación constantes. A los miembros de mi comité tutor, la Dra. Luisa Isaura Falcón Álvarez y al Dr. Luis David Alcaraz Peraza, por su disposición, y sus comentarios durante el proyecto, siempre dirigidos a mejorar la calidad del mismo. A los tres, gracias por compartir su conocimiento y sus ideas, y por contribuir a la formación de mi pensamiento biológico, su trabajo es un ejemplo y una fuente de inspiración para mi quehacer científico.

A los profesores miembros de mi jurado, quienes hicieron comentarios muy pertinentes para mejorar el manuscrito de la tesis.

Agradecimientos a título personal

Al Dr. Antonio Lazcano por darme una oportunidad en el Laboratorio de Origen de la Vida, y porque gracias a su apoyo incondicional he crecido profesional y personalmente. De nuevo agradezco a mi tutor, el Dr. Arturo Becerra, por su calidad humana y por la confianza que me ha tenido, porque su gusto por el estudio de la evolución se contagia. Es un privilegio tener a ambos como maestros, son excelentes, gracias por enseñarme a plantear las preguntas más interesantes y retadoras concernientes al origen y la evolución de la biota.

A mis colegas del laboratorio, los macacos y las macacas, quienes de alguna u otra forma contribuyeron al éxito de este proyecto. Gracias por compartir su conocimiento.

A mi familia, porque su cariño y su esfuerzo es fundamental en mi vida. Sin ustedes nada de esto sería posible.

A Raziel, mi compañero, mi cómplice, mi mejor amigo, mi colega, y tantas cosas más, quien en las buenas y en las malas me acompaña, gracias, este logro también es tuyo.

A mi país, mi roto México, le deseo tiempos mejores, paz y amor

A mis abuelas, Elsa y Josefina

Índice

Lista de figuras y tablas

Figura anexo 3. Mapa de calor del número total de proteínas presentes en cada categoría funcional.

Figura anexo 4. Proteínas de la subunidad 30S ribosomal presentes en las reconstrucciones ancestrales.

Figura anexo 5. Proteínas de la subunidad 50S ribosomal presentes en las reconstrucciones ancestrales.

Figura anexo 6. Enzimas aminoacil-tRNA sintetasas presentes en las reconstrucciones ancestrales.

Tabla 1. Distintas reconstrucciones del catálogo génico del LCA.

Tabla 2. Comparación entre las especies modernas y las reconstrucciones ancestrales.

Tabla 3. Enzimas con altos niveles de conservación al interior de los catálogos ancestrales.

Tabla 4. Resultados del análisis pangenómico.

Tabla 5. Tamaño de las secuencias proteínicas conservadas al interior de los catálogos ancestrales de los distintos linajes.

Tabla anexo 1. Especies a partir de cuyo genoma se realizó la agrupación de posibles ortólogos, utilizando el algoritmo BDBH con el programa *Get-homologues*.

Tabla anexo 2. Categorías funcionales propuestas en la base de datos COG.

Resumen

El último ancestro común (LCA, por sus siglas en inglés) representa la etapa de evolución biológica previa a la diversificación de los dominios primarios, una población que se estima habitó la Tierra hace más de 3,500 millones de años. No obstante la falta de información sobre las posibles condiciones ambientales, es factible aproximarnos a la naturaleza del LCA ya que, en teoría, las características asociadas al nodo basal de la filogenia universal representan al ancestro, o bien, bajo la lógica de que los rasgos homólogos a los tres dominios de la vida (Bacteria, Archaea y Eukarya) corresponden al LCA. Gracias al desarrollo de bases de datos y de herramientas de análisis bioinformático, se ha aplicado dicho razonamiento para proponer el catálogo génico/proteínico del LCA. Aunque las distintas reconstrucciones publicadas tienen semejanzas importantes, difieren significativamente en cuanto al contenido genético que proponen tendría el ancestro, lo que ha generado discusiones en torno a su biología. Además, el método de reconstrucción ancestral basado en la comparación directa de genomas completos no se ha aplicado a casos de ancestros más recientes, lo que significa que no se cuenta con un análisis cuyos resultados se puedan comparar con los estudios genómicos del LCA. Por esta razón, el objetivo general del presente trabajo fue reconstruir los genomas ancestrales de los principales linajes de Archaea y Bacteria, bajo la hipótesis de que la reconstrucción de dichos catálogos ancestrales fungirá como un control metodológico para el caso del LCA. Para este fin, primero se seleccionó una muestra representativa de especies de vida libre y se descargó su proteoma completo del NCBI, lo que resultó en una base de datos de 785 procariontes, 642 pertenecientes a 12 phyla de Bacteria (Actinobacteria, Firmicutes, Bacteroidetes, Cyanobacteria, Deinococcus–Thermus, Thermotogae, Chloroflexi, Aquificae, Verrucomicrobia–Planctomycetes, Spirochaetes, Acidobacteria y Proteobacteria) y 143 de tres phyla de Archaea (Euryarchaeota, Crenarchaeota, Thaumarchaeota). Después, con el programa *Get-homologues* se identificaron las proteínas homólogas al interior de cada linaje, las cuales se clasificaron funcionalmente. En total se reconstruyeron 19 catálogos ancestrales, los cuales se caracterizaron por contener procesos y/o rutas incompletas, familias proteínicas de distinta antigüedad, y por la presencia de un sesgo doble. La

comparación de los catálogos con las especies modernas evidenció que se recupera un porcentaje muy bajo de los ancestros, lo que sugirió un sesgo cuantitativo, pero además no todas las funciones celulares están igualmente conservadas, lo que sugirió también un sesgo cualitativo. Los factores que afectan el tamaño de las reconstrucciones ancestrales son el número de especies en la muestra, la diversidad fisiológica de los organismos, la varianza tanto en el tamaño del genoma como del pangenoma, y la distancia filogenética. Se observó que conforme se reduce el tamaño de los catálogos, se excluyen elementos asociados al metabolismo y se mantienen aquellos relacionados con procesos informativos. El presente trabajo evidenció los límites y alcances de la comparación de la estructura primaria de las proteínas (secuencia de aminoácidos), ya que los resultados explican de manera satisfactoria la naturaleza de las distintas reconstrucciones del LCA, las cuales exhiben el doble sesgo descrito. Se concluye que el LCA no era un organismo simple.

Palabras clave: LCA, último ancestro común, reconstrucción ancestral, genómica comparada, análisis pangenómico

Abstract

The Last Common Ancestor (LCA) represents to a population during the early evolutionary stage prior the diversification of the primary cellular domains. Since geological evidence of the environmental conditions, as well as a fossil record of this epoch is scarce or absent, the nature of the LCA must be inferred from the homologous traits, i.e. genes/proteins, found among its descendants. Published estimates of the genetic complement of the LCA, although not strictly comparable, are remarkably dissimilar, noticeably lacking a methodological control to compare with. In this work, we reconstructed the ancestral genome catalogue of diverse prokaryotic lineages, in order to explore the scope and limitations of the direct comparative genomics approach.

Introducción

El último ancestro común

El concepto de un ancestro universal a todos los seres vivos, fue propuesto de manera visionaria por Charles Darwin en su libro *El origen de las especies* (1859), cuando planteó una extrapolación hacia el pasado profundo de su teoría de evolución por ancestría común. Dicha predicción se confirmó más de un siglo después cuando Woese y Fox (1977), mediante la comparación de secuencias de RNA de la subunidad pequeña del ribosoma (16S/18S), construyeron la primera filogenia universal y descubrieron la existencia de tres dominios celulares, denominados Bacteria, Archaea y Eukarya (Woese *et al.,* 1990), los cuales representan líneas de descendencia a partir de un ancestro común.

Como consecuencia de su origen monofilético, actualmente los tres dominios se asemejan en rasgos esenciales de la expresión de la información genética, así como en la forma en la que producen su energía. Pero también es un hecho que, por el tiempo de divergencia, han acumulado diferencias (Becerra *et al.,* 2007). Justamente por las diferencias en las maquinarias de transcripción y de traducción, en un inicio Woese propuso que los tres dominios son descendientes directos de una comunidad de *progenotes*, entidades simples con genomas de RNA, que se caracterizan por que su genotipo no está claramente separado del fenotipo (Woese, 1987; Woese, 1998). Sin embargo, esta primera noción del ancestro común rápidamente se abandonó, pues tanto el estudio del código genético (Fitch y Upper, 1987), como un análisis de los rasgos homólogos entre los dominios (Lazcano *et al.,* 1992), coincidieron en que el LCA debió ser un organismo tan complejo como un procarionte moderno.

El LCA, o la población de dichos organismos ancestrales, debe verse como resultado del proceso evolutivo que inició muy temprano en la Tierra, es decir que le antecede un período de evolución molecular y biológica cuya duración desconocemos. Debido a que comúnmente se confunde el problema del origen de la vida con el del LCA, es importante establecer la temporalidad relativa del LCA.

Temporalidad relativa del LCA

Existe un consenso en el sentido de que la vida microbiana ya poblaba la Tierra hace 3,500 millones de años (Ma) y posiblemente antes (Altermann y Kazmierczak, 2003; Knoll *et al.,* 2016), periodo en el que el planeta se encontraba pleno de microorganismos, tales como reductores de azufre, fotótrofos anoxigénicos y posiblemente metanógenos (Canfield, 2006). Si el LCA representa la etapa previa a la diversificación de los dominios primarios, los cuales actualmente se reconoce que son Bacteria y Archaea (Williams *et al.,* 2013), entonces el LCA debió haber existido antes de la fecha mencionada (Figura 1). En el contexto planetario, se plantea que hubo un bombardeo masivo de cuerpos originados en asteroides planetésimos, que se extendió hasta hace aproximadamente 3,850 Ma (Brasier *et al.,* 2006), el cual pudo haber limitado la existencia de la biota.

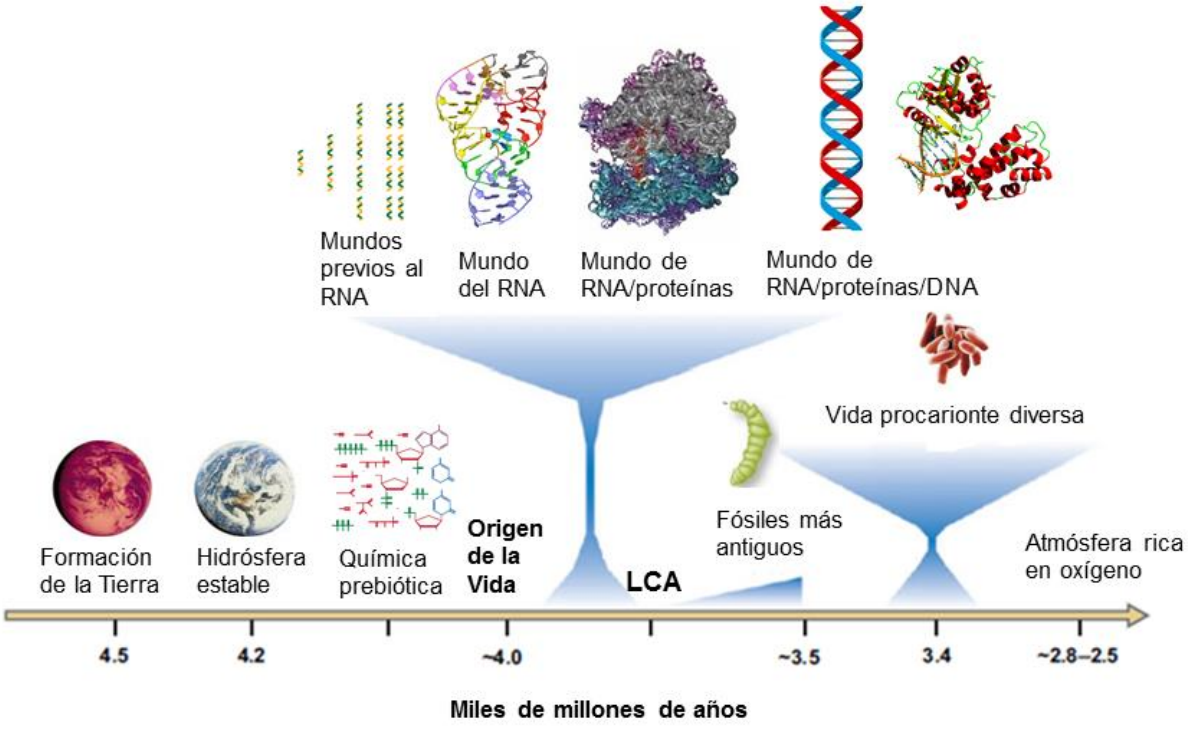

Figura 1. Temporalidad relativa del último ancestro común (LCA). Modificada de Joyce, 2002 y Becerra *et al.,* 2007.

A diferencia de lo expuesto en la Figura 1, se puede suponer que el bombardeo tardío no fue catastrófico para la vida. Recientemente se ha sugerido que el LCA estaría situado en el periodo Hadeano, hace aproximadamente 4250 Ma (Cantine y Forunier, 2018), lo que a su vez colocaría al origen de la vida como un evento muy temprano una vez formada la Tierra.

De cualquier manera, una serie de eventos precedieron al LCA, tales como la acumulación de agua líquida y compuestos orgánicos en la Tierra primitiva, el origen de la vida, así como el origen y la diversificación de las primeras células (Figura 1). Aquí vale la pena mencionar la definición de Fitch y Upper (1987) del término cenancestro (sinónimo de LCA), como "el ancestro más reciente de los organismos que están vivos hoy en día", porque si bien hubo seres vivos que le antecedieron e incluso que le fueron contemporáneos, todos los organismos modernos son descendientes del LCA.

Debido a la falta de evidencias de la Tierra primitiva, la investigación contemporánea de la biología del LCA se basa en la posibilidad de reconstruir su catálogo de genes, a partir del estudio de sus descendientes.

Reconstrucción del genoma del LCA

Desde el punto de vista cladístico, los estados de un caracter asociados a los nodos ancestrales, llamados plesiomorfías, representan el fenotipo del ancestro. Es decir, que el conjunto de rasgos que se puedan inferir en el nodo ancestral de la filogenia universal representan al LCA (Becerra *et al.,* 2007; Delaye y Becerra 2012).

Por lo anterior, el problema de enraizar el árbol universal para darle polaridad a los caracteres cobró suma importancia en el estudio del LCA, cuestión que se resolvió gracias al análisis de genes parálogos cuya duplicación fue un evento muy antiguo, previo a la diversificación de los dominios primarios. La raíz del árbol universal sitúa al LCA en la rama de Bacteria, como muestran tanto las subunidades alfa y beta de las ATPasas tipo F (Gogarten *et al.,* 1989), como los factores de elongación EF-G y EF-Tu (Iwabe *et al.,* 1998).

Sin embargo, hay que tener precaución, pues los grupos que resultan cladísticamente antiguos en las filogenias, no necesariamente presentan rasgos primitivos (Islas *et al.,* 2003). Sin duda, enraizar al dominio Bacteria o al dominio Archaea son temas de investigación no resueltos del todo y con implicaciones importantes en el estudio del LCA.

Otra aproximación a la reconstrucción ancestral es la caracterización directa, que se basa en el principio de la biología comparada, la cual permite deducir las características del ancestro de un linaje a través de la detección de rasgos homólogos (así sean anatómicos, morfológicos, fisiológicos, ultraestructurales, moleculares, etc.). Esta lógica simple se puede extender a la comparación de genomas completos, bajo la premisa de que los genes homólogos a los tres dominios, por el principio de parsimonia, estarían presentes en el LCA (Lazcano *et al.,* 1992; Delaye y Becerra, 2012).

Con lo dicho anteriormente podemos afirmar que el catálogo génico del LCA se puede reconstruir mediante cladística molecular, genómica comparada, o una combinación de ambos métodos. A continuación, se describen las distintas estimaciones que se han propuesto del catálogo de genes del LCA (Tabla 1).

Una vez que hubo dos genomas completamente secuenciados (*Mycoplasma genitalium* y *Haemophilus influenzae*), Mushegian y Koonin (1996) los compararon para identificar genes ortólogos. Definieron a los ortólogos como genes en dos especies, transmitidos por descendencia vertical a partir de un ancestro común, y que codifican para la misma función (Mushegian y Koonin, 1996). Construyeron un catálogo mínimo de genes que extrapolaron al LCA y concluyeron que éste tendría un genoma de RNA. Este trabajo ha sido fuertemente criticado debido a que *Haemophilus influenzae* y *Mycoplasma genitalium* son especies parásitas y presentan pérdidas secundarias de genes, lo que claramente origina un sesgo en la reconstrucción ancestral (Becerra *et al.,* 1997).

El trabajo de Kyrpides y colaboradores (1999) consistió en buscar genes homólogos a los de *Methanococcus jannaschii* (el primer genoma arqueano secuenciado) en los dominios Bacteria y Eukarya, aplicando el método comparativo. El catálogo ancestral resultante presentó 246 funciones bioquímicas únicas (reacciones enzimáticas) que

asociaron al LCA, y concluyeron que éste ya presentaba metabolismo, así como un sistema genético similar al de los organismos modernos. Este ejercicio evidenció que la metodología comparativa tiene limitaciones, pero que sin embargo provee de una estimación razonable del mínimo de las funciones del LCA (Kyrpides *et al.,* 1999).

Tabla 1. Distintas reconstrucciones del catálogo génico del LCA.

Autores	Catálogo génico del LCA
Mushegian y Koonin, 1996	256 genes
Kyrpides *et al.,* 1999	324 proteínas
Koonin, 2003	63 genes (proteínas)
Mirkin *et al.,* 2003	600 genes (COGs)
Harris *et al.,* 2003	80 COGs
Delaye *et al.,* 2005	115 dominios proteínicos (Pfam)
Yang *et al.,* 2005	49 plegamientos (superfamilias de SCOP)
Sobolevsky y Trifonov, 2006	20 motivos (octapéptidos en proteínas)
Ouzounis *et al.,* 2006	1000 genes con 561 – 661 secuencias (proteínas) o categorías funcionales
Ranea *et al.,* 2006	140 dominios proteínicos ancestrales (CATH)
Tuller *et al.,* 2010	784 proteínas ancestrales
Kim y Caetano-Anollés, 2011	70 – 152 dominios (superfamilias)
Kannan *et al.,* 2013	517 – 507 genes
Weiss *et al.,* 2016	335 familias proteínicas

Acrónimos: COGs (*Clusters of Orthologous Groups of proteins*; Tatusov *et al.,* 2000); Pfam (*Protein domain families*, Bateman *et al.,* 2004); SCOP (*Structural Classification of Proteins*, Murzin *et al.,* 1995); CATH (*Protein Structure Classification*, Orengo *et al.,* 1997).

En 2003 Koonin se basó en la metodología comparativa para buscar genes ubicuos en 100 genomas completamente secuenciados, lo que resultó en solamente 63 proteínas universales, la mayoría relacionadas con la traducción, y muy pocas con la transcripción y la replicación. Por lo reducido del catálogo génico universal, el autor concluyó que el LCA sería simple, al grado de no contar con un sistema genético y replicativo basado en DNA (Koonin 2003).

De forma casi simultánea, el mismo laboratorio publicó otra reconstrucción del LCA con resultados muy distintos. Su metodología se basó primero en construir árboles para los conjuntos de proteínas ortólogas contenidas en los COGs, y segundo en proponer un escenario, basado en el criterio de máxima parsimonia, que te reconcilia dichos árboles con la filogenia de las especies, tomando en cuenta la pérdida de genes, la transferencia horizontal y la aparición de nuevos COGs. Así, los autores concluyeron que el LCA tendría un poco menos de 600 proteínas, número casi suficiente para mantener a un organismo vivo (Mirkin *et al.,* 2003). Debido a que no les interesaba describir a detalle las funciones del LCA, ellos enfatizaron la idea de que la transferencia horizontal de genes, así como la pérdida de éstos, son eventos muy frecuentes en la evolución de los procariontes (Mirkin *et al.,* 2003).

En el mismo año Harris y sus colegas identificaron aquellos COGs, cuya filogenia tuviera la misma topología que la filogenia universal, es decir, aquella construida con el gen del RNA ribosomal, y que por lo tanto podrían haber estado en el LCA (Harris *et al.,* 2003). De los 80 COGs universales que encontraron, solamente 50 cumplieron con el criterio descrito, pero bastaron para proponer un LCA con un sistema de transcripción y un ribosoma eficiente, con funciones asociadas a la membrana y con la capacidad de sintetizar moléculas de DNA (Harris *et al.,* 2003).

Aplicando el enfoque de la caracterización directa, Delaye y colaboradores propusieron un catálogo de 115 dominios proteínicos, una vez que compararon 20 genomas completamente secuenciados de organismos provenientes de los tres dominios de la vida (Delaye *et al.,*2005). Sus resultados apoyan la noción de un LCA con un sistema genético

equivalente al de un procarionte moderno, e hicieron énfasis en que dentro del conjunto de proteínas conservadas la mayoría se relacionan con el metabolismo de RNA.

Utilizando otra metodología, que se basó en detectar la presencia o ausencia de dominios en genomas completos, pero cuyo análisis se hizo a nivel de superfamilias SCOP, Yang y su equipo propusieron un catálogo de casi 50 plegamientos comunes en los 174 genomas que analizaron (Yang *et al.,* 2005), inventario suficiente para proponer a un LCA genética y estructuralmente muy sofisticado.

En un trabajo posterior, Sobolevsky y Trifonov (2006) identificaron pequeñas secuencias de aminoácidos, octapéptidos, que se encuentran universalmente conservados. Éstos forman módulos o regiones conservadas en la estructura de diversas proteínas. Los autores concluyeron que alrededor de 21 octapéptidos podrían haber estado en LCA (Sobolevsky y Trifonov, 2006), algunos de los cuales se relacionan con funciones como la traducción y la transcripción.

Una nueva reconstrucción del LCA, realizada por Ouzounis y sus colegas (2006), estableció en 1000 genes el contenido del ancestro, que representan aproximadamente 600 categorías funcionales. El análisis se basó en 184 genomas completamente secuenciados, para los cuales se identificaron familias proteínicas comunes. Posteriormente construyeron una filogenia universal y determinaron aquellas proteínas presentes en el nodo ancestral, utilizando un algoritmo que considera eventos de pérdida de genes a lo largo del árbol (Ouzounis *et al.,* 2006). Coincidieron con otros estudios en que el LCA sería tan complejo genéticamente como un procarionte moderno de vida libre.

En el mismo año se publicó otra reconstrucción del LCA, en la cual se proponen 140 dominios proteínicos como ancestrales (Ranea *et al.,* 2006). Para llegar a dicho resultado, Ranea y su equipo buscaron secuencias representativas, correspondientes a las superfamilias de dominios estructurales de la base de datos CATH, en 114 genomas completamente secuenciados pertenecientes a los tres dominios de la vida. Aquellas familias distribuidas en el 90% de los genomas se consideraban ancestrales. Los autores reconocen la versatilidad de las superfamilias proteínicas, en el sentido de que algunas

muestran una mayor diversidad genética y funcional, como es el caso de las relacionadas con el metabolismo. Además, concluyen que el LCA sería genéticamente complejo, con rasgos muy semejantes a los organismos modernos (Ranea *et al.,* 2006).

Con una metodología un tanto novedosa, Tuller y sus colegas (2010) utilizaron la información de la coevolución de las proteínas para resolver ambigüedades en cuanto al contenido de genes en lo nodos ancestrales, y propusieron un LCA con 784 proteínas y por lo tanto, con un genoma y un funcionamiento similar al de los procariontes modernos (Tuller *et al.,* 2010).

Con la idea de que la estructura de las proteínas está más conservada que su secuencia primaria, Kim y Caetano-Anollés (2011) utilizaron estructuras de las superfamilias (FSFs) de la base de datos SCOP presentes en 420 organismos, para construir un árbol filogenético e inferir las características del nodo ancestral, considerando eventos de pérdida y de ganancia. El resultado es un mínimo de 70 dominios estructurales y un máximo de 152 para el LCA, aunque de manera sorprendente concluyen que tendría un genoma de RNA (Kim y Caetano-Anollés 2011).

Para determinar qué familias de ortólogos presentes en diversas especies, podrían haber estado en el LCA, Kannan y sus colegas (2013), construyeron una filogenia universal y aplicaron un modelo de máxima verosimilitud que considera eventos de pérdida y ganancia de ortólogos a lo largo de los nodos del árbol. Sus estimaciones fueron de poco más de 500 genes para el LCA, pero no profundizaron en la interpretación hacia la naturaleza del LCA (Kannan *et al.,* 2013).

La reconstrucción más reciente del LCA es la de Weiss y su equipo (2016), y se basó en las filogenias de diferentes familias proteínicas, las cuales se consideraban ancestrales básicamente si recuperaban la topología de dos dominios (Archaea y Bacteria) y si se presentaban en al menos dos linajes de cada dominio, lo que resultó en un ancestro de 335 familias proteínicas (Weiss *et al.,* 2016). Sin embargo, este trabajo ha sido fuertemente criticado por la falta de precaución en su metodología, la cual presenta falsos positivos y

falsos negativos, y por errores tan garrafales como confundir el concepto de progenote con el del LCA (Gogarten y Deamer, 2016).

El conjunto de familias de genes/proteínas sumamente conservadas que emergen de las distintas estimaciones del LCA, sugieren que éste poseía un genoma de DNA y un metabolismo equivalente en términos de complejidad genética y funcional al de un procarionte moderno (Becerra *et al*., 2007). Sin embargo, llama la atención el hecho de que las estimaciones del contenido del catálogo ancestral, aunque no son estrictamente comparables, son numéricamente muy disímiles. Además, las distintas reconstrucciones no siempre recuperan las mismas familias proteínicas. Si bien se recuperan genes (o proteínas) que participan en procesos biológicos esenciales, es notable que éstos no están conservados en su totalidad, es decir, los catálogos presentan siempre rutas metabólicas o funciones incompletas (Becerra *et al*., 2007), lo que limita las interpretaciones sobre el metabolismo del LCA. Esta falta de consenso incluso ha motivado que exista una base de datos exclusivamente para los resultados de las distintas reconstrucciones del LCA (Goldman *et al*., 2013). A pesar de esta ambigüedad, no faltan propuestas en la literatura que afirman que el LCA sería metanógeno (Wong *et al*., 2007; Weiss *et al*., 2016).

Es un hecho que el desarrollo de la genómica del LCA ha generado muchos debates respecto de su naturaleza. Dichas controversias se ven reflejadas en la diversidad de nombres que se han propuesto (cf. Becerra *et al*., 2007), como progenote (Woese y Fox, 1977), cenancestro (Fitch y Upper, 1987), LUCA (Kyrpides *et al*., 1999), último ancestro celular universal (Philippe y Forterre, 1999), ancestro universal (Doolittle, 2000), última comunidad común (Line, 2002), ancestro común más reciente (Zhaxybayeva y Gogarten, 2004) o urancestro (Kim & Caetano-Anollés, 2011), pero también en la diversidad de los temas que se han elaborado y que se discuten en la literatura, tales como el nivel de complejidad, el tipo de material genético, de membrana y de metabolismo, así el posible hábitat y la temperatura en la que se pudo haber originado el LCA (Woese, 1998; Lazcano *et al.,* 1992; Becerra *et al*., 2007; Wächstershäuser, 2003; Peretó *et al*., 2004; Lazcano 2011, y referencias de **Tabla 1**).

Factores que afectan la inferencia ancestral

La diversidad en los resultados del contenido genético del LCA se debe en principio a diferencias en las bases de datos y en las metodologías utilizadas (Becerra *et al*., 2007). Por ejemplo, existen diferencias en los procedimientos para identificar a los ortólogos o en los métodos de inferencia filogenética. Otro factor a considerar es la variación en las tasas de sustitución de las diferentes proteínas (Becerra *et al*., 2007), la cual puede originar artefactos en las inferencias filogenéticas.

Aunado a lo anterior, la genómica del LCA ha revelado una serie de procesos (intrínsecos a la evolución de los genomas y de las proteínas) que pueden generar ruido en las reconstrucciones ancestrales. Los más evidentes son la pérdida de genes y la transferencia horizontal (Becerra *et al*., 1997; Zhaxybayeva y Doolittle, 2011). Por ejemplo, si la transferencia lateral de genes fue un evento muy común en el pasado, entonces el catálogo del ancestro se puede sobreestimar (Delaye y Becerra, 2012), al grado de construir un LCA totipotencial. Por el contrario, si los eventos de pérdida fueron muy intensos, entonces se corre el riesgo de subestimar el catálogo y por lo tanto las funciones del ancestro (Delaye y Becerra, 2012).

Definición del pangenoma y núcleo genómico

La aplicación de métodos de genómica comparada a cepas de la especie *Streptococcus agalactiae* reveló la existencia del pangenoma y del núcleo genómico (Tettelin *et al*., 2005). El pangenoma es la totalidad de genes presentes en un conjunto de cepas bacterianas. Se compone de los genes universales, es decir, que están presentes en todas las cepas y que se denominan núcleo genómico estricto. También contiene genes casi universales, presentes en algunas cepas, denominados núcleo genómico accesorio o relajado, el cual incluye también a las proteínas que son exclusivas de cada cepa y que son la mayoría (Tettelin *et al*., 2005). Existen diversos análisis del pangenoma de distintas cepas y de ciertos géneros (revisados por ejemplo en Rouli *et al*., 2015; Vernikos et al., 2015), y se ha visto que los pangenomas pueden ser de tamaño finito o infinito (cerrados o abiertos)

(Lefebure *et al*., 2010). En una interpretación más amplia del concepto, se ha estimado que el pangenoma del dominio Bacteria es abierto (Lapierre y Gogarten, 2009). Desde el punto de vista evolutivo, el núcleo genómico de todos los organismos modernos es una consecuencia de la naturaleza conservativa de la evolución (Lapierre y Gogarten, 2009) y de su herencia común a partir del LCA, como lo sugieren las distintas reconstrucciones que se han realizado (Tabla 1).

