UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

BIOLOGÍA EVOLUTIVA

EL PAPEL DE LA DUPLICACIÓN GÉNICA EN LA EVOLUCIÓN DE LA FUNCIÓN ENZIMÁTICA

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN CIENCIAS

PRESENTA:

ALEJANDRO ALBERTO ÁLVAREZ LUGO

TUTOR PRINCIPAL DE TESIS: DR. ARTURO CARLOS II BECERRA BRACHO
FACULTAD DE CIENCIAS, UNAM

COMITÉ TUTOR: DR. JESÚS AGUIRRE LINARES
INSTITUTO DE FISIOLOGÍA CELULAR, UNAM

COMITÉ TUTOR: DR. MARIO ALBERTO MARTÍNEZ NÚÑEZ
UNIDAD ACADÉMICA DE CIENCIAS Y TECNOLOGÍA, UNAM

CIUDAD UNIVERSITARIA, CD. MX. **NOVIEMBRE, 2023**

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

BIOLOGÍA EVOLUTIVA

EL PAPEL DE LA DUPLICACIÓN GÉNICA EN LA EVOLUCIÓN DE LA FUNCIÓN ENZIMÁTICA

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN CIENCIAS

PRESENTA:

ALEJANDRO ALBERTO ÁLVAREZ LUGO

TUTOR PRINCIPAL DE TESIS: DR. ARTURO CARLOS II BECERRA BRACHO
FACULTAD DE CIENCIAS, UNAM

COMITÉ TUTOR: DR. JESÚS AGUIRRE LINARES
INSTITUTO DE FISIOLOGÍA CELULAR, UNAM

COMITÉ TUTOR: DR. MARIO ALBERTO MARTÍNEZ NÚÑEZ
UNIDAD ACADÉMICA DE CIENCIAS Y TECNOLOGÍA, UNAM

CIUDAD UNIVERSITARIA, CD. MX. **NOVIEMBRE, 2023**

COORDINACIÓN GENERAL DE ESTUDIOS DE POSGRADO

COORDINACIÓN DEL POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

OFICIO: CGEP/CPCB/ FC/0788/2023

ASUNTO: Oficio de Jurado

M. en C. 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 **19 de junio de 2023** se aprobó el siguiente jurado para el examen de grado de **DOCTOR EN CIENCIAS** del estudiante **ÁLVAREZ LUGO ALEJANDRO ALBERTO** con número de cuenta **307521493** con la tesis titulada: **"El papel de la duplicación génica en la evolución de la función enzimática",** realizada bajo la dirección del **DR. ARTURO CARLOS II BECERRA BRACHO:**

Presidente:	**DRA. GLORIA SOBERON CHÁVEZ**
Vocal:	**DRA. ROCIO JETZABEL ALCÁNTARA HERNÁNDEZ**
Vocal:	**DR. LUIS DAVID ALCARAZ PERAZA**
Vocal:	**DR. DIEGO CLAUDIO CORTEZ QUEZADA**
Secretario:	**DR. JESÚS AGUIRRE LINARES**

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

A T E N T A M E N T E
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Ciudad Universitaria, Cd. Mx., a 02 de octubre de 2023

COORDINADOR DEL PROGRAMA

POSGRADO EN CIENCIAS BIOLOGICAS
COORDINACIÓN

DR. ADOLFO GERARDO NAVARRO SIGÜENZA

c. c. p. Expediente del alumno

AGNS/AAC/GEMF/EARR/ipp

AGRADECIMIENTOS INSTITUCIONALES

Al Posgrado en Ciencias Biológicas por todo el apoyo y oportunidades de crecimiento académico y profesional recibidos a lo largo de mi formación científica.

Al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCyT) por otorgarme el apoyo económico (núm. 747513) a lo largo de los cuatro años de mi proyecto de Doctorado.

Al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) por apoyarme económicamente a través de los proyectos de investigación BV100218 y IN214421.

A mi tutor, el Dr. Arturo Carlos II Becerra Bracho, por aceptar dirigir mi proyecto de doctorado y la presente tesis. A los miembros de mi comité tutoral, los Dres. Jesús Aguirre Linares y Mario Alberto Martínez Núñez, por todo su apoyo académico a lo largo de mi proyecto de investigación de doctorado y durante la elaboración de esta tesis. A los tres, mi más sincero agradecimiento por todas sus sugerencias y cuestionamientos a lo largo de estos cuatro años de mi formación académica.

AGRADECIMIENTOS PERSONALES

A mi tutor de licenciatura, maestría y doctorado, el Dr. Arturo Becerra. Gracias por todo el apoyo y conocimientos que me ha brindado a lo largo de todos estos años, así como por su invaluable amistad y su infinita paciencia.

Al Dr. Antonio Lazcano por su amistad, por compartir su conocimiento y sus valores éticos y profesionales y por haber sido quien, en primera instancia, me motivó a adentrarme en el maravilloso campo del origen y evolución temprana de la vida.

Mi más sincero e infinito agradecimiento a todos mis amigos, amigas y colegas del Laboratorio de Origen de la Vida (Ricardo Hernández, José Campillo, Rodrigo Jácome, Coral Cruz-González, Wolfgang Cottom, Ingrid Miranda, Adrián Cruz, Abelardo Aguilar, Hilda Palacios, Grisel Córdova, Ervin Silva, Mario Rivas, Alberto Vázquez, Israel Muñoz y Alonso Quintero). Sé que fueron tiempos difíciles debido a una situación que nadie pudo prever, pero a pesar de todo, siempre estuvieron presentes (aunque sea a la distancia) para apoyarme en todo lo que necesitara.

DEDICATORIA

A la memoria de mi abuela, Judith Vázquez Miranda

ÍNDICE

RESUMEN

La duplicación génica es un proceso de gran relevancia para la evolución biológica debido a que posibilita el surgimiento de nuevos genes y funciones, lo cual entonces se interrelaciona con la expansión del metabolismo celular. En los procariontes, distintos procesos metabólicos poseen una sobrerrepresentación de genes duplicados; esto ha llevado a proponer que este fenómeno podría tener un papel importante en la diversificación de estos organismos y que podría estar asociado con la adaptación a diferentes condiciones ambientales.

El objetivo principal de este trabajo fue evaluar el posible papel de la duplicación de genes en la evolución de las enzimas procariontes, así como determinar si había grupos específicos de estas en los que hubiera una sobrerrepresentación de enzimas duplicadas y si esto estaba asociado de alguna manera al estilo de vida de los organismos. Partiendo de una muestra representativa de proteomas procariontes hicimos análisis bioinformáticos para identificar aquellas secuencias que pudieran haber surgido por medio de un evento de duplicación.

Mediante estos análisis hallamos que entre una cuarta parte y la mitad de enzimas metabólicas podrían ser producto de la duplicación génica. Además, algunos grupos de organismos filogenéticamente distantes pero con estilos de vida comunes poseen aspectos genómicos y bioquímicos similares, estos últimos en términos de grupos de enzimas con una proporción de duplicados parecida.

A nivel enzimático, tres clases presentan una mayor proporción de duplicados: las oxidorreductasas, las transferasas y las isomerasas. Para todas hallamos que solo unas cuantas de sus respectivas subclases concentran la mayoría de las duplicaciones, aunque solo para las isomerasas fue posible hacer una exploración más profunda. Tanto en arqueas como en bacterias, la mayoría de las duplicaciones están asociadas al metabolismo de carbohidratos, pero solo en bacterias hallamos un número considerable de isomerasas duplicadas relacionadas con la modificación de bases nitrogenadas en el RNA. La mayoría de estas últimas, que pertenecen a la superfamilia de pseudouridina sintasas, parecen haberse originado a partir de eventos de duplicación. Además, una de estas enzimas, RluD, posee múltiples copias en una gran diversidad de organismos.

Nuestros análisis sugieren que la duplicación de genes en procariontes podría ser una manera de hacer frente a condiciones ambientales específicas. Por otra parte, aunque en algunos casos podría haber cierta relación entre el total de enzimas en una categoría y la proporción de duplicados, en otras parece haber distintos factores involucrados, como el papel que desempeñan en el contexto celular. Finalmente, pensamos que podríamos tener un mejor

entendimiento del por qué ciertos grupos de enzimas poseen una sobrerrepresentación de duplicados si consideramos clasificaciones hechas con un sentido evolutivo, como son las familias y superfamilias de proteínas.

ABSTRACT

Gene duplication is a highly important process for biological evolution because it enables the emergence of new genes and new functions, which is related to the expansion of cellular metabolism. In prokaryotes, different metabolic processes show an overrepresentation of duplicated genes; this has led to the proposal that this phenomenon might have an important role in the diversification of these organisms and that it might be related to adaptation in changing environments.

The main goal of this study was to evaluate the possible role of gene duplication in the evolution of prokaryotic enzymes, and also to determine whether there were specific groups with an overrepresentation of duplicated enzymes and if this was somehow related to their lifestyle. Starting from a representative sample of prokaryotic proteomes we performed bioinformatic analyses to identify those sequences that might have arisen by a duplication event.

By means of these analyses we found that between one quarter and a half of metabolic enzymes could have been originated by gene duplication. We also identified several phylogenetically-distant groups of organisms with similar lifestyles which have alike genomic and biochemical traits, the later in terms of enzyme groups with akin duplicates ratio.

At the enzymatic level, there are three classes with a high ratio of duplicates: oxidoreductases, transferases, and isomerases. Within them, just a few subclasses concentrate most duplicates but only for isomerases a deeper exploration could be made. For both the archaea and bacteria, most duplications are related to carbohydrate metabolism but only for bacteria we found a significantly high number of duplicated isomerases related to RNA-nucleobase modification. Most of these, which belong to the pseudouridine synthase superfamily, seem to have evolved from duplication events. Besides, one of these enzymes, RluD, has multiple copies in a wide variety of organisms.

Our analyses suggest that gene duplication in prokaryotes could be a way to cope with specific environmental conditions. On the other hand, although in some cases a relationship between the number of enzymes and the number of duplicates could be established, there are others in which different factors could be involved, such as their role in the cellular context. Ultimately, we think that a better understanding of why certain enzyme groups show an overrepresentation of duplicates could be achieved if we take into account evolutionary-based classifications such as protein families and superfamilies.

INTRODUCCIÓN

Duplicación génica: Generalidades

Una de las preguntas centrales en los campos de la Biología Evolutiva y la Evolución Molecular es ¿cómo surgen nuevos genes en los diferentes organismos que conforman la biodiversidad? A lo largo de varias décadas se ha encontrado que no hay un mecanismo exclusivo para la formación de nuevos genes sino que dicho proceso puede ocurrir a través de mecanismos tan diferentes como la duplicación génica (Kaessmann 2010), el origen *de novo* a partir de una secuencia que puede ser codificante o no-codificante (Weisman 2022), la fusión de genes (Zhou et al. 2022), el barajeo de dominios (Kawashima et al. 2009) y el transporte lateral u horizontal de genes (Boucher et al. 2003). Un mecanismo adicional que da origen a nuevos genes es la domesticación de proteínas de elementos transponibles, en la cual ciertas proteínas involucradas en la inserción de elementos móviles son cooptadas o reclutadas por el hospedero para algún proceso celular diferente (Feschotte & Pritham 2007; Jangam *et al.* 2017).

De todos estos mecanismos, la duplicación génica es quizá el mejor estudiado. No solo porque los primeros reportes de fragmentos cromosómicos duplicados datan de hace casi un siglo (Sturtevant 1925; Muller 1936) sino que, además, es quizá el más fácil de detectar por medio de métodos bioinformáticos. Pero a pesar de ser tan común, existen diversos procesos moleculares capaces de originar una copia idéntica de un gen específico, los cuales se resumen en la Tabla 1.