Planteamiento del problema y objetivos

El estudio del último ancestro común es central para entender la evolución biológica temprana en la Tierra, en particular, el periodo que antecede directamente a la diversificación de los dominios celulares modernos. La generación de bases de datos de genomas completos, aunado al desarrollo de herramientas bioinformáticas, han permitido el planteamiento de metodologías que tienen como objetivo reconstruir el catálogo de genes que tendría el LCA. Sin embargo, las diversas propuestas que se han realizado, si bien tienen coincidencias importantes, difieren de manera drástica en cuanto al contenido genético/proteínico del LCA, lo que ha originado discusiones importantes en torno a su naturaleza. Esta disparidad en los resultados de las estimaciones, se explican por i) la intensidad de la pérdida de genes y de la transferencia lateral en el pasado, ii) la aplicación de metodologías y criterios un tanto diferentes, iii) la cantidad y la calidad de la información disponible en cada estudio. Aunado a lo anterior, el método de caracterización directa en el contexto de la reconstrucción ancestral nunca se ha aplicado a un caso que no sea el LCA, lo que significa que no se cuenta con un análisis control cuyos resultados se puedan comparar con los estudios genómicos del LCA.

Debido a las notables diferencias en los resultados de las distintas estimaciones del catálogo ancestral del LCA (Tabla 1), en el presente trabajo se propone la hipótesis de que las reconstrucciones de los genomas ancestrales de los principales linajes de los dominios Archaea y Bacteria, fungirán como un control metodológico para el caso del LCA, en el

sentido de que permitirán evaluar el alcance y las limitaciones de la comparación de la estructura primaria de las proteínas.

Objetivo general

- Reconstruir el catálogo ancestral de los principales linajes procariontes, mediante el método de genómica comparada, para contrastar los resultados con el caso de la reconstrucción del LCA.

Objetivos particulares

- Construir una base de datos de genomas completos (proteomas) pertenecientes a especies procariontes de vida libre.
- Detectar las proteínas homólogas al interior de cada linaje con el programa *Get-homologues* (Contreras-Moreira y Vinuesa, 2013).
- Clasificar funcionalmente las familias proteínicas al interior de los catálogos ancestrales.
- Contrastar los resultados con las especies modernas y con las reconstrucciones del LCA.

Material y Métodos

Características de la muestra

No todas las especies cuyo genoma está completamente secuenciado y disponible (hasta diciembre de 2016) se consideraron en el presente análisis. Se excluyeron organismos con un estilo de vida endosimbionte y endoparásito, debido a que causan un sesgo en las reconstrucciones ancestrales (Becerra *et al*., 1997). También se excluyeron los patógenos o comensales obligados, de tal forma que los organismos que conforman la muestra son bacterias y arqueas de vida libre.

Debido a que el objetivo era reconstruir al ancestro de cada grupo, no se consideraron linajes que tuvieran menos de cinco especies completamente secuenciadas, lo que dejó fuera de la muestra grupos de reciente descubrimiento, como por ejemplo Melainabacteria, Kiritimatiellaeta, Cloacimonetes, Peregrinibacteria, Thermodesulfobacteria, Dyctioglomi, Nanoarchaeota, Nanohaloarchaeota, Korarchaeota, Bathyarchaeota o Lokiarchaeota.

Descarga de las secuencias

Una vez que se hizo la selección final de las especies a estudiar, se descargó el archivo correspondiente a su proteoma, en formato FASTA, a partir de la base de datos del NCBI (*National Center for Biotechnology Information*, https://www.ncbi.nlm.nih.gov/genome/). Dicho archivo contiene las secuencias de aminoácidos de todas las proteínas que posee cada organismo de la muestra. La descarga de las secuencias se realizó durante el periodo de agosto de 2015 hasta diciembre de 2016.

Cabe mencionar que la base de datos del NCBI posee varios sesgos, uno de ellos consiste en la presencia de linajes, en particular de Bacteria, con muchas especies completamente secuenciadas. Debido a esta sobrerepresentación de ciertos grupos, fue necesario eliminar la redundancia al interior de la muestra, por lo que se eligió solamente a un representante por especie. En el caso de los grupos más numerosos que fueron

Firmicutes, Gammaproteobacteria y Betaproteobacteria, fue necesario seleccionar a un representante por género.

Construcción de los catálogos ancestrales

Para comparar las secuencias de cada proteoma y detectar las proteínas conservadas al interior de cada linaje, se utilizó el programa *Get-homologues* (Contreras-Moreira y Vinuesa, 2013). Dicho paquete informático funciona en dos etapas: primero hace un BLASTp (Altschul *et al*., 1997) para evaluar la similitud entre las secuencias, y posteriormente se agrupan aquellas que son potencialmente ortólogas. Los parámetros establecidos para el BLASTp fueron una cobertura (*query coverage*) del 75%, un valor de *e* menor a $1x10^{-5}$, además de que se evaluó el alineamiento con la matriz BLOSUM62 (Henikoff y Henikoff, 1992).

El algoritmo utilizado para agrupar a los posibles ortólogos se denomina, por sus siglas en inglés, BDBH o BBH (*bidirectional best hits* o *reciprocal best hit*) y funciona bajo el supuesto de que los ortólogos son todos los pares de genes compartidos entre dos especies, que son más parecidos entre ellos, que con el resto de los genes presentes en otras especies. Por lo tanto, para que dos proteínas se consideren homólogas, digamos A y B, entonces A debe representar el mayor puntaje al evaluar el alineamiento con B y viceversa.

A continuación se describe cómo se implementa el algoritmo BDBH en el programa *Get-homologues* (Vinuesa y Contreras-Moreira, 2015). Para agrupar a los ortólogos potenciales es requisito seleccionar un proteoma de referencia (secuencia *query* o genoma semilla). El proceso comienza identificando los parálogos internos (*inparalogues*), definidos como BDBH dentro de un mismo proteoma, los cuales se van a descartar. Posteriormente se compara la secuencia *query* o genoma semilla con otro proteoma, se identifican los BDBH y se almacenan, proceso que se repite con cada uno del resto de los genomas (Figura 2). Una desventaja del proceso es que solamente aquellas proteínas presentes en el genoma semilla pueden estar en el núcleo conservado, es decir que, si una secuencia no se presenta en el genoma de referencia entonces no hay forma de que se incluya en el núcleo

conservado, lo que sugiere que la selección del genoma semilla sesga el resultado del análisis.

Una ventaja del método usado es que permite identificar posibles homólogos presentes en un menor número de especies. Por ejemplo, si el total de especies en la muestra se representa como *n,* entonces es posible detectar proteínas que se encuentran en casi todas las especies, digamos en el conjunto *n-1* o *n-2*. A este catálogo de proteínas casi universales se le denomina núcleo genómico relajado (s*oft* o *relaxed core*), y el valor de corte que lo define es arbitrario. Por ejemplo Kaas y sus colegas establecen el corte en un 95%, es decir, si la proteína se encuentra en al menos el 95% de las cepas, entonces lo consideran parte del núcleo relajado (Kaas *et al*., 2012). En el presente trabajo se exploraron, además del núcleo genómico estricto, aquellas proteínas presentes en el total de la muestra menos una especie (el conjunto *n-1*).

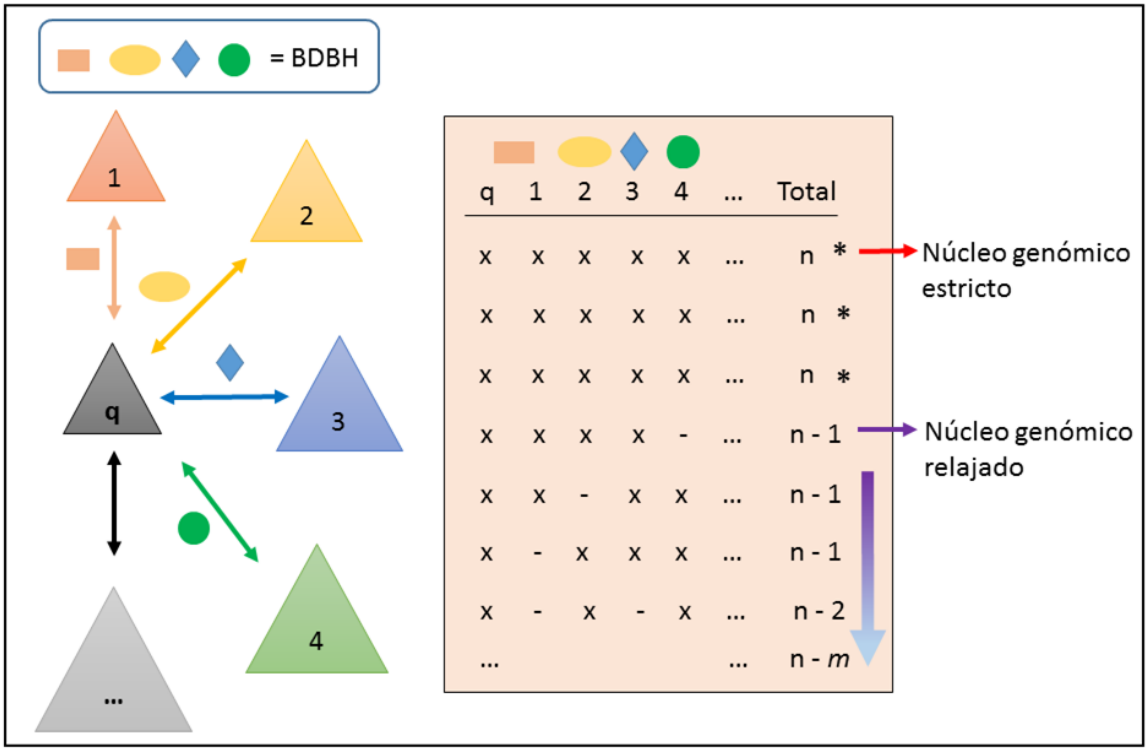

Figura 2. Algoritmo BDBH en *Get-homologues*. El genoma semilla (q) se compara con cada uno del resto de los genomas y los BDBH se almacenan. Al final los conjuntos que contienen al menos una

proteína por genoma representan el núcleo genómico estricto, aunque ciertamente el criterio puede ser más laxo, para explorar lo que se conoce como el núcleo genómico relajado.

En general los genomas con la menor cantidad de proteínas fungieron como la semilla para iniciar el agrupamiento (Anexo 1), salvo cuando en el ensamblaje no se proporcionaba la información suficiente para realizar la clasificación funcional (ver siguiente sección).

Clasificación funcional de las proteínas conservadas

Las proteínas presentes en los catálogos ancestrales se clasificaron, utilizando como referencia las categorías funcionales propuestas en la base de datos COG, las cuales se enlistan en el Anexo 2. Esto fue posible gracias a un archivo con terminación ".ptt", el cual contiene una tabla que asocia el identificador del gen con su clave COG. Por esta razón las especies elegidas como semilla para el agrupamiento de ortólogos mediante BDBH necesariamente tenían que contar con dicho archivo. La extracción de la información a partir de la tabla en texto plano se realizó con un programa desarrollado en el Laboratorio de Origen de la Vida de la Facultad de Ciencias de la UNAM.

Es importante aclarar que a partir de septiembre de 2016, el NCBI dejó de integrar el archivo *.ptt* a las carpetas de los ensamblajes de los nuevos genomas, por lo que solamente se pueden consultar en la siguiente dirección FTP, que corresponde a la base de datos antigua del NCBI: ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/.

Para complementar el análisis funcional, se asociaron los identificadores de cada gen con sus respectivas entradas en la base de datos PATHWAY y EC de KEGG (*Kyoto Encyclopedia of Genes and Genomes*, Kanehisa *et al*., 2017). La visualización de la clasificación funcional, así como también el resto de las gráficas de los resultados, se hicieron con la paquetería R (R Core Team, 2015).

Análisis de composición genómico a nivel de linaje

El genoma nuclear y el pangenoma de cada grupo se calculó también con ayuda del programa *Get-homologues*, el cual se basa en el trabajo de Tettelin y colaboradores (2005), quienes para calcularlo plantearon la idea de un experimento de muestreo aleatorio. Éste consiste en tomar un proteoma, luego otro proteoma y así sucesivamente, y en cada paso se contabilizan tanto las proteínas que se agregan al catálogo genético del grupo, como también aquellas que son comunes entre las especies. Dicho muestreo se repite varias veces cambiando el orden de los proteomas seleccionados, los cuales se eligen al azar (Tettelin *et al*., 2005). El núcleo genómico y el pangenoma se estimaron en *Get-homologues* con las funciones de extrapolación propuestas en el análisis de Tettelin *et al.* (que son esencialmente funciones de decaimiento exponencial), y finalmente se graficaron.

Árbol de referencia y distancia filogenética

Para disponer de un árbol que evidenciara las relaciones filogenéticas entre las especies de la muestra, primero se construyó el alineamiento del gen 16S rRNA. Para esto se descargó la paquetería *Arb-software* versión 6.0.3 y la base de datos no redundante SILVA SSU versión 128 (Quast *et al*., 2013), a partir de la cual se seleccionaron las secuencias de las especies de interés, y posteriormente se extrajo un archivo con el alineamiento. Para aquellas especies que no se encontraban en la base de datos, se descargó la secuencia de dicho gen directamente del NCBI (https://www.ncbi.nlm.nih.gov/) y se alineó con la herramienta de *Arb* en línea (https://www.arb-silva.de/aligner/).

El archivo del alineamiento del gen 16S rRNA de todas las especies de la muestra, se trabajó con el programa trimAL versión 1.2 (Capella-Gutierrez *et al*., 2009), para descartar regiones no informativas. Esta nueva versión del alineamiento se utilizó para calcular un árbol filogenético mediante máxima verosimilitud con la paquetería IQ-TREE (Nguyen *et al.*, 2015), bajo el modelo de evolución denominado GTR+I+G4 y utilizando la opción del *bootstrap* rápido (Minh *et al*., 2013). Para la visualización del árbol resultante se utilizó la herramienta en línea *Interactive Tree of Life* (Letunic y Bork, 2011, http://itol.embl.de/).

Posteriormente se cuantificó la distancia filogenética (PD *sensu* Faith, 1992) entre las especies al interior de cada linaje (Fórmula 1), utilizando el valor numérico de la longitud de las ramas del árbol (L_b) previamente construido. Debido a que esta métrica se obtiene sumando un valor que representa diferencia filogenética, se puede interpretar como la cantidad de historia evolutiva asociada a un linaje (Tucker *et al.*, 2017). Con ayuda del programa Excel se realizó la gráfica de los valores de distancia filogenética, y las estimaciones del núcleo genómico provenientes del análisis pangenómico.

$$\mathrm{PD} = \sum_{b \in Bt} L_b$$

Fórmula 1. Distancia filogenética. Donde b es una rama, L_b su longitud y Bt es el conjunto de ramas que conectan a las especies de un linaje en una filogenia.

Resultados

Base de datos de genomas completos

La muestra a partir de la cual se reconstruyeron los catálogos ancestrales contiene 785 proteomas, que corresponden a 143 especies de Archaea, distribuidas en los grupos Euryarchaeota, Crenarchaeota y Thaumarchaeota, así como 642 especies de Bacteria, pertenecientes a 12 linajes distintos: Actinobacteria, Firmicutes, Bacteroidetes, Cyanobacteria, Deinococcus–Thermus, Thermotogae, Chloroflexi, Aquificae, Verrucomicrobia–Planctomycetes, Spirochaetes, Acidobacteria y Proteobacteria (Figura 3). La lista de especies de la base de datos final, su clave de identificación, así como el ensamblaje a partir del cual se obtuvo el proteoma, se pueden consultar en la liga https://goo.gl/s9baKP.

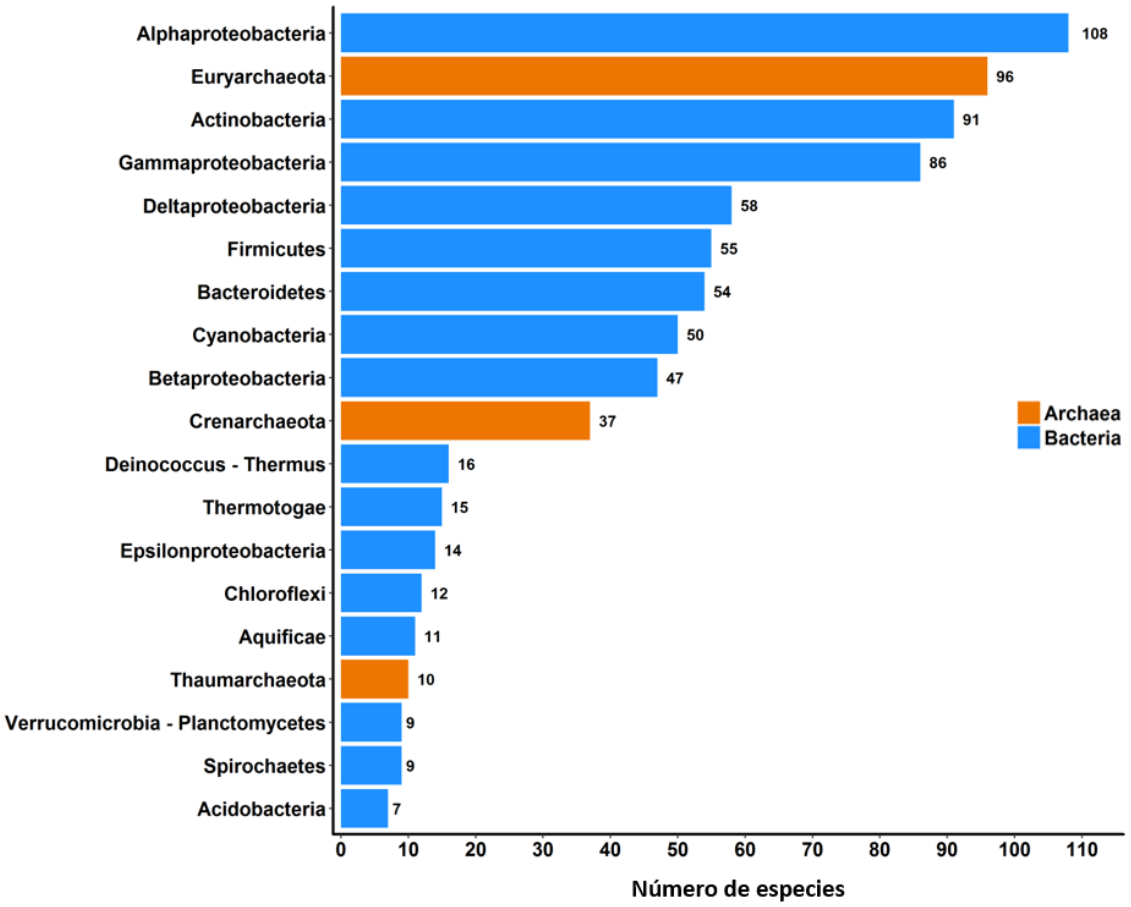

Figura 3. Grupos de Archaea y Bacteria a partir de los cuales se reconstruyeron los respectivos catálogos ancestrales. Se indica el total de especies presentes en cada linaje.

Reconstrucciones ancestrales

En total se reconstruyeron 19 catálogos ancestrales mediante la metodología descrita, pues como se observa en la Figura 3, el phylum Proteobacteria se analizó a nivel de clase. En la Tabla 2 se muestra el número de proteínas presentes en las reconstrucciones ancestrales, tanto en el núcleo estricto como en el núcleo relajado. Además, se comparan los resultados con los organismos modernos de cada linaje, en cuyo caso se calculó el promedio y la desviación estándar del número de proteínas contenidas en los genomas, así como también se identificaron aquellas especies con la menor y la mayor cantidad de proteínas. El porcentaje de reconstrucción se define como la proporción que representan el núcleo estricto y el relajado, respecto del promedio de cada grupo.

Tabla 2. Comparación entre las especies modernas y las reconstrucciones ancestrales.

Muestra		Características de los proteomas de las especies modernas (número de proteínas)				Reconstrucciones ancestrales		
Grupo	N	Promedio	Desviación estándar	Mínimo	Máximo	Núcleo estricto	Núcleo relajado	Porcentaje de reconstrucción
Archaea								
Euryarchaeota	96	2376	706	1283	4540	70	116	2.94 - 4.88
Crenarchaeota	37	1988	481	1345	2978	149	212	7.49 - 10.66
Thaumarchaeota	10	2319	707	1730	3565	385	823	16.60 - 35.48
Bacteria								
Actinobacteria	91	5253	2051	1964	10022	68	128	1.29 - 2.43
Firmicutes	55	2921	947	1217	5947	93	151	3.18 - 5.16
Bacteroidetes	54	3563	1204	1753	7192	191	269	5.36 - 7.54
Cyanobacteria	50	4191	1252	1882	6609	343	557	8.18 - 13.29
Deinococcus – Thermus	16	2597	501	1908	3636	453	702	17.44 - 27.03
Thermotogae	15	1953	200	1750	2572	489	660	25.03 - 33.79
Chloroflexi	12	3370	1187	1580	4977	233	323	6.91 - 9.58
Aquificae	11	1673	153	1497	1981	495	599	29.58 - 35.80
Verrucomicrobia – Planctomycetes	9	4092	1496	2138	7136	244	334	5.96 - 8.16
Spirochaetes	9	3236	630	2264	4219	211	287	6.52 - 8.86
Acidobacteria	7	4492	1723	2227	7826	411	556	9.14 - 12.37
Alphaproteobacteria	108	4036	949	2543	7042	143	270	3.54 - 6.68
Gammaproteobacteria	86	3622	888	1572	6202	175	276	4.83 - 7.62
Deltaproteobacteria	58	4334	1901	1677	9470	97	146	2.23 - 3.36
Betaproteobacteria	47	4153	1315	2255	8091	256	447	6.16 - 10.76
Epsilonproteobacteria	14	2323	377	1730	3138	533	639	22.94 - 27.50

N = total de proteomas en cada grupo. $Porcentaje\ de\ reconstrucción = \left(\frac{núcleo}{promedio}\right) * 100$.

Como se puede ver en la Tabla 2, el número de proteínas conservadas resultó distinto en cada linaje. Los grupos con menos proteínas en el núcleo estricto son Actinobacteria (68), Euryarchaeota (70), Firmicutes (93) y Deltaproteobacteria (97). En el otro extremo están los grupos Epsilonproteobacteria (533), Aquificae (495), Thermotogae (489) y Deinococcus - Thermus (453) con las reconstrucciones ancestrales de mayor tamaño. Evidentemente en todos los grupos el número de proteínas conservadas aumenta en el núcleo relajado, siendo Thaumarchaeota el grupo más destacado, pues se agregaron 438 proteínas al relajar el criterio de inclusión en el catálogo ancestral (Figura 4).

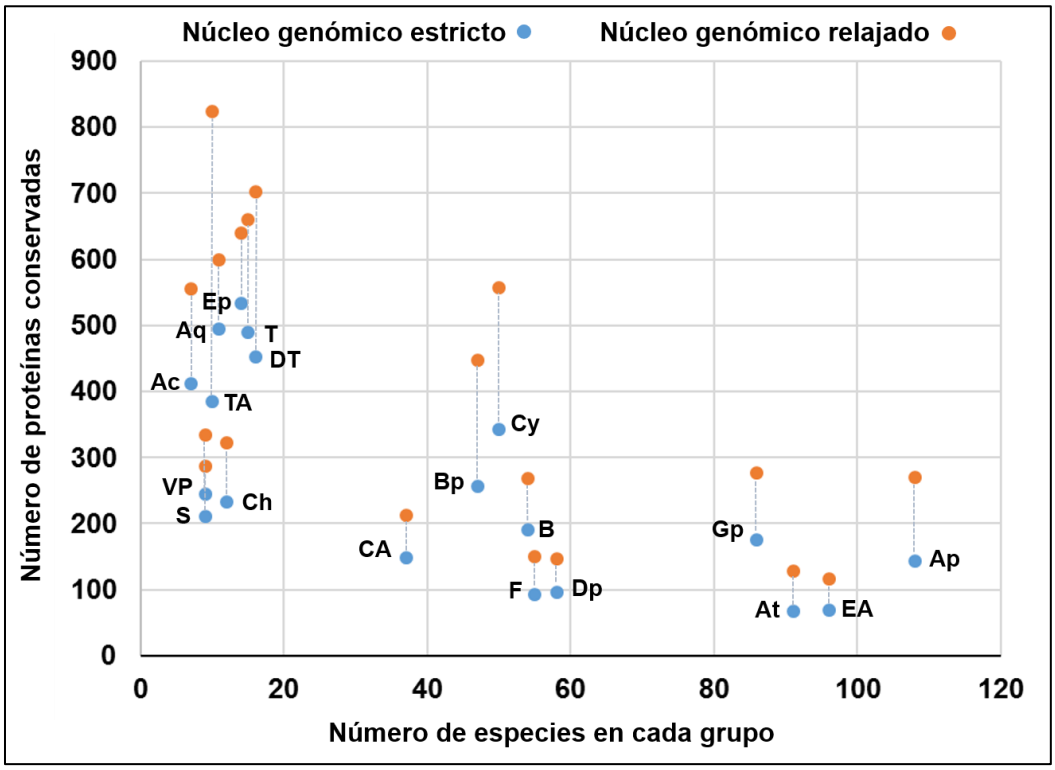

Figura 4. Tamaño de los catálogos ancestrales en términos del número total de proteínas conservadas presentes en cada uno. Grupos de Bacteria: Acidobacteria (Ac), Spirochaetes (S), Verrucomicrobia-Planctomycetes (VP), Aquificae (Aq), Chloroflexi (Ch), Epsilonproteobacteria (Ep), Thermotogae (T), Deinococcus-Thermus (DT), Betaproteobacteria (Bp), Cyanobacteria (Cy), Bacteroidetes (B), Firmicutes (F), Deltaproteobacteria (Dp), Gammaproteobacteria (Gp), Actinobacteria (At), Alphaproteobacteria (Ap). Linajes de Archaea: Thaumarchaeota (TA), Crenarchaeota (CA) y Euryarchaeota (EA).

El porcentaje de reconstrucción es una medida de qué tanto se reconstruye del genoma ancestral, tomando como referencia el promedio del número de proteínas en los organismos modernos (Tabla 2). Los resultados muestran que las reconstrucciones no superan el 40%, en el mejor de los casos se reconstruye hasta un 35%, como en los grupos Thaumarchaeota, Aquificae y Thermotogae, número que disminuye si se considera solamente al núcleo genómico estricto (Figura 5).

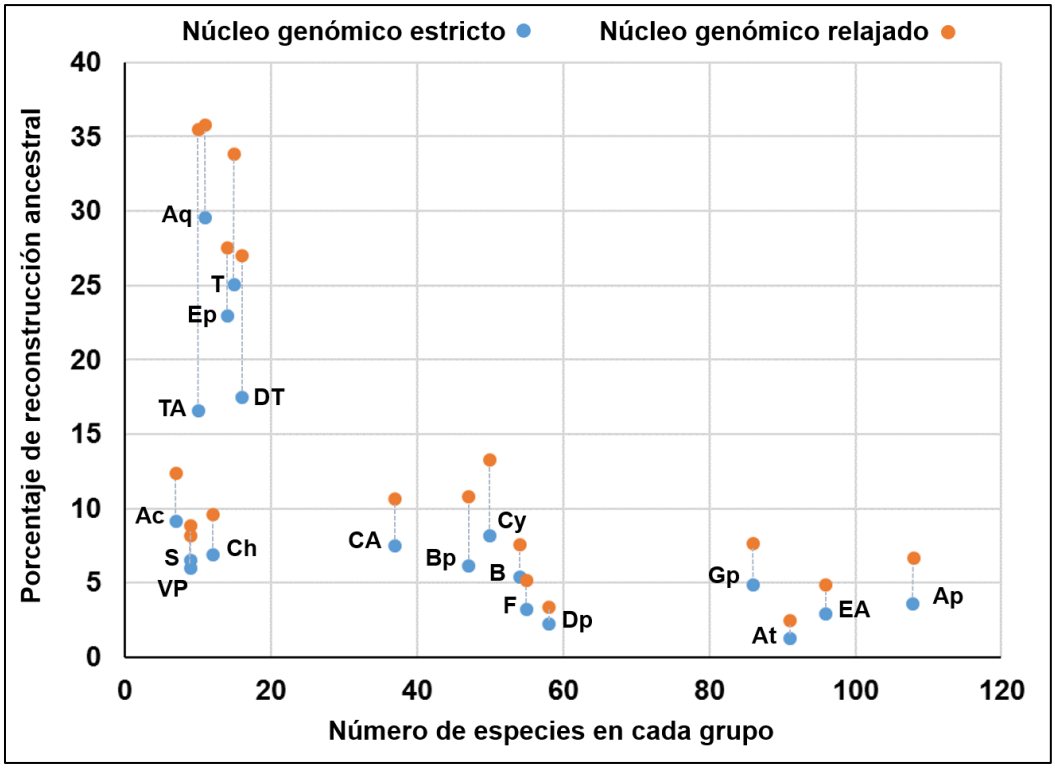

Figura 5. Tamaño de los catálogos ancestrales en términos del porcentaje de reconstrucción ancestral. Los grupos Aquificae y Thermotogae tienen el mayor porcentaje, no así las reconstrucciones con mayor número de proteínas (ver figura anterior). La mayoría de los grupos no supera el 10% de reconstrucción para el núcleo genómico estricto. Las abreviaturas de los grupos son las mismas que en la Figura 4.

En general, los grupos para los que se reconstruyó un menor porcentaje del genoma ancestral son los que presentan mayor cantidad de especies al interior de la muestra, como Euryarchaeota, Actinobacteria, Alphaproteobacteria y Gammaproteobacteria. Sin embargo,

el hecho de que haya otros grupos con reconstrucciones pequeñas como Firmicutes, Deltaproteobacteria y Crenarchaeota, sugiere que hay otros factores que influyen en el tamaño de la reconstrucción ancestral.