Los genes originados a partir de un evento de duplicación, sin importar el mecanismo involucrado ni el número de copias resultantes, se conocen como *parálogos* (Henikoff *et al.* 1997). La formación de estos, a diferencia de los genes ortólogos, no coincide con un evento de especiación sino que pueden surgir en cualquier momento previo o posterior a este.

Mientras que ciertos mecanismos de duplicación génica en pequeña escala pudieran no ocurrir de manera homogénea a lo largo del genoma, aquellos a gran escala, como la poliploidización, dan como resultado una copia íntegra de este. En algunos casos, particularmente en plantas, este proceso puede ser la base para un posterior evento de especiación (van de Peer *et al.* 2021). Sin embargo, lo más común es que la mayoría de los

genes recientemente duplicados se pierdan en el corto o mediano plazo (Gout & Lynch 2015; Hooper & Berg 2003; Lynch & Conery 2003). Esto podría ocurrir debido a que, para la mayoría de estos genes, la selección natural puede relajarse durante este periodo. Posteriormente, la selección purificadora podría ser la responsable del proceso de fijación en la mayoría de los genes que no fueron eliminados (Lynch & Conery 2000; 2003).

Tabla 1. Principales mecanismos involucrados en la duplicación de genes (a partir de Hahn 2009 y Reams & Roth 2015).

Mecanismo	Breve descripción
Entrecruzamiento desigual	Apareamiento entre dos cromosomas, que no necesariamente deben ser homólogos, en el que en uno de ellos se elimina una secuencia que posteriormente será reemplazada con una copia durante la meiosis o mitosis. La pérdida de dicha región se debe a un apareamiento incorrecto entre ambos cromosomas. Parece ser el mecanismo más común para el surgimiento de un nuevo gen por duplicación.
Transposición duplicativa	Ocurre directamente sobre el DNA y puede ser de dos tipos: *a)* *Recombinación homóloga no alélica*. La recombinación entre este tipo de regiones puede propiciar que se dupliquen las regiones contiguas; no es raro que dichas regiones se dupliquen varias veces. *b)* *Unión de extremos no homólogos*. Se da en regiones donde no hay segmentos de DNA repetitivo ni grandes regiones de secuencias homólogas
Duplicación por inversión en tándem	Una secuencia palindrómica en el extremo 3' es la que inicia el proceso. La hebra molde ahora pasa a ser la hebra opuesta gracias a la intervención de una secuencia palindrómica adicional. Al replicarse esta región, se obtiene un total de tres copias contiguas. De estas, la copia central se encuentra invertida.
Retrotransposición	Se da a partir de la transcripción reversa de un fragmento de mRNA maduro y se inserta directamente en el genoma como cDNA. Los genes duplicados resultantes poseen la cola de poli-Adeninas, pero carecen de intrones y de las regiones reguladoras. Debido a ello, su expresión posterior suele ser poco probable.

Poliploidización	Duplicación completa del genoma que inicialmente resulta en que haya dos copias por cada gen aunque, a largo plazo, se pierden entre un 70 y 90% de los genes duplicados.

De esta forma, la duplicación génica puede ser vista como un proceso dinámico que está moldeada por un balance entre el surgimiento de genes y la eliminación de la mayoría de estos. En general, se considera que la aparición de genes duplicados ocurre a una tasa relativamente alta, de alrededor de 0.01/genes/Ma, los cuales poseen una vida media de alrededor de 4 millones de años (de acuerdo con datos obtenidos de diferentes eucariontes) (Lynch & Conery 2000; 2003). Como se mencionó previamente, la selección purificadora parece operar sobre la mayoría de los duplicados que no se pierden, pero en algunos casos, varios de estos pueden experimentar evolución neutral desde el principio (Lynch & Conery 2000). Por otra parte, la evolución por medio de selección direccional parece ser poco común y solo operaría sobre una minoría de genes recientemente duplicados (Lynch & Conery 2003).

El destino de los genes duplicados que se fijan en el genoma

Para aquellos genes duplicados que logren mantenerse en el genoma, existen distintos caminos que pueden tomar. Susumu Ohno, uno de los pioneros en analizar el papel de la duplicación génica en la evolución biológica, consideraba que, eventualmente, la adquisición de una nueva función era el único destino de los parálogos (Ohno 1970). En muchos casos, esta nueva función puede entenderse como la capacidad de una enzima paráloga de llevar a cabo una reacción diferente a la de la copia, como ha ocurrido con distintas maltasas fúngicas que pertenecen a la misma familia (Voordeckers *et al.* 2012).

La neofuncionalización no implica necesariamente la adquisición de una nueva actividad enzimática. La nueva función también puede darse a nivel de cambios en la expresión génica en una o más de las copias, como ocurre con varios miembros de la familia chalcona sintasa, los cuales se expresan en distintos tejidos y momentos del desarrollo de las plantas (Durbin *et al.* 2000). En otras ocasiones, puede ocurrir un cambio en la función en dos o más de estos genes en comparación con sus parálogos en algún organismo diferente, lo que da lugar a un fenómeno conocido como *barajeo de funciones*. Tal es el caso de los parálogos del gen *hox1*. En el ratón

encontramos tres parálogos de este gen (*hoxa*1, *hoxb*1, y *hoxd*1), a diferencia del pez cebra, en donde el gen *hoxb1* pasó por un evento adicional de duplicación, dando lugar a los genes *hoxb*1a y *hoxb*1b. En este caso, *hoxb*1b es quien lleva a cabo la función que *hoxa*1 desempeña en el ratón, mientras que el parálogo equivalente a este último, *hoxa*1a, ha adquirido una función diferente a nivel de sitio de expresión (McClintock *et al.* 2001).

En algunas ocasiones es posible que, previamente al evento de duplicación, el o los genes involucrados lleven a cabo más de una función. En el caso particular de los que codifican enzimas, esto puede entenderse como la capacidad de catalizar más de una reacción, lo cual se conoce como promiscuidad catalítica, y que puede implicar que la enzima en cuestión lleve a cabo reacciones diferentes sobre uno o más sustratos o la misma reacción sobre sustratos diferentes (Copley 2003). Sin embargo, la subfuncionalización también puede darse al nivel de la expresión génica. Algunos ejemplos de partición de funciones como resultado de un evento de duplicación los encontramos en un los receptores de hidrocarburos de arilos *ahr1α* and *ahr1β* en el eucarionte *Xenopus laevis* (Freeburg *et al.* 2016), en varias bombas de resistencia a múltiples fármacos en la bacteria *Burkholderia cepacia* (Perrin *et al*. 2017), y un grupo de varias proteínas de respuesta al estrés en *Arabidopsis thaliana* las cuales, en comparación con sus ancestros, presentan una reducción en el número de respuestas a condiciones estresantes pero una mayor especificidad en las que conservan, lo cual parece deberse a la pérdida de ciertos elementos reguladores al nivel del DNA (Zou *et al.* 2009). En todos estos ejemplos la división y posterior incremento en la especificidad de funciones se da a nivel regulatorio; más adelante se mencionarán algunos ejemplos específicos de subfuncionalización a nivel catalítico.

La conservación de la función, tanto a nivel metabólico como regulador, de señalización, etc. parece ser el destino más común para los genes duplicados. Esto no solamente ocurre con los genes codificantes sino también con los no codificantes (por ejemplo, transcritos de rRNA); en ambos casos, la mera presencia de copias adicionales puede ser suficiente para conferirle una ventaja a la célula (Zhang 2003).

Que uno o más parálogos con la misma función que el gen original se conserven en el genoma puede deberse principalmente a dos motivos (aunque hay otros que se mencionarán en la sección posterior): 1) se requiere una dosis adicional de determinado transcrito o proteína; 2) la o las copias adicionales actuarán como un tipo de “refuerzo molecular” en caso de que la copia original se vea comprometida (Kuzmin *et al.* 2021). Por ejemplo, en el caso de las histonas de la

familia H4 de muchas especies de eucariontes, varios de sus miembros conservan la misma función que, en este caso, es mantenida gracias a la selección purificadora (Piontkivska *et al*. 2002). Por su parte, en la levadura *Saccharomyces cerevisiae* existen ciertas proteínas glicolíticas cuyos parálogos, originados por medio de un evento de WGD, conservan la misma función. Esto parece conferirle una ventaja al organismo en términos de un mayor flujo metabólico en la vía glicolítica, llevando así a un aumento en la velocidad a la que se fermenta la glucosa (Conant & Wolfe 2007), lo cual podría ser la razón de que estas copias, idénticas en términos funcionales, se hayan conservado en el genoma.

Un escenario diferente en el que los parálogos pueden conservar la misma función es cuando estos adquieren una función de respaldo en caso de que haya algún problema con la expresión de alguna copia. Aparentemente, los parálogos más eficientemente podrían rescatar una función son aquellos cuya regulación transcripcional difiere bastante de la copia que se encuentra comprometida, lo cual sugiere un tipo de *trade-off* entre la capacidad de rescate de las copias y su expresión diferencial (Kafri *et al*. 2005). Esto se ha podido comprobar a partir de un estudio de mutagénesis en *S. cerevisiae*, en el cual se mutó una de las copias de varios grupos de parálogos y se evaluó la expresión de las copias no mutadas. A partir de esto se pudo observar que los parálogos que lograban rescatar la función eran aquellos cuyas condiciones de expresión eran las más diferentes a las de la copia mutada (Kafri *et al*. 2005). Por su parte, en las bacterias también parece haber casos en los que el rescate de función ayuda a preservar los parálogos. Muchos organismos de los phyla Firmicutes y Epsilonproteobacteria poseen varias copias de la enzima RluD, una pseudouridina sintasa que modifica tres posiciones del rRNA. De todas las enzimas que conforman esta familia, RluD es la única cuya inactivación resulta en efectos celulares deletéreos (Gutgsell *et al*. 2001; Liiv *et al*. 2005; Ofengand *et al*. 2001) al no poderse ensamblar correctamente el ribosoma. Por ello, se ha propuesto que las copias adicionales de esta enzima podrían rescatar esta importantísima función en caso de que la que se expresa normalmente esté inactivada (Álvarez-Lugo & Becerra 2023).

Duplicación génica y su relación con condiciones ambientales cambiantes: el caso de los ecoparálogos.

En los últimos años se han descubierto, tanto en arqueas como en bacterias, varios ejemplos de genes parálogos que llevan a cabo la misma función *in vivo* pero cuya expresión depende de las condiciones ambientales en un momento específico, motivo por el cual se les han denominado *ecoparálogos* (Sanchez-Perez *et al*. 2008). La mayoría de los ejemplos descritos tienen que ver con fluctuaciones en parámetros como temperatura, salinidad y disponibilidad de nutrientes. Por su parte, existen otros casos en los que la expresión de una u otra copia está asociada con procesos de patogenicidad pero, en última instancia, esta expresión diferencial también depende de variables químicas como la concentración de determinado ion. En la Tabla 2 se presenta una lista de duplicados que podríamos considerar ecoparálogos y que se han descrito en la literatura.

Tabla 2. Ejemplos de ecoparálogos en arqueas y bacterias.