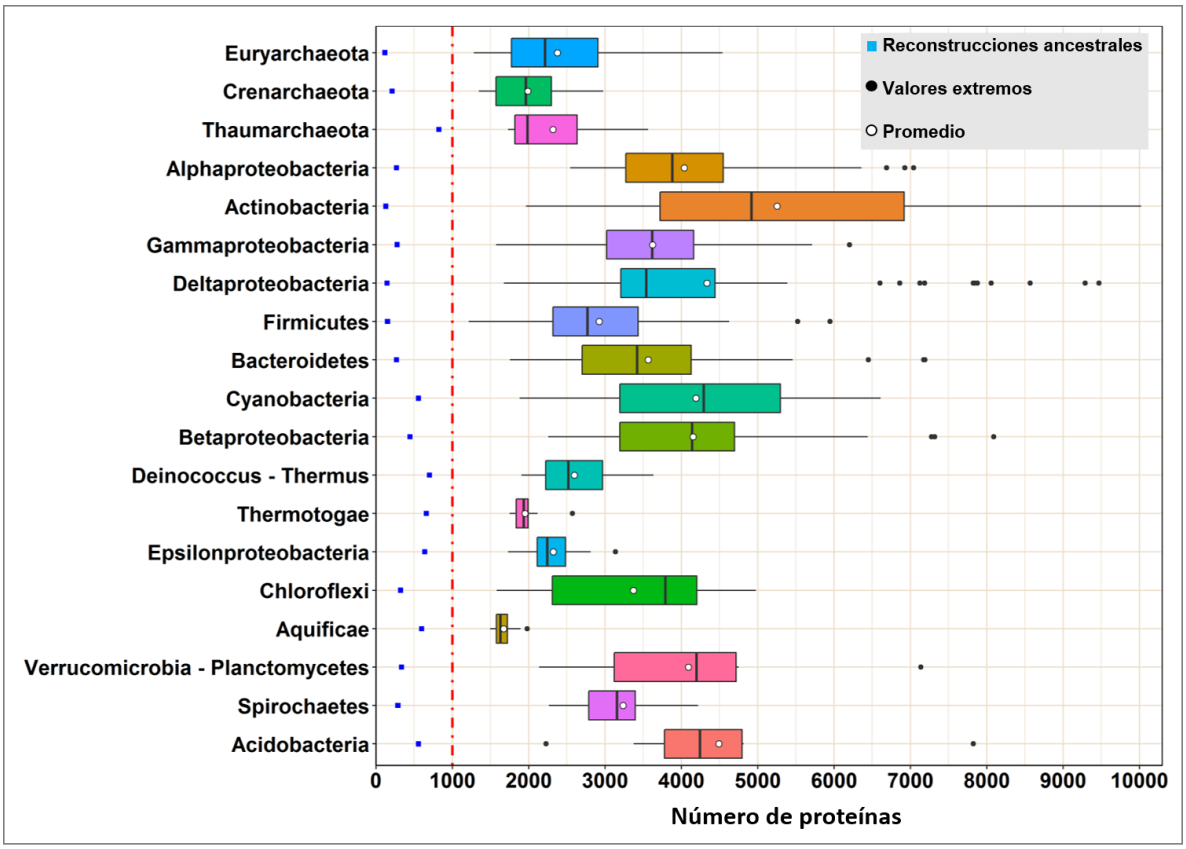

Figura 6. Número de proteínas contenidas en los genomas modernos de los distintos linajes, comparado con el tamaño de las reconstrucciones ancestrales, las cuales no superan el límite de 900 proteínas (línea roja en 1000 proteínas).

Considerando que las especies en la base de datos son de vida libre, es interesante notar que el genoma de menor tamaño es un firmicute con 1217 proteínas (*Lactobacillus sanfranciscensis* TMW 1.1304), seguido por un representante de Euryarchaeota con 1283 proteínas (*Methanothermus fervidus* DSM 2088) y una especie del grupo Crenarchaeota con 1345 proteínas (*Desulfurococcus mucosus* DSM 2162). En el otro extremo sobresale una actinobacteria que posee 10,022 proteínas (*Streptomyces bingchenggensis* BCW-1), seguida por una deltaproteobacteria que presenta 9470 proteínas (*Archangium gephyra*).

El promedio del número de proteínas en los genomas varía desde 1673 en Aquificae, hasta 5253 en Actinobacteria, que es el grupo que presenta la mayor variación en el tamaño del genoma, como se puede ver en la desviación estándar (DE) de 2051 proteínas. Le sigue el grupo de Deltaproteobacteria con una DE de 1901 proteínas. Es notable que los grupos con menor DE en lo que se refiere al número de proteínas en el genoma son Aquificae con 153 proteínas, y Thermotogae con 200 proteínas, así como Crenarchaeota con 481 proteínas (Figura 6).

Clasificación funcional en las categorías de COG

Al interior de los 19 catálogos ancestrales hubo un total de 7495 conjuntos de proteínas conservadas, cuya distribución en las clases funcionales más generales de la base de datos COG se observa en la Figura 7. La categoría mejor representada fue la de metabolismo, seguida de la de almacenamiento y procesado de la información genética. Los linajes que, en proporción, más recuperaron de la categoría de metabolismo fueron Cyanobacteria y Verrucomicrobia-Planctomycetes. Por el contrario, al interior de Firmicutes y Euryarchaeota, la categoría de metabolismo estuvo reducida, mientras que la de almacenamiento y procesado de la información genética fue la más abundante (Figura 7).

La categoría de pobremente caracterizado presentó un número importante de proteínas que no cuentan con una asignación COG (414), en particular en el grupo Epsilonproteobacteria (140) y en Thaumarchaeota (98) en su núcleo relajado. También incluye a las dos categorías funcionales (R y S) de proteínas que, si bien tienen una asignación COG, no están caracterizadas de manera satisfactoria. La suma de las proteínas no clasificadas (414), más las que presentan una predicción general de su función (595), más las proteínas de función desconocida (336), da un total de 1345 proteínas pobremente caracterizadas (colores rojo y anaranjado de la Figura 7), lo que representa 17.94% de las proteínas conservadas al interior de los catálogos ancestrales.

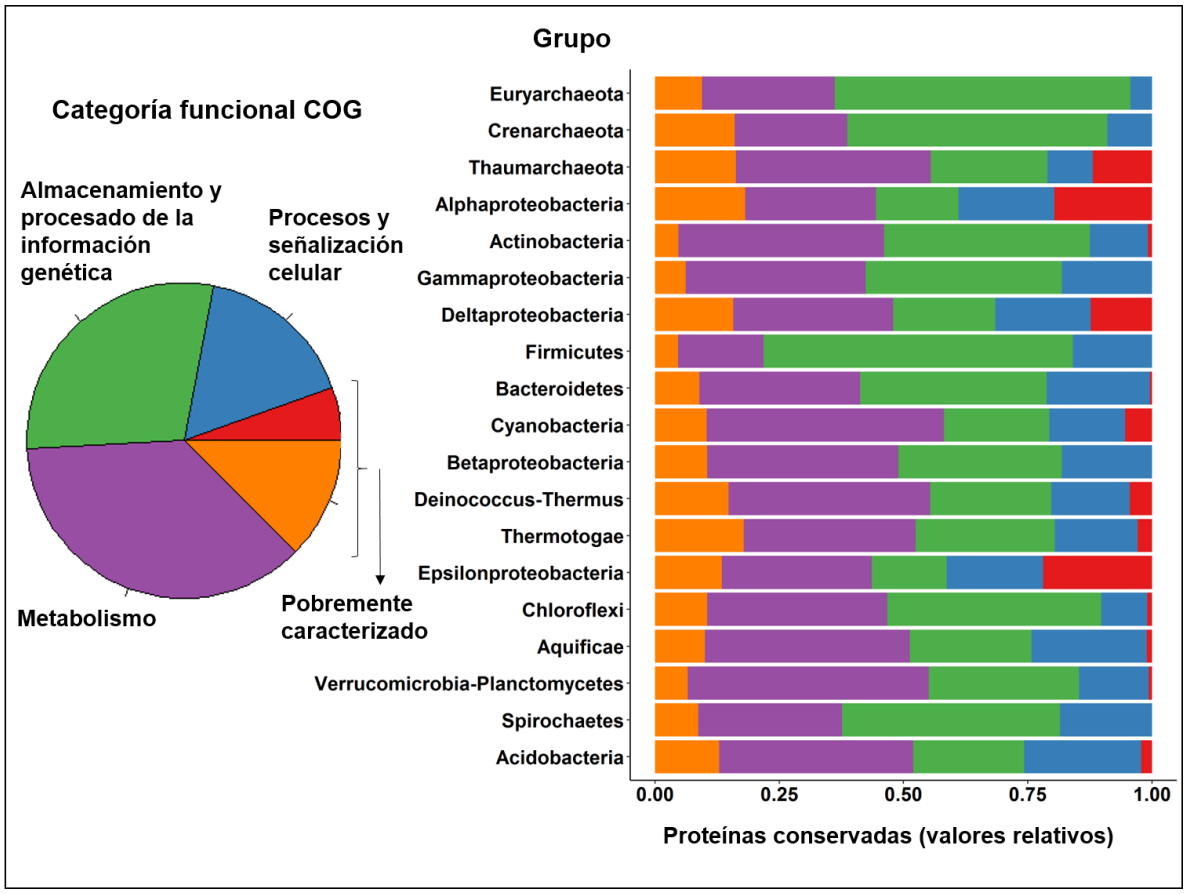

Figura 7. Clasificación de las proteínas conservadas en las categorías funcionales generales de COG. Se muestra la distribución del total (izquierda), así como al interior de las reconstrucciones ancestrales de cada linaje (derecha). El color rojo representa proteínas sin asignación COG y el naranja representa proteínas cuya asignación funcional es mala.

La categoría de proteínas sin asignación COG (-), es resultado de la variación en la calidad de la anotación de los genomas, en particular de aquellos que fungieron como el genoma semilla. Las reconstrucciones con una cantidad importante de proteínas que no se pudieron clasificar fueron las de Epsilonproteobacteria, Thaumarchaeota, Alphaproteobacteria, Deinococcus–Thermus, Cyanobacteria, Deltaproteobacteria y Thermotogae, principalmente (color rojo en Figura 7). En el otro extremo, hubo grupos en los que sí se contaba con una asignación COG para la totalidad de las secuencias del proteoma, como en Euryarchaeota, Crenarchaeota, Gammaproteobacteria, Firmicutes,

Bacteroidetes, Betaproteobacteria y Spirochaetes, para los cuales todas las proteínas se pudieron clasificar en una categoría funcional COG.

La Figura 8 muestra con un mayor detalle la distribución de las proteínas conservadas al interior de las categorías funcionales. Los valores se normalizaron por linaje de tal forma que resaltan las casillas con una conservación mayor y/o menor al promedio de cada grupo (ver el mapa de color sin normalizar en el Anexo 3). Tanto en el núcleo estricto como en el relajado, se observa que la conservación al interior de los catálogos ancestrales no es homogénea, en el sentido de que las proteínas se presentan mayoritariamente en unas pocas categorías funcionales. De hecho, si bien la Figura 7 muestra que a nivel general la categoría con más proteínas fue la de metabolismo, en realidad al interior de cada reconstrucción siempre se recupera más de la categoría de traducción, biogénesis y estructura del ribosoma (J), la cual presentó 1437 conjuntos de proteínas conservadas, lo que representó un 19% del total. Al interior de dicha categoría se detectaron una serie de proteínas que se sabe están muy conservadas, como es el caso de las proteínas estructurales del ribosoma, tanto de la subunidad 30S (Figura anexo 4) como de la 50S (Figura anexo 5), así como las familias de las aminoacil-tRNA sintetasas (Figura anexo 5).

Después de la categoría de traducción, las categorías funcionales restantes se pueden agrupar en dos casos: aquellas para las que no se reconstruye prácticamente nada (V, Q, B, A, U, D, P, T, N, en Figura 8), y aquellas con niveles medios de conservación (E, H, R, M, L, C, S, G, K, I, O, F en Figura 8), entre las que destaca el metabolismo y transporte de aminoácidos (E), el metabolismo y transporte de coenzimas (H), la biogénesis de la membrana, pared o envoltura celular (M), la conversión y producción de energía (C), así como la categoría de predicción general de la función (R). Si se analiza el comportamiento del núcleo relajado, se agregan a las categorías ya mencionadas las de metabolismo de nucleótidos (F), de lípidos (I), de carbohidratos (G), la transcripción (K), la modificación postraduccional de proteínas (O), y la de función desconocida (S), todas con niveles intermedios de conservación. Una descripción de las proteínas presentes en cada categoría funcional, así como la base de datos de los resultados se pueden consultar en las siguientes ligas https://goo.gl/XmkxDs y https://goo.gl/sUHPQs, respectivamente.

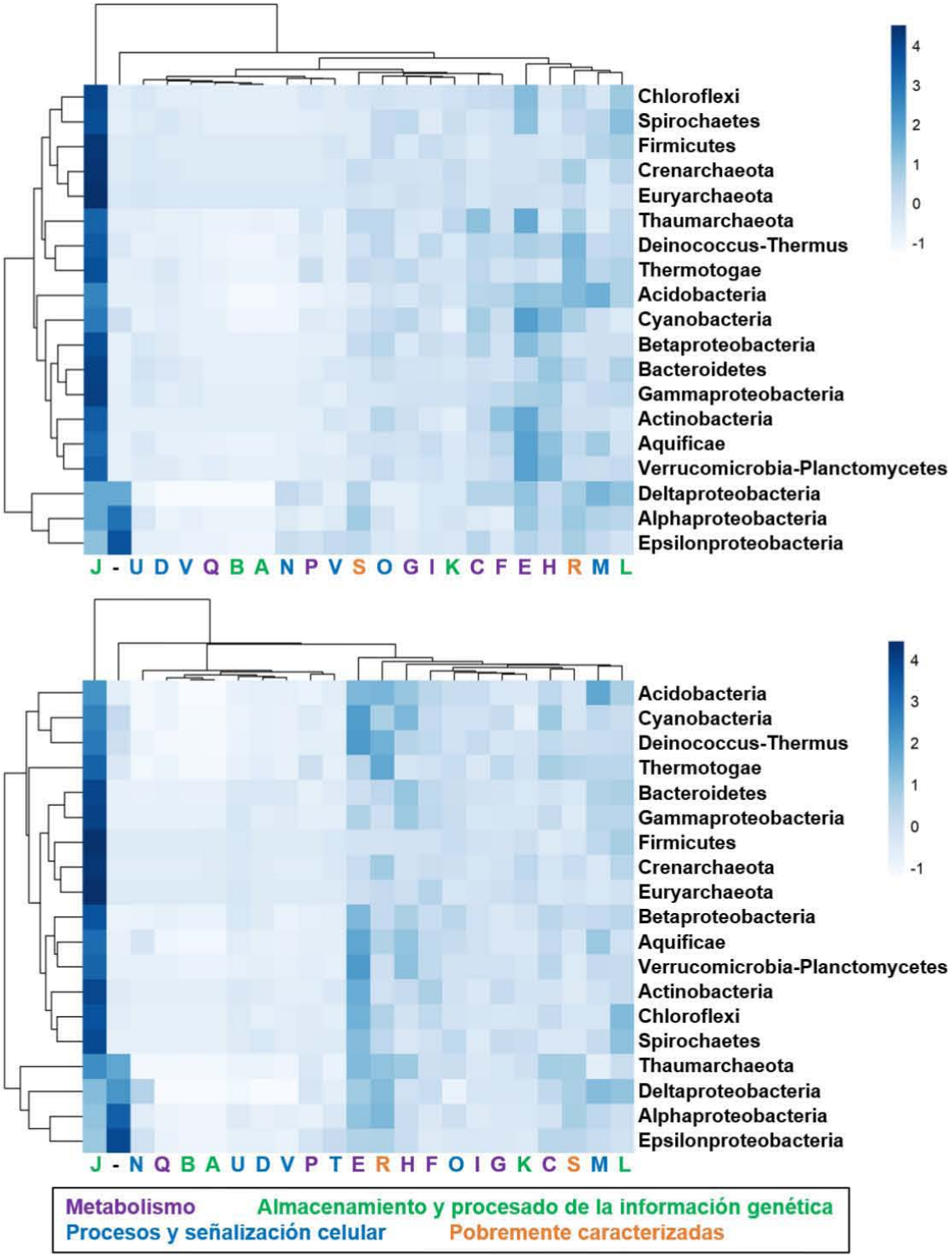

Figura 8. Clasificación funcional de las proteínas presentes en los catálogos ancestrales. Los valores están normalizados por linaje. Se muestra el núcleo estricto (arriba) y el núcleo relajado (abajo). Las categorías son: [J] Traducción, biogénesis y estructura del ribosoma, [-] Proteínas sin asignación COG, [P] Metabolismo y transporte de iones inorgánicos, [Q] Biosíntesis, transporte y catabolismo

de metabolitos secundarios, [B] Estructura y dinámica de la cromatina, [A] Modificación y procesamiento del RNA, [U] Tráfico intracelular, transporte de vesículas y secreción, [D] Control del ciclo celular, división celular y segregación cromosómica, [V] Mecanismos de defensa, [T] Mecanismos de transducción de señales, [N] Movilidad celular, [E] Metabolismo y transporte de aminoácidos, [R] Predicción general de la función, [H] Metabolismo y transporte de coenzimas, [M] Biogénesis de la membrana, pared o envoltura celular, [C] Conversión y producción de energía, [S] Función desconocida, [L] Replicación, recombinación y reparación, [G] Metabolismo y transporte de carbohidratos, [K] Transcripción, [I] Metabolismo y transporte de lípidos, [F] Metabolismo y transporte de nucleótidos, [O] Modificación postraduccional, chaperonas. Los linajes y las categorías se agruparon con un método jerárquico basado en la similitud promedio no ponderada (*WPGMA* por sus siglas en inglés).

Cabe mencionar que la mayoría de las proteínas sin asignación COG, así como aquellas contenidas en las categorías de pobremente caracterizado (R y S), corresponden a proteínas anotadas como hipotéticas (las cuales son proteínas o marcos abiertos de lectura identificados de manera automática, que carecen de evidencia *in vivo* de su función). Sin embargo, una fracción sí está identificada y es interesante que se pueden asociar a funciones como traducción, modificación postraduccional de proteínas, replicación y reparación, señalización celular, procesos de transporte a través de la membrana, metabolismo de carbohidratos, biosíntesis de lípidos y pared celular, e incluso a cuestiones energéticas como proteínas asociadas a la fotosíntesis o a la transferencia de electrones. Es decir, si clasificáramos dichas proteínas, se incorporarían a la categoría de traducción, así como a aquellas con niveles intermedios de conservación.

Otra forma de representar los datos al interior de los catálogos es normalizar las categorías funcionales, de tal manera que resalten las casillas donde la conservación es mayor y/o menor que el promedio de cada categoría. En la Figura 9 se observa que las reconstrucciones de Aquificae, Thermotogae, Deinococcus-Thermus y Acidobacteria sobresalen del resto, pues presentan más proteínas conservadas en la mayoría de las categorías funcionales. Dichos catálogos junto con los de Epsilonproteobacteria,

Thaumarchaeota y Cyanobacteria fueron los de mayor tamaño, todos con más de 300 proteínas conservadas (ver Figura 4).

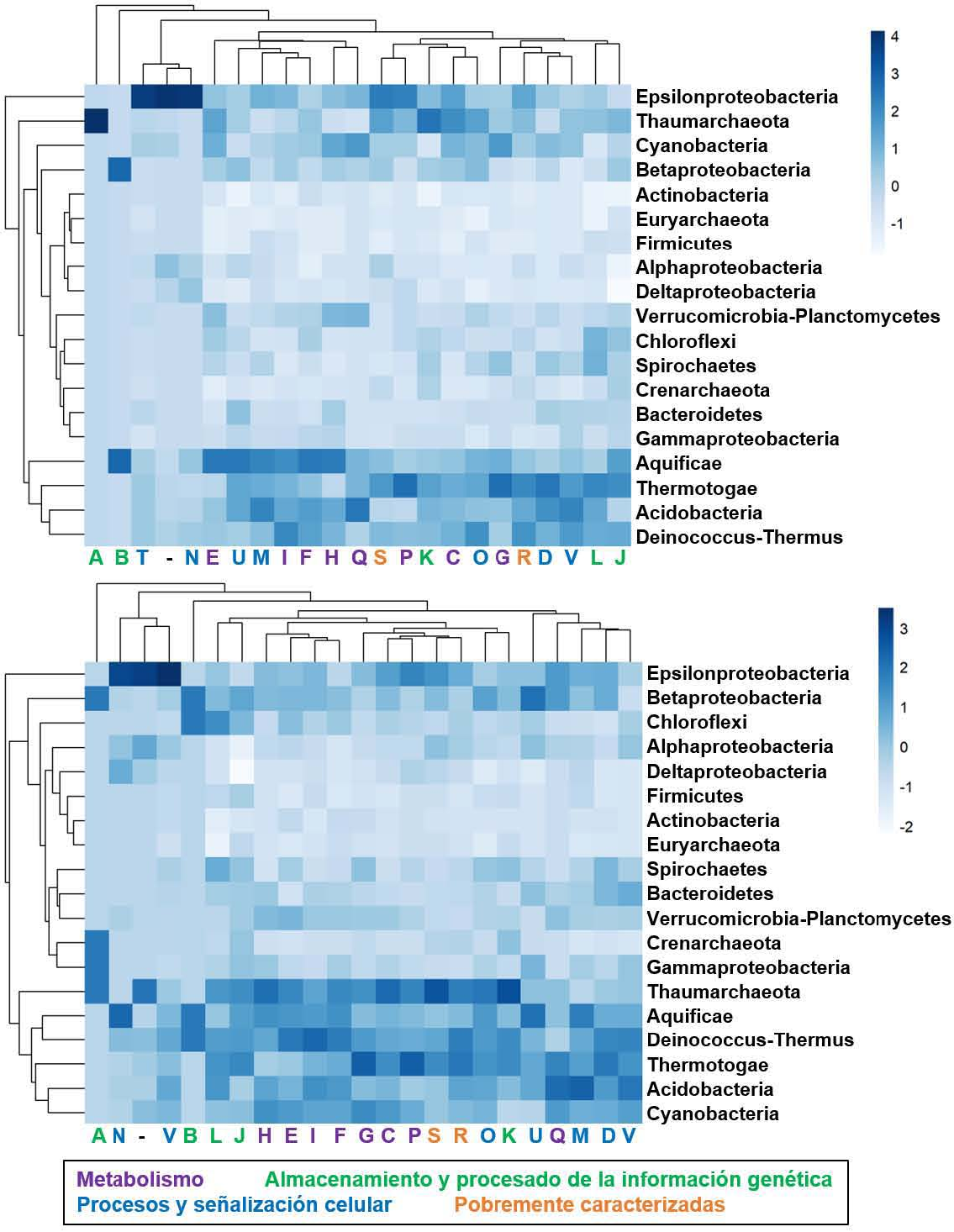

Figura 9. Clasificación funcional de las proteínas presentes en los catálogos ancestrales. Los valores están normalizados por categorías funcionales. Se muestra el núcleo estricto (arriba) y el núcleo

relajado (abajo). Las categorías son: [J] Traducción, biogénesis y estructura del ribosoma, [-] Proteínas sin asignación COG, [P] Metabolismo y transporte de iones inorgánicos, [Q] Biosíntesis, transporte y catabolismo de metabolitos secundarios, [B] Estructura y dinámica de la cromatina, [A] Modificación y procesamiento del RNA, [U] Tráfico intracelular, transporte de vesículas y secreción, [D] Control del ciclo celular, división celular y segregación cromosómica, [V] Mecanismos de defensa, [T] Mecanismos de transducción de señales, [N] Movilidad celular, [E] Metabolismo y transporte de aminoácidos, [R] Predicción general de la función, [H] Metabolismo y transporte de coenzimas, [M] Biogénesis de la membrana, pared o envoltura celular, [C] Conversión y producción de energía, [S] Función desconocida, [L] Replicación, recombinación y reparación, [G] Metabolismo y transporte de carbohidratos, [K] Transcripción, [I] Metabolismo y transporte de lípidos, [F] Metabolismo y transporte de nucleótidos, [O] Modificación postraduccional, chaperonas. Los linajes y las categorías se agruparon jerárquicamente con el método de similitud promedio no ponderada.

Una tendencia que se observa en la Figura 9 es que conforme los catálogos se reducen, la conservación se va perdiendo en todas las funciones, como se ve para las reconstrucciones menores a 200 proteínas; Euryarchaeota, Actinobacteria, Firmicutes, Deltaproteobacteria, Alphaproteobacteria y Crenarchaeota, las cuales presentan muy pocas proteínas en la mayoría de las categorías funcionales, incluyendo aquellas relacionadas con el almacenamiento y procesado de la información genética.

Clasificación en las categorías funcionales particulares de KEGG

Las proteínas conservadas al interior de los distintos catálogos ancestrales se asociaron a su correspondiente categoría funcional de la base de datos PATHWAY de KEGG. En la Figura 10 se muestra un mapa de calor con la distribución resultante de 4345 proteínas, las cuales representan 58% del total. El resto no se incluyeron pues presentan una menor calidad en la anotación y, por lo tanto, en la información disponible, lo cual no quiere decir necesariamente que no pertenezcan a alguna de las categorías funcionales.

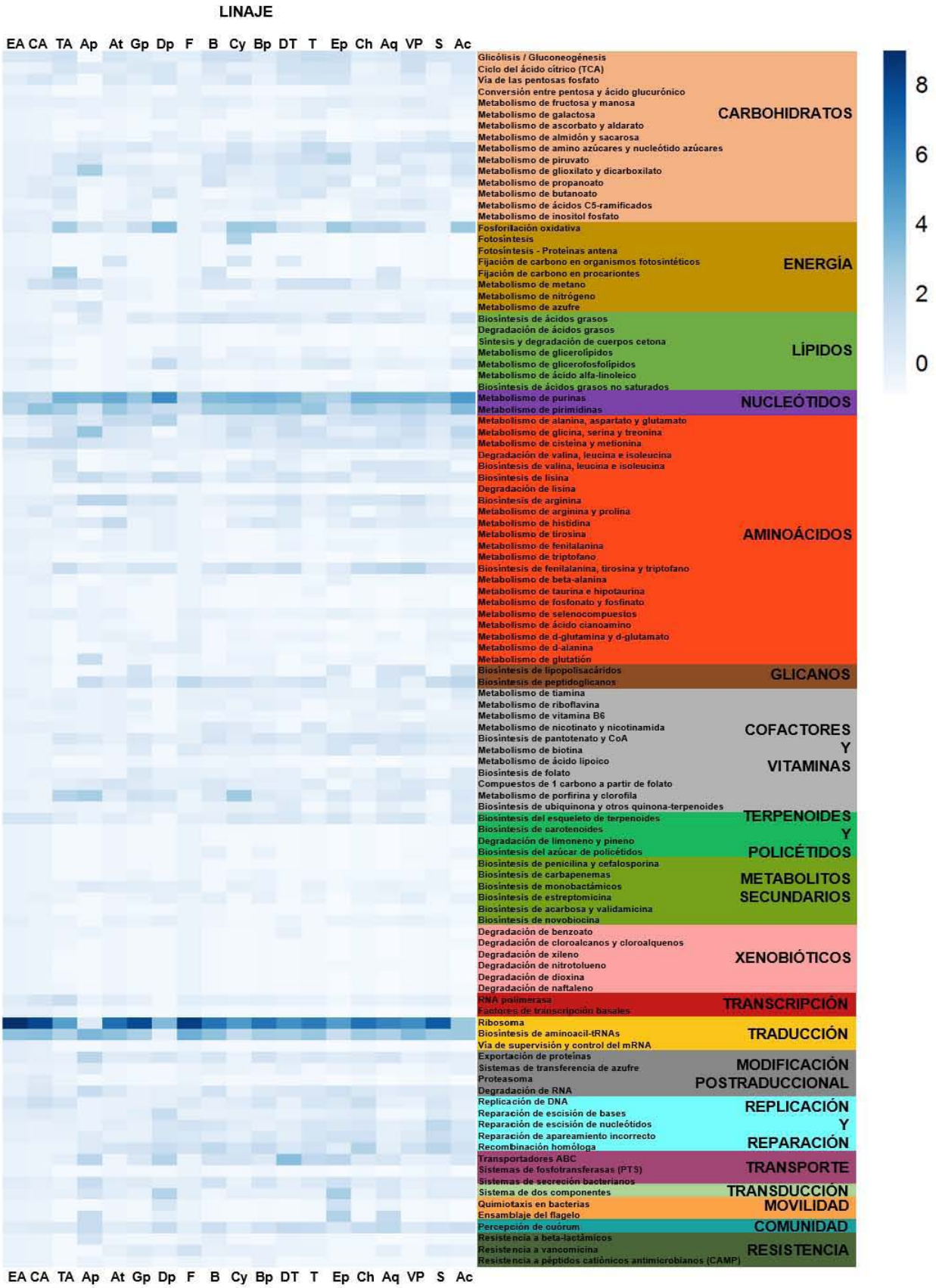
LINAJE
EA CA TA Ap At Gp Dp F B Cy Bp DT T Ep Ch Aq VP S Ac
Glicólisis / Gluconeogénesis
Ciclo del ácido cítrico (TCA)
Vía de las pentosas fosfato
Conversión entre pentosa y ácido glucurónico
Metabolismo de fructosa y manosa
Metabolismo de galactosa
Metabolismo de ascorbato y aldarato
Metabolismo de almidón y sacarosa
Metabolismo de amino azúcares y nucleótido azúcares
Metabolismo de piruvato
Metabolismo de glioxilato y dicarboxilato
Metabolismo de propanoato
Metabolismo de butanoato
Metabolismo de ácidos C5-ramificados
Metabolismo de inositol fosfato
CARBOHIDRATOS
Fosforilación oxidativa
Fotosíntesis
Fotosíntesis - Proteínas antena
Fijación de carbono en organismos fotosintéticos
Fijación de carbono en procariontes
Metabolismo de metano
Metabolismo de nitrógeno
Metabolismo de azufre
ENERGÍA
Biosíntesis de ácidos grasos
Degradación de ácidos grasos
Síntesis y degradación de cuerpos cetona
Metabolismo de glicerolípidos
Metabolismo de glicerofosfolípidos
Metabolismo de ácido alfa-linoleico
Biosíntesis de ácidos grasos no saturados
LÍPIDOS
Metabolismo de purinas
Metabolismo de pirimidinas
NUCLEÓTIDOS
Metabolismo de alanina, aspartato y glutamato
Metabolismo de glicina, serina y treonina
Metabolismo de cisteína y metionina
Degradación de valina, leucina e isoleucina
Biosíntesis de valina, leucina e isoleucina
Biosíntesis de lisina
Degradación de lisina
Biosíntesis de arginina
Metabolismo de arginina y prolina
Metabolismo de histidina
Metabolismo de tirosina
Metabolismo de fenilalanina
Metabolismo de triptofano
Biosíntesis de fenilalanina, tirosina y triptofano
Metabolismo de beta-alanina
Metabolismo de taurina e hipotaurina
Metabolismo de fosfonato y fosfinato
Metabolismo de selenocompuestos
Metabolismo de ácido cianoamino
Metabolismo de d-glutamina y d-glutamato
Metabolismo de d-alanina
Metabolismo de glutatión
AMINOÁCIDOS
Biosíntesis de lipopolisacáridos
Biosíntesis de peptidoglicanos
GLICANOS
Metabolismo de tiamina
Metabolismo de riboflavina
Metabolismo de vitamina B6
Metabolismo de nicotinato y nicotinamida
Biosíntesis de pantotenato y CoA
Metabolismo de biotina
Metabolismo de ácido lipoico
Biosíntesis de folato
Compuestos de 1 carbono a partir de folato
Metabolismo de porfirina y clorofila
Biosíntesis de ubiquinona y otros quinona-terpenoides
COFACTORES Y VITAMINAS
Biosíntesis del esqueleto de terpenoides
Biosíntesis de carotenoides
Degradación de limoneno y pineno
Biosíntesis del azúcar de policétidos
TERPENOIDES Y POLICÉTIDOS
Biosíntesis de penicilina y cefalosporina
Biosíntesis de carbapenemas
Biosíntesis de monobactámicos
Biosíntesis de estreptomicina
Biosíntesis de acarbosa y validamicina
Biosíntesis de novobiocina
METABOLITOS SECUNDARIOS
Degradación de benzoato
Degradación de cloroalcanos y cloroalquenos
Degradación de xileno
Degradación de nitrotolueno
Degradación de dioxina
Degradación de naftaleno
XENOBIÓTICOS
RNA polimerasa
Factores de transcripción basales
TRANSCRIPCIÓN
Ribosoma
Biosíntesis de aminoacil-tRNAs
Vía de supervisión y control del mRNA
TRADUCCIÓN
Exportación de proteínas
Sistemas de transferencia de azufre
Proteasoma
Degradación de RNA
MODIFICACIÓN POSTRADUCCIONAL
Replicación de DNA
Reparación de escisión de bases
Reparación de escisión de nucleótidos
Reparación de apareamiento incorrecto
Recombinación homóloga
REPLICACIÓN Y REPARACIÓN
Transportadores ABC
Sistemas de fosfotransferasas (PTS)
Sistemas de secreción bacterianos
TRANSPORTE
Sistema de dos componentes
TRANSDUCCIÓN
Quimiotaxis en bacterias
Ensamblaje del flagelo
MOVILIDAD
Percepción de cuórum
COMUNIDAD
Resistencia a beta-lactámicos
Resistencia a vancomicina
Resistencia a péptidos catiónicos antimicrobianos (CAMP)
RESISTENCIA
EA CA TA Ap At Gp Dp F B Cy Bp DT T Ep Ch Aq VP S Ac
8
6
4
2
0

Figura 10. Clasificación funcional basada en las categorías de PATHWAY, KEGG. Se muestran las vías metabólicas presentes en las distintas reconstrucciones ancestrales. La escala representa el número de proteínas presentes en cada vía metabólica, normalizado por linaje. Linajes de Archaea: Euryarchaeota (EA), Crenarchaeota (CA), Thaumarchaeota (TA). Linajes de Bacteria: Alphaproteobacteria (Ap), Actinobacteria (At), Gammaproteobacteria (Gp), Deltaproteobacteria (Dp), Firmicutes (F), Bacteroidetes (B), Cyanobacteria (Cy), Betaproteobacteria (Bp), Deinococcus-Thermus (DT), Thermotogae (T), Epsilonproteobacteria (Ep), Chloroflexi (Ch), Aquificae (Aq), Verrucomicrobia - Planctomycetes (VP), Spirochaetes (S), Acidobacteria (Ac).