Organismo	Parámetro	Descripción	Referencia
Bacteria			
Salinibacter ruber	Salinidad	Existen varios conjuntos de parálogos cuyos miembros difieren en su potencial electrostático, tanto en la superficie como en el sitio activo. Esto resulta en diferente halofilicidad para cada una de estas enzimas.	Sanchez-Perez *et al*. 2008
Lactobacillus johnsonii	Distintas condiciones de cultivo	Esta especie posee tres copias de la enzima enolasa (*eno1-3*). En las condiciones de cultivo utilizadas no se transcribió la enzima *eno2*. Aún no se ha caracterizado el papel biológico específico de estos tres parálogos.	Antikainen *et al*. 2007
Bacillus subtilis	Disponibilidad de nutrientes en distintas etapas del ciclo celular	Esta especie posee dos copias de la proteína de unión a cadena sencilla de DNA, involucradas en un mecanismo de reparación llamado “respuesta SOS”. Una copia se expresa en la fase de crecimiento exponencial, mientras que la otra se expresa durante la fase estacionaria si los nutrientes son escasos.	Linder *et al*. 2004
Distintas especies de bacterias	–	Poseen rutas redundantes para la degradación de compuestos aromáticos; cada una de estas podría estar adaptada a condiciones ambientales específicas	Pérez-Pantoja *et al*. 2016
Archaea			
Haloferax volcanii	Salinidad y temperatura	Este organismo tiene dos chaperoninas del grupo II. A pesar de que existe expresión diferencial dependiente de temperatura y concentración de sal, si una es inactivada por mutagénesis, el fenotipo resultante no muestra efecto deletéreo alguno.	Kapatai *et al*. 2006
Haloarcula marismortui	Salinidad	Posee dos parálogos, *fla*A2 y *fla*B, que participan en la biosíntesis de arquelina, el monómero que constituye el arquelo	Syutkin *et al*. 2014

		o flagelo arqueano. *fla*A2 se expresa a altas temperaturas y baja salinidad, mientras que *fla*B lo hace cuando la salinidad es alta.	
H. marismortui	Temperatura	En este organismo existen tres copias de operones de rRNA (A-C). La secuencia de los operones A y C es prácticamente idéntica, mientras que la del operón B difiere bastante en la región promotora. Este último se expresa en mayor concentración a altas temperaturas y, por el contrario, casi no se expresa si la temperatura es baja.	López-López *et al*. 2007
Asociados a patogenicidad (solamente bacterianos)			
Photorhabdus luminescens	Concentración de Mg^{2+}	Estos organismos son simbiontes de nemátodos del género *Heterorhabditis*, los cuales infectan larvas de insectos al secretar grandes números de *P. luminiscens* en la hemolinfa de estos. *P. luminiscens* posee dos copias del gen *ail* (a*il1*P1 and a*il2*P1),el cual codifica una proteína asociada con la virulencia. La expresión de cada una de estas copias podría estar relacionada con el hospedero de la bacteria en un momento determinado.	Mouammine *et al*. 2014
Pseudomonas aeruginosa	Concentración de zinc	El gen *dks*A, involucrado en la regulación de muchos otros (y que en la mayoría de las bacterias de este grupo se encuentra como unicopia), posee dos parálogos con la misma función: *dks*A1 y *dks*A2. El primero se expresa constitutivamente, mientras que el segundo lo hace cuando el zinc es escaso.	Fortuna *et al*. 2022

La duplicación génica y su posible papel en el origen y evolución del metabolismo: hipótesis de evolución retrógrada e hipótesis del reclutamiento enzimático (Patchwork)

Como se mencionó anteriormente, la duplicación génica no es el único proceso que permite la aparición de nuevos genes y nuevas funciones, pero sí uno de los más importantes y, posiblemente, de los más antiguos. Particularmente, pudo haber jugado un papel crucial en la evolución de muchas de las enzimas y rutas que constituyen el metabolismo celular. De hecho, las que quizá son las dos hipótesis de evolución metabólica más aceptadas, tienen como piedra angular a la duplicación de genes (Scossa & Fernie 2020). A continuación se hará una breve descripción de cada una de ellas.

El modelo de evolución retrógrada propuesto por Norman Horowitz (1945) establece que, en etapas tempranas, los primeros seres vivos tomarían del medio los compuestos necesarios para su supervivencia. Dado que estos se habrían originado en etapas prebióticas, eventualmente llegaría un punto en el que uno de esos compuestos se encontrara cada vez en cantidades menores. Esto provocaría una presión selectiva para que una enzima catalizara la síntesis de dicho producto a partir de su precursor inmediato. Evidentemente, el sustrato para dicha reacción también se acabaría en algún punto, por lo que operaría una nueva presión selectiva para producir dicho compuesto, y así sucesivamente hasta llegar a un precursor inicial (Fani 2012). Esta hipótesis no solo es importante porque conecta las primeras etapas de la evolución biológica con la química prebiótica sino que, además, considera a la selección natural como la fuerza que dirigió la evolución del metabolismo ancestral (Becerra 2021). Además, implica que las enzimas resultantes del evento de duplicación sean altamente específicas, es decir, que lleven a cabo una sola reacción.

En cambio, la hipótesis de evolución metabólica por reclutamiento enzimático, más popularmente conocida como “Hipótesis de Patchwork”, establece que en las primeras etapas del metabolismo existía un conjunto reducido de enzimas con una especificidad muy baja, lo cual les permitía llevar a cabo una gran diversidad de reacciones, aunque no de la manera más óptima (Fani & Fondi 2009). De acuerdo con esta hipótesis, desarrollada de manera independiente por Martynas Yčas (1974) y Roy Jensen (1976), las enzimas recientemente duplicadas no necesariamente tendrían que encontrar una función en la misma ruta sino que, probablemente, habrían desempeñado un papel en alguna vía metabólica diferente. A este proceso se le conoce como “reclutamiento” enzimático (Jensen 1976).

Es probable que, en las primeras etapas de evolución metabólica, dos o más de las enzimas disponibles pudieran sintetizar el mismo sustrato, lo cual provocaría cierto nivel de redundancia funcional. A su vez, el producto de una o más enzimas sería el sustrato de otra u otras más, lo cual eventualmente podría conectar a distintas enzimas y así conformar una especie de proto-metabolismo con rutas metabólicas muy poco específicas. Tras una serie de eventos de duplicación, algunas de las enzimas de una ruta en particular podrían adquirir mayor especificidad, lo cual haría que la ruta fuera más eficiente. Por otra parte, las parálogos generados en dichos episodios podrían optimizar alguna otra función y así ir integrando, poco a poco, una ruta metabólica diferente (Rison & Thornton 2002).

A diferencia de la hipótesis por evolución retrógrada, la evolución por reclutamiento presupone que las enzimas ancestrales eran bastante promiscuas catalíticamente hablando (Copley 2003). Adicionalmente, considera a la subfuncionalización como el destino más común de las enzimas duplicadas. Precisamente, el hecho de que una enzima posea más de una actividad para la que utilice el mismo sitio activo implica la existencia de un conflicto que impide que cualquiera de estas se lleve a cabo de una manera eficiente. De acuerdo con un modelo evolutivo conocido como Escape del Conflicto Adaptativo (Des Marais & Rausher 2008), el impedimento de una enzima para optimizar alguna de sus actividades podría resolverse mediante un evento de duplicación, el cual resultaría en la subfuncionalización de las dos copias. Después de este punto, un relajamiento en la selección purificadora podría dirigir la evolución de ambas (Sikosek *et al*. 2012).

El metabolismo celular se ha diversificado de tal manera que sería imposible atribuir su evolución a una sola de las dos hipótesis presentadas. De hecho, la evidencia sugiere que el escenario más plausible involucra una combinación de los modelos de evolución retrógrada y por reclutamiento enzimático (Díaz-Mejía *et al*. 2007). Aun así, este último parece predominar en comparación con la evolución secuencial de enzimas. Evidencias del modelo de “Patchwork” las encontramos en procesos como el ciclo de Krebs (Meléndez-Hevia *et al*. 1996), algunas rutas de biosíntesis de aminoácidos como la del triptófano (Xie *et al*. 2003), leucina (Fondi *et al*. 2007), lisina y arginina (Fondi *et al*. 2007; Velasco *et al*. 2002), así como en el ciclo de la urea (Takiguchi *et al*. 1989) y la biosíntesis de quinato y shikimato, las cuales pertenecen al metabolismo secundario en plantas (Carrington *et al*. 2018). Por su parte, los casos de evolución retrógrada solamente se han detectado en unas cuantas rutas, como en cuatro genes involucrados en la

fijación de nitrógeno (Fani *et al*. 2000) y en un par de genes (*his*A y *his*F*)* que participan en la biosíntesis de histidina (Fani *et al*. 1994).

ARTÍCULO REQUISITO

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 July 2021
doi: 10.3389/fgene.2021.641817

The Role of Gene Duplication in the Divergence of Enzyme Function: A Comparative Approach

Alejandro Álvarez-Lugo[1,2] *and Arturo Becerra*[2]*

[1] Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México, Mexico City, Mexico, [2] Facultad de Ciencias, Universidad Nacional Autónoma de México, Mexico City, Mexico

Gene duplication is a crucial process involved in the appearance of new genes and functions. It is thought to have played a major role in the growth of enzyme families and the expansion of metabolism at the biosphere's dawn and in recent times. Here, we analyzed paralogous enzyme content within each of the seven enzymatic classes for a representative sample of prokaryotes by a comparative approach. We found a high ratio of paralogs for three enzymatic classes: oxidoreductases, isomerases, and translocases, and within each of them, most of the paralogs belong to only a few subclasses. Our results suggest an intricate scenario for the evolution of prokaryotic enzymes, involving different fates for duplicated enzymes fixed in the genome, where around 20–40% of prokaryotic enzymes have paralogs. Intracellular organisms have a lesser ratio of duplicated enzymes, whereas free-living enzymes show the highest ratios. We also found that phylogenetically close phyla and some unrelated but with the same lifestyle share similar genomic and biochemical traits, which ultimately support the idea that gene duplication is associated with environmental adaptation.

Keywords: gene duplication, enzymatic classes, paralogous enzymes, enzyme evolution, function class

OPEN ACCESS

Edited by:
Jorge Humberto Ramirez-Prado, Scientific Research Center of Yucatán (CICY), Mexico

Reviewed by:
Gabriel Moreno-Hagelsieb, Wilfrid Laurier University, Canada
Hong-Yu Zhang, Huazhong Agricultural University, China

***Correspondence:**
Arturo Becerra
abb@ciencias.unam.mx

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: *15 December 2020*
Accepted: *21 June 2021*
Published: *14 July 2021*

Citation:
Álvarez-Lugo A and Becerra A (2021) The Role of Gene Duplication in the Divergence of Enzyme Function: A Comparative Approach. Front. Genet. 12:641817. doi: 10.3389/fgene.2021.641817

INTRODUCTION

Gene duplication is one of the most important mechanisms that lead to the appearance of new genes and new functions (Ohno, 1970) in both prokaryotes (Serres et al., 2009; Wang and Chen, 2018) and eukaryotes (Maere et al., 2005; Panchy et al., 2016). There are distinct categories of duplications: those that comprise one or few genes (small-scale duplication; SSD) and those that comprise many genes (large-scale duplication; LSD) or even the entire genome (whole-genome duplication; WGD). SSDs have been widely documented in both prokaryotes and eukaryotes (Conant and Wagner, 2002). On the other hand, LSDs, specifically WGDs, for a long time had been considered to be an exclusively eukaryotic trait, but recent evidence strongly suggests that it is much prevalent in prokaryotes than we have previously thought (Pecoraro et al., 2011; van de Peer et al., 2017) and that it might be a way to cope with extreme environmental conditions (Soppa, 2017).

Theoretically, almost every gene has a similar probability of being duplicated, but not all are equally retained (McGrath et al., 2014). Most duplicated genes are eventually silenced in the short term (Lynch and Conery, 2000), and those that remain can either retain the original function (Zhang, 2003) or acquire a new one, either by subfunctionalization (a subdivision of an ancestral, often generalist function) or neofunctionalization (acquisition of a novel function) (Walsh, 2003). Besides providing the raw material for the emergence of new gene functions, gene duplication also

seems to play an essential role in the adaptation of organisms to different environments (Gevers et al., 2004; Bratlie et al., 2010; Kondrashov, 2012) and in more complex processes like species diversification and increases in biological complexity (van de Peer et al., 2009).