De la Figura 10 se puede extraer información que con la clasificación COG no fue tan evidente. Por ejemplo, el hecho de que el metabolismo de nucleótidos tiene niveles muy importantes de conservación, casi análogos a los observados en la categoría de traducción. Otras funciones con niveles de conservación importantes al interior de los catálogos ancestrales fueron: la biosíntesis de diversos aminoácidos, la replicación y reparación del DNA y del RNA, la biosíntesis de ácidos grasos, los sistemas ABC de transporte y lo relacionado con exportar las proteínas a las membranas, además de la percepción de quórum. En lo que respecta a los linajes, las reconstrucciones ancestrales de menor tamaño fueron Firmicutes, Deltaproteobacteria, Actinobacteria y Euryarchaeota, y son catálogos muy incompletos en los que se recuperan muy pocas funciones, particularmente traducción y metabolismo de nucleótidos. En algunos grupos, como Deltaproteobacteria, sobresale la fosforilación oxidativa, así como la fotosíntesis en Cyanobacteria, y la biosíntesis de peptidoglicanos que es exclusiva de grupos bacterianos. En los grupos con reconstrucciones más numerosas como Epsilonproteobacteria, Aquificae, Thermotogae y Deinococcus-Termus, se agregan a los catálogos ancestrales proteínas relacionadas con diversas funciones metabólicas.

Si bien muchas de las proteínas tuvieron solamente una asignación funcional, lo común es que una misma proteína se asocie a diversas rutas metabólicas o funciones (ver Figura 11). El extremo fueron dos proteínas relacionadas hasta con 18 categorías distintas, como fue el caso de la enzima propanoil-CoA C-aciltransferasa, exclusiva del dominio Archaea (grupos Thaumarchaeota y Crenarchaeota), que es una tiolasa del catabolismo de

ácidos grasos, encargada de catalizar la formación del intermediario acetil-CoA y el ácido cólico. La base de datos indica que también participa en la degradación de otros compuestos como butanoato, propanoato, piruvato, glioxilato, dicarboxilato, benzoato, valina, isoleucina, leucina, lisina, así como en la fijación de carbono.

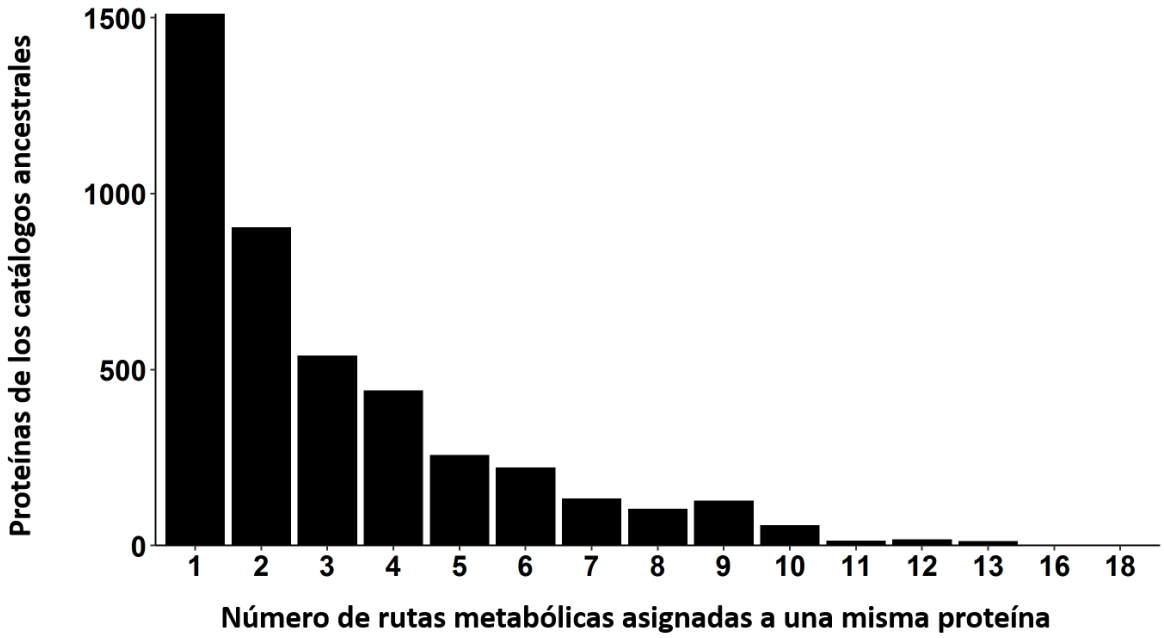

Figura 11. Redundancia en la anotación funcional en la base de datos PATHWAY de KEGG, para las proteínas presentes en los catálogos ancestrales.

La otra proteína con alta redundancia funcional fue NAD^+ aldehído deshidrogenasa, una oxidorreductasa presente en una amplia gama de funciones como la vía de la glucólisis / gluconeogénesis, la degradación de ácidos grasos, de valina, de leucina, de isoleucina, de lisina, el metabolismo de ascorbato y aldarato, de arginina, de prolina, de histidina, de triptófano, de beta-alanina, de glicerolípidos y de piruvato, además de otras funciones en apariencia más modernas y eucariontes (como por ejemplo la biosíntesis de hormonas en insectos). Las familias de enzimas acetil-CoA acetiltransferasa y las aminotransferasas de clase I y II, son otros ejemplos de proteínas conservadas y con una redundancia funcional muy elevada.

Clasificación de las enzimas en los catálogos ancestrales

Para cada grupo de secuencias conservadas, el identificador del genoma semilla se asoció a la base de datos KEGG ENZYME, lo que permitió asignar un número EC a 3629 proteínas, es decir que al menos un 48% del total del catálogo ancestral presenta una actividad catalítica. Como se observa en la Figura 12, en lo que se refiere al número total de proteínas conservadas, la categoría de transferasas (EC2) es la más abundante, seguida por las categorías de ligasas (EC6), hidrolasas (EC3) y oxidorreductasas (EC1).

Al normalizar los valores por linaje, fue todavía más evidente que las transferasas son las enzimas mejor conservadas en todos los catálogos ancestrales, y en conjunto participan en 16 categorías funcionales COG. El caso de las polimerasas es muy interesante porque en todos los catálogos hubo al menos una polimerasa conservada, ya sea de RNA, de DNA (2.7.7.6, 2.7.7.7) o ambas.

En el grupo de las ligasas (EC6) sobresalieron las aminoacil-tRNA sintetasas, pero también muchas otras enzimas que participan en el metabolismo de coenzimas, de nucleótidos, de lípidos, de aminoácidos, en la síntesis de la pared celular y en la replicación. La categoría de las hidrolasas (EC3) también presentó niveles de conservación importante en los catálogos, con proteínas como peptidil-tRNA hidrolasas, nucleasas y proteasas, enzimas que participan en la síntesis de la pared celular, en el metabolismo de carbohidratos, subunidades de la F0F1 ATP sintetasa, helicasas, la partícula de reconocimiento señal (srp54) y otras enzimas de la biosíntesis de nucleótidos y de coenzimas. De hecho, a pesar de que los grupos de oxidorreductasas, liasas e isomerasas tienen una menor conservación, contienen enzimas muy importantes, las cuales se distribuyen en todas las categorías funcionales de COG.

Ningún número enzimático se presentó en la totalidad de los catálogos ancestrales, mas algunos mostraron niveles de conservación semejantes a las proteínas ribosómicas, casos que se muestran en la Tabla 3.

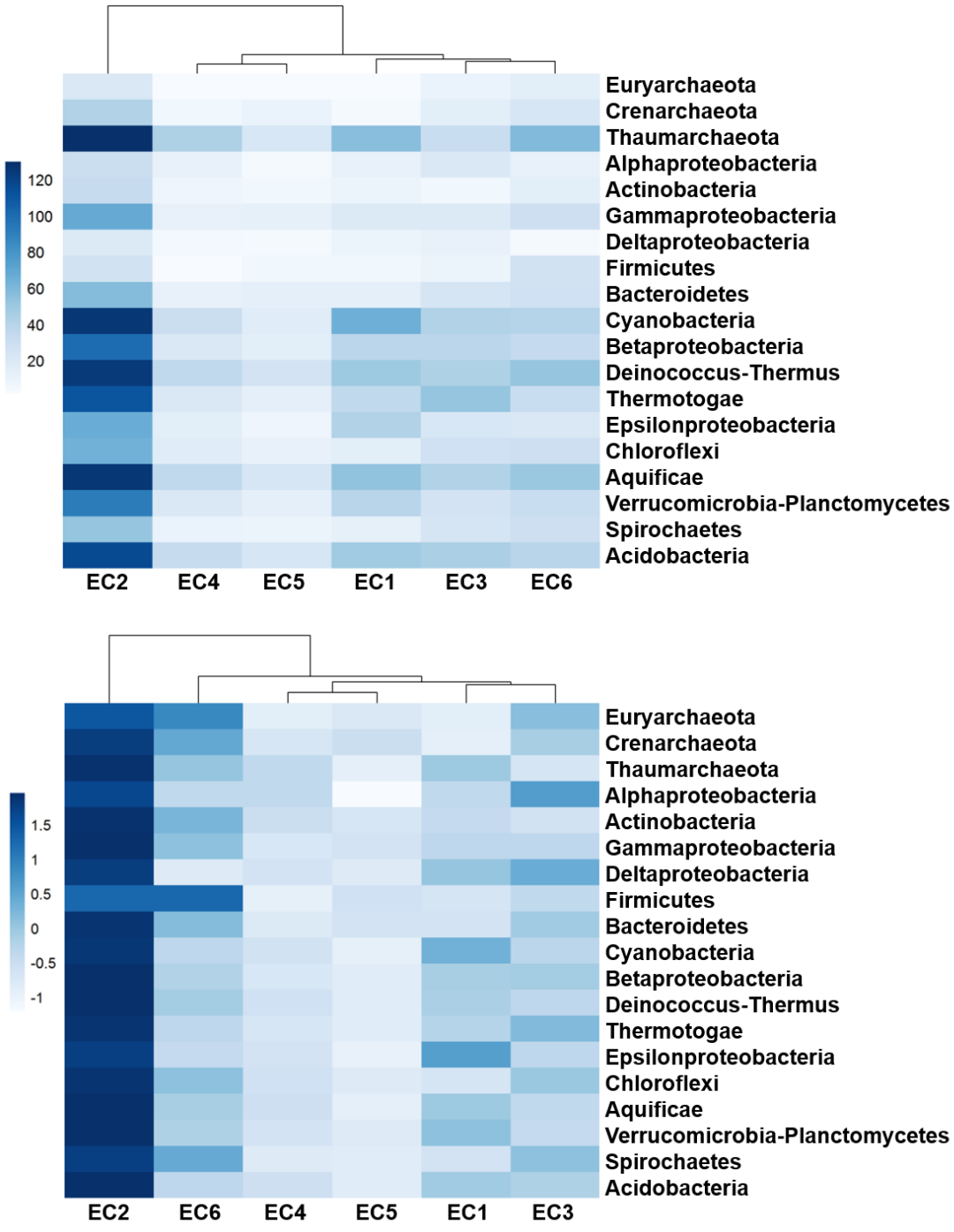

Figura 12. Clasificación de las enzimas presentes en las reconstrucciones ancestrales. EC1 = oxidorreductasas, EC2 = transferasas, EC3 = hidrolasas, EC4 = liasas, EC5 = isomerasas, EC6 = ligasas. Arriba se muestra el número total de proteínas, en la parte inferior los valores se normalizaron por linaje.

Tabla 3. Enzimas con altos niveles de conservación al interior de los catálogos ancestrales. Se muestra el número de taxa en los que fueron identificadas.

EC	Nombre	Taxa	Anotación funcional
1.1.1.100	3-Oxoacil-(proteína transportadora de acil) reductasa	14	Biosíntesis de ácidos grasos
1.5.1.5 (3.5.4.9)	Metilentetrahidrofolato deshidrogenasa ($NADP^+$) Metilentetrahidrofolato ciclohidrolasa (bifuncional)	13	Metabolismo de coenzimas Fijación de carbono Interconversión del tetrahidrofolato Síntesis de purinas, de histidina y de metionina
2.7.4.22	Uridilato cinasa (UMP cinasa)	15	Metabolismo de nucleótidos Síntesis de novo de pirimidinas
2.7.6.1	Ribosa-fosfato pirofosfocinasa	15	Metabolismo de nucleótidos Vía de las pentosas fosfato[1]
2.7.2.3	Fosfoglicerato cinasa	14	Glicólisis
2.7.7.3	Fosfopantetein adenililtransferasa	14	Síntesis de coenzima A
2.3.1.234	N^6-L-treonilcarbamoiladenin sintasa	14	Traducción[2]
2.5.1.31	UDP difosfato sintetasa	13	Síntesis de peptidoglicanos
2.7.4.6	Nucleósido difosfato cinasa (UDP cinasa)	13	Metabolismo de nucleótidos Síntesis de purinas y de pirimidinas
2.5.1.6	S-adenosilmetionin sintetasa	13	Metabolismo de coenzimas Metabolismo de aminoácidos (metionina)
2.6.1.1	Transaminasa piridoxal-fosfato	13	Interconversión de L-aspartato a oxaloacetato[3]
3.6.4.12	Helicasa de DNA[4], dependiente de ATP	16	Replicación, reparación, recombinación y transcripción
3.4.11.18	Metionin aminopeptidasa	15	Traducción[5]
3.6.1.66	XTP/dITP difosfatasa	14	Metabolismo de nucleótidos Síntesis de purinas
4.2.1.11	Fosfopiruvato hidratasa	15	Glucólisis
5.3.1.1	triosafosfato isomerasa	15	Metabolismo de carbohidratos Glucólisis/Gluconeogénesis Vía de las pentosas fosfato

5.4.99.23	Pseudouridin sintetasa	14	Traducción[6]
5.99.1.3	DNA topoisomerasa	14	Replicación y reparación[7]
6.1.1.X	tRNA ligasas	>11	Traducción[8]
6.3.5.2	GMP sintetasa	15	Metabolismo de purinas Síntesis de novo de GMP Metabolismo de glutamina
6.3.4.2	CTP sintetasa	14	Metabolismo de pirimidinas Biosíntesis *de novo* de CTP Metabolismo de glutamina
6.3.5.5	Carbamoil-fosfato sintetasa	13	Metabolismo de aminoácidos[9] Metabolismo de nucleótidos[9]

[1] El producto de la reacción, el pirofosforibosil fosfato o PRPP, se conecta con la biosíntesis de purinas, de histidina, de coenzimas nucleotídicas como NAD^+ y con la biosíntesis de novo de UMP.

[2] Modifica ciertos tRNAs en la adenina 37, donde agrega L-treonilcarbamoiladenilato.

[3] Puede actuar también en L-tirosina, L-fenilalanina y L-triptofano.

[4] Puede actuar en RNA también.

[5] Remueve la metionina N-terminal de los péptidos nacientes.

[6] Modifica diversas uridinas en el RNA de la subunidad grande del ribosoma.

[7] Utiliza la hidrólisis de ATP para romper y/o reincorporar la doble hélice, además de que puede agregar un giro negativo a dicha molécula.

[8] Ver Figura anexo 6.

[9] Uno de sus sustratos es la L-glutamina y forma carbamoil-fosfato, un intermediario en la ruta de arginina y de pirimidinas, así como en la síntesis *de novo* de UMP.

Estimación del núcleo genómico y el pangenoma

El análisis de remuestreo aleatorio con el programa *Get-homologues*, permitió estimar el núcleo genómico de cada linaje, así como su pangenoma (Tabla 4). Las estimaciones más bajas fueron para Euryarchaeota y Firmicutes con 114 y 161 genes respectivamente, mientras que en los grupos con menos especies la cantidad de proteínas conservadas aumentó hasta cinco veces. En todos los casos se observó un comportamiento asintótico del núcleo genómico, el cual decae rápidamente para luego estabilizarse (Figura 13). Por el contrario, el pangenoma no se estabiliza ni en los grupos mejor representados

(Figura 13). Aparentemente en Bacteria los pangenomas son de mayor tamaño que en Archaea, aunque en Thermotogae y Aquificae el tamaño del pangenoma fue conservador, comparado con el resto de los linajes.

Tabla 4. Resultados del análisis pangenómico. Se muestran las estimaciones del núcleo genómico y del pangenoma, así como el número de especies en cada linaje (N).

Grupo	N	Núcleo genómico	Pangenoma
Euryarchaeota	96	114	>30,000
Crenarchaeota	37	202	>12,000
Thaumarchaeota	10	584	>7,000
Alphaproteobacteria	108	283	>50,000
Actinobacteria	91	181	>50,000
Gammaproteobacteria	86	285	>40,000
Deltaproteobacteria	58	166	>50,000
Firmicutes	55	161	>30,000
Bacteroidetes	54	291	>35,000
Cyanobacteria	50	528	>30,000
Betaproteobacteria	47	405	>30,000
Deinococcus-Thermus	16	560	>9,000
Thermotogae	15	553	>5,000
Epsilonproteobacteria	14	653	>7,000
Chloroflexi	12	316	>12,000
Aquificae	11	553	>4,000
Verrucomicrobia-Planctomycetes	9	317	>16,000
Spirochaetes	9	289	>10,000
Acidobacteria	7	527	>12,000

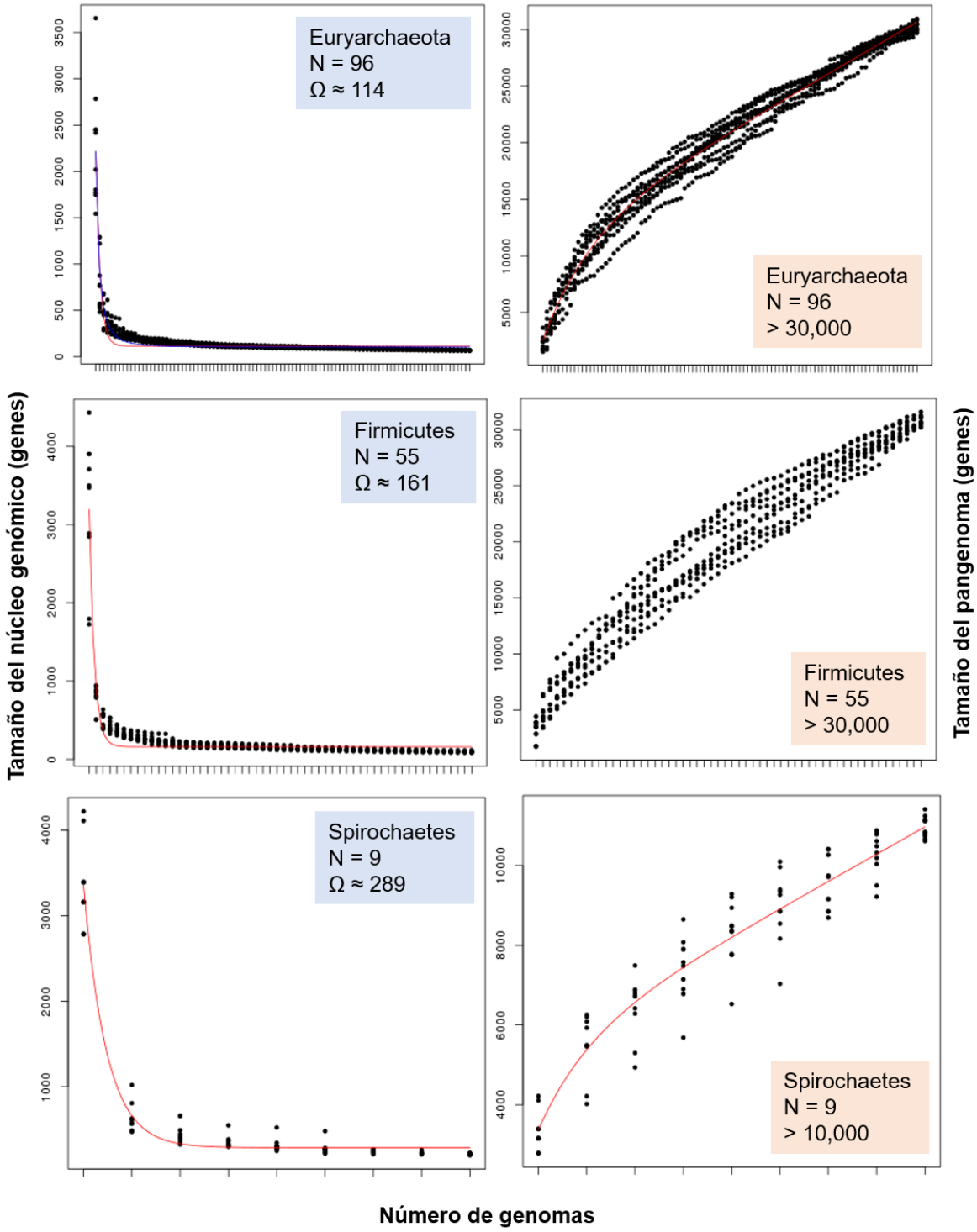

Figura 13. Gráficas representativas del comportamiento del núcleo genómico y del pangenoma. Omega representa la estimación del núcleo genómico. La totalidad de las gráficas se puede consultar en https://goo.gl/szEv6p.

Árbol de referencia y distancia filogenética

La filogenia que se construyó contiene a las 785 especies de la muestra. En la topología resultante se separan los tres linajes de Archaea de los 16 grupos de Bacteria, los cuales también se distinguen claramente entre ellos (Figura 14).

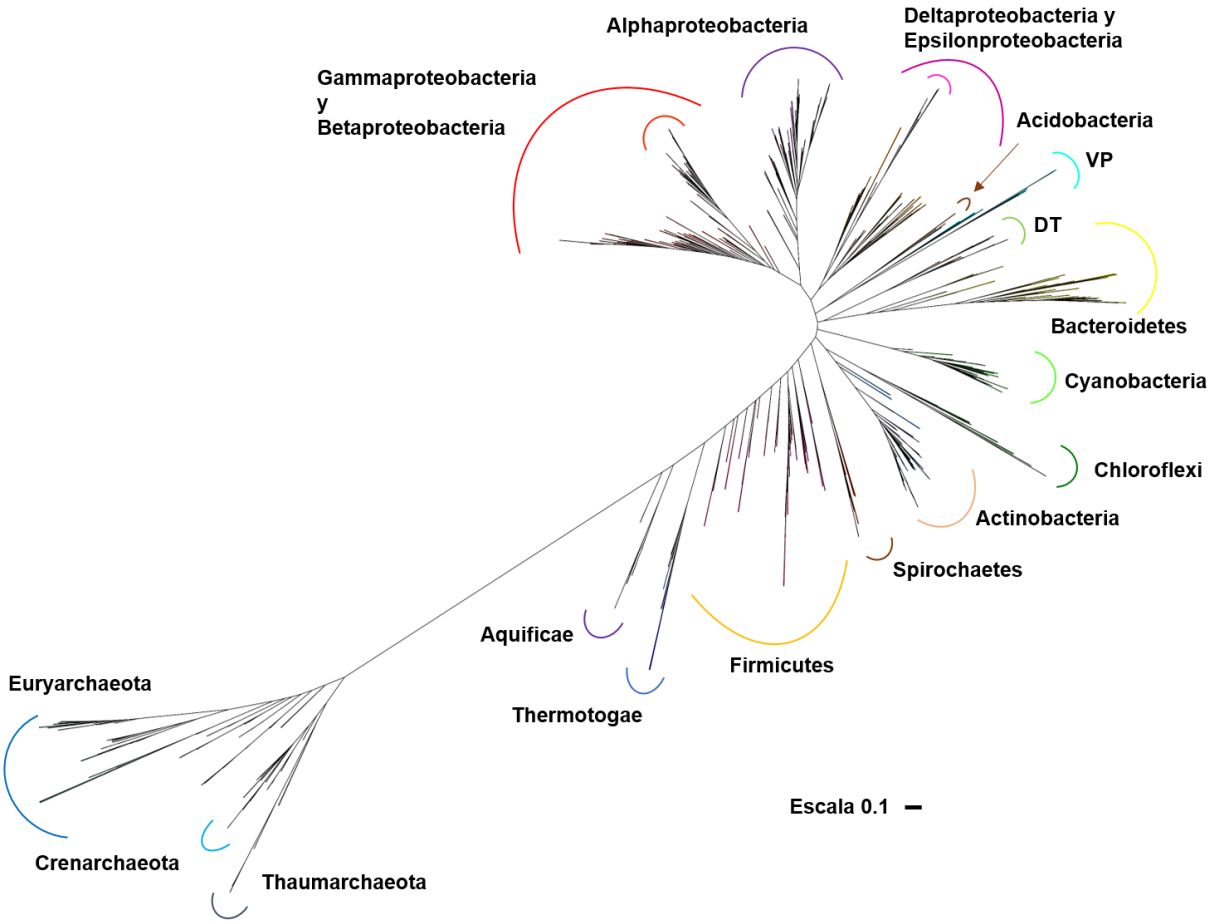

Figura 14. Árbol construido con 16S rRNA donde se muestran las relaciones filogenéticas entre las especies de Archaea y Bacteria contenidas en la muestra. VP = Verrucomicrobia-Planctomycetes, DT = Deinococcus-Thermus. Gammaproteobacteria contiene a Betaproteobacteria, así como también Deltaproteobacteria contiene a Epsilonproteobacteria.

Utilizando como base el árbol previamente construido, se calculó la distancia filogenética (PD) al interior de cada linaje. Dicha medida cabe aclarar, depende de la cantidad de especies, aunque por ejemplo el grupo Alphaproteobacteria con 108 especies,

presentó una menor distancia filogenética comparado con Euryarchaeota, incluso cuando éste tiene 96 especies.

En la Figura 15 se puede ver que el núcleo genómico disminuye conforme aumenta la distancia filogenética de las especies en la muestra. Los grupos Euryarchaeota y Firmicutes son los linajes con mayor PD, lo que sugiere que dicha métrica refleja acertadamente lo que de manera intuitiva se observa en la filogenia (Figura 14). En ese sentido, Crenarchaeota tiene una reconstrucción ancestral pequeña, respecto a lo que se esperaría por la distancia filogenética de sus especies, mientras que Cyanobacteria presenta un catálogo ancestral con una conservación más alta a lo esperado, considerando que se comparó un número importante de especies (50).