Gene duplication has been a widespread mechanism in the evolution of metabolism. The Patchwork model (Yčas, 1974; Jensen, 1976), which is perhaps the most accepted model for metabolic evolution, suggests that gene duplication may have played a crucial role at the dawn of metabolism. At this stage, ancient enzymes probably lacked substrate or reaction specificity, allowing them to catalyze different reactions involving more than one substrate. Over time, one or more of these ancestral activities could have become so important that the ancestral enzyme could not have carried them out in the most efficient way. Thus, a duplication event involving such an enzyme could have led to a new copy with increased specificity. According to this model, throughout evolution, different metabolic pathways could have been assembled from the recruitment of newly evolved enzymes (Lazcano and Miller, 1999; Schmidt et al., 2003; Caetano-Anollés et al., 2009; Fani and Fondi, 2009; Becerra, 2021). Evidence of episodes of gene duplication leading to the enrichment of metabolic functions is found in both ancient and recent metabolic innovations. For example, it has been suggested that around three billion years ago, in a period known as the Archaean genetic expansion, gene duplication contributed to the appearance of new genes involved in respiratory and electron-transport pathways (David and Alm, 2011). It also seems to have fostered the expansion of many secondary metabolic pathways in plants (Weng et al., 2012; Moghe et al., 2017). Moreover, even for recently evolved pathways, like the mandelate pathway in several *Pseudomonas* species (Petsko et al., 1993), there is compelling evidence suggesting that some of the enzymes involved may have arisen by gene duplication.

It is now generally assumed that early life could have done well with a very limited set of enzymatic functions (Goldman et al., 2012), which could serve as a starting point for the evolution of new functions through scenarios involving gene duplication and other mechanisms like domain combinations, which could also lead to the appearance of functions other than catalytic activity (Bashton and Chothia, 2007). It has also been suggested that an interplay between the patchwork and the retrograde evolution model (Horowitz, 1945) is more likely than either of the two separately (Díaz-Mejía et al., 2007). Today, we can observe the outcome of these processes in the great functional diversity found within families and superfamilies of enzymes, at the levels of catalytic machinery, substrate specificity, and reaction chemistry (Bartlett et al., 2003; Furnham et al., 2016), though it is more common to see a greater substrate diversity within a single superfamily (Todd et al., 2001). Additionally, it is quite common to see drastic functional changes across the evolutionary history of enzymes. This is illustrated by the fact that changes in enzymes' primary function (defined by the first digit of the Enzyme Commission number) have been observed between every enzymatic class, though some are more frequent than others (Furnham et al., 2012; Martínez Cuesta et al., 2015). But ultimately, what seems to be more important for the appearance of new functions is the inherent capacity of an enzyme to accept different substrates and/or perform different reactions (known as substrate and catalytic promiscuity, respectively) and its ability to evolve new functions in a changing environment (Tyzack et al., 2017).

The current enzyme classification system, which assigns a unique four-digit number for each enzyme, is exclusively based on the biochemical activities performed by each enzyme and groups them in terms of reaction similarity (McDonald et al., 2015), and not by evolutionary-related members. It was established during the early 60s by the International Commission on Enzymes from the International Union of Biochemistry and Molecular Biology (Tipton and Boyce, 2000). Until the first half of 2018, the classification remained without significant changes and consisted of six enzymatic (EC) classes, divided into different sub and sub-subclasses (McDonald and Tipton, 2014). However, in the second half of 2018, a new enzymatic class was added: the translocases (EC 7). A statement made in the ExplorEnz database (McDonald et al., 2009) highlighted the importance of a group of enzymes whose main function is the movement of ions or molecules from one side of biological membranes to the other. Many of these perform a different reaction as a means of achieving the movement of substances across membranes.

In this work, we try to analyze the role of gene duplication in the diversification of enzymatic functions across the enzymatic classes of the Enzyme Commission (EC) classification, including the recently proposed translocases (EC 7). We further explore the possible link between organisms' lifestyle and specific patterns of retention of duplicated enzymes. Besides, due to recent proposals of a two-domain view of life, which suggests that eukaryotes do not constitute a separate domain but are part of the Archaea domain (Williams et al., 2013; Doolittle, 2020), we decided only to consider prokaryotic organisms, which as a group possess a much wider biochemical repertoire than that for eukaryotes.

MATERIALS AND METHODS

Proteomes Analyzed

The complete set of prokaryotic proteomes was downloaded from the KEGG Database (Kanehisa and Goto, 2000). We selected a sample of non-redundant, representative proteomes based on criteria reported elsewhere (Martínez-Núñez et al., 2013, 2015). Altogether, we analyzed 655 bacterial and 90 archaeal proteomes (**Supplementary Data Sheet 2**). These belong to organisms whose genome has been completely sequenced, except for those from the phyla Bathyarchaeota and Lokiarchaeota, which come from metagenomic sequences.

Identification of Within-Genome Paralogous Sequences

For this work, the criteria for defining paralogous proteins included an *E*-value cutoff of 10e-07 and query coverage $\geq$70%. We performed an *all against all* BlastP search (Altschul et al., 1997) for each of the 745 proteomes from the sample. Different Perl *ad hoc* scripts were used to filter the BlastP results and retain only those sequences that fulfilled the above criteria.

Identification of Enzymes

Once we filtered the BlastP results, we extracted the IDs from the proteomes and paralogous data sets and crosschecked them with the FTP files downloaded from the KEGG Database. Additionally, the online tool db2db, from the bioDBnet resource (Mudunuri et al., 2009), was used to corroborate the enzyme codes (EC numbers) for all the paralogous-enzyme sequences. These are taken directly from the KEGG database. We considered all the sequences for which we obtained, at least, the first digit from the EC number, which refers to the general function of the enzyme (Tian and Skolnick, 2003; Concu and Cordeiro, 2019). Sequences for which we did not obtain an EC number were excluded from the subsequent analysis. EC codes from translocases (EC 7) had not been properly updated in the db2db tool. To solve this problem, we identified which enzymes had changed their EC code and manually updated them.

Ratio of Paralogous Enzymes

We counted the number of enzymes and sorted them into one of the seven enzymatic classes for each of the proteomes and their respective paralogous data sets. The ratio of paralogous enzymes per class was defined as the ratio between the number of paralogous enzymes and the number of enzymes found within the proteome. In sum, we obtained seven different ratios per organism.

Statistical Analysis

Non-parametric Kruskal–Wallis tests were used to evaluate the difference between paralogous enzymes for all enzymatic classes, followed by Dunn tests with the Bonferroni adjustment to identify those classes which differed significantly. Additionally, Spearman's test was used to evaluate the relationship between the number of proteins and the number of enzymes, and a number of different regression analyses were also performed. In all cases, statistical significance was set at $p \leq 0.05$. All statistical analyses were done with the R programming language (R Core Team, 2020) in the RStudio software (RStudio Team, 2020).

Lifestyles Identification

After selecting our representative sample, we assigned the lifestyle to each of the organisms in our set. Such lifestyles are free-living, extremophile, pathogen, and intracellular. We relied on data from Martínez-Núñez et al. (2013) and the prokaryotic metadatabase BacDive (Reimer et al., 2019), accessed through specific entries for each organism in the NCBI Taxonomy Browser[1], which has specific entries for each strain.

RESULTS

The Relationship Between Enzymes, Proteins, and Genome Size Follows a Power-Law Distribution

Before analyzing paralogous enzymes' content, we inspected the relationship between enzymes, proteins, and genome size. Visually, it seemed that there was a linear relationship between each pair of those variables. However, regression analyses revealed that a power-law function was the best that explained our data (**Figure 1**). This makes more sense for the relationship between enzymes and proteins, and for enzymes and genome size (**Figures 1A,B**), because there are different kinds of proteins (i.e., regulatory, structural, etc.) encoded in genes. So, as genomes grow, one does not necessarily expect that organisms accumulate a higher ratio of enzymes because that would imply that many more regulatory proteins would be needed to regulate those enzymes (Koonin and Wolf, 2008). However, one would expect a linear relationship between the number of proteins and the genome size. As **Figure 1C** shows, this is not precisely the case due to, perhaps, the organisms with the smallest genomes (lower-left part of the figure).

As in the previous point, we performed the same analysis with the sample divided by its lifestyle. The results are shown in **Supplementary Figures 1–4**. We found the same trend for variable comparison for free-living and pathogen organisms, as in **Figure 1** (i.e., a power-law distribution) (**Supplementary Figures 1, 3**). Surprisingly, for extremophile organisms, this was not the case. In all cases, we found a linear relationship between each pair of variables (**Supplementary Figure 2**). It is interesting to note that this is perhaps the most homogenous group of organisms concerning genome size (most of them have a genome under six megabases (Mb), and none of them has a genome less than 1 Mb). Finally, we found a trend like that of the extremophiles for intracellular organisms, with one exception. Linear regression is what best explains the relationship between the number of enzymes and the number of proteins and genome size, although this is not the case for the relationship between proteins and genome size, which follows a power-law distribution (**Supplementary Figure 4**).

The Ratio of Paralogous Enzymes Also Follows a Power-Law Distribution

The ratio of paralogous enzymes within each proteome was calculated by dividing the number of paralogous enzymes identified in each proteome by the same proteome's total number of enzymes. We defined as "enzymes" all those sequences that had assigned the first number of the EC code, which indicates the general function of the enzyme. We considered the ratio instead of the total number of enzymes because there was such a disparity across organisms' whole sample. So, this was a way to eliminate the bias associated with such disparity and homogenize the data. As shown in **Figure 2**, the relation between those variables follows a power-law distribution ($R^2 = 0.68$). It is noteworthy that such a ratio is less than 0.6 for most organisms (less than ten organisms have a higher ratio; their number of enzymes goes from 1000 to 2000).

The Ratio of Paralogous Enzymes Differs Between the Different Enzymatic Classes

We performed a Kruskal–Wallis test to evaluate if there was any difference in the ratio of paralogs between different enzymatic classes. The *P*-value was statistically significant ($P \leq 2.2e\text{-}16$), and

[1]https://www.ncbi.nlm.nih.gov

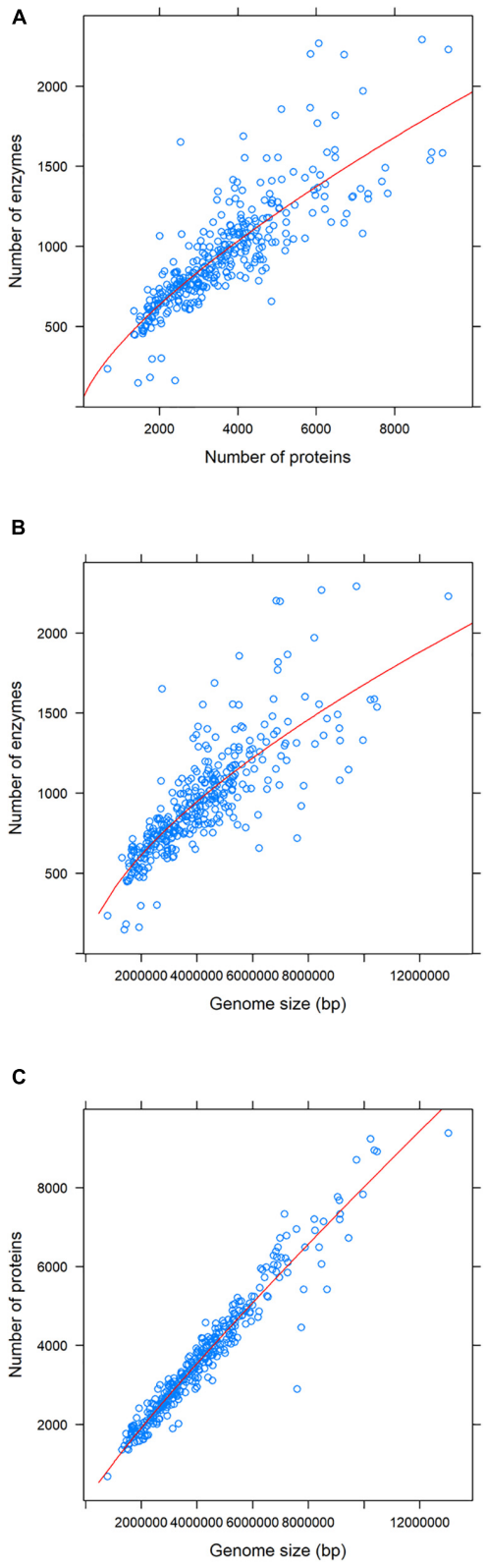


FIGURE 1 | Relation between the enzyme and protein content, and the genome size. For each pair of variables, a power-law equation is the one that best explains the distribution of the data. The equations and *R*-squared values are as follows: **(A)** $y = 2.51x^{0.72}$; $R^2 = 0.79$, **(B)** $y = 0.04x^{0.66}$; $R^2 = 0.76$; **(C)** $y = 0.004x^{0.9}$; $R^2 = 0.95$.

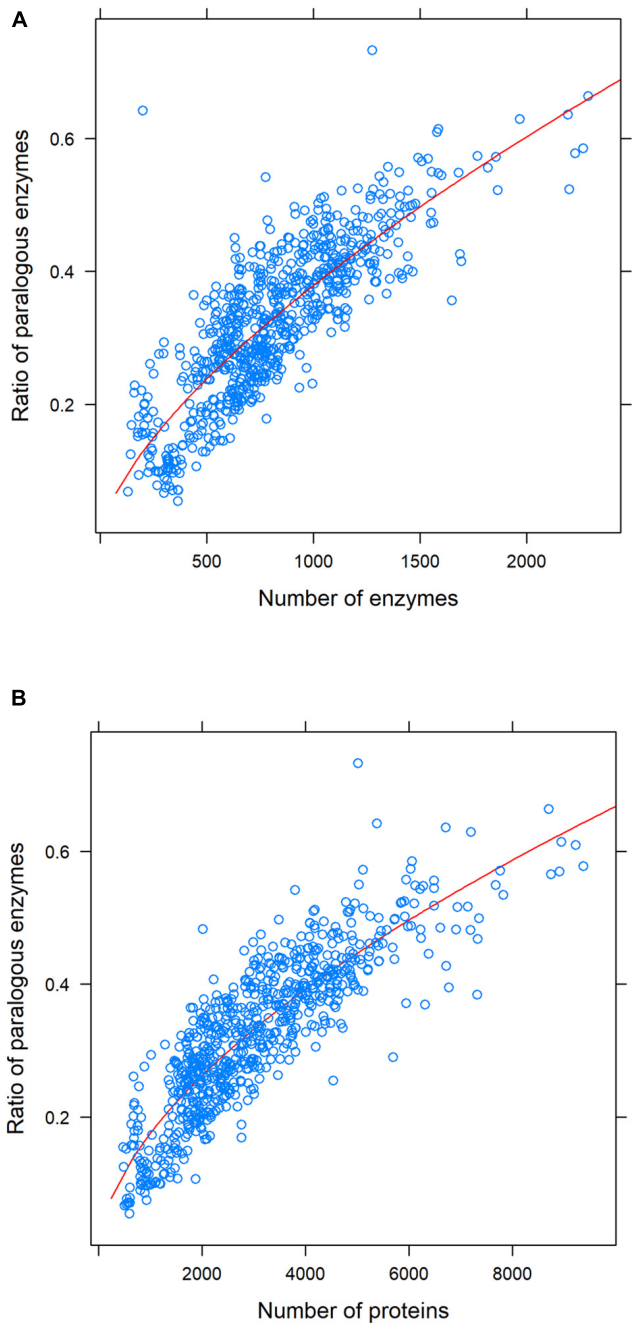


FIGURE 2 | Ratio of paralogous enzymes as a function of the number of enzymes **(A)** and the number of proteins **(B)**, for each individual genome. The power-law equation for each adjustment and the *R*-squared value are as follows: **(A)** $y = 0.003x^{0.67}$; $R^2 = 0.68$; **(B)** $y = 0.003x^{0.58}$; $R^2 = 0.75$. Note that the *R*-squared value is higher when the number of proteins is considered ($R^2 = 0.75$) instead of the number of enzymes ($R^2 = 0.68$).

so we then performed a *post hoc* Dunn test with the Bonferroni adjustment in order to identify between which classes there was a significant difference (**Figure 3A** and **Supplementary Table 1**). The α value was set at 0.05, and the *P*-value at α/2 (*P*-value = 0.025). Overall, we found three enzymatic classes whose ratio of paralogs differed significantly from all the others:

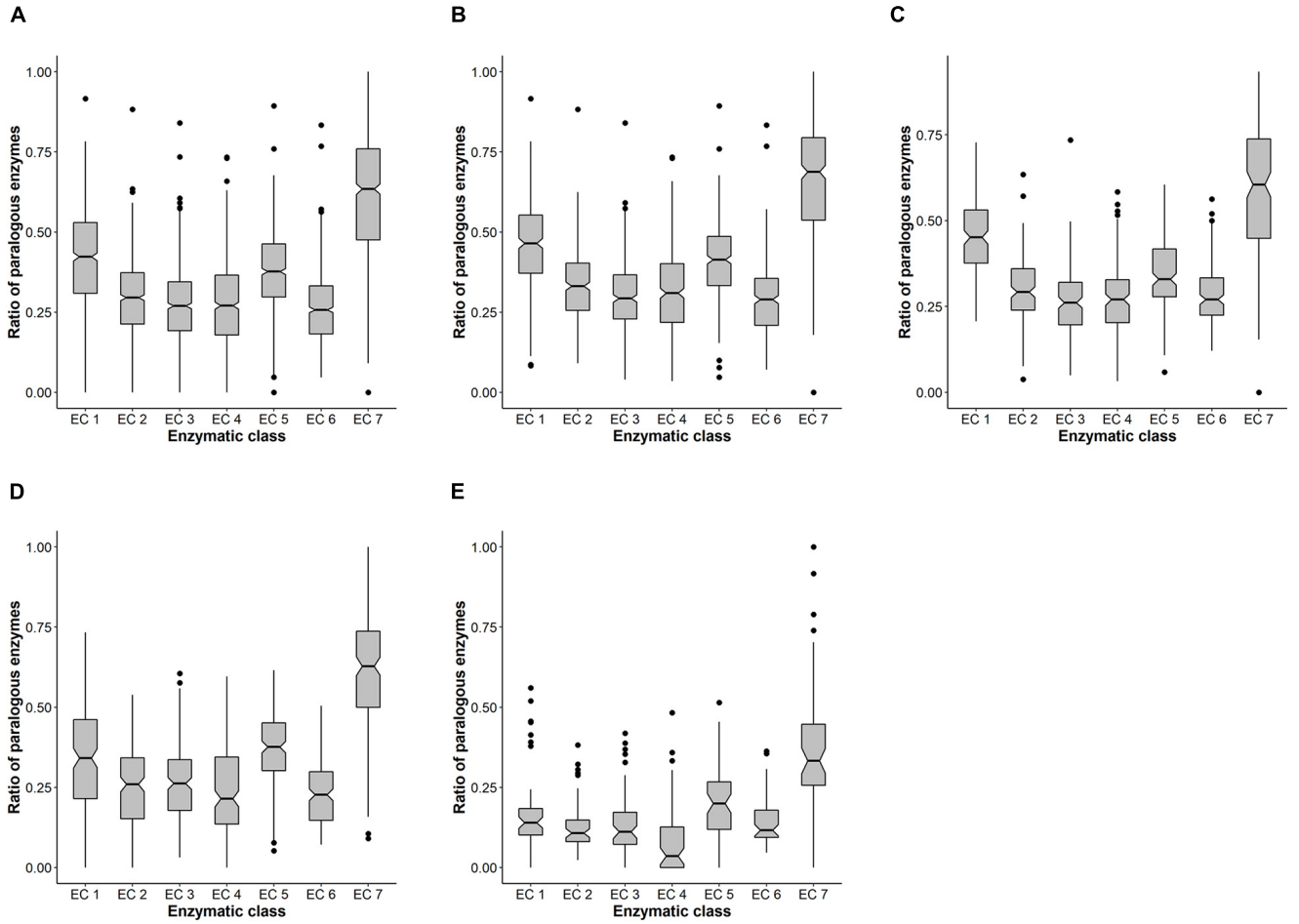


FIGURE 3 | Ratio of paralogous enzymes in prokaryotic organisms. In panel **(A)**, the ratio of paralogous enzymes is plotted for the whole sample and is separated by enzymatic class. For panels **(B–E)**, we also plot the ratio of paralogous enzymes for each enzymatic class but separated by the four different lifestyles in which we sorted our initial sample: **(B)** free-living organisms; **(C)** extremophiles; **(D)** pathogenic (non-intracellular) organisms; **(E)** intracellular organisms (both endosymbiont and intracellular pathogens).

the Oxidoreductases, the Isomerases, and the recently created Translocases.

We then wondered if this trend was found in different sub-samples of prokaryotic organisms or if we were detecting significant differences due to the large dataset we were considering. It has been previously reported that the number of enzymes differs significantly among different lifestyles of organisms (Martínez-Núñez et al., 2015), so we decided to investigate if the same thing also happened regarding the number of paralogs. To do so, we reclassified our sample into four sub-samples. These correspond to different lifestyles: free-living, extremophile, pathogen, and intracellular. Each organism's lifestyle was identified using the bacterial metadatabase Bac*Dive* (Reimer et al., 2019). We performed a Kruskal–Wallis test for each of the four sub-samples, and we found significant differences in all cases. Afterward, we performed a Dunn test and obtained similar results to those of the whole sample (**Figures 3B–E** and **Supplementary Table 2**). In summary, the trend we found in the whole sample, regarding those classes with a significantly higher ratio of paralogs, is also found no matter the organisms' lifestyle. Isolated exceptions are found in extremophiles between classes EC 1 and EC 7, for which there are no significant differences (**Figure 3C**); in pathogens, between EC 1 and EC 5 (they do not differ significantly) (**Figure 3D**); and in intracellular organisms (**Figure 3E**), for which the ratio of paralogous oxidoreductases and isomerases is underrepresented.

The Ratio of Paralogous Enzymes Differs Among Lifestyles

As was noted previously, we found that some enzymatic classes have significantly higher ratios than others within each lifestyle and that this pattern, if not the same, was quite similar within each of the four lifestyles that we considered. We also wanted to know if there were any differences in the ratio of paralogs *among* the different lifestyles. A Dunn test with the Bonferroni adjustment was performed for the whole dataset to evaluate whether paralogous enzymes' overall ratio was either the same or different when comparing the four lifestyles. The α value was set at 0.05, and *P*-value at α/2 (*P*-value = 0.025). As it is

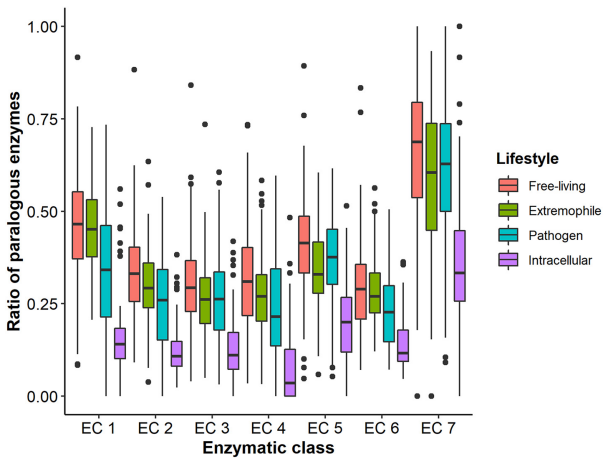


FIGURE 4 | Boxplot showing the comparison of the average paralogous-enzymes ratio across the different lifestyles. Black dots represent the outliers that are found within each lifestyle. Interestingly, the lifestyle within which we found the highest number of outliers was in the intracellular organisms.

shown in **Supplementary Table 3** and **Supplementary Figure 5**, we found significant differences among each lifestyle, and the highest ratio is found for the free-living organisms, followed by the extremophiles (both over 30%), then pathogens (less than 30 but over 20%) and, finally, intracellular organisms (less than 20%) (**Supplementary Table 4**).