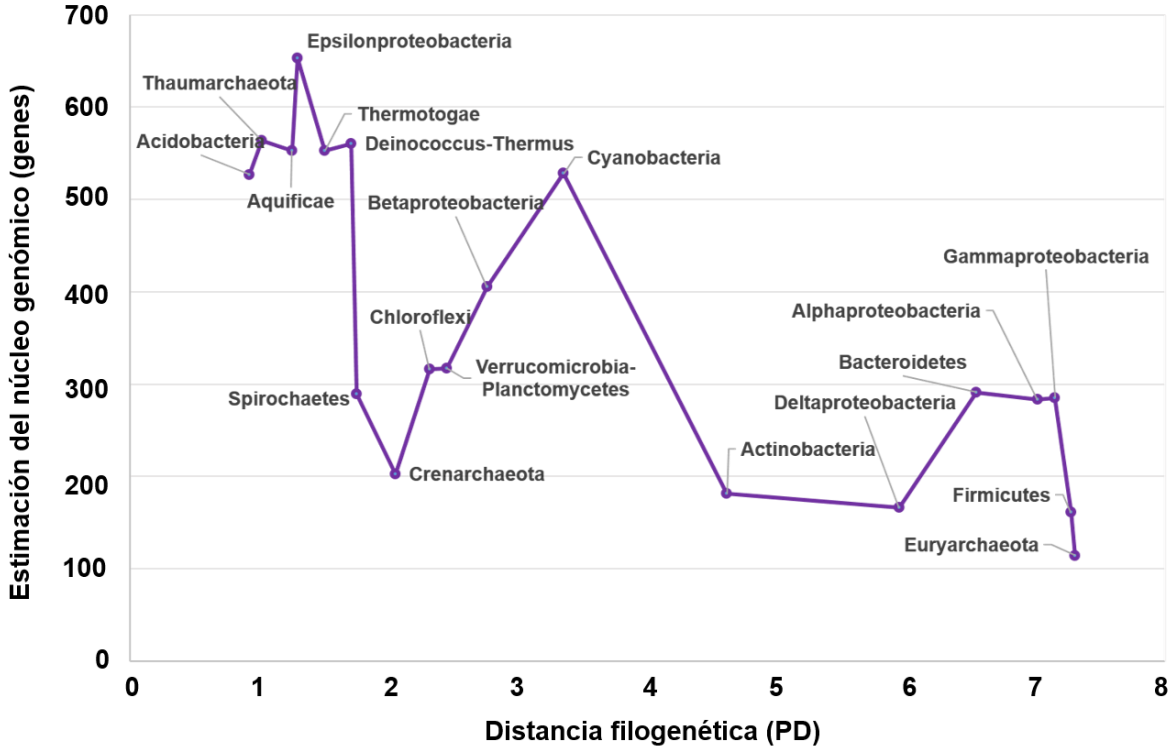

Figura 15. Se muestra la distancia filogenética y el tamaño del núcleo genómico para cada linaje.

Tamaño de las proteínas conservadas

Finalmente, un aspecto que fue posible examinar fue la longitud de las secuencias proteínicas conservadas al interior de las reconstrucciones ancestrales, cuya distribución a

nivel de Dominio se muestra en la Figura 16. Las proteínas de menor tamaño tuvieron 37 aminoácidos, mientras que las más grandes hasta 1650, intervalo que fue muy semejante al interior de todos los catálogos (ver Tabla 5).

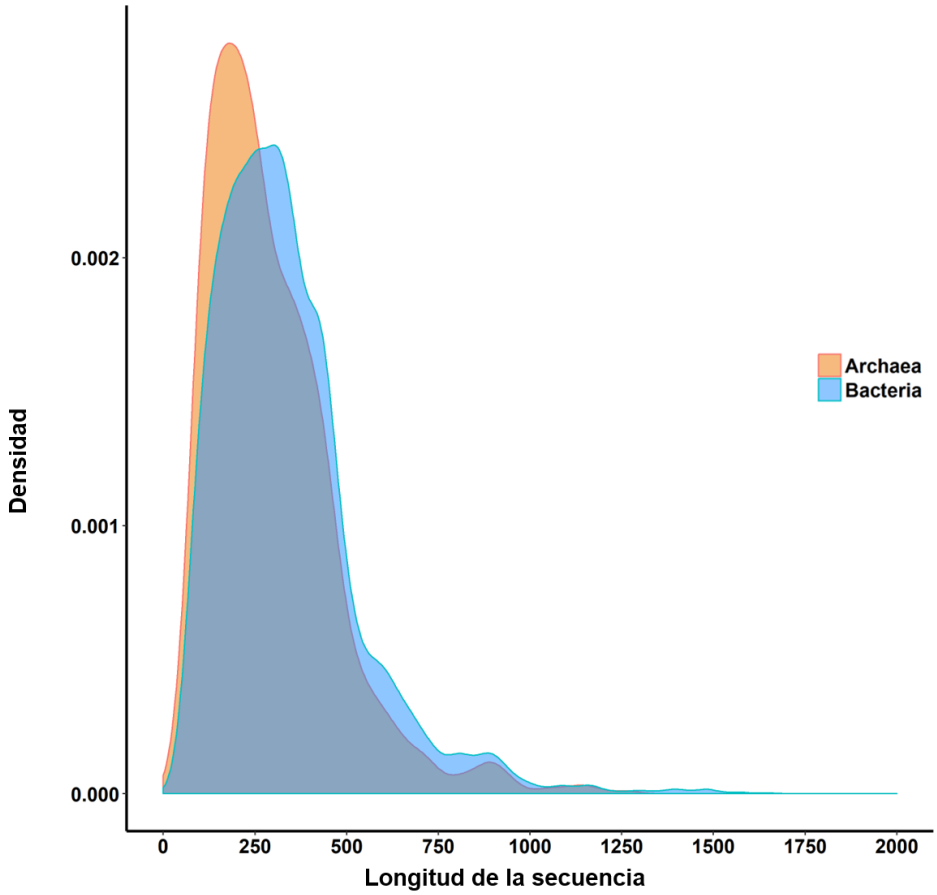

Figura 16. Distribución de la longitud de las secuencias al interior de los catálogos ancestrales. Se muestran ambos dominios procariontes, Archaea en naranja y Bacteria en azul.

Se detectó un dato atípico de 3460 aminoácidos en el catálogo del linaje Alphaproteobacteria, el cual representa una familia proteínica inconsistente, porque solamente la secuencia semilla tiene dicha longitud y las restantes son más pequeñas. Este tipo de conjuntos se detectaron de manera recurrente en la reconstrucción de Alphaproteobacteria, pero no en el resto de los linajes.

Tabla 5. Tamaño de las secuencias proteínicas conservadas al interior de los catálogos ancestrales de los distintos linajes.

Grupo	Mínimo	Mediana	Media	DE	Máximo
Euryarchaeota	60	255	304	177	1036
Crenarchaeota	52	284	320	183	1165
Thaumarchaeota	55	251	285	171	1263
Alphaproteobacteria	49	292	349	294	3460
Actinobacteria	85	336	354	197	1287
Gammaproteobacteria	51	291	326	199	1342
Deltaproteobacteria	49	298	320	179	900
Firmicutes	72	295	333	199	1221
Bacteroidetes	72	328	360	207	1484
Cyanobacteria	59	310	338	191	1367
Betaproteobacteria	51	280	318	193	1409
Deinococcus-Thermus	49	303	330	182	1524
Thermotogae	44	298	328	189	1650
Epsilonproteobacteria	37	257	292	177	1070
Chloroflexi	37	305	335	199	1272
Aquificae	49	305	330	188	1574
Verrucomicrobia-Planctomycetes	80	339	368	198	1479
Spirochaetes	59	349	377	214	1433
Acidobacteria	49	338	379	214	1479

DE = desviación estándar.

La distribución de la longitud de las secuencias conservadas al interior de cada linaje, presenta el mismo patrón que se observa a nivel de Dominio (Figura 16), en el cual hay un pico alrededor de la media, que es de 293 (DE = 174) aminoácidos en Archaea y de 337 (DE = 200) en Bacteria, seguido por otro pico más pequeño, entre los 800 y los 1000 aminoácidos (Figura 16).

Una revisión más detallada mostró que las secuencias de menor longitud, corresponden a proteínas anotadas como hipotéticas, proteínas ribosomales como L34, L36, L33, L32, L30, L24, S14, L29, L35, S21, L31, L28, la subunidad e'' de la RNA polimerasa dependiente de DNA (RpoE2) de Archaea, la subunidad SecE de la proteína translocasa, y la subunidad omega de la RNA polimerasa. En el otro extremo, las proteínas de mayor tamaño fueron las subunidades beta', beta, A'/A'' y B, de la RNA polimerasa dependiente de DNA, la DNA polimerasa III, la enzima glutamato sintasa (NADPH y ferredoxina), una proteína SMC de segregación del cromosoma, un factor de transcripción y de reparación, la subunidad grande de la carbamoil fosfato sintetasa, y algunas tRNA sintetasas, entre otras. Éstas representan el segundo pico de la distribución de las familias proteínicas conservadas (Figura 16).

Discusión

Aspectos relacionados con la muestra

Con el propósito de contar con un control metodológico para la reconstrucción del LCA, en el presente trabajo se aplicó el método de caracterización directa para reconstruir los genomas ancestrales de diversos linajes procariontes, a través de la comparación de proteomas completos. En ese sentido, lo primero que se hizo fue seleccionar a las especies que conformaron la base de datos para, posteriormente, descargar su proteoma a partir del repositorio de genomas completos del NCBI.

Como ya se mencionó, las diferencias que exhiben las distintas reconstrucciones publicadas del catálogo génico del LCA, se deben a las distintas bases de datos que se utilizaron (Becerra *et al*., 2007), las cuales se describen en la Tabla 1. De manera natural, con el aumento en la cantidad de información, también las bases de datos y las herramientas han evolucionado. Basta comparar la propuesta de Mushegian y Koonin, quienes en 1996 utilizaron los dos genomas bacterianos secuenciados al momento, con el último trabajo publicado del LCA, que es un análisis filogenético de 6.1 millones de secuencias de genes (Weiss *et al*., 2016). Es innegable que esta tendencia se mantiene (Land *et al*., 2015; Mukherjee *et al*., 2017), lo que indica que seguirá siendo un factor a considerar en el estudio del LCA.

Es un hecho que, en el presente trabajo, el contenido de la muestra depende en principio de los genomas completamente secuenciados disponibles en la base de datos del NCBI. Dicho catálogo comprende una selección sesgada y limitada del mundo procarionte, pues no están igualmente representados los distintos linajes. Algunos grupos de Bacteria como Proteobacteria, Firmicutes, Bacteroidetes y Actinobacteria, así como Euryarchaeota, están sobrerrepresentados, porque la base de datos contiene en su mayoría a especies cultivables en el laboratorio, que por lo general son de interés clínico, agropecuario o biotecnológico. Dicha tendencia está cambiando rápidamente con el desarrollo de las técnicas de metagenómica y de secuenciación en una sola célula (Lasken y McLean, 2014), gracias a las cuales actualmente se han descrito al menos 89 phyla de Bacteria y 20 de

Archaea, y se estima la existencia hasta de 1500 en Bacteria (Solden *et al*., 2016). Aunado a esta limitante intrínseca de la base de datos, en el presente estudio se estableció el criterio de solo incluir a linajes con más de 5 especies completamente secuenciadas, por lo que al final se reconstruyeron los catálogos ancestrales de tres linajes de Archaea y de 12 taxa de Bacteria. Este filtro dejó fuera muchos de los grupos de reciente descubrimiento, como es el caso de los phyla candidatos (Brown *et al*., 2015), los cuales han expandido de manera significativa la visión del árbol de la vida, al grado de que se ha propuesto que los linajes procariontes para los que no se conocen representantes cultivables, contienen la mayor diversidad biológica (Hug *et al*., 2016). A pesar de estos sesgos iniciales, los 19 linajes de la muestra tienen la virtud de que son los que mejor se conocen hasta ahora, lo cual representa una ventaja al interpretar los catálogos ancestrales.

El punto más relevante relacionado con la selección de la muestra es que solamente se incluyeron procariontes de vida libre, ya que las especies con genomas reducidos subestiman la inferencia del contenido del catálogo ancestral de genes, debido a la gran cantidad de pérdidas secundarias que presentan (Becerra *et al*., 1997). En un principio se descartaron aquellas especies que habitan el interior de la célula de su hospedero, ya sea endoparásitas o endosimbiontes, pero este criterio resultó laxo y fue necesario descartar también a los patógenos obligados, a los comensales obligados, y a organismos relacionados con un hospedero y que poseen menos de 1300 proteínas (límite para especies de vida libre publicado por Islas *et al*., 2004). Dichos procariontes por lo general presentan pérdidas secundarias como consecuencia de la interacción que establecen con el hospedero. La muestra final, es decir la base de datos de genomas completos utilizada en el presente trabajo, contiene 785 especies de vida libre. Sin embargo, la selección de la muestra no fue un ejercicio trivial, en parte por las diferencias en la calidad de la información asociada a cada secuencia (Land *et al*., 2015), pero sobre todo por la enorme diversidad de las interacciones que establecen los procariontes (ver por ejemplo Braga *et al*., 2016). Adicionalmente, ahora se sabe que hay bacterias y arqueas de vida libre, generalmente marinas, con genomas reducidos, los cuales tienden a estabilizarse alrededor de las 1300 proteínas, y pueden sostener metabolismos como heterotrofía aerobia y

anaerobia, respiración del azufre y metanogénesis (Martínez-Cano *et al*., 2014). En la base de datos hubo ejemplos como *Lactobacillus sanfranciscensis, Desulfurococcus mucosus y Methanothermus fervidus*, los cuales presentaron 1217, 1345 y 1283 proteínas respectivamente.

Al interior de la muestra, los grupos con mayor variación en el tamaño de su proteoma fueron Actinobacteria y Proteobacteria (Figura 6). El filo Actinobacteria contiene bacterias Gram positivas de elevado contenido G+C, la mayoría se conocen como Actinomycetes y se asemejan morfológicamente a hongos filamentosos (Willey *et al*., 2008). Este fenotipo, incluyendo su ciclo de vida y su dependencia al oxígeno, posiblemente es la razón por la cual son el grupo con más proteínas en promedio en sus genomas. Las proteobacterias son un grupo de bacterias Gram negativas, ubicuas, que se caracterizan por ser uno de los linajes más diversos, presentan una notable variedad morfológica, metabólica, ecológica y reproductiva (Willey *et al*., 2008). En estos grupos, los organismos con mayor número de proteínas en sus proteomas fueron *Streptomyces bingchenggensis* y *Archangium gephyra*, ambos habitantes de suelo.

En el otro extremo encontramos a los grupos Aquificae y Thermotogae, los cuales comparten la característica de que sus miembros son bacterias que viven a temperaturas muy elevadas. Los genomas relativamente pequeños son característicos de termófilos extremos (Counts *et al*., 2017), lo que explica la poca variabilidad en cuanto al tamaño del proteoma, tendencia que se observa en menor medida en Crenarchaeota, pues también una proporción importante de los géneros de la muestra son termófilos.

Aspectos relacionados con la metodología

El presente trabajo tuvo como objetivo secundario aplicar la herramienta *Get-homologues* para la reconstrucción ancestral. Consideramos que los valores utilizados para el análisis de similitud entre las secuencias, que fueron una cobertura de la secuencia *query* del 75% y un valor de *e* menor a 1×10^{-5}, son estrictos y permiten descartar falsos positivos,

aunque la matriz utilizada para evaluar el alineamiento, que fue BLOSUM62, ciertamente limita la búsqueda de homólogos muy lejanos.

Para agrupar a las secuencias ortólogas se utilizó el algoritmo BDBH implementado en *Get-homologues*. Hay estudios que reportan que el método de BDBH es robusto para detectar ortólogos reales en organismos procariontes (Wolf y Koonin, 2012; Dalquen y Dessimoz, 2013), y en general este fue el caso. Cabe mencionar que dicho proceso en *Get-homologues* demanda mucha memoria y esto impuso una limitante en el análisis.

Como se describió en el capítulo de material y métodos, es indispensable seleccionar un proteoma semilla a partir del cual se construyen los grupos de secuencias presuntamente homólogas (familias proteínicas). En el presente análisis se eligieron los proteomas de menor tamaño, aunque también se realizaron controles con los de mayor tamaño, y en ambos los resultados tienden a valores semejantes del contenido del núcleo genómico. Dicho valor además coincidió con la estimación hecha en el análisis pangenómico (Tabla 4), todo lo cual sugiere que los resultados son robustos, aunque es posible que a pesar de que numéricamente hay congruencia, la identidad de las proteínas sea un tanto distinta (Kannan *et al*., 2013) como consecuencia de la utilización de secuencias semilla diferentes.

Una revisión de las familias proteínicas, reveló algunos conjuntos en los que la longitud de la secuencia semilla era mucho mayor que el resto de las secuencias al interior del grupo, las cuales incluso podían ser proteínas con una anotación muy distinta. Esto podría deberse a que no se modificó el parámetro de cobertura de la secuencia sujeto (*subject*), lo que finalmente sí tuvo un impacto en los resultados. Sin embargo, dichos conjuntos con incongruencias fueron pocos y se detectaron sobre todo en los linajes con mayor cantidad de especies en la muestra, en particular en Alphaproteobacteria. La naturaleza misma de las proteínas, las cuales se componen de unidades estructurales y funcionales llamadas dominios (Chothia *et al*., 2003), explica los conjuntos atípicos descritos anteriormente, pues aunque algunas proteínas solo presentan un dominio, la mayoría son

multidominio (Vogel *et al*., 2004), y éstos pueden tener historias evolutivas complejas e incluso distintas de la proteína completa (Vogel *et al*., 2004).

No sobra decir que dichas inconsistencias se detectaron mediante una revisión visual de la calidad de los resultados, la cual se realizó de manera aleatoria en algunas familias proteínicas. Dicha revisión visual puede ser un reto cuando se trabaja con mucha información, mas no implica necesariamente una revisión exhaustiva y es un ejercicio fundamental en un análisis bioinformático. Tan solo en la reconstrucción más reciente del LCA, los autores concluyeron que éste podía adquirir nitrógeno vía nitrogenasas NifD (Weiss *et al*., 2016), sin embargo, se le ha criticado que al interior de los dos conjuntos de proteínas que ellos interpretaron como NifD, ninguna secuencia tiene la firma característica de aminoácidos que le permite unirse al cofactor metálico (FeMoco) responsable de la actividad catalítica (McGlynn, comentario en https://www.ncbi.nlm.nih.gov/pubmed/27562259). Esto implica que en realidad el análisis no recuperó ninguna nitrogenasa NifD verdadera, y lo grave es que a partir de dicho resultado los autores hacen afirmaciones contundentes sobre una característica del LCA (Weiss *et al*., 2016). Ello muestra que si bien establecer la función de una secuencia en un genoma moderno es complicado, se debe tener mucha pecacución para extrapolarlo a los ancestros.

Contraste de las reconstrucciones ancestrales con las especies modernas

En la Tabla 2 se observa que el número de proteínas o mejor dicho de familias proteínicas presentes en cada reconstrucción ancestral fue variable. Los catálogos con menos proteínas conservadas fueron Actinobacteria, Euryarchaeota, Firmicutes y Deltaproteobacteria. Los linajes Actinobacteria y Euryarchaeota son dos de los grupos con más especies en la muestra (Figura 3), mientras que éste último, junto con Firmicutes y Deltaproteobacteria, comparten la característica de ser grupos fisiológicamente diversos. Aunado a esto, Firmicutes y Euryarchaeota son de los grupos con mayor distancia filogenética al interior de la base de datos (Figura 15).

Euryarchaeota es un grupo metabólicamente diverso (Willey *et al*., 2008). En la muestra analizada hubo géneros de arqueas metanógenas, halófilas extremas, termoacidófilas que se caracterizan por la ausencia de pared celular, termófilas extremas con metabolismo del azufre y también reductoras de sulfato. El linaje Firmicutes contiene bacterias Gram positivas con bajo contenido de G+C, y también es un grupo diverso (Willey *et al*., 2008). La muestra contiene organismos heterótrofos, que pueden ser aerobios o anaerobios, totales o facultativos y en algunas ocasiones forman endoesporas. Hay metilótrofos, acetógenos, homoacetógenos, bacterias del ácido láctico, bacilos quimioheterótrofos, e incluso hay especies fotoheterótrofas. Varios géneros presentan una versatilidad metabólica notable y pueden convertirse en quimiolitótrofos de manera facultativa. Son mesófilos pero también viven en ambientes extremos como el fondo del mar, hay firmicutes termófilos, termoacidófilos, alcalitermófilos, termohalófilos y halófilos. En las Deltaproteobacteria hay organismos quimiorganótrofos con dos estilos de vida radicalmente distintos, unas tienen hábito depredador y ciclos de vida complejos, y otras son anaerobias con metabolismo del azufre (oxidan compuestos orgánicos mediante la reducción del sulfuro o del sulfato).

Los linajes con las reconstrucciones más numerosas fueron Deinococcus-Thermus, Thermotogae, Aquificae y Epsilonproteobacteria. No son los linajes con el menor número de especies en la muestra (Figura 3), pero tienen valores bajos de distancia filogenética (Figura 15). De todas las estimaciones, el catálogo ancestral más numeroso contiene 533 familias proteínicas en el caso del núcleo estricto y corresponde a Epsilonproteobacteria, mientras que dicho límite aumentó hasta 823 en el núcleo relajado del linaje Thaumarchaeota (Tabla 2, Figura 4).

Aquificae y Thermotogae son bacterias Gram negativas hipertermófilas. Las primeras son quimiolitoautótrofas microaerofílicas, y las segundas son quimioheterótrofas anaerobias. El grupo Deinococcus-Thermus contiene bacterias aerobias, heterótrofas, algunas en la muestra son termófilas, y todas comparten una organización y composición característica en su envoltura celular, que las dota de una resistencia increíble a la desecación y a la radiación (Willey *et al*., 2008). Epsilonproteobacteria es el grupo más

reducido de Proteobacteria al interior de la muestra, y contiene organismos quimiolitótrofos, tanto aerobios como anaerobios, así como un par de heterótrofos. Aquí es importante aclarar que hay muchas otras especies de Epsilonproteobacteria cuyo genoma ha sido completamente secuenciado, pero su biología sugiere que han establecido interacciones persistentes y complejas con linajes eucariontes, lo que ha modificado de manera importante sus genomas, razón por la cual no las consideré para el análisis. Por ejemplo, dicha tendencia se ha documentado extensivamente en *Helicobacter pylori* (Whalen y Massida, 2015).

Es un hecho que mientras más especies contiene un linaje, se detectan menos familias proteínicas conservadas (Figura 4), pues como ya se ha reportado, el número de genes compartidos está en función del número de cepas analizadas (Tettelin *et al*., 2005). Además, por definición el núcleo relajado contiene más proteínas que el estricto (Tabla 2). En ese sentido, en los resultados se observó la tendencia de que conforme más especies se incorporan a la muestra, entonces la diferencia entre el núcleo estricto y el relajado se hace menor (Figura 5). El grupo Thaumarchaeota destacó de los demás porque se incorporaron más del doble de proteínas al relajar el criterio de inclusión al catálogo ancestral (Figura 4). Al interior de la muestra dicho linaje contiene arqueas aerobias, mesófilas, en su mayoría acuáticas, todas quimiolitoautótrofas que oxidan el amoníaco. Adicionalmente, el grupo presentó una distancia filogenética baja comparado con otros (Figura 15).

En términos del número de familias proteínicas conservadas, aunque existe variación al interior de los catálogos ancestrales, hay que señalar que ninguno supera las 900 proteínas, número que disminuye a 300 en los grupos mejor muestreados (Figura 4). Esta cantidad se encuentra por debajo del mínimo de proteínas presentes en los genomas de las especies modernas (Figura 6), y del límite planteado de 1300 proteínas para especies de vida libre (Islas *et al*., 2004; Martínez-Cano *et al*., 2014). Dicho de otro modo, no hay bacterias o arqueas modernas de vida libre con un número tan reducido de proteínas, lo que sugiere que existe un sesgo cuantitativo muy importante en las reconstrucciones ancestrales.

Un resultado original del trabajo es el porcentaje de reconstrucción ancestral, cuyo valor máximo fue de 36% para el núcleo relajado o de 30% si consideramos al núcleo estricto (Tabla 2). Sin embargo, estos números son optimistas porque la Figura 5 muestra cómo la mayoría de las estimaciones no supera el 15%, número que disminuye a 10% para el núcleo estricto. Esto no significa que los ancestros tuvieran solamente el 10% del contenido proteínico de las especies modernas, sino que las reconstrucciones están subestimadas y que en realidad se recupera muy poco del contenido genómico de los ancestros. Visto así, los catálogos de Aquificae, Thermotogae, Epsilonproteobacteria, Deinococcus-Thermus y Thaumarchaeota son la excepción a la regla de que las estimaciones ancestrales no superan valores de reconstrucción de entre el 10% y el 15%. Estos grupos tienen tamaño de muestra intermedios, valores bajos de distancia filogenética (Figura 15) y son metabólicamente homogéneos comparados con otros linajes de la muestra.

Un primer resultado a discutir de la clasificación funcional en las categorías de COG, es que el 18% de las familias proteínicas conservadas están pobremente caracterizadas, ya sea porque i) no tienen un COG asignado, ii) pertenecen a la categoría COG de función desconocida, o iii) pertenecen a la categoría COG de predicción general de la función. Al interior de los tres conjuntos mencionados la mayoría son proteínas cuya actividad se desconoce por completo, lo que sugiere que ignoramos el papel de una gran cantidad de proteínas cuya presencia en los catálogos ancestrales advierte que son importantes. Un resultado semejante se obtuvo en la construcción de un genoma mínimo artificial, donde a partir de la reducción de *Mycoplasma genitalium* se llegó a JCVI-syn3.0, un organismo con 473 genes, de los cuales 149 son de función desconocida pero esenciales (Hutchison *et al*., 2016), como podría ser el caso de las proteínas al interior de las diferentes reconstrucciones ancestrales.

Aunado a lo anterior, es pertinente recordar que hay un grado de certeza asociado a la asignación funcional de las proteínas. En un extremo se encuentran aquellas que cuentan con evidencia experimental, pero ciertamente son minoría. Se sabe que por la forma como se anotan los genomas, se producen y se replican errores que van desde mala ortografía en los nombres de las proteínas hasta anotaciones inconsistentes. Por ejemplo,

se originan varios nombres para el producto de un mismo gen, o bien hay proteínas que no tienen homólogos aparentes y se denominan hipotéticas (Richardson y Watson, 2013). Estos casos se detectaron al interior de las familias proteínicas y son inherentes a la base de datos utilizada. El punto es que si no existe un consenso sobre la función de un conjunto de secuencias, entonces es menos lo que se puede afirmar al respecto de la posible función en el ancestro.

La clasificación en las categorías funcionales generales de COG reveló que la mayoría de las familias proteínicas conservadas se relacionan con el metabolismo (Figura 7). Dicha tendencia fue muy evidente en Cyanobacteria (Figura 7), resultado interesante porque es el único linaje capaz de realizar fotosíntesis oxigénica gracias a dos fotosistemas y a pigmentos como la clorofila *a* (Willey *et al*., 2008), además de contar con una diversidad morfológica importante e incluso multicelularidad en diversos taxa. Por si fuera poco, se reconoce su presencia en la Tierra primitiva y se postula como el linaje responsable del gran evento oxidativo (Schirrmeister *et al*., 2015). En el presente estudio, las cianobacterias tuvieron una reconstrucción ancestral grande respecto de otros grupos con un tamaño de muestra similar (Figura 15), lo que se debe a que su metabolismo está muy conservado.

Un segundo grupo con una gran proporción de proteínas metabólicas en la reconstrucción ancestral fue Verrucomicrobia-Planctomycetes. En la muestra éste contiene solamente nueve representantes, todos heterótrofos, algunos aerobios y otros anaerobios, que son Gram negativo y tienen una pared celular peculiar, presentan sistemas endomembranosos bien desarrollados y una forma de división con gemación característica (Rivas-Marín *et al*. 2016). En el otro extremo se identificaron los linajes Firmicutes y Euryarchaeota, para los que la categoría de almacenado y procesado de la información genética fue, en proporción, la más numerosa (Figura 7). De hecho, se observa la tendencia general de que conforme disminuye el tamaño de las reconstrucciones ancestrales, entonces se recuperan menos cuestiones asociadas al metabolismo y más proteínas relacionadas con el procesado de la información genética.

Al analizar con mayor detalle la clasificación funcional, resultó evidente que la categoría con más proteínas conservadas fue la de traducción, biogénesis y estructura del ribosoma (Figura anexo 3). En los catálogos reducidos, como por ejemplo los grupos Deltaproteobacteria, Alphaproteobacteria o Actinobacteria, aparentemente disminuye la cantidad de proteínas conservadas a lo largo de todas las categorías funcionales, incluyendo la de traducción (Figura 9). Al normalizar los datos por linaje, se reafirma que, en proporción, la categoría de traducción es la más conservada al interior de todos los catálogos ancestrales (Figura 8), como ya lo había dicho Woese (1987). La identidad de las proteínas dentro de dicha categoría explica los niveles de conservación, pues ninguna bacteria o arquea moderna puede prescindir de las proteínas estructurales (Figuras anexo 4 y anexo 5), de biosíntesis, de maduración o de reciclaje del ribosoma, además del conjunto de enzimas que hacen posible la traducción, como son los factores de iniciación, de elongación, de liberación de la cadena peptídica, las familias de las aminoacil-tRNA sintetasas (Figura anexo 6), y otras proteínas que modifican directamente a los RNAs y a péptidos recién sintetizados. Los resultados sugieren que los ancestros de los linajes estudiados ya contaban con la maquinaria enzimática que les permitía llevar a cabo una traducción de tipo contemporánea.

La presencia de las familias de proteínas ribosomales y de las aminoacil-tRNA sintetasas en todos los catálogos ancestrales, puede verse como un control interno del análisis, pues se sabe que están sumamente conservadas. Por ejemplo, se ha publicado que las siguientes proteínas de la subunidad pequeña: S2-S5, S7-S15, S17 y S19 y de la subunidad grande: L1-L6, L10-L15, L17, L18, L22-L24, L29 y L30 están universalmente distribuidas (Korobeinikova *et al*., 2012), de las cuales, en el presente trabajo, todas menos seis (S2, S14, S17, L17, L29, L30) exhibieron niveles muy elevados de conservación (Figuras anexo 4 y anexo 5). Estas proteínas tienen contactos directos con el RNA y participan en la formación del sitio activo y el ensamblaje correcto del ribosoma (Korobeinikova *et al*., 2012). Además de las ya mencionadas, se identificaron otras siete proteínas ribosómicas con niveles importantes de conservación (L9, L16, L20, L21, L25, L27), de las cuales se sabe que la L16 y la L27 presentan contactos con el tRNA, así como la L20 con otras proteínas (Korobeinikova

et al., 2012), lo que sugiere que son importantes para el correcto funcionamiento del ribosoma.

Las aminoacil-tRNA sintetasas, otro caso de conservación al interior de todos los linajes, son las responsables de establecer el código genético e interactúan directamente con el tRNA. La presencia de dichas familias proteínicas en los catálogos ancestrales, refuerza la noción de que los ancestros ya contaban con ribosomas funcionales y con la capacidad de llevar a cabo el proceso de traducción, lo que confirma que éste es un rasgo muy antiguo, compartido entre los dominios Bacteria y Archaea. Además, la presencia de proteínas conservadas cuya longitud es mayor a 600 aminoácidos (Figura 16, Tabla 5), sugiere que los ribosomas de los ancestros tenían una capacidad de síntesis equiparable al de las especies modernas.

Uno de los resultados más notables del trabajo, es que la conservación dentro de los catálogos ancestrales no es homogénea al interior de las diferentes categorías funcionales (Figuras 8, 9 y 10). Es decir que después de la traducción, son pocas las categorías que se recuperan en las reconstrucciones, entre ellas destacan la de transporte y metabolismo de aminoácidos, transporte y metabolismo de coenzimas, la de replicación, recombinación y reparación, además de la biogénesis de la membrana, pared o envoltura celular. Lo anterior sugiere la presencia de un sesgo cualitativo en las reconstrucciones ancestrales.

De acuerdo a la clasificación COG, la categoría más numerosa en los catálogos ancestrales relacionada con el metabolismo, fue la de transporte y metabolismo de aminoácidos (Figuras 8 y 9). Todas las células modernas, sean bacterias o arqueas, poseen la habilidad de obtener dichos compuestos, ya sea a partir del medio o a través de su síntesis. Las proteínas conservadas presentes en las reconstrucciones, que son principalmente transportadoras y enzimas relacionadas con el catabolismo y la biosíntesis de los 20 aminoácidos canónicos, sugiere que los ancestros de los linajes estudiados ya contaban con dicha capacidad. De forma interesante, muchas de las proteínas conservadas tienen otras funciones asignadas, por ejemplo, participan en la biosíntesis de más de un aminoácido, o incluso en la síntesis de otros compuestos importantes como coenzimas,

carbohidratos o pirimidinas. Esto podría ser un reflejo del bricolaje, proceso mediante el cual se han ido ensamblando las rutas metabólicas (Lazcano y Miller, 1999; Peretó, 2011).

Los catálogos ancestrales presentan un número importante de proteínas que participan en el metabolismo de diversas coenzimas, por ejemplo, de NAD+ y de FAD+, de las vitaminas B1 (tiamina), B2 (riboflavina), B5 (pantotenato), B6 (piridoxal fosfato), B7 (biotina), B9 (folato), B12 (cobalamina), de la coenzima A, de isoprenoides (también presentes en la categoría funcional de metabolismo y transporte de lípidos), quinonas, la coenzima M, el grupo hemo, el ácido lipoico, la clorofila y el cofactor de molibdeno, entre otros. Como se mencionó en el capítulo de resultados, no todas están presentes en todos los linajes, de hecho, existe una sobrerepresentación de las enzimas de los taxa Aquificae y Cyanobacteria. Sin embargo, los niveles de conservación que presentan, así como su participación en reacciones esenciales del metabolismo general, sugieren que las familias de proteínas que participan en el metabolismo de coenzimas son muy antiguas, y que los ancestros de los linajes modernos de arqueas y bacterias ya requerían y utilizaban coenzimas. Es interesante que muchas se sintetizan a partir de nucleótidos, de aminoácidos, o de una combinación de ambos, por ejemplo, el NAD+, el FAD+ y la coenzima A, se forman a partir de ATP (Berg *et al*., 2002), mientras que los anillos de porfirinas, el grupo hemo y la clorofila, a partir de aminoácidos (Berg *et al*., 2002).

La categoría de replicación, recombinación y reparación, contiene principalmente familias de proteínas como DNA y RNA helicasas, topoisomerasas, DNA polimerasas, DNA primasas, proteínas de reparación, procesamiento y recombinación del material genético, así como también una variedad de nucleasas y de ligasas. De nuevo, las funciones que cumplen dichas enzimas son esenciales para el funcionamiento de cualquier arquea o bacteria moderna, por lo que es sensato concluir que los ancestros ya contaban con genomas de DNA, y con el conjunto de enzimas que les permiten, no sólo acceder a la información genética, sino también mantener su integridad.

El ensamblaje de la pared, la membrana y la envoltura celular, fue la categoría COG restante con niveles importantes de conservación. Contiene principalmente proteínas que

participan en la biosíntesis de compuestos como lípidos, lipopolisacáridos, glicolípidos, lipoproteínas, peptidoglicanos, y el intermediario UDP-acetilglucosamina, así como también enzimas de la síntesis de carbohidratos, proteínas asociadas a la membrana externa, y transportadores, entre otras. Sin embargo, pude notar que al interior de los catálogos había más familias proteínicas involucradas en el metabolismo de la envoltura celular, por ejemplo, en la categoría de tráfico intracelular, transporte de vesículas y secreción, se detectaron translocasas, proteínas de la membrana interna, insertasas, más transportadores y partículas de reconocimiento señal, entre otras. También aporta a esta función la categoría de transporte y metabolismo de lípidos, pues contiene enzimas del metabolismo de fosfolípidos, de glicerofosfolípidos, y de ácidos grasos. La conservación de este conjunto de enzimas en las reconstrucciones, sugiere que los ancestros de los linajes estudiados ya presentaban la capacidad de sintetizar y mantener una envoltura celular funcional. La pared celular define la forma del procarionte y es esencial para sobrevivir en cualquier medio, e incluso se le considera un rasgo de valor adaptativo (Kysela *et al*., 2016). Si bien la forma como se organizan las membranas, la pared celular y la capa-S es variable, lo que hace de la envoltura celular un rasgo sofisticado y por lo general distintivo de los linajes procariontes (Sutcliffe, 2010; Albers y Meyer, 2011). No hay duda de que, así como es esencial para la supervivencia de todas las bacterias y las arqueas modernas, lo era para sus ancestros.

Una segunda clasificación basada en las categorías funcionales de KEGG (Figura 10), reveló nuevos patrones que con la primera clasificación no fueron tan evidentes, aunque tuvo la desventaja de que solamente se pudo trabajar con el 58% de las familias proteínicas conservadas. Aún así, hubo congruencia con los resultados obtenidos a partir de las categorías COG, en el sentido de que después de la traducción, presentan niveles medios de conservación, la biosíntesis de diversos aminoácidos y de coenzimas, así como también la replicación, la reparación, la recombinación y los procesos asociados a la biosíntesis de la envoltura celular, confirmando así el sesgo cualitativo en las reconstrucciones.

El resultado más relevante de aplicar el esquema clasificatorio KEGG a los catálogos ancestrales, fue que el metabolismo de nucleótidos presentó niveles de conservación

similares a los observados para la función de traducción (Figura 10). Es decir, el metabolismo de purinas y de pirimidinas es universal, y se encuentra excepcionalmente conservado al interior de los catálogos ancestrales, lo que reafirma la idea de ancestros capaces de sintetizar y modificar ácidos nucleicos, además de que habla de un mundo donde el RNA y los ribonucleótidos eran relevantes (Delaye *et al*., 2005). Si bien el resultado anterior se podría sospechar, dado que la categoría COG de metabolismo de nucleótidos presentó niveles medios de conservación en el núcleo relajado de las reconstrucciones (Figura 8), la señal se diluye debido a que son familias proteínicas asociadas a múltiples funciones. De hecho, al analizar el contenido de la categoría se detectó que la mayoría de las enzimas están asignadas también a otras funciones, como el metabolismo de ciertos aminoácidos, coenzimas, carbohidratos, en la replicación y la reparación del DNA, e incluso algunas tienen la palabra bifuncional en el nombre asignado. Además, al interior de otras categorías funcionales fue común identificar a enzimas asociadas a su vez al metabolismo de nucleótidos. Como ya se mencionó, este patrón es de esperarse si se considera que las diferentes rutas metabólicas tienen distinta antigüedad, y que las enzimas preexistentes en un momento dado constituyen la materia prima para el ensamblaje de nuevas vías (Peretó, 2011).

Un aspecto que se reconoce en los trabajos bioinformáticos, es la redundancia que existe a nivel de la anotación funcional de las proteínas, como se observó en mayor medida con la clasificación de KEGG. Es decir, que una secuencia está asociada, además de a las categorías generales, a más de una ruta metabólica o función (Figura 11). Tal fue el caso de enzimas como propanoil-CoA C-aciltransferasa (exclusiva de reconstrucciones de Archaea), NAD+ aldehído deshidrogenasa, acetil-CoA acetiltransferasa y aminotransferasas de clase I y clase II, entre muchas otras. La presencia de dichas familias proteínicas con multiplicidad funcional en los catálogos ancestrales, sugiere que, así como son muy importantes para las especies modernas, también lo eran para los ancestros. Sin embargo, queda por resolver cuál de todas las funciones que actualmente exhiben corresponde mejor a la ancestral.

Además, la clasificación KEGG facilitó la detección de los transportadores ABC al interior de los catálogos ancestrales (Figura 10), puesto que en la clasificación COG están

distribuidos en múltiples categorías. Este resultado sugiere que son familias proteínicas muy antiguas, que están vinculadas a la presencia de nucleótidos como el ATP. Se sabe que presentan una gran diversidad funcional en cuanto al sustrato que transportan (Wilkens, 2015), y su presencia en las reconstrucciones nos permite sugerir que se diversificaron muy temprano en la evolución, lo que explica su presencia en los ancestros de los linajes de las bacterias y arqueas modernas. Aunado a esto, las familias proteínicas conservadas más conspicuas en la categoría de energía fueron ATPasas, tanto ATPasas transportadoras de protones como ATPsintasas rotoras, lo que indica la presencia de dichos sistemas en los ancestros.

Otro patrón que se hizo visible fue que la función de percepción de quorum está conservada en la mayoría de los linajes. Esto sugiere que la comunicación entre procariontes es fundamental, no sólo entre organismos modernos sino también para los ancestros, los cuales muy posiblemente ya requerían la habilidad de comunicarse y actuar como un grupo (sociabilizar), tal vez en espacios como biofilms o estromatolitos.

Ya se mencionó la notable conservación al interior de las reconstrucciones ancestrales, de las familias proteínicas asociadas con la producción y el mantenimiento de la envoltura celular. Aunado a esto, gracias al segundo esquema clasificatorio se detectó que las diferencias entre Archaea y Bacteria, en cuanto a la composición y la biosíntesis de sus membranas y de la pared, están conservadas al interior de los catálogos ancestrales. Tal es el caso de la biosíntesis de peptidoglicanos, que se presentó de forma exclusiva en los catálogos de los linajes bacterianos (Figura 10), mientras que en los ancestros de arqueas se detectaron enzimas de la biosíntesis de isoprenoides (ruta del mevalonato) y de lípidos exclusivos de Archaea. Este resultado advierte que los ancestros de los linajes modernos ya heredaron la envoltura celular propia de cada dominio procarionte, rasgos que por tanto se establecieron hace mucho tiempo.

Finalmente, podemos decir que el esquema de clasificación KEGG aportó un mayor nivel de granularidad al análisis funcional, al menos en ciertos casos. Por ejemplo, utilizando el esquema de COG, se pudo notar que la conversión y producción de energía se mantiene

dentro de las categorías con niveles medios a bajos de conservación (Figuras 8 y 9), lo que implica que en los catálogos se recuperan también cuestiones asociadas con procesos energéticos. Pero, además, con las categorías KEGG fue posible determinar que la fosforilación oxidativa está muy conservada, principalmente en Deltaproteobacteria, y en Thaumarchaeota, Cyanobacteria, Betaproteobacteria y Epsilonproteobacteria, entre otros linajes. Así como también diversas familias proteínicas en el catálogo de Cyanobacteria están relacionadas con la fotosíntesis y otras en Thaumarchaeota con la fijación de carbono (Figura 10). Otro ejemplo es la categoría de transporte y metabolismo de coenzimas, la cual presentó niveles elevados de conservación (Figuras 8 y 9), a lo que se puede agregar que el metabolismo de porfirinas y clorofilas está muy conservado particularmente en los linajes Thaumarchaeota, Alphaproteobacteria y Cyanobacteria (Figura 10), o que la elevada conservación detectada en Aquificales no se concentra en una categoría, sino en el metabolismo de diversas coenzimas.

Otra vertiente del uso de KEGG fue la clasificación enzimática, cuyo primer resultado consistió en que el 48% del total de las familias proteínicas conservadas, en todos los catálogos ancestrales, son enzimas. Dicha estimación es a la baja pues podría sumarse una fracción del ≈18% de las proteínas pobremente caracterizadas, pero además porque al interior de la base de datos se identificaron muchos casos, que no contaban con EC y cuya asignación funcional corresponde a una enzima. Por lo tanto, es muy posible que el número de enzimas esté subestimado.

Aunado a esto, se observó que la conservación se distribuye principalmente en la clase de transferasas (EC2), seguida por enzimas de tipo ligasas (EC6), hidrolasas (EC3) y oxidorreductasas (EC1) (Figura 12). Las transferasas son la clase enzimática más abundante en *Escherichia coli* (Ouzounis y Karp, 2000), y de manera sobresaliente, las enzimas de este tipo presentes en los catálogos ancestrales están distribuidas en todas las categorías funcionales COG. Es decir, que transferir un grupo funcional de un sustrato a otro es una función bioquímica básica para cualquier célula. Por el contrario, la categoría de ligasas es la menos abundante en *E. coli* (Ouzounis y Karp, 2000), y es la segunda mejor conservada en los catálogos. Este resultado se debe, en parte, a que contiene a las familias de las

enzimas aminoacil-tRNA sintetasas, las cuales están muy conservadas (Figura anexo 6), aunque ciertamente también contiene a otras enzimas. De hecho, en todas las clases funcionales (incluidas las liasas EC4 e isomerasas EC5) hubo enzimas con niveles muy elevados de conservación, incluso comparables con las proteínas ribosómicas (Tabla 3), las cuales participan en funciones como el metabolismo de nucleótidos, de coenzimas, de aminoácidos, la glucólisis y gluconeogénesis, la biosíntesis de ácidos grasos y de peptidoglicanos, la maduración de RNA (tRNA, rRNA), la traducción, la replicación, la reparación, la recombinación y la transcripción. Esto demuestra la presencia de familias de enzimas muy antiguas, que probablemente ya estaban en los ancestros.

Sin duda, la segunda clasificación aportó datos que complementaron a la primera, contribuyendo así a una mejor interpretación de las funciones conservadas al interior de las reconstrucciones ancestrales. En ambos esquemas se observó la tendencia general de que mientras más numerosa sea la reconstrucción ancestral, entonces contendrá más familias proteínicas relacionadas con el metabolismo, tanto energético como de los distintos biopolímeros. De hecho, conforme el tamaño del catálogo ancestral se reduce, entonces las proteínas que se detectan conservadas se relacionan principalmente con el procesamiento de la información genética. Tal es el caso por ejemplo de Firmicutes, linaje cuya reconstrucción es pequeña y contiene muy pocas familias proteínicas asociadas al metabolismo (Figuras 8, 9 y 10). Sin embargo, es notable que el resto de los catálogos ancestrales, incluso si son reducidos, contienen proteínas que participan en el metabolismo (Figuras 8, 9 y 10). Si bien esto depende de los niveles de astringencia que se establecen para considerar una proteína homóloga a otra, y por lo tanto para incluirla en el catálogo ancestral, es un hecho que hay una pérdida de señal importante a nivel de la secuecia de las proteínas. Además, en las reconstrucciones de menor tamaño fue muy evidente que se recuperan procesos esenciales incompletos, y lo mismo sucede, aunque en menor medida, en los catálogos más numerosos, los cuales presentan una conservación heterogénea al interior de las diversas categorías funcionales (Figuras 8, 9 y 10). Las observaciones anteriores indican que conforme el catálogo ancestral se hace más nutrido, entonces los procesos se van “completando”, y las reconstrucciones se asemejan más a las especies

modernas. No obstante, los catálogos ancestrales siempre presentan funciones esenciales o rutas metabólicas fragmentadas, patrón que se exacerba conforme se reduce su tamaño al mínimo. Estos resultados indican que los catálogos ancestrales están subestimados de forma importante. Dicho patrón se puede originar si las proteínas evolucionan a ritmos diferentes, y sugiere también que la pérdida de genes es un proceso central en la evolución del genoma de los organismos procariontes.

La estimación del tamaño del núcleo genómico (Tabla 4, Figura 13) mediante el análisis de remuestreo aleatorio, arrojó valores similares a los obtenidos en los núcleos relajados de los catálogos ancestrales (Tabla 2). Tal como se ha descrito anteriormente (Tettelin *et al*., 2005), el núcleo genómico se comporta como una curva asintótica que decae muy rápido y luego se estabiliza, conforme se agregan más genomas a la comparación. En el caso del presente estudio, se estabilizó en valores que rondan un intervalo de entre ≈650 a ≈110 genes, dependiendo del linaje. En el mismo orden de magnitud, estudios previos establecieron el núcleo genómico de Archaea en ≈300 (Makarova y Koonin, 2003) y el de Bacteria en ≈250 familias proteínicas (Lapierre y Gogarten, 2009). A nivel de género, se ha visto que el tamaño del núcleo genómico puede variar desde 500 hasta 2800 proteínas (Vernikos *et al*., 2015). Como ya se discutió, estos números no se deben interpretar como el tamaño real del genoma ancestral de dichos linajes. Se ha calculado que en Bacteria, aproximadamente 8% del genoma de cualquier especie corresponde al núcleo genómico (Lapierre y Gogarten, 2009). En el presente estudio, el porcentaje de reconstrucción por lo general no superó el 15%, y también se puede interpretar como una medida del núcleo genómico al interior de los genomas procariontes.

Por otro lado, el análisis mostró que la mayoría de los linajes tienen un pangenoma abierto (ver Figura 13), lo cual es consistente con lo publicado para el dominio Bacteria (Lapierre y Gogarten, 2009), así como también resultó ser el caso de Archaea. Esto significa que en teoría el pangenoma seguirá creciendo, conforme más especies se agreguen al análisis. Los pangenomas de mayor tamaño se encontraron en grupos bacterianos como Actinobacteria, Alphaproteobacteria, y Deltaproteobacteria y superan los 50,000 genes, mientras que los de menor tamaño corresponden, de manera interesante, a los linajes

hipertermófilos Aquificae y Thermotogae, apenas superando las 4000 y 5000 proteínas respectivamente. Dicha variación podría sugerir que no todos los pangenomas se han expandido de manera tan drástica.

Los organismos de los linajes Aquificae y Thermotogae gustan de habitar sitios con temperaturas altas. Sus características como genomas pequeños (Counts *et al*., 2017), poca variación en el tamaño de su proteoma (Figura 6), un nivel de conservación muy elevado del núcleo genómico (Tabla 2, Figuras 4 y 5), pangenomas relativamente pequeños (Tabla 4), y por ende una diversidad metabólica baja, sugieren que dichos linajes están adaptados genómicamente a la termofilia extrema. Dicha condición impone un régimen de estabilidad en el genoma, que a su vez se ve reflejado en el pangenoma. Los resultados muestran que este par de taxa son excepcionales, ya que generalmente las especies de un linaje han tenido la oportunidad de colonizar una mayor cantidad de nichos, y por ende se han diversificado sus posibilidades metabólicas y su pangenoma, y en ese sentido son más versátiles.

Si se considera que el pangenoma contiene genes prescindibles, no universales, que son responsables de la diversidad de un taxón (Vernikos *et al*., 2015), entonces *a priori* podemos suponer que los grupos con un mayor pangenoma, poseen una mayor diversidad metabólica. Los resultados de las reconstrucciones muestran la tendencia general de que a mayor diversidad metabólica o pangenoma, menor es el tamaño del catálogo ancestral. Y sucede lo contrario en los grupos con pangenomas reducidos, los cuales destacan por tener reconstrucciones ancestrales relativamente más numerosas.

La descripción general de las proteínas conservadas al interior de las reconstrucciones ancestrales, evidenció una característica muy interesante, y es que las familias proteínicas, si bien son muy antiguas, no necesariamente se originaron en el mismo periodo (Becerra *et al*., 2007). Dicho de otro modo, en los catálogos ancestrales algunas proteínas son más antiguas, respecto de otras cuyo origen se antoja más reciente. Por ejemplo, algunas secuencias anotadas como virales o relacionadas a antibióticos, podrían tener un origen relativamente reciente. Las proteínas que participan en la fotosíntesis

oxigénica son exclusivas de los ancestros de Cyanobacteria y, de hecho, un criterio para establecer la antigüedad relativa de proteínas es si requieren o no del oxígeno molecular (Rivas *et al*., 2018). El hábito hipertermófilo favoreció una conservación muy marcada de las proteínas en los linajes Thermotogae y Aquificae (Figura 9). Otras familias proteínicas son incluso más antiguas, de épocas del LCA o pre-LCA, como las aminoacil-tRNA ligasas, los factores de elongación, las ATP sintetasas, los transportadores ABC, y lo relacionado con el ensamblaje y el funcionamiento de la maquinaria ribosomal (Becerra *et al*., 2007; Delaye *et al*., 2005; Cantine y Fournier, 2018).

La observación de que los catálogos ancestrales presentan proteínas cuya edad relativa difiere, se refuerza con los resultados del análisis del tamaño de las secuencias al interior de las reconstrucciones. La mayoría de las proteínas conservadas con una longitud menor a 70 aminoácidos son proteínas de las subunidades del ribosoma, las cuales posiblemente se originaron muy temprano en la evolución celular, durante las etapas finales del establecimiento del ribosoma (Petrov *et al*., 2015). Por analogía, dentro del conjunto de las proteínas conservadas, las cuales exhiben una variabilidad importante respecto a su longitud (Figura 16), las más pequeñas podrían ser más antiguas que el resto de las secuencias de mayor tamaño (o por lo menos son candidatas a serlo). Aunado a esto, las secuencias más largas sugieren que los ancestros de bacterias y arqueas ya contaban con un sistema de traducción muy eficiente, que permitía la síntesis de enzimas de gran tamaño y/o multiméricas.

En resumen, la comparación de las reconstrucciones ancestrales de los diversos linajes con las especies modernas, demostró que son varios los factores que afectan el tamaño de los catálogos ancestrales. Entre ellos se pudieron identificar el tamaño de la muestra, es decir, el número de especies consideradas (Tettelin *et al*., 2005), la diversidad fisiológica de los organismos al interior de la muestra, la varianza tanto en el tamaño del genoma como del pangenoma (reflejo de la diversidad metabólica), y la distancia filogenética, entendida como la cantidad de historia evolutiva asociada a un linaje.

De manera consistente, en todas las reconstrucciones se detectó un sesgo doble. Un sesgo cuantitativo, porque las familias proteínicas conservadas en realidad representan solamente una fracción de los genomas de los ancestros, y un sesgo cualitativo, debido a que la conservación no es homogénea al interior de las categorías funcionales. En pocas palabras, el método comparativo proporciona una estimación mínima, de algunas funciones, presentes en los genomas de los ancestros de los linajes procariontes, lo cual no quiere decir que los ancestros contaban únicamente con dichos elementos, sino que representa el alcance de la comparación de la estructura primaria de las proteínas. Como consecuencia, los catálogos ancestrales compartieron la característica de que contienen procesos celulares incompletos, ya sean informativos o metabólicos. Finalmente, si bien todas las familias proteínicas son muy antiguas, y se postula su presencia en los ancestros, es evidente que no todas tienen la misma edad relativa, así como que no están sujetas a los mismos mecanismos de evolución al interior del genoma.

Contraste de las reconstrucciones ancestrales con el caso del LCA

El objetivo de construir y analizar los catálogos ancestrales de los linajes principales de Archaea y Bacteria, es que los resultados, las tendencias y los patrones observados, son directamente comparables con el caso de la reconstrucción del LCA. Debido a que el LCA representa la etapa previa a la diversificación de los dominios primarios, entonces precede a los ancestros de los linajes procariontes modernos. Esta lógica permite evaluar el alcance y las limitaciones de la genómica comparada, como herramienta para reconstruir de forma directa a los genomas ancestrales, y en particular al LCA.

En el presente trabajo se hizo patente que los factores que afectan a las reconstrucciones ancestrales, tienen una relación de proporcionalidad inversa con el tamaño del catálogo ancestral. Es decir, cuando se compara a una muestra de organismos con el objetivo de reconstruir el genoma de su ancestro, mientras mayor es el tamaño de la muestra, y/o mayor diversidad metabólica contiene, y/o mayor distancia filogenética hay entre las especies, entonces menor es el número de familias proteínicas comunes o

conservadas. Por consecuencia, dado que el LCA es el ancestro de la biota moderna, engloba a toda la biodiversidad terrestre y representa el nodo más alejado de la filogenia universal, entonces su reconstrucción inevitablemente resulta en un catálogo reducido. De hecho, el método comparativo arroja, por lo general, estimaciones menores (ver Kyrpides *et al*., 1999; Koonin, 2003; Delaye *et al.,* 2005; Yang *et al.,* 2005; Sobolevsky y Trifonov, 2006; Ranea *et al.,* 2006 en Tabla 1), respecto de metodologías que combinan la genética comparada con un enfoque filogenético (ver Harris *et al.,* 2003; Ouzounis *et al.,* 2006; Kim y Caetano-Anollés 2011 en Tabla 1), o con aquellas aproximaciones basadas totalmente en cladística molecular (ver Mirkin *et al.,* 2003, Tuller *et al.,* 2010; Kannan *et al.,* 2013; Weiss *et al.,* 2016 en Tabla 1).

A pesar de las diferencias numéricas entre los catálogos (Tabla 1), llama la atención que ninguna de las reconstrucciones que se han propuesto del LCA se ajusta al límite de vida libre de los procariontes modernos, el cual ronda las 1300 proteínas por genoma (Islas *et al*., 2004; Martínez-Cano *et al*., 2014). En los resultados, las diferencias numéricas entre las especies modernas y las reconstrucciones ancestrales, demostraron que el método comparativo tiene un sesgo cuantitativo importante. Ningún catálogo ancestral de los linajes modernos de Archaea y Bacteria, supera un porcentaje de reconstrucción del 30%, éste es en promedio del 15%, y una tercera parte están entre el 1% y el 5% (Tabla 2). Si el LCA se remonta más atrás en el tiempo, se puede suponer que las reconstrucciones mediante genómica comparada, es decir, las propuestas de 63 genes (Koonin, 2003), 115 dominios proteínicos (Delaye *et al.,* 2005), 49 plegamientos (Yang *et al.,* 2005), 20 motivos (Sobolevsky y Trifonov, 2006), y 140 dominios ancestrales (Ranea *et al.,* 2006), no superan dichos porcentajes. Visto así, el método comparativo permite establecer un mínimo de genes o proteínas que podrían tener los ancestros, pero es claro que la cantidad de información que se recupera es limitada, y que los catálogos representan solamente una fracción del genoma ancestral y de las funciones presentes en el LCA.

El segundo sesgo que se detectó en el análisis fue cualitativo, y se refiere a que la conservación al interior de las categorías funcionales no es homogénea. Nótese que debido al sesgo cuantitativo, ya en los catálogos ancestrales de mayor tamaño, las rutas

metabólicas se presentan incompletas, tendencia que se exacerba en los catálogos reducidos. Pero, de manera sorprendente, este patrón también se observa en procesos esenciales, como en las proteínas ribosomales o en las tRNA ligasas (Figuras anexo 4, anexo 5 y anexo 6). Lo anterior sugiere que hay procesos que conllevan a subestimar los catálogos ancestrales, como pueden ser tasas de evolución distintas o la pérdida de genes, fenómeno común en la evolución de los procariontes. Más aún, los resultados mostraron que conforme disminuye el tamaño de las reconstrucciones ancestrales, se recuperan menos familias proteínicas asociadas al metabolismo, y más relacionadas con el procesado de la información genética. Estas observaciones explican la naturaleza de las propuestas del LCA (Tabla 1; Becerra *et al*., 2007), las cuales siempre contienen procesos biológicos incompletos, y recuperan casi de forma exclusiva elementos que participan en la traducción, la transcripción, la replicación, la reparación del DNA, y en menor medida proteínas asociadas a membrana (Becerra *et al*., 2007), lo anterior no implica que el LCA contaba solamente con dichas funciones, como suele interpretarse, sino que es lo que el método permite detectar.

A pesar de las limitaciones y de los sesgos descritos en la reconstrucción directa de ancestros mediante genómica comparada, las proteínas que se identificaron al interior de los catálogos ancestrales permitieron establecer que los ancestros de los distintos linajes de Archaea y Bacteria exhibían características equivalentes a las de las especies procariontes modernas. Ciertamente, conforme nos vamos más atrás en el tiempo, las reconstrucciones se hacen más pequeñas y la conservación se concentra en ciertas funciones, aunque de manera sorprendente algunas familias proteínicas datan de épocas del LCA o pre-LCA (Becerra *et al*., 2007). Sin embargo, a partir de las funciones presentes en los ancestros de los grupos bacterianos y arqueanos, es razonable concluir que su ancestro, el LCA, también tenía la naturaleza típica de un procarionte (Becerra *et al*., 2007). Los resultados sugieren que el LCA ya contaba con una envoltura celular funcional y la capacidad de sintetizarla, con un genoma de DNA y la capacidad de mantenerlo y replicarlo, con un sistema muy eficiente de traducción mediante ribosomas, así como con proteínas estructurales y enzimas metabólicas, muchas de las cuales eran complejas, ya sea grandes,

multiméricas o con cierto grado de especificidad, además de que ocupaban coenzimas derivadas mayormente de nucleótidos.

Una polémica que se estableció en el campo de estudio del LCA es la naturaleza de su material genético. Los catálogos de los ancestros de los linajes modernos, indican que éstos ya contaban con genomas de DNA y con toda la maquinaria proteínica para el mantenimiento de los genes. Si los ancestros de Bacteria y Archaea ya poseían genomas de DNA, la conclusión más parsimoniosa es que el LCA ya contaba con dicho rasgo, una idea a la que han llegado todos los análisis del LCA (Tabla 1), con algunas excepciones (Mushegian y Koonin, 1996; Kim y Caetano-Anollés 2011). La genómica comparada sugiere que el DNA se estableció como la molécula que almacena la información genética antes de la divergencia entre Bacteria y Archaea, y antes del LCA (Lazcano *et al*., 1992).