A similar approach was taken to compare each class among the four lifestyles. Although we obtained similar results to those when we analyzed the ratio without separating it by enzymatic classes, we think some exceptions are worth mentioning. These are listed below and shown in **Figure 4** and **Supplementary Table 5**.

1. ***Oxidoreductases.*** This is one of the classes with the highest ratio values, mainly for free-living and extremophile organisms (both have a ratio higher than 40%), but there are no significant differences among them. This is the only case for this class in which ratios are not statistically significant. Their corresponding paralogs-ratio is higher than in pathogens and intracellular organisms.
2. ***Transferases.*** For this class, the ratios follow the same trend as in the whole dataset. We did not find non-significant differences.
3. ***Hydrolases.*** This class exhibits lower paralogous-enzymes ratios than the oxidoreductases and transferases. The highest ratio corresponds to free-living organisms and is roughly 30%. Extremophiles and pathogens have very similar values (26–27%), whose difference is non-significant. The intracellular have a ratio of less than 15%.
4. ***Lyases.*** For this class, the difference between the ratios was always significant. The ratio for free-living organisms is slightly higher than 30%, followed by extremophiles and pathogens (between 20 – 30%). Intracellular organisms possess the lowest ratio, which is lower than 10%. It is noteworthy that this is the lowest ratio in this group of organisms.
5. ***Isomerases.*** This is one of the enzymatic classes in which we found some of the highest ratios. For free-living organisms, the ratio is slightly higher than 40%, followed by the pathogens (37%), the extremophiles (35%), and intracellular organisms (20%). This ratio was exceptionally high for this last group and is only surpassed by that of translocases. For this enzymatic class, the only non-significant difference was found between extremophiles and pathogens.
6. ***Ligases.*** In this case, none of the ratios is higher than 30%, although in free-living, extremophile and pathogen organisms are higher than 20%. This ratio is slightly less than 15% in the intracellular organisms. For extremophile and free-living organisms, there are no significant differences.
7. ***Translocases.*** This recently created enzymatic class exhibits the highest ratios of paralogous enzymes. For all the groups but intracellular organisms, such a ratio is well over 50%, and the difference is non-significant only between pathogens and extremophiles. Even the intracellular organisms have a high ratio, slightly fewer than 40%.

Taking these results together, we can argue that the extremophiles represent perhaps the most interesting group in terms of their paralogous-enzyme content. They seem to be in-between the free-living and pathogenic organisms, sometimes very close to one or the other. This is reflected by the fact that the only five cases in which we found similar, non-significant ratios involved the extremophiles. There were non-significant differences between extremophiles and pathogens in three such cases, and the other two, between free-living organisms and extremophiles. For the intracellular organisms, the ratio difference was always the lowest (and always significantly) for each of the seven enzymatic classes.

Detailed Exploration of the Parologous Enzymes Ratio

Our data clearly show an overrepresentation of paralogous enzymes within oxidoreductases, isomerases, and translocases. However, considering only the enzymes' general function gives us scarce information about the patterns found within each class. This is important because there is an unequal number of subcategories within each enzymatic class, inherent to the Enzyme Commission classification system (**Table 1**). Furthermore, if we want to get a complete picture of the reasons underlying the high ratio of paralogs within these categories, a deeper analysis breaking down each category could be quite useful.

We identified the number of paralogs within each of the subclasses from the above-mentioned enzymatic classes for our whole dataset. Given that this was an exploratory analysis, we considered that the average value for each individual phylum could be a good starting point. So, we averaged the number of paralogous enzymes for each subclass, and we report the values *per* phylum for each of them. The results are separated

TABLE 1 | Number of subcategories and entries for each enzymatic class.

Enzymatic class	EC code	No. of subclasses	No. of sub-subclasses	No. of enzymes
Oxidoreductases	EC 1	26	148	1798
Transferases	EC 2	10	38	1900
Hydrolases	EC 3	13	66	1360
Lyases	EC 4	8	17	677
Isomerases	EC 5	7	19	310
Ligases	EC 6	6	12	203
Translocases	EC 7	6	10	90

Data as of November 2020, taken from the ExplorEnz Database (McDonald et al., 2009).

by enzymatic classes and are presented as different heatmaps (**Figure 5** and **Supplementary Figure 6**).

Separating the data into three different heatmaps allows us to make direct comparisons within each enzymatic class. The maximum number of paralogous oxidoreductases (about 44 in Betaproteobacteria) exceeds the same value for the isomerases (about 12 in several phyla). Besides, within each enzymatic class, there is also a significant disparity in the average number of paralogs. The most extreme cases are oxidoreductases, within which subclasses EC 1.1 and EC 1.2 are the ones with the highest values, followed by EC 1.3 and EC 1.8, but to a much lesser degree. For isomerases, the subclass with the highest numbers of paralogs is EC 5.4, followed by EC 5.1. However, unlike oxidoreductases, the difference between isomerases' subclasses is less than between oxidoreductases' subclasses. Finally, for translocases, we found the highest ratio of paralogs for subclass EC 7.1, followed by EC 7.2. For many phyla, both subclasses exhibit similar values, though there are some cases in which EC 7.1 exceeds considerably EC 7.2.

Phylogenetically and Lifestyle-Related Phyla Share Similar Genomic and Biochemical Traits

One of the main questions at the beginning of this study was whether similar organisms would share similar ratios of paralogous enzymes in terms of their phylogenetic position or lifestyle. To address this question, we performed a principal component analysis (PCA). Overall, we considered 11 variables: genome size, number of proteins, number of paralogs, number of enzymes, and the ratio of paralogous enzymes for each enzymatic class (EC 1–EC 7). As a first approach, we decided to perform this analysis with the mean values for each of these variables *per* phylum instead of individual organisms. This was due mainly to two reasons: (1) we wanted to know if there was a global pattern that might show clear differences among different phyla, and (2) given the great variation that we found for each of the eleven variables, considering individual organisms maybe would have been counterproductive, and general patterns much harder to identify. Besides, most phyla are grouped into a broader category: the superphylum. This way, it is easier to identify similarity patterns between different phyla. The only exceptions that were considered as individual phyla were the Aquificae, Thermotogae, and Spirochaetes (**Supplementary Table 6**) due to their lack of assignment to a superphylum. The results from the PCA are depicted in **Figure 6**. We decided to exclude the phylum Lokiarchaeota from the present analysis because it considerably skewed the rest of the data points (data not shown). Given that the proteome assembled for this phylum lacks a proper annotation, we think its removal from the analysis is well justified. As shown in **Figure 6**, the two main components explain the variation of nearly 80% of our data (PC1 = 67.7%; PC2 = 11.4%).

By taking the current approach, in which we considered the mean values *per* phylum for each of the variables, we found several interesting clusters of different phyla. The most striking result is that some phyla seem to be clustered by their lifestyle, while its phylogenetic closeness more clearly clusters others. As examples of the first type of clustering, we distinguish two main groups. One is formed by phyla whose majority of members lives in extreme or anoxygenic conditions and includes the following: Deinococcus–Thermus, Chlorobi, Aquificae, Thermotogae, Deferribacteres, Dictyoglomi, Nitrospirae, and the archaeal phylum Euryarchaeota (**Figure 6**, numbers 15, and 28–34). All of these belong to different superphyla. The other cluster comprises phyla in which many of its members undergo genome shrinkage due to an intracellular lifestyle. These are: Tenericutes, Elusimicrobia, and Bacteroidetes (**Figure 6**, numbers 11, 25, and 26). We also found two other clusters comprising closely related phyla that do not necessarily share the same lifestyle. The most remarkable case can be seen on top of the plot (**Figure 6**, numbers 35–37, and 39), including the Crenarchaeota, Thaumarchaeota, Korarchaeota, and Bathyarchaeota phyla. These are not only phylogenetically close, but they are all included within the TACK group of Archaea (Guy and Ettema, 2011; Spang et al., 2017). Finally, we also found that most of the proteobacteria phyla group together (**Figure 6**, numbers 1–3, and 5–7). The only proteobacteria phylum which is far from this group is the Epsilonproteobacteria (**Figure 6**, number 4) and is shown in the lower-right portion of the plot.

DISCUSSION

Most Paralogous Genes in Prokaryotes Are Likely to Arise by SSD Events

The issue of LSDs and polyploidy in prokaryotes has only raised concerns until very recently. Given that in this analysis we did not make a distinction between paralogs originated by SSDs or WGDs, it could be argued that our results might be biased in some respects. Nonetheless, we do not consider this to be a severe issue.

Polyploidy does not appear to be unusual in prokaryotes (Soppa, 2011), but unlike eukaryotes, ploidy level in Bacteria and Archaea may vary depending on environmental conditions like growth rate, growth phase, among others (Breuert et al., 2006; Soppa, 2017). Besides, there does not seem to be a correlation between the ploidy level and factors such as growth temperature or lifestyle, as occurs in proteobacteria (Pecoraro et al., 2011). Having multiple genome copies could confer prokaryotes with protection against double helix breaks or serve as a phosphate