La traducción es el proceso mejor conservado a nivel de secuencia primaria, lo cual es congruente con la idea de que el LCA ya contaba con ribosomas muy eficientes, y con un flujo de la información genética equivalente al de un procarionte moderno. De manera interesante, al interior de los catálogos ancestrales, el metabolismo de nucleótidos resultó casi tan conservado como la traducción. Esto indica que los ancestros de arqueas y bacterias tenían la capacidad de sintetizar y modificar nucleótidos, y a su vez apoya la noción de un LCA equipado con RNA, DNA y sus derivados.

Aunque el caso del metabolismo de nucleótidos resultó sobresaliente, al interior de los catálogos ancestrales se identificaron diversas proteínas asociadas al metabolismo de aminoácidos, de coenzimas, de carbohidratos y de lípidos. Es difícil suponer un LCA de naturaleza procarionte sin dichas capacidades. Los resultados mostraron cómo las funciones asociadas al metabolismo desaparecen de los catálogos ancestrales conforme más atrás nos vamos en la filogenia universal, lo que explica su ausencia de las distintas reconstrucciones del LCA (ver Becerra *et al*., 2007). Esta tendencia puede deberse a la historia evolutiva de las familias proteínicas, las cuales no pertenecen al núcleo genómico y son más propensas a eventos como la pérdida de genes, la sustitución no ortóloga o la acumulación de mutaciones a nivel de secuencia, las cuales diluyen la señal de homología.

Por lo tanto, para detectar enzimas del metabolismo que pudieron haber estado en el LCA, además de mejorar la calidad de las bases de datos, es necesario el uso de metodologías diseñadas específicamente para reconocer homología lejana. Un ejemplo es la comparación de estructuras terciarias, como ya se ha reportado para el catabolismo de purinas (Rivas *et al*., 2018).

La descripción detallada de los catálogos ancestrales evidenció que una buena proporción de las familias proteínicas se relacionan de alguna u otra manera con la envoltura celular. Es razonable suponer un LCA con una envoltura funcionalmente semejante a la de sus descendientes modernos, como lo sugiere la presencia de ATP sintetasas y transportadores ABC en las distintas reconstrucciones del LCA. Por lo tanto, la propuesta de un LCA con una membrana del tipo bicapa fosfolipídica (Lombard *et al*., 2012a) es congruente con los resultados. La observación de que las diferencias que exhiben bacterias y arqueas en su membrana están conservadas al interior de los catálogos ancestrales, tal como se había descrito en Archaea (Lombard *et al*., 2012b), sugiere que dicho rasgo se estableció muy temprano. Debido al sesgo cualitativo inherente a las reconstrucciones ancestrales, no es posible postular más detalles del tipo de envoltura que tendría el LCA, aunque por ejemplo se ha sugerido que tendría ambos tipos de fosfolípidos en sus membranas (Peretó *et al*., 2004; Jain *et al*., 2014). Cabe notar que los procesos energéticos de la célula están asociados a las membranas, por lo que es notable el vínculo entre el tipo de membrana y de envoltura y el metabolismo general de la célula, todo lo cual desaparece de los catálogos ancestrales conforme más atrás nos vamos en la filogenia universal. Esto sugiere que entender la divergencia entre Archaea y Bacteria a partir del LCA, nos exige comprender el origen de las diferencias en su envoltura celular, y a su vez, esto se conecta con el tipo de metabolismo que tendrían tanto los ancestros, como el LCA.

Una pregunta que surge con los resultados de la tesis es: ¿porqué al comparar distintos genomas procariontes en búsqueda de elementos conservados, emerge un patrón de conservación diferencial? La respuesta es la gran variabilidad que exhiben las especies procariontes al interior de sus genomas. En los años recientes, el descubrimiento y estudio del pangenoma ha hecho patente la enorme diversidad genética de las cepas bacterianas,

por ejemplo, en *Escherichia coli* asciende a casi 100,000 proteínas (Land *et al*., 2015), mientras que la bacteria marina *Prochlorococcus,* tiene un pangenoma estimado de 85,000 genes, aun cuando su genoma es de los más pequeños para especies de vida libre (Biller *et al*., 2014). Ello sugiere que los análisis pangenómicos a nivel de linaje realizados en el presente estudio (Tabla 4) podrían estar subestimados de forma importante, puesto que los más numerosos apenas superaron los 50,000 genes (Actinobacteria, Alphaproteobacteria y Deltaproteobacteria). Esto nos habla de que el espacio muestral bajo el que estamos buscando genes o proteínas que posiblemente estaban en el LCA no ha sido explorado por completo. Si el pangenoma es un registro histórico de las especies o de los linajes, y su naturaleza es infinita, claramente queda mucha diversidad procarionte por conocer, lo que podría afectar nuestras interpretaciones del LCA.

Sin embargo, a pesar de que los pangenomas pueden ser abiertos, todos los genomas procariontes presentan un núcleo genómico, es decir, un conjunto de genes que están sorprendentemente conservados, que son muy antiguos, que se transmiten de manera vertical y que soportan al resto del genoma (Lapierre y Gogarten, 2009), por lo que ninguna célula moderna puede prescindir de dicho núcleo. Los catálogos ancestrales de los linajes de Bacteria y Archaea, así como las distintas reconstrucciones del LCA, confirman que el núcleo genómico data de épocas pre-LCA, pues tanto el LCA como los ancestros de los linajes modernos ya contaban con dicho rasgo, lo que demuestra la naturaleza conservativa de la evolución (Lapierre y Gogarten, 2009). Precisamente, la genómica comparada es un método que permite revelar dicho núcleo proteínico conservado, el cual claramente se reduce, cuantitativa y cualitativamente, conforme más atrás nos vamos en la filogenia universal.

Adicionalmente al núcleo genómico, el genoma procarionte se conforma de genes que son prescindibles, en el sentido de que se pueden perder o sustituir con el paso del tiempo, además de que pueden movilizarse mediante transferencia horizontal. Se conocen como el genoma accesorio y son justamente los responsables de las características distintivas de los taxa procariontes (Lapierre y Gogarten, 2009). Así, los genomas tanto de bacterias como de arqueas, son una especie de “recipiente *casi* vacío”, que puede adquirir

y/o perder genes como respuesta a los retos que impone el ambiente. Dicha versatilidad genómica podría ser la razón del éxito evolutivo de los descendientes directos del LCA, puesto que les ha permitido habitar prácticamente cualquier lugar del planeta desde hace más de 3,000 Ma. Sin embargo, los resultados del presente análisis sugieren que mientras más tiempo pasa, disminuye la probabilidad de detectar la señal del núcleo accesorio y de los genes que definieron las características metabólicas de los ancestros. En principio estos límites de la genómica comparada, se pueden superar con metodologías que construyen modelos conciliando las filogenias moleculares y la historia de las especies, así como mediante la comparación de estructuras terciarias, la cual ha demostrado que puede detectar homología lejana en proteínas relacionadas con el metabolismo (Rivas *et al*., 2018).

A pesar de que la genómica comparada no puede ir más allá del periodo de evolución celular en el que la traducción ya estaba operando (Becerra *et al*., 2007; Rivas *et al*., 2018), pues dicha etapa representa el límite temporal del análisis, algunos resultados de las reconstrucciones ancestrales permiten hacer ciertas inferencias hacia el pasado, como a continuación se discute.

Si la traducción ya estaba operando, entonces al LCA le antecede el periodo durante el cual se originó la maquinaria de síntesis proteínica, es decir, el ribosoma. Se ha propuesto un modelo de evolución por acreción, que favoreció la coevolución de las proteínas y el RNA en las células (Petrov *et al*., 2015). Dicha propuesta comienza con el RNA como actor principal, el cual evitaba su degradación al formar estructuras secundarias, tipo tallo asa, estabilizadas con cationes metálicos. En un inicio, lo que sería la subunidad grande y la chica evolucionaban de manera independiente, la mayor presentaba actividad catalítica no específica, pues favorecía la condensación de oligómeros de aminoácidos y oxiácidos sin un código, mientras que la pequeña posiblemente se unía a hebras sencillas de RNA. Posteriormente, ambas subunidades se asocian, con participación de los tRNAs, los cuales reclutaban también a oligómeros de RNA (proto-mRNA), lo que resultó en una catálisis más eficiente y con productos de mayor longitud. Con el tiempo, se refuerza la integración de ambas subunidades y de los tRNAs y mRNAs, formando una ribozima eficiente, que

eventualmente desarrollará la habilidad decodificante. Una vez formada la región decodificadora, el código genético se fue optimizando y el ribosoma entró en una etapa de "proteinización" (Petrov *et al*., 2015).

Entonces, la evolución del ribosoma en el contexto de sistemas celulares más sencillos, en los que el RNA tendría un papel predominante (Joyce, 2002), derivó en células donde las proteínas adquirieron roles cada vez más centrales. Durante dicho proceso, el rRNA, al ser parte integra del ribosoma, se quedó "congelado" en el tiempo (Petrov *et al*., 2015). La idea de que la evolución del ribosoma favoreció la formación de proteínas de mayor tamaño, es consistente con los resultados de los catálogos ancestrales, ya que las proteínas más pequeñas conservadas corresponden a proteínas ribosómicas, las cuales al igual que el rRNA están "congeladas" al interior de las subunidades. Ahora bien, un mayor tamaño de las proteínas propició a su vez la aparición de nuevas funciones. Por ejemplo, una función inaccesible para el RNA es la capacidad de atravesar las membranas lipídicas (Fournier y Gogarten, 2007). Este momento fue crítico para la sustitución de los procesos mediados por RNA, y resultó en la "proteinización" de los procesos celulares (Fournier y Gogarten, 2007; Petrov *et al*., 2015). El hecho de que al interior de los catálogos ancestrales, tanto del LCA (ver Becerra *et al*., 2007) como de los principales linajes procariontes, se identifican una gran cantidad de proteínas conservadas que interactúan de manera directa con el RNA, que utilizan enzimas derivadas de ribonucleótidos, o que toman parte en el metabolismo de nucleótidos, se puede interpretar como un eco del mundo del RNA, el cual posiblemente antecedió al mundo de DNA, RNA y proteínas (Delaye *et al*., 2005; Becerra *et al*., 2007).

Woese propuso el concepto de temperatura evolutiva y la definió como el número de cambios que acepta un sistema genético (Woese, 1998). Es claro que los rasgos que hoy representan el núcleo genómico de todo procarionte se "congelaron" hace muchísimo tiempo, pero ¿por qué? Una posible explicación para el proceso más conservado, la traducción, es que por ser un mecanismo que depende de diversos genes, y de la coordinación directa de proteínas y de RNAs, está sujeta a presiones de selección muy fuertes, las cuales no han permitido cambios mayores a nivel de secuencia, y menos a nivel

de estructura terciaria. Sin embargo, ni la replicación, ni la transcripción, ni el metabolismo cumplen con esta condición, lo que sugiere que la proteinización de los procesos celulares trajo consigo diversas formas de evolución, tanto al nivel de las proteínas, como del genoma.

Al interior de los catálogos ancestrales de los linajes principales de Archaea y Bacteria fue posible detectar a las polimerasas, lo cual sugiere que son proteínas muy antiguas, sin embargo, conforme los catálogos se reducen, las polimerasas se salen del núcleo conservado, lo que explica su ausencia en las reconstrucciones del LCA (ver Becerra *et al*., 2007). De manera interesante son las proteínas conservadas de mayor tamaño presentes en las reconstrucciones ancestrales, por lo que es lógico suponer que dichas secuencias han acumulado mutaciones, siempre y cuando no afecten su estructura y por lo tanto su función. De hecho, análisis estructurales han demostrado la homología de las proteínas que contienen el dominio *palm*, como las DNA polimerasas I y II, pol Y, reverso transcriptasas, RNA polimerasas virales dependientes de RNA, RNA polimerasas dependientes de DNA de mitocondrias y virales, así como primasas y proteínas prim-pol presentes en arqueas y eucariontes (Steitz, 1999; Iyer *et al*., 2005; Becerra *et al*., 2007; Guilliam *et al*., 2015; Jácome *et al*., 2015). Se ha hipotetizado que dicho dominio representa un componente de una polimerasa ancestral, que funcionaba como replicasa y transcriptasa en etapas tempranas de evolución celular (Becerra *et al*., 2007). Es posible que después del surgimiento del dominio catalítico poco específico, al aumentar el tamaño de la secuencia, se originaron o incorporaron nuevos dominios funcionales, los cuales permitieron una mayor especificidad de la enzima y/o aportaron un papel regulatorio. De hecho, la estructura de las enzimas multidominio RNAP demuestra que existe un núcleo conservado en las RNAP de bacterias, arqueas y eucariontes, lo cual apoya la noción de un aparato de transcripción que ganó especificidad conforme se agregaron módulos funcionales (Korkhin *et al*., 2009). En suma, las polimerasas representan familias proteínicas muy antiguas, que se diversificaron muy temprano (quizá a la par de la proteinización del ribosoma), e ilustran cómo se estableció la naturaleza modular de las proteínas como el motor de su evolución.

Cabe mencionar que existen otros plegamientos, además del motivo de reconocimiento de RNA (RRM) del dominio *palm*, presentes en los dominios catalíticos de algunas polimerasas de ácidos nucleicos. Por ejemplo el TOPRIM (topoisomerasa-primasa) de las DNA-G primasas de bacterias, primasas pequeñas de arqueas y bacterias, topoisomerasas de tipo 1 y 2, y nucleasas de la familia RecR (Aravind *et al*., 1998), el plegamiento doble-psi-barril-beta presente en la polimerasa de RNA dependiente de RNA (RdRP), en la polimerasa de DNA dependiente de RNA (DdRP), y en una polimerasa del sistema de RNAi en eucariontes (Iyer *et al*., 2003; Salgado *et al*., 2006) y por último el dominio presente en la DNA pol III bacteriana y en las transferasas terminales de DNA o RNA independientes de templado (super familia X) (Lamers *et al*., 2006). Se ha propuesto que dichos dominios también son muy antiguos, quizá de etapas del LCA (Iyer *et al*., 2003; Salgado *et al*., 2006), y ejemplifican otro punto relacionado con la evolución proteínica. Y es que cuando una proteína realiza la misma función que otra, puede haber redundancia y/o una sustitución no ortóloga cuando la segunda termina por excluirse o perderse del genoma, lo que implica que a lo largo de la evolución las proteínas se pueden sustituir mientras las funciones celulares estén cubiertas.

Así como la replicación y la transcripción, el metabolismo también está basado en proteínas con actividad catalítica. De manera similar a las polimerasas, se ha propuesto que en etapas tempranas las enzimas eran poco específicas, y que las rutas se fueron ensamblando a partir de otras preexistentes, en un proceso denominado reclutamiento enzimático (Jensen, 1976; Lazcano y Miller, 1999; Peretó, 2011). Es importante aclarar que el origen del metabolismo es un problema distinto al de establecer el tipo de metabolismo que tendría el LCA, además de que son eventos temporalmente separados. Es razonable suponer que en etapas tempranas, durante la proteinización de los procesos celulares, aumentó también la complejidad metabólica (Peretó, 2011). Los resultados sugieren que, así como los ancestros de los linajes de arqueas y bacterias, el metabolismo del LCA era tan complejo como el de sus descendientes modernos. Sin embargo, establecer el metabolismo del LCA se complica porque las enzimas metabólicas, al igual que ocurre con las polimerasas,

evolucionan a nivel de secuencia, de módulos, y a nivel de genoma pueden transferirse horizontalmente, así como perderse o sustituirse.

Lo expuesto anteriormente permite proponer una secuencia relativa de la evolución de las proteínas a partir del mundo de RNA (Figura 17). Inicialmente las proteínas eran pequeños péptidos y cumplían con alguna función, muy posiblemente relacionada con el RNA (Blanco et al., 2018; Vázquez-Salazar y Lazcano 2018). La evolución del ribosoma fue un evento clave que tuvo como consecuencias: i) el establecimiento de la naturaleza modular de las proteínas, ii) la posterior proteinización de los procesos celulares, y iii) el origen del núcleo genómico del LCA y sus descendientes. El establecimiento del ribosoma permitió la formación de polipéptidos de mayor tamaño y de naturaleza modular, lo que a su vez les dio acceso a nuevas funciones como integrarse a las membranas, mayor especificidad metabólica, o papeles estructurales, de regulación, de señalización, etc. Este periodo se caracterizó por la diversificación y expansión de muchas de las familias proteínicas que observamos actualmente, gracias a mecanismos como la duplicación de genes, el cual tuvo un papel muy importante en diversos ámbitos celulares, como se ha documentado para los transportadores ABC, las ATP sintetazas, las polimerasas, los factores de elongación, en la fijación de CO_2, el metabolismo de nitrógeno y en otras rutas biosintéticas (Becerra y Lazcano, 1998; Becerra *et al*., 2007; Rivas *et al*., 2018), aunque muy posiblemente también participaron otros mecanismos como mutaciones que cambian los marcos de lectura (Delaye *et al*., 2008).

El modelo debe explicar el desarrollo evolutivo de los genomas de DNA, lo cual posiblemente sucedió en el mundo de RNA y proteínas. Pero, sin duda alguna, como consecuencia de la diversificación proteínica, éstas adquirieron los papeles centrales y entonces se proteinizaron todos los procesos celulares (Fournier y Gogarten, 2007; Petrov *et al*., 2015). En este período se dio una expansión temprana de los genomas (Becerra *et al*., 2007), gracias también a la presencia de enzimas más eficientes que mantenían un genoma más complejo, lo que a su vez hizo posible la aparición de nuevos mecanismos de evolución del genoma, como la sustitución no ortóloga, la pérdida de genes y la transferencia horizontal. La recurrencia de dichos procesos a través del tiempo, es decir, la

movilidad diferencial de los genes al interior de los genomas individuales, los cuales pueden expandirse o reducirse dependiendo de las circunstancias, originó un patrón en el que no todos los genomas presentan el mismo contenido, salvo por el núcleo genómico. En otras palabras, la expansión temprana de los genomas y su posterior evolución, la diversificación de los linajes, significó el origen del genoma accesorio y de los pangenomas. Todos estos elementos ya estaban presentes en la etapa del LCA. Por ejemplo, se ha propuesto que no todos los genes modernos tienen sus ancestros en el LCA, pues podría haber algunos que se originaron en otros linajes y llegaron al LCA mediante transferencia horizontal (Fournier *et al*., 2015). En resumen, los genomas modernos constituidos por un núcleo genómico y un genoma accesorio, así como también los procesos que actualmente moldean a los genomas, se establecieron hace mucho tiempo y preceden al LCA.

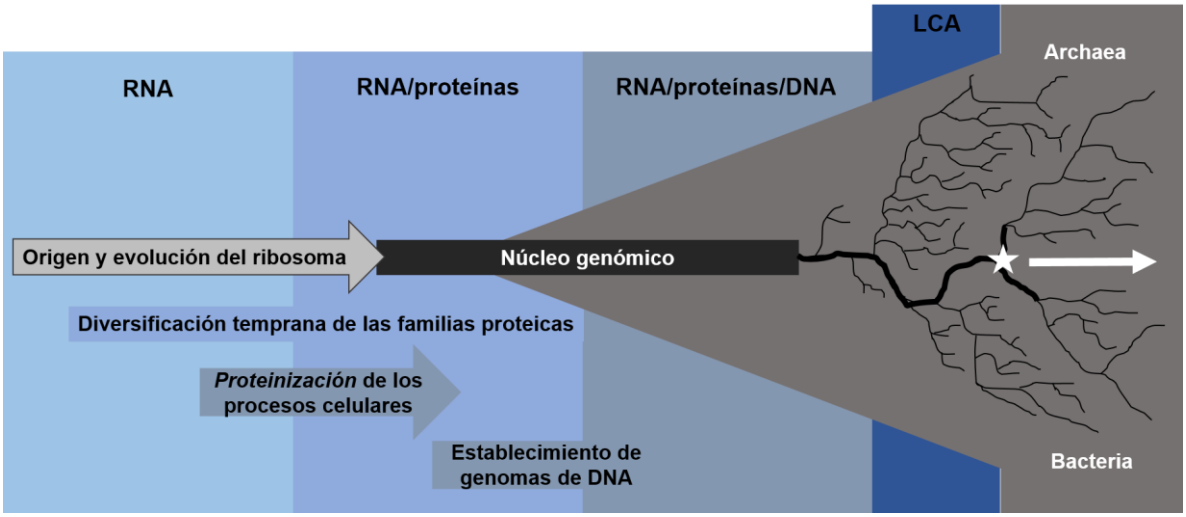

Figura 17. Modelo propuesto del origen y evolución de los genomas modernos.

Conclusiones

Sin lugar a dudas, los catálogos génicos/proteínicos ancestrales de los linajes principales de Archaea y Bacteria, aportaron elementos de análisis y discusión para una mejor comprensión del problema de la reconstrucción ancestral del LCA, fungiendo así como un control metodológico, y estableciendo claramente los límites y alcances de la comparación de la estructura primaria de las proteínas, como criterio para la reconstrucción de ancestros. A continuación se describen las conclusiones generales del trabajo.

Independientemente de la metodología aplicada para la reconstrucción del LCA, los análisis bioinformáticos ocupan las bases de datos disponibles. La naturaleza de cualquier reconstrucción ancestral depende, en principio, de la base de datos que utilice. Éstas siempre presentan un sesgo a raíz de que es imposible secuenciar el genoma completo de todos y cada uno de los organismos procariontes. Es posible que existan proteínas antiguas, quizá ancestrales, que desconocemos por falta de muestreo.

La anotación funcional de las secuencias presentes en las bases de datos tiene un grado de incertidumbre asociado, salvo por aquellas que cuentan con evidencia experimental. Por esta razón, la búsqueda de proteínas homólogas debe basarse en la comparación, ya sea de la secuencia o de la estructura, y nunca en la anotación. Y por lo mismo, toda propuesta sobre la posible función ancestral de una familia proteínica, deberá tratarse como una hipótesis hasta que más líneas de evidencia la respalden.

Aunque no era un objetivo del trabajo, la selección de la muestra confirmó reportes previos de que los genomas más pequeños de procariontes de vida libre contienen entre 1200 y 1300 proteínas (Islas *et al*., 2004; Martínez-Cano *et al*., 2017).

Se construyeron un total de 19 catálogos ancestrales de distintos linajes de Archaea y Bacteria, que presentaron las siguientes características:

- Ninguno superó las 900 proteínas conservadas, bajo el criterio relajado del núcleo genómico, o 600 proteínas conservadas en el caso del núcleo genómico estricto.
- El porcentaje de reconstrucción ancestral, es una estimación de qué tanto podemos conocer del genoma ancestral suponiendo que fuese como sus descendientes, y fue

menor a 15%, salvo por algunas excepciones (Aquificae, Thermotogae, Epsilonproteobacteria, Deinococcus-Thermus y Thaumarchaeota).

- De acuerdo a la clasificación basada en COG, la conservación al interior de los catálogos se concentró en la función de traducción, y en menor medida en el metabolismo y transporte de aminoácidos, de coenzimas, la replicación, recombinación y reparación, y la biogénesis de la membrana, pared o envoltura celular.
- En una segunda clasificación basada en el esquema de KEGG, la conservación se presentó principalmente en la traducción (ribosoma y biosíntesis de aminoacil-tRNAs), seguida por el metabolismo de nucleótidos.
- Las funciones celulares, así como las rutas metabólicas, siempre se presentan fragmentadas o incompletas, patrón que se intensifica si disminuye el tamaño del catálogo.
- Conforme disminuye el tamaño de las reconstrucciones ancestrales, éstas contienen menos proteínas asociadas a procesos metabólicos, mientras que la conservación se concentra en proteínas vinculadas al procesado de la información genética.

Dichas tendencias permiten concluir que las reconstrucciones ancestrales presentan un doble sesgo, uno cuantitativo y otro cualitativo, en el sentido de que se recupera muy poco, de algunas -pocas- funciones. En otras palabras, la comparación de la estructura primaria de las proteínas como un criterio para la reconstrucción ancestral, permite recuperar solamente una fracción del genoma ancestral, principalmente lo relacionado con cuestiones informativas (núcleo genómico). Esto indica que a nivel primario de las proteínas hay una falta de señal, y que conforme retrocedemos en el tiempo disminuye la probabilidad de detectar los aspectos metabólicos de los ancestros.

El tamaño de las reconstrucciones ancestrales depende, además de la base de datos y de la metodología aplicada, de factores como el tamaño de la muestra (número de especies y varianza en el tamaño del genoma), la diversidad fisiológica/metabólica del grupo (relacionada con la varianza del pangenoma), y la distancia filogenética entre las especies consideradas. Si alguno de dichos factores aumenta, entonces el catálogo

ancestral se reduce. Este resultado explica porqué las reconstrucciones del LCA tienden a ser pequeñas, y su contenido está sesgado hacia procesos informativos (traducción).

A pesar del doble sesgo presente en las reconstrucciones, los ancestros de los linajes modernos de Bacteria y Archaea ya tenían la naturaleza típica de un procarionte y lo mismo se puede concluir del LCA, el cual contaba con un genoma de DNA, una envoltura celular funcional y un metabolismo basado en enzimas. El LCA no era un organismo simple.

Es cierto que mayor información ha significado más debates en torno a la naturaleza del LCA (Becerra *et al*., 2007), pero sin lugar a dudas, su estudio ha contribuido a profundizar en el origen y evolución de los genomas procariontes, así como de los procesos que los moldean.

Los genomas con un núcleo genómico y un genoma accesorio se establecieron antes del LCA y de la divergencia de los dominios primarios. La aparición de la traducción significó el origen del núcleo genómico, mientras que el núcleo accesorio fue la consecuencia natural de la evolución de los genomas y de la movilidad y pérdida diferencial de los genes en los distintos linajes.

Referencias

- Albers SV, Meyer BH. 2011. The archaeal cell envelope. *Nature Reviews in Microbiology*. 9:414-26.
- Altermann W, Kazmierczak J. 2003. Archaean microfossils: a reappraisal of early life on Earth. *Research in Microbiology*. 154:611–617.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 25:3389-3402.
- Aravind L, Leipe DD, Koonin EV. 1998. Toprim – a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucleic Acids Research*. 26:4205-4213.
- Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. 2004. The Pfam Protein Families Database. *Nucleic Acids Research*. 32:138–141.
- Becerra A, Delaye L, Islas S, Lazcano A. 2007. Very early stages of biological evolution related to the nature of the last common ancestor of the three major cell domains. *Annual Review of Ecology, Evolution and Systematics*. 38:361–79.
- Becerra A, Islas S, Leguina JI, Silva E, Lazcano A. 1997. Polyphyletic gene losses can bias backtrack characterizations of the cenancestor. *Journal of Molecular Evolution*. 45:115-8.
- Becerra A, Lazcano A. 1998. The role of gene duplication in the evolution of purine nucleotide salvage pathways. *Origins of Life and Evolution of Biospheres*. 28:539-53.
- Berg JM, Tymoczko JL, Stryer L. 2002. Biochemistry. 5th edition. New York: W H Freeman. Sección 25.5. https://www.ncbi.nlm.nih.gov/books/NBK22576/.
- Biller SJ, Berube PM, Berta-Thompson JW, Kelly L, Roggensack SE, Awad L, Roache-Johnson KH, Ding H, Giovannoni SJ, Rocap G, Moore LR, Chisholm SW. 2014.

Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *Scientific Data*. 1:140034.

- Blanco C, Bayas M, Yan F, Chen IA. 2018. Analysis of evolutionarily independent protein-RNA complexes yields a criterion to evaluate the relevance of prebiotic scenarios. *Current Biology*. 28:526-537.e5.
- Braga RM, Dourado MN, Araujo WL. 2016. Microbial interactions: ecology in a molecular perspective. *Brazilian Journal of Microbiology*. 47:86-98.
- Brasier M, McLoughlin N, Green O, Wacey D. 2006. A fresh look at the fossil evidence for early Archaean cellular life. *Philosophical Transactions of the Royal Society. Biological Sciences*. 361:887–902.
- Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*. 523:208-11.
- Canfield DE. 2006. Biochemistry, gas with an ancient history. *Nature*. 440:426–7.
- Cantine MD, Fournier GP. 2018. Environmental adaptation from the origin of life to the Last Universal Common Ancestor. *Origins of Life and Evolution of Biospheres*. 48:35-54.
- Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 25:1972-1973.
- Chothia C, Gough J, Vogel C, Teichmann SA. 2003. Evolution of the protein repertoire. *Science*. 300:1701-3.
- Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology*. 79:7696-701.

- Counts JA, Zeldes BM, Lee LL, Straub CT, Adams MWW, Kelly RM. 2017. Physiological, metabolic and biotechnological features of extremely thermophilic microorganisms. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*. 9.
- Dalquen DA, Dessimoz C. 2013. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. *Genome Biology and Evolution*. 5:1800-1806.
- Darwin C. 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: John Murray. p. 415.
- Delaye L, Becerra A, Lazcano A. 2005. The last common ancestor: what's in a name. *Origins of Life and Evolution of Biospheres*. 35:537-554.
- Delaye L, Becerra A. 2012. Cenancestor: the last universal common ancestor. *Evolution: Education and Outreach.* 5:382-388.
- Delaye L, Deluna A, Lazcano A, Becerra A. 2008. The origin of a novel gene through overprinting in *Escherichia coli*. *BMC Evolutionary Biology*. 8:31.
- Doolittle WF. 2000. The nature of the universal ancestor and the evolution of the proteome. *Current Opinion in Structural Biology*. 10:355–358.
- Faith DP. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation*. 61:1–10.
- Fitch WM, Upper K. 1987. The phylogeny of tRNA sequences provides evidence of ambiguity reduction in the origin of the genetic code. *Cold Spring Harbor Symposium. Quantitative Biology*. 52:759–67.
- Fournier G, Andam CP, Gogarten JP. 2015. Ancient horizontal gene transfer and the last common ancestors. *BMC Evolutionary Biology.* 15:70.
- Fournier GP, Gogarten JP. 2007. Signature of a primitive genetic code in ancient protein lineages*. Journal of Molecular Evolution*. 65:425-36.
- Gogarten JP, Deamer D. 2016. Is LUCA a thermophilic progenote? *Nature Microbiology*. 1:16229.

- Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ. 1989. Evolution of the vacuolar H + -ATPase; implications for the origin of eukaryotes. *Proceedings of the National Academy of Sciences of the USA*. 86:6661–5.
- Goldman AD, Bernhard TM, Dolzhenko E, Landweber LF. 2013. LUCApedia: a database for the study of ancient life. *Nucleic Acids Research*. 41:D1079-82.
- Guilliam TA, Keen BA, Brissett NC, Doherty AJ. 2015. Primase-polymerases are a functionally diverse superfamily of replication and repair enzymes. *Nucleic Acids Research*. 43:6651-6664.
- Harris JK, Kelley ST, Spiegelman GB, Pace NR. 2003. The genetic core of the universal ancestor. *Genome Research*. 13:407–12.
- Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the USA*. 89:10915-10919.
- Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. *Nature Microbiology.* 1:16048.
- Hutchison CA 3rd, Chuang RY, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, Gill J, Kannan K, Karas BJ, Ma L, Pelletier JF, Qi ZQ, Richter RA, Strychalski EA, Sun L, Suzuki Y, Tsvetanova B, Wise KS, Smith HO, Glass JI, Merryman C, Gibson DG, Venter JC. 2016. Design and synthesis of a minimal bacterial genome. *Science*. 351:aad6253.
- Islas S, Becerra A, Luisi PL, Lazcano A. 2004. Comparative genomics and the gene complement of a minimal cell*. Origins of Life and the Evolution of Biosphere*. 34:243-56.
- Islas S, Velasco AM, Becerra A, Delaye L, Lazcano A. 2003. Hyperthermophily and the origin and earliest evolution of life*. International Microbiology*. 6:87–94.
- Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of

duplicated genes. *Proceedings of the National Academy of Sciences of the USA*. 86:9355–9.

- Iyer LM, Koonin EV, Aravind L. 2003. Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Structural Biology*. 3:1-10.1186/1472-6807-3-1.
- Iyer LM, Koonin EV, Leipe DD, Aravind L. 2005. Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. *Nucleic Acids Research*. 33:3875-3896.
- Jácome R, Becerra A, Ponce de León S, Lazcano A. 2015. Structural analysis of monomeric RNA-dependent polymerases: Evolutionary and therapeutic implications. *PLoS ONE.* 10:e0139001.
- Jain S, Caforio A, Driessen AJ. 2014. Biosynthesis of archaeal membrane ether lipids. *Frontiers in Microbiology*. 5:641.
- Jensen RA. 1976. Enzyme recruitment in evolution of new function. *Annual Review of Microbiology*. 30:409-425.
- Joyce GF. 2002. The antiquity of RNA-based evolution. *Nature*. 418:214-21.
- Kaas RS, Friis C, Ussery DW, Aarestrup FM. 2012. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics*. 13:577.
- Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*. 4;45:D353-D361.
- Kannan L, Li H, Rubinstein B, Mushegian A. 2013. Models of gene gain and gene loss for probabilistic reconstruction of gene content in the last universal common ancestor of life*. Biology Direct*. 8:32.
- Kim KM, Caetano-Anollés G. 2011. The proteomic complexity and rise of the primordial ancestor of diversified life. *BMC Evolutionary Biology*. 11:140.

- Knoll AH, Bergmann KD, Strauss JV. 2016. Life: the first two billion years. *Philosophical Transactions of the Royal Society. Biological Sciences*. 371:20150493.
- Koonin EV. 2003. Comparative genomics, minimal gene set and the last universal common ancestor. *Nature Reviews Microbiology*. 1:127-136.
- Korkhin Y, Unligil UM, Littlefield O, Nelson PJ, Stuart DI, Sigler PB, Bell SD, Abrescia, NGA. 2009. Evolution of complex RNA polymerases: The complete archaeal RNA polymerase structure. *PLoS Biology.* 7:e1000102.
- Korobeinikova AV, Garber MB, Gongadze GM. 2012. Ribosomal proteins: structure, function, and evolution. *Biochemistry (Moscow).* 77:562-74.
- Kyrpides N, Overbeek R, Ouzounis C. 1999. Universal protein families and the functional content of the last universal common ancestor. *Journal of Molecular Evolution*. 49:413–23.
- Kysela DT, Randich AM, Caccamo PD, Brun YV. 2016. Diversity Takes Shape: Understanding the Mechanistic and Adaptive Basis of Bacterial Morphology. *PLoS Biology.* 14:e1002565.
- Lamers MH, Georgescu RE, Lee SG, O'Donnell M, Kuriyan J. 2006. Crystal structure of the catalytic alpha subunit of *E. coli* replicative DNA polymerase III. *Cell*. 126:881-892.
- Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*. 15:141-61.
- Lapierre P, Gogarten JP. 2009. Estimating the size of the bacterial pan-genome. *Trends in Genetics*. 25:107-10.
- Lasken RS, McLean JS. 2014. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nature Reviews Genetics*. 15:577-84.

- Lazcano A, Fox GE, Oró J. 1992. Life before DNA, the origin and early evolution of early Archean cells. En: The Evolution of Metabolic Function, ed. Mortlock RP. Boca Raton, FL. CRC Press, pp. 237–95.
- Lazcano A, Miller SL. 1999. On the origin of metabolic pathways. *Journal of Molecular Evolution*. 49:424-31.
- Lazcano A. 2011. Comparative genomics and early cell evolution. En: Gargaud M, López-García P, Martin H (Eds.). Origins and Evolution of Life: An Astrobiological Perspective. pp 259-269. Cambridge University Press, New York.
- Lefebure T, Bitar PD, Suzuki H, Stanhope MJ. 2010. Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. *Genome Biology and Evolution*. 2:646–655.
- Letunic I, Bork P. 2011. Interactive Tree of Life v2: Online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research*. 39:475–8.
- Line MA. 2002. The enigma of the origin of life and its timing. *Microbiology*. 148:21-7.
- Lombard J, López-García P, Moreira D. 2012a. The early evolution of lipid membranes and the three domains of life. *Nature Reviews Microbiology*. 10:507-15.
- Lombard J, López-García P, Moreira D. 2012b. Phylogenomic investigation of phospholipid synthesis in archaea. *Archaea*. 2012:630910.
- Makarova KS, Koonin EV. 2003. Comparative genomics of archaea: how much have we learned in six years, and what's next? *Genome Biology*. 4:115.
- Martínez-Cano DJ, Reyes-Prieto M, Martínez-Romero E, Partida-Martínez LP, Latorre A, Moya A, Delaye L. 2014. Evolution of small prokaryotic genomes. *Frontiers in Microbiology*. 5:742.
- Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution*. 30:1188-1195.

- Mirkin BG, Fenner TI, Galperin MY, Koonin EV. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology*. 3:2.
- Mukherjee S, Seshadri R, Varghese NJ, Eloe-Fadrosh EA, Meier-Kolthoff JP, Göker M, Coates RC, Hadjithomas M, Pavlopoulos GA, Paez-Espino D, Yoshikuni Y, Visel A, Whitman WB, Garrity GM, Eisen JA, Hugenholtz P, Pati A, Ivanova NN, Woyke T, Klenk HP, Kyrpides NC. 2017. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature Biotechnology*. 35:676-683.
- Murzin AG, Brenner SE, Hubbard TJP, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*. 247:536-540.
- Mushegian AR, Koonin EV. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences USA*. 93:10268-10273.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Molecular Biology and Evolution*. 32:268-274.
- Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH—a hierarchic classification of protein domain structures. *Structure*. 5:1093–1108.
- Ouzounis AC, Kunin V, Darzentas N, Goldovsky L. 2006. A minimal estimate for gene content of the last universal common ancestor- exobiology from a terrestrial perspective. *Research in Microbiology*. 157:57–68.
- Ouzounis CA, Karp PD. 2000. Global properties of the metabolic map of *Escherichia coli*. *Genome Research*. 10:568-576.
- Peretó J, López-García P, Moreira D. 2004. Ancestral lipid biosynthesis and early membrane evolution. *Trends in Biochemical Sciences*. 29:469–77.

- Peretó J. 2011. Origin and evolution of metabolisms. En: Gargaud M, López-García P, Martin H (Eds.). Origins and Evolution of Life: An Astrobiological Perspective. pp 270-290. Cambridge University Press, New York.
- Petrov AS, Gulen B, Norris AM, Kovacs NA, Bernier CR, Lanier KA, Fox GE, Harvey SC, Wartell RM, Hud NV, Williams LD. 2015 History of the ribosome and the origin of translation. *Proceedings of the National Academy of Sciences of the USA*. 112:15396-401.
- Philippe H, Forterre P. 1999. The rooting of the universal tree of life is not reliable. *Journal of Molecular Evolution*. 49:509–23.
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies, J, Glöckner, FO. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*. 41:590–6.
- R Core Team. 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org.
- Ranea AG, Sillero A, Thorton MJ, Orengo AC. 2006. Protein superfamily evolution and the last universal common ancestor (LUCA). *Journal of Molecular Evolution*. 63:513–25.
- Richardson EJ, Watson M. 2013. The automatic annotation of bacterial genomes. *Briefings in Bioinformatics*. 14:1-12.
- Rivas M, Becerra A, Lazcano A. 2018. On the early evolution of catabolic pathways: a comparative genomics approach. I. The cases of glucose, ribose, and the nucleobases catabolic routes. *Journal of Molecular Evolution*. 86:27-46.
- Rivas-Marín E, Canosa I, Devos DP. 2016. Evolutionary cell biology of division mode in the bacterial Planctomycetes-Verrucomicrobia-Chlamydiae superphylum. *Frontiers in Microbiology*. 7:1964.
- Rouli L, Merhej V, Fournier P-E, Raoult D. 2015. The bacterial pangenome as a new tool for analyzing pathogenic bacteria. *New Microbes and New Infections*. 7:72-85.

- Salgado PS, Koivunen MRL, Makeyev EV, Bamford DH, Stuart DI, Grimes JM. 2006. The structure of an RNAi polymerase links RNA silencing and transcription. *PLoS Biology*. 4:e434.
- Schirrmeister BE, Gugger M, Donoghue PC. 2015. Cyanobacteria and the Great Oxidation Event: evidence from genes and fossils. *Palaeontology.* 58:769-785.
- Sobolevsky Y, Trifonov EN. 2006. Protein modules conserved since LUCA. *Journal of Molecular Evolution*. 63:622–34.
- Solden L, Lloyd K, Wrighton K. 2016. The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Current Opinion in Microbiology*. 31:217-26.
- Steitz TA. 1999. DNA polymerases: structural diversity and common mechanisms. *Journal of Biological Chemistry*. 274:17395-8.
- Sutcliffe IC. 2010. A phylum level perspective on bacterial cell envelope architecture. *Trends in Microbiology*. 18:464-70.
- Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*. 28:33-36.
- Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the USA*. 102:13950-5.

- Tucker CM, Cadotte MW, Carvalho SB, Davies TJ, Ferrier S, Fritz SA, Grenyer R, Helmus MR, Jin LS, Mooers AO, Pavoine S, Purschke O, Redding DW, Rosauer DF, Winter M, Mazel F. 2017. A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews of the Cambridge Philosophical Society*. 92:698-715.

- Tuller T, Birin H, Gophna U, Kupiec M, Ruppin E. 2010. Reconstructing ancestral gene content by coevolution. *Genome Research*. 20:122-132.

- Vázquez-Salazar A, Lazcano A. 2018. Early life: Embracing the RNA world. *Current Biology*. 28:R220-R222.

- Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. *Current Opinion in Microbiology*. 23:148-54.

- Vinuesa, P, Contreras-Moreira, B. 2015. Robust identification of orthologues and paralogues for microbial pan-genomics using GET_HOMOLOGUES: A case study of pIncA/C plasmids. En Bacterial Pangenomics. Methods and Protocols. Volume 1231 of the series Methods in Molecular Biology. pp 203-232.

- Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. 2004. Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology*. 14:208-16.

- Wächstershäuser G. 2003. From pre-cells to Eukarya – a tale of two lipids. *Molecular Microbiology*. 47:13–22.

- Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, Martin WF. 2016. The physiology and habitat of the last universal common ancestor. *Nature Microbiology*. 1:16116.

- Whalen MB, Massidda O. 2015. *Helicobacter pylori*: enemy, commensal or, sometimes, friend? *The Journal of Infection in Developing Countries*. 9:674-8.

- Wilkens S. 2015. Structure and mechanism of ABC transporters. *F1000Prime Reports*. 7:14.

- Willey JM, Sherwood LM, Woolverton, CJ. 2008. Prescott, Harley, and Klein´s Microbiology. 7th edition. McGraw Hill. pp.539-570.
- Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*. 504:231-6.
- Woese C, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the USA*. 74:5088-5090.
- Woese C. 1998. The universal ancestor. *Proceedings of the National Academy of Sciences of the USA*. 95:6854-6859.
- Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the USA*. 87:4576–4579.
- Woese CR. 1987. Bacterial evolution. *Microbiological Reviews*. 51:221-271.
- Wolf YI, Koonin EV. 2012. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biology and Evolution*. 4:1286-1294.
- Wong JT, Chen J, Mat WK, Ng SK, Xue H. 2007. Polyphasic evidence delineating the root of life and roots of biological domains. *Gene*. 403:39-52.
- Yang S, Doolittle RF, Bourne PV. 2005. Phylogeny determined by protein domain content. *Proceedings of the National Academy of Sciences USA*. 102:373–78.
- Zhaxybayeva O, Doolittle WF. 2011. Lateral gene transfer. *Current Biology*. 21:R242-R246.
- Zhaxybayeva O, Gogarten PJ. 2004. Cladogenesis, coalescence and the evolution of the three domains of life. *Trends in Genetics*. 20:182–7.

Anexos

Anexo 1

Tabla anexo 1. Especies a partir de cuyo genoma se realizó la agrupación de posibles ortólogos, utilizando el algoritmo BDBH con el programa *Get-homologues*.

Grupo	**Genoma semilla (*query sequence*)**
Archaea	
Thaumarchaeota	*Nitrosopumilus maritimus* SCM1
Crenarchaeota	*Desulfurococcus mucosus* DSM 2162
Euryarchaeota	*Methanothermus fervidus* DSM 2088
Bacteria	
Acidobacteria	*Chloracidobacterium thermophilum* B
Actinobacteria	*Acidimicrobium ferrooxidans* DSM 10331
Aquificae	*Aquifex aeolicus* VF5
Bacteroidetes	*Chlorobium phaeovibrioides* DSM 265
Chloroflexi	*Dehalococcoides mccartyi* 195
Cyanobacteria	*Prochlorococcus marinus* subsp. marinus str. CCMP1375
Deinococcus - Thermus	*Thermus thermophilus* HB8
Firmicutes	*Lactobacillus sanfranciscensis* TMW 1.1304
Spirochaetes	*Leptospira biflexa* serovar Patoc strain 'Patoc 1 (Paris)'
Thermotogae	*Fervidobacterium nodosum* Rt17-B1
Verrucomicrobia - Planctomycetes	*Akkermansia muciniphila* ATCC BAA-835
Alphaproteobacteria	Candidatus *Puniceispirillum marinum*
Betaproteobacteria	*Methylobacillus flagellatus* KT
Gammaproteobacteria	*Idiomarina loihiensis* L2TR
Deltaproteobacteria	*Hippea maritima* DSM 10411
Epsilonproteobacteria	*Nautilia profundicola* AmH

Anexo 2

Tabla anexo 2. Categorías funcionales propuestas en la base de datos COG. Tomado de ftp://ftp.ncbi.nih.gov/pub/COG/COG/fun.txt.

ALMACENAMIENTO Y PROCESADO DE LA INFORMACIÓN
[J] Traducción, biogénesis y estructura del ribosoma
[A] Modificación y procesamiento del RNA
[K] Transcripción
[L] Replicación, recombinación y reparación
[B] Estructura y dinámica de la cromatina
PROCESOS CELULARES Y SEÑALIZACIÓN
[D] Control del ciclo celular, división celular y segregación cromosómica
[Y] Estructura nuclear
[V] Mecanismos de defensa
[T] Mecanismos de transducción de señales
[M] Biogénesis de las membrana/pared/envoltura celular
[N] Movilidad celular
[Z] Citoesqueleto
[W] Estructuras extracelulares
[U] Tráfico intracelular, transporte de vesículas y secreción
[O] Modificación postraduccional, chaperonas
METABOLISMO
[C] Conversión y producción de energía
[G] Metabolismo y transporte de carbohidratos
[E] Metabolismo y transporte de aminoácidos
[F] Metabolismo y transporte de nucleótidos
[H] Metabolismo y transporte de coenzimas
[I] Metabolismo y transporte de lípidos
[P] Metabolismo y transporte de iones inorgánicos
[Q] Biosíntesis, transporte y catabolismo de metabolitos secundarios
POBREMENTE CARACTERIZADAS
[R] Predicción general de la función
[S] Función desconocida

Anexo 3

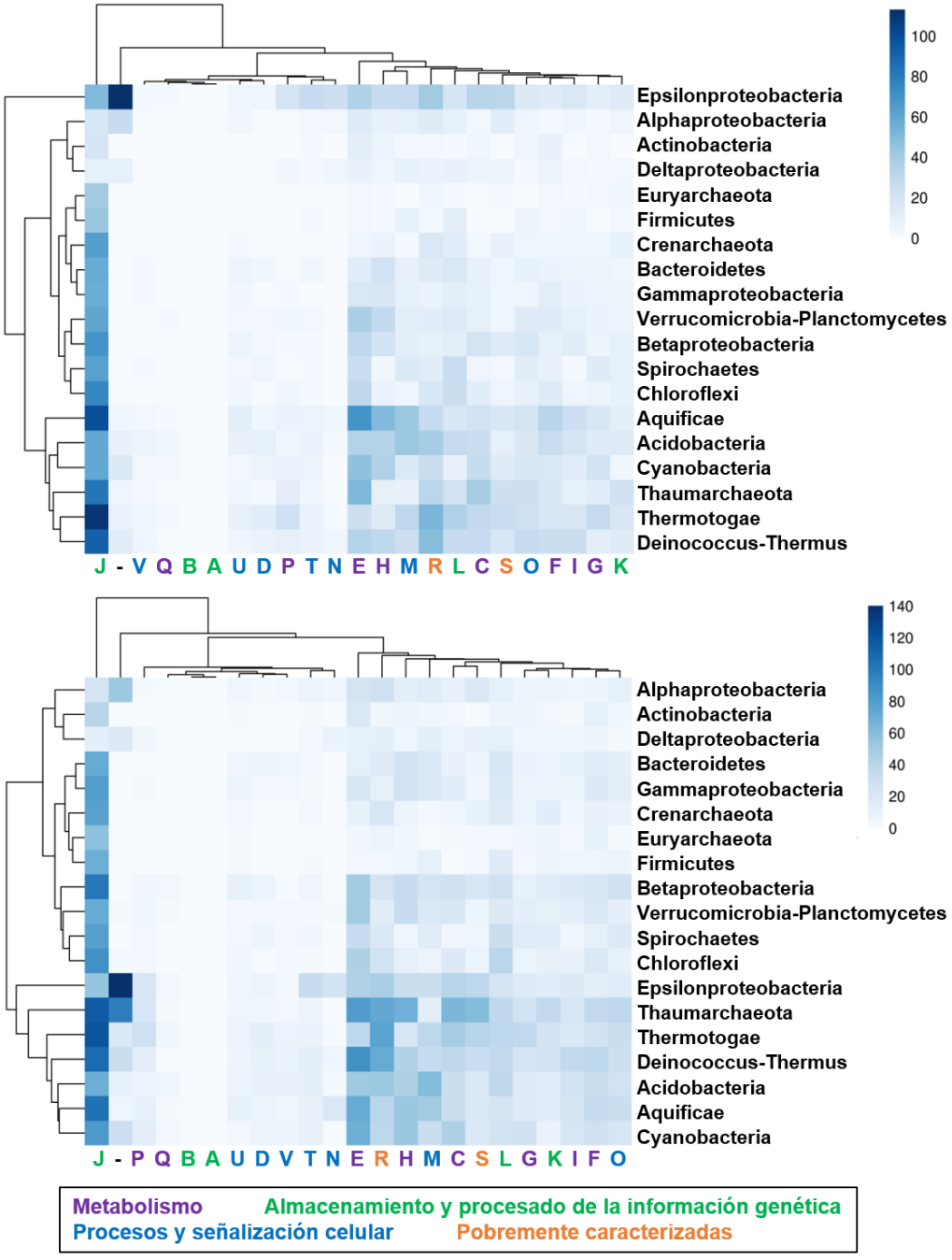

Figura anexo 3. Clasificación funcional de las proteínas al interior de los catálogos ancestrales. El color representa el número total de proteínas presentes tanto en el núcleo estricto (arriba), como en el núcleo relajado (abajo). Las categorías son las mismas que en las figuras 8 y 9, así como el método de agrupación utilizado.

Anexo 4

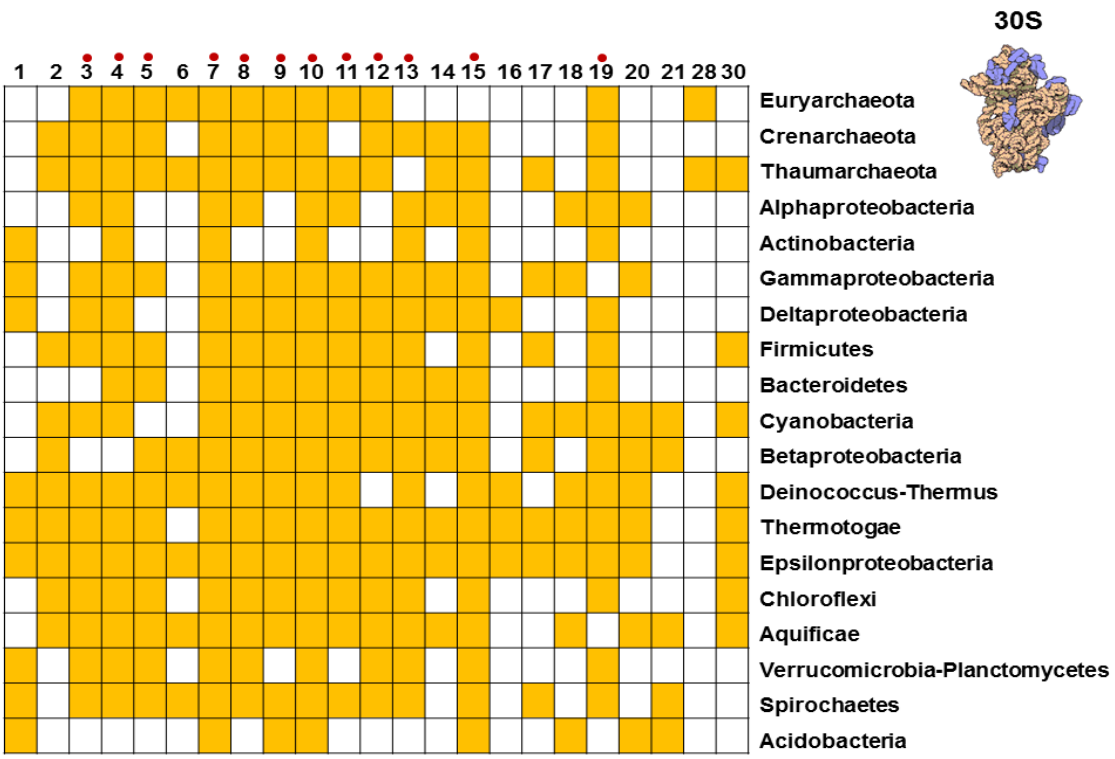

Figura anexo 4. Proteínas de la subunidad 30S ribosomal presentes en las reconstrucciones ancestrales. Se muestran las proteínas de la subunidad pequeña ribosomal (columnas) identificadas al interior de los catálogos ancestrales de los distintos linajes (filas). El color amarillo indica presencia y el blanco ausencia. El punto rojo señala familias proteicas con niveles elevados de conservación, puesto que se identificaron en más de 11 linajes.

Anexo 5

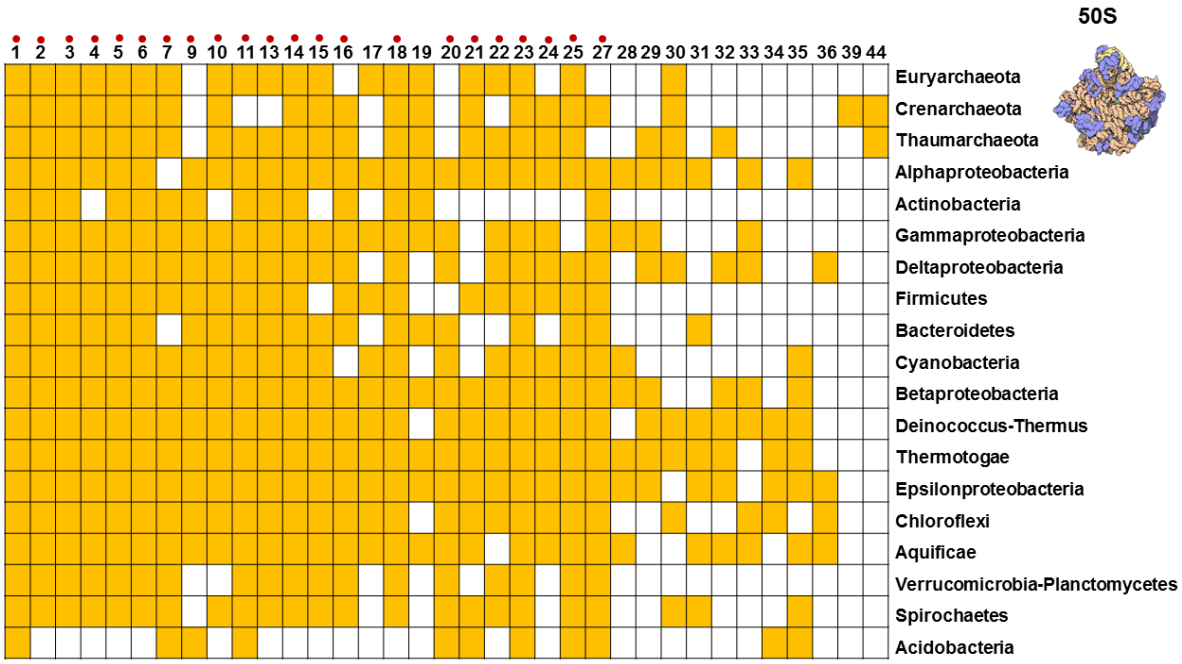

Figura anexo 5. Proteínas de la subunidad 50S ribosomal presentes en las reconstrucciones ancestrales. Se muestran las proteínas de la subunidad grande ribosomal (columnas) identificadas al interior de los catálogos ancestrales de los distintos linajes (filas). El color amarillo indica presencia y el blanco ausencia. El punto rojo señala familias proteicas con niveles elevados de conservación, ya que se identificaron en más de 11 linajes.

Anexo 6

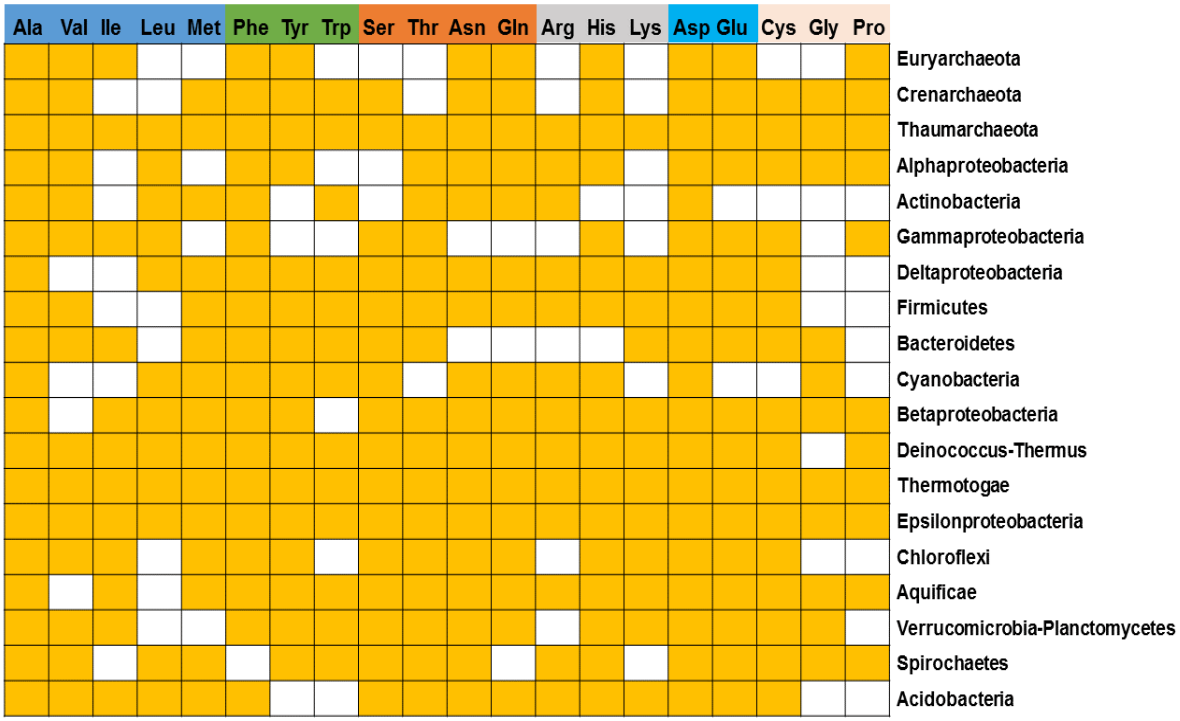

Figura anexo 6. Enzimas Aminoacil-tRNA sintetasas presentes en las reconstrucciones ancestrales. En las columnas se muestran las enzimas identificadas al interior de los linajes (filas) por el nombre del aminoácido que cargan. El color naranja indica presencia, mientras que el color blanco es ausencia. De izquierda a derecha se indican primero los aminoácidos cuyas cadenas laterales son hidrofóbicas (Ala-Met), aromáticas (Phe-Trp), polares no cargadas (Ser-Gln), cargadas positivamente (Arg-Lys), cargadas negativamente (Asp-Glu) y los casos especiales (Cys-Pro).