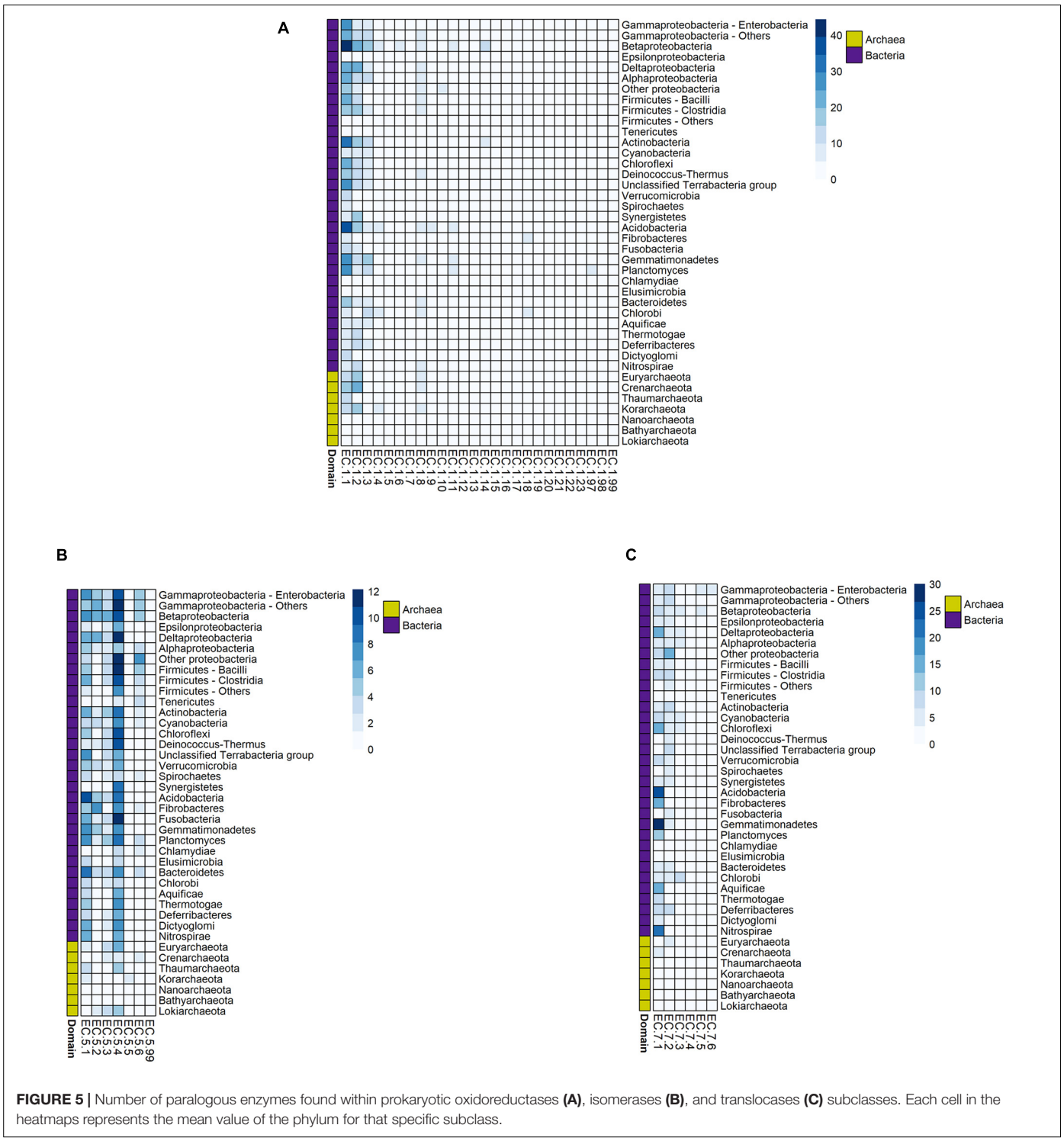


FIGURE 5 | Number of paralogous enzymes found within prokaryotic oxidoreductases **(A)**, isomerases **(B)**, and translocases **(C)** subclasses. Each cell in the heatmaps represents the mean value of the phylum for that specific subclass.

reserve in phosphate-poor environments (van de Peer et al., 2017). Other benefits could be a reduction in the rate of spontaneous mutations and a way of regulating gene expression (Pecoraro et al., 2011). It has also been reported that in some of the biggest bacteria, which in many cases also have one of the largest genomes known to date, having multiple genome copies in specific parts of the cell can serve as a means of optimizing the production of locally required proteins (for example, transporters in the cell periphery) (Angert, 2012). Besides, some cultivated, monoploid bacteria could undergo one or more WGD events due to the lack of selective pressures under laboratory conditions (Soppa, 2017).

It is plausible that several organisms from our sample, either or not cultivated, have one or more copies of their entire genomes but as the evidence suggests, different genome copies are not joined together into a single chromosome but separated from

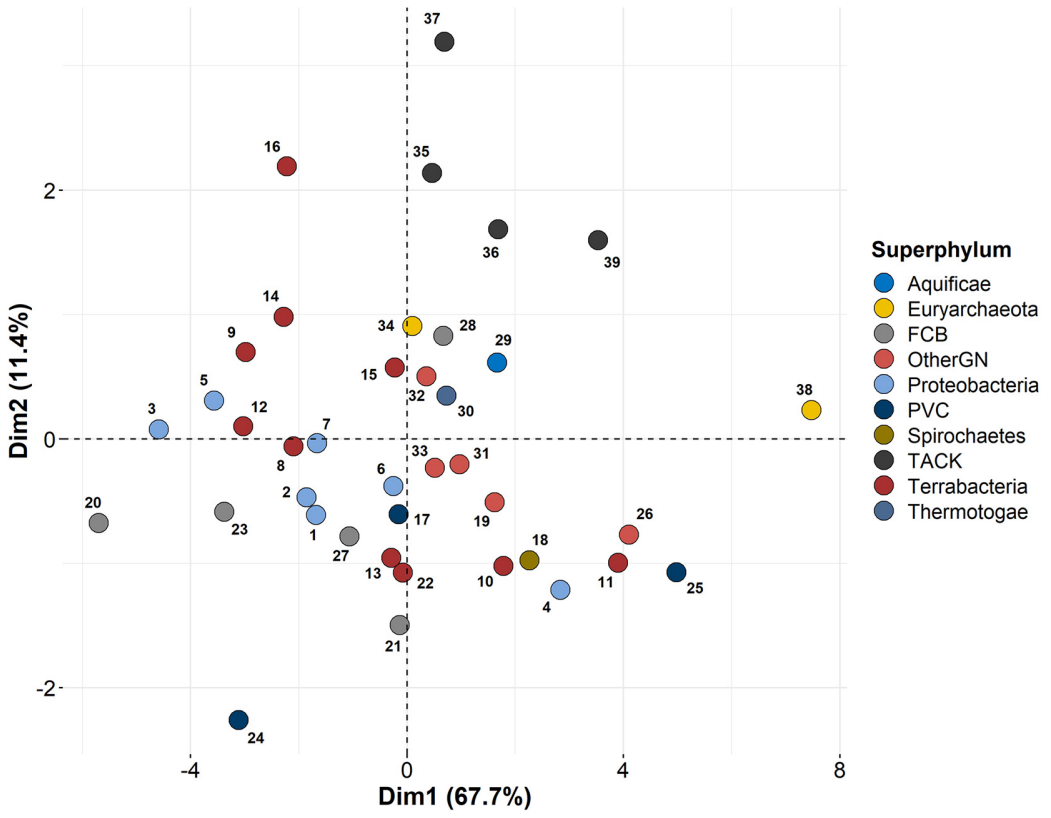


FIGURE 6 | Principal component analysis, considering the mean values for the following 11 variables: genome size, number of proteins, number of enzymes, and number of paralogous proteins, plus the average ratio of paralogous enzymes for each of the seven enzymatic classes. Each circle represents a single phylum, according to the KEGG Organisms Database, and those grouped into the same superphylum are depicted with the same color. The only exceptions that aren't included into a supergroup but considered as individual phyla are the Aquificae, Thermotogae, and Spirochaetes. Single phyla are indicated by the following numbers: (1) Gammaproteobacteria-Enterobacteria, (2) Gammaproteobacteria-Others, (3) Betaproteobacteria, (4) Epsilonproteobacteria, (5) Deltaproteobacteria, (6) Alphaproteobacteria, (7) Other proteobacteria, (8) Firmicutes-Bacilli, (9) Firmicutes-Clostridia, (10) Firmicutes-Others, (11) Tenericutes, (12) Actinobacteria, (13) Cyanobacteria, (14) Chloroflexi, (15) Deinococcus-Thermus, (16) Unclassified Terrabacteria Group, (17) Verrucomicrobia, (18) Spirochaetes, (19) Synergistetes, (20) Acidobacteria, (21) Fibrobacteres, (22) Fusobacteria, (23) Gemmatimonadetes, (24) Planctomyces, (25) Chlamydia, (26) Elusimicrobia, (27) Bacteroidetes, (28) Chlorobi, (29) Aquificae, (30) Thermotogae, (31) Deferribacteres, (32) Dictyoglomi, (33) Nitrospirae, (34) Euryarchaeota, (35) Crenarchaeota, (36) Thaumarchaeota, (37) Korarchaeota, (38) Nanoarchaeota, (39) Bathyarchaeota.

each other and distributed along the cytoplasm. On the other hand, genes originated by SSDs are maintained in the bacterial chromosome until they become non-functional or acquire a function. So, when a prokaryotic genome is sequenced, it is highly likely that the obtained set of genes correspond to those located in a single genome copy and, therefore, would include only those paralogs originated by SSDs.

The presence of additional genome copies could have an impact on different kinds of studies, such as those that measure total amounts of DNA, RNA or proteins. But in our case, we think it is safe to say that we are only considering paralogous genes that are the product of SSD events, though the possibility of including in some cases ohnologs cannot be absolutely discarded.

A Power-Law Function Explains the Relationship Between Proteins, Enzymes and Genome Size

When evaluating the relationship between proteins, enzymes, and genome size in the whole sample, we identified that the function that best fits each pair of variables was a power-law function. The most obvious cases are shown in **Figures 1A,B**, and involve the number of enzymes. Not all of the proteins within each genome have a catalytic function (some can be regulatory or structural proteins), and it has been shown that as prokaryotic genomes increase in size, there is an exponential growth of transcription factors (van Nimwegen, 2003) and that the opposite happens for enzymes (the larger the genome, the lower the number of enzymes/genome-size ratio) (Martínez-Núñez et al., 2013). We could say that as genomes increase their size, they also increase their protein content almost in the same proportion, which indirectly tells us that prokaryotic genomes are mainly composed of coding DNA (Koonin and Wolf, 2008). **Figure 1C** shows this trend, which closely resembles a linear relationship though fitting to a power-law distribution.

Regarding the ratio of paralogous enzymes, we found that it follows a power-law distribution when plotting it against the number of enzymes (R^2 = 0.68) (**Figure 2**). For most organisms, such a ratio is between 0.2 and 0.4, which means that around

20–40% of their enzymes have at least one paralog. Congruently, most organisms with ratios lower than 0.2 are intracellular. This seems to reflect the genome reduction that happens in both endosymbionts (Wernegreen, 2015) and intracellular parasites (Sakharkar et al., 2004). It has also been shown that many intracellular organisms lose many enzymes (Price and Wilson, 2014; Manzano-Marín and Latorre, 2016). We found that the ratio of paralogous enzymes seems to reach a plateau at about 0.6. Only seven organisms exceed this value (six free-living and one extremophile), and 42 out of more than 700 organisms have a ratio higher than 0.5. One possible explanation is that there are probably more paralogs that we are not detecting with the chosen criteria. However, given that we are considering a representative sample of prokaryotes (which includes early and recently diverged lineages and some of the organisms with the largest genomes), this seems unlikely. Another more likely explanation considers the essentiality of the enzymes' function. Although almost every gene can undergo duplication, not all of them possess the same likelihood of being retained. For example, in the eukaryote *Caenorhabditis elegans*, essential genes duplicate less often than non-essential ones but are more likely to be retained over more extended periods (Woods et al., 2013). It also has been noted that changes in the dosage of specific genes could lead to strong deleterious effects (Rice and McLysaght, 2017). However, many duplicated genes could persist if a higher gene dosage is advantageous for the organism (Kondrashov et al., 2002). Thus, one possibility is that some of the enzymes for which we found no paralogs carry out functions for which an increased dosage would result in a disruption of the metabolic flux, which in turn could compromise cell integrity. Another possibility is that, for any given query sequence, one or more of the targets are not enzymes. These are commonly known as pseudoenzymes (Jeffery, 2020). For example, Belitsky (2004) has shown that a pyridoxal 5′-phosphate(PLP)-dependent transcriptional regulator from *Bacillus subtilis* belongs to the same superfamily of a kind of PLP-dependent aminotransferases. A similar case occurs with protein kinases, which comprise one of the most diverse microbial enzyme superfamilies in terms of structure and function (Kannan et al., 2007). Phylogenetic analyses reveal that pseudokinases (that is, proteins with a kinase domain but without catalytic activity) are widely distributed throughout the tree of life (mainly in eukaryotes and bacteria) and have a pivotal, non-catalytic role in signaling processes (Kwon et al., 2019).

High Levels of Promiscuity and Evolvability Within Oxidoreductases May Explain Their High Ratio of Paralogs

After identifying the ratio of paralogous enzymes for each enzymatic class, we noticed no clear relationship between this and the abundance of such enzymes in the genome. If this were so, one would expect that classes containing many enzymes would also show the highest ratio of paralogs. However, for the three more abundant classes (**Table 1**), only the oxidoreductases have a high ratio of paralogs (around 0.41), which is significantly higher than that of transferases (0.29) and hydrolases (0.27) (**Figure 3**). One possible explanation for this is the tremendous functional diversity within the oxidoreductases, which is reflected in the number of subdivisions within this class (**Table 1**). The oxidoreductases have the greatest number of subclasses amongst all enzymatic classes, and the same is observed when considering sub-subclasses. By comparing this with what is observed for the transferases, which is the class with the highest number of enzymes (**Table 1**), we can see that oxidoreductases' subclasses exceed those of translocases by a factor of 2.6, whereas for sub-subclasses, it is by a factor of 3.9.

One possible explanation for why we see so much functional diversity within the oxidoreductases, which we think might also account for the high ratio of paralogs within this class, has to do with enzyme promiscuity. Promiscuous enzymatic activities are those physiologically irrelevant reactions that an enzyme can perform in addition to its native activity (Copley, 2003, 2017), and can be of two kinds: substrate promiscuity (Copley, 2020) and catalytic promiscuity (O'Brien and Herschlag, 1999). Many oxidoreductases are known to exhibit promiscuous activities of both kinds (Biegasiewicz et al., 2018; Sellés-Vidal et al., 2018); for example, the alcohol dehydrogenase of *Thermus* sp. ATN1 (TADH), which can synthesize both chiral alcohols and carboxylic acids (Höllrigl et al., 2008).

Within this enzymatic class, the highest ratios of paralogous enzymes are mainly found in subclass EC 1.1, and in subclasses EC 1.2, EC 1.3, and EC 1.8, but to a lesser degree (**Figure 5A**). They act upon different functional groups of their substrates; however, one common feature of these subclasses is that they contain many enzymes that utilize NAD(P)H as a cofactor. Altogether, they are the subclasses that contain the highest numbers of enzymes utilizing this cofactor, according to the CoFactor database (Fischer et al., 2010), and most of them adopt the same fold: the Rossmann fold. Phylogenetic analyses have shown that there is a common origin for proteins that share this fold, and it is likely to have been present even before the last universal common ancestor (LUCA) (Laurino et al., 2016), making it one of the most ancient protein folds (Bukhari and Caetano-Anollés, 2013; Edwards et al., 2013). Rossmann-fold proteins are also known to show high levels of evolvability, i.e., the ability to adopt new functions and to accommodate sequence changes along evolutionary time (Tóth-Petróczy and Tawfik, 2014). This capacity, along with their high levels of promiscuity (Sellés-Vidal et al., 2018), may provide an advantage for the organism (Khersonsky and Tawfik, 2010) but could also compromise the native activity of the enzyme, leading to detrimental effects. Thus, gene duplication and further optimization of the secondary function through selection could improve the new activity leading to two paralogous enzymes (Force et al., 1999).

Unique Paralogous-Gene Retention Patterns Within the Isomerases

For the isomerases, we identified two subclasses with a high ratio of paralogs: EC 5.1 and EC 5.4 (**Figure 5B**). The intramolecular transferases' subclass (EC 5.4) is also the one with the highest number of unique entries among all isomerases' subclasses.

Within it, there also exist clusters of enzymes with similar chemistries, as represented by oxidosqualene cyclases, RNA-pseudouridine synthases, and carbon mutases (Martínez Cuesta et al., 2016). Oxidosqualene cyclases comprise the biggest group of isomerases catalyzing the same kind of reaction, but although there is substantial evidence of gene duplication within this group of enzymes (Xue et al., 2012; Dahlin et al., 2016; Busta et al., 2020), the paralogous isomerases that we found are unlikely to belong to it. This is because oxidosqualene cyclases are involved in sterols and triterpenes biosynthesis, a typical eukaryotic pathway. It has been identified in several bacterial groups (Wei et al., 2016), but it is more widely considered to be a trait associated with the transition from prokaryotes to eukaryotes (Chen et al., 2007). Thus, it is more likely that paralogous enzymes belonging to this subclass are associated with different biochemical roles. It is also possible that their paralogs perform functions other than isomerization, considering that isomerases are a unique class in which changes of the primary function along their evolutionary history are widespread (Martínez Cuesta et al., 2014, 2015).

Evidence of the previous point is found within the racemases and epimerases (EC 5.1), which is the other subclass for which we found an overall high number of paralogous sequences (**Figure 5B**). It contains different members belonging to the subfamily of short-chain dehydrogenases/reductases (SDR), which also includes oxidoreductases (EC 1) and lyases (EC 4); all their members act upon nucleoside diphosphate (NDP) sugars (Martínez Cuesta et al., 2014). Furthermore, as it occurs with the oxidoreductases' subclasses with more paralogs, all members of the SDR subfamily share the Rossmann fold (Jörnvall et al., 1995). It thus seems likely that, as it happens with oxidoreductases, the high evolvability of enzymes with this fold (Tóth-Petróczy and Tawfik, 2014) may explain the high number of paralogs. Additional support for this comes from several bacterial strains in which there have been identified different gene-duplication events within the SDR subfamily (Serres et al., 2009).

Paralogous Translocases Reflect Adaptation to Different Environmental Conditions

Overall, translocases make up a unique enzymatic class because all its members come from other enzymatic classes. There are 90 different entries identified in the ExplorEnz database (McDonald et al., 2009) as of November 2020, and it is noteworthy that more than half of these entries (around 50) used to be included in a single hydrolases' sub-subclass: EC 3.6.3, which contains enzymes acting on acid anhydrides to catalyze the transmembrane movement of substances. Most of these enzymes are ABC transporters, which constitute one of the most ancient protein superfamilies, are represented throughout both prokaryotes and eukaryotes (Saurin et al., 1999), and most likely were present in the Last Common Ancestor (Davidson et al., 2008). Within the ABC superfamily, there have been many duplication events (Saier and Paulsen, 1999; Higgins, 2001), which may be one of the reasons why we observe a high ratio of paralogous translocases (0.62), which indeed is the highest of all classes (**Figure 3**).

Throughout all prokaryotic diversity, ABC transporters are equally essential and classified into two main groups: uptake and efflux systems. The former plays a very important role in the nutrition of organisms because they allow direct acquisition of nutrients (Ren and Paulsen, 2005; Nicolás et al., 2007). On the other hand, efflux ABC transporters are involved in the exporting of molecules that are toxic to the organism (Nicolás et al., 2007; El-Awady et al., 2017). In the present study, we found that free-living organisms possess the highest ratio of paralogous translocases (0.67), followed by pathogens (0.62), extremophiles (0.59), and finally, intracellular organisms (0.38) (**Figure 4**). The only case in which we didn't find significant differences was between pathogens and extremophiles. For both lifestyles, ABC transporters play a crucial role, though due to different reasons. Extremophiles usually live in environments where nutrients are scarce, so having a high ratio of paralogous transporters must be a good strategy for the uptake of both organic molecules and ions (Albers et al., 2001). Pathogens, rely on different kinds of transporters (including the ABC-type) to ensure the uptake of nutrients necessary for pathogenesis (Tanaka et al., 2018), and in some cases, different types of ABC transporters are active at different stages of it (Murphy et al., 2016). Again, for this group of organisms, having many paralogous translocases seems to be an adaptation for the kind of environment in which they live.

However, for intracellular organisms, we also expected a high ratio of paralogs for this class of enzymes, given the fact that they depend mainly on the uptake of nutrients from the host. Although it is the highest ratio compared to the other enzyme classes within the group, this is not the case compared to the ratios found in other lifestyles. One reason that may account for this could have to do with the kind of intracellular organisms that we considered. When comparing different groups of these organisms, Ren and Paulsen (2005) found that those associated with plants and soil environments have many more transporters than other intracellular organisms. However, in our present study, only four plant symbionts were considered, which could explain why we found a relatively low ratio of paralogous transporters compared to the other lifestyles. Nonetheless, such a ratio is still significantly higher than that of the other categories (**Figure 4**), which indirectly shows the importance of this class of enzymes for the intracellular lifestyle (Rodriguez and Smith, 2006).

In terms of subclasses, we found the highest ratio of paralogous translocases within subclass EC 7.1 (**Figure 5C**), which contains enzymes that catalyze the movement of protons across membranes. Of these, only a few contain the ATP-binding domain, so it seems unlikely that most of the paralogs found within this subclass belong to the ABC transporters. Nonetheless, many of these paralogous proteins could be involved in ATP biosynthesis. One remarkable example is the ATP synthase (EC 7.1.2.2), which is widely distributed across prokaryotes. It has been postulated that a series of several gene duplication events may have occurred earlier in the evolution of this family (Cross and Taiz, 1990), and in fact, more than one copy of ATP synthase has been found in different prokaryotic organisms (Klenk et al., 1997; Ruppert et al., 2001). Thus, many of these copies could have retained their original function, which may be related to

an additional dosage requirement and would provide a benefit in terms of gene expression, given the importance of this enzyme. That this is a common trend across many distinct prokaryotic groups could be interpreted as a means of adaptation to different environments (Cross and Müller, 2004).

Phylogenetic Proximity and Lifestyle Are Reflected in the Content of Paralogous Enzymes

Despite performing a PCA with the mean values for each phylum instead of considering each organism separately, we found different clusters of phylogenetically and lifestyle-related phyla. This was very interesting, given the high heterogeneity that exists within many different phyla. The most significant cluster comprises phyla associated with extreme environments and includes five bacterial and one archaeal phylum (**Figure 6**). Among these, we found two of the bacterial phyla known to have diverged earlier in bacterial evolution: Aquificae and Thermotogae. The other ones are considered lately diverging groups. This clustering suggests that there might be some genomic and biochemical constraints for organisms that inhabit hyperthermophilic environments. This notion of common features concerning lifestyles is also shown in a smaller cluster, comprising Tenericutes, Chlamydia, and Elusimicrobia phyla. All of them include many obligate intracellular organisms, which are known to have reduced genomes and incomplete metabolic pathways, as mentioned above. Although it is not known if intracellular organisms of different phyla share losses of the same (or very functionally similar) enzymes, most of them usually retain proteins involved in the uptake and internalization of organic nutrients (Saier and Paulsen, 1999) and some inorganic ions (Wandersman and Delepelaire, 2004).

Besides the clustering of phyla that share a similar lifestyle, we also found two cases of phylogenetically close phyla that cluster together. The first one comprises members of the so-called TACK group, which includes different phyla belonging to the Archaea domain. The Crenarchaeota, Thaumarchaeota, Korarchaeota, and Bathyarchaeota belong to this archaeal group (Guy and Ettema, 2011), though it includes additional phyla for which there are no fully sequenced genomes. Although belonging to the same phylogenetic group, each of these four phyla lives in different environmental conditions (Spang et al., 2017). We also found that almost all proteobacterial phyla cluster near each other in the PCA plot (**Figure 6**; numbers 1–3 and 5–7). As it is shown, this cluster also seems to include non-proteobacterial phyla, which we think might be due to the great physiological diversity found within the Proteobacteria as a single group (Woese, 1987), as well as the sharing of environmental conditions with other phyla like Actinobacteria and Verrucomicrobia, particularly regarding soil bacteria (Janssen et al., 2002). The only proteobacteria phylum that is far from this cluster is the Epsilonproteobacteria (**Figure 6**; no. 4). Recently, it has been proposed that this phylum might not be related to the other proteobacteria but constitutes an independent, monophyletic group (Waite et al., 2017). This might be reflected in genomic and biochemical traits, as our analysis suggests.

CONCLUSION

In this study, we analyzed the ratio of paralogous enzymes according to the EC classification system established by the IUBMB almost 60 years ago, and that had remained without major changes until the second half of the year 2018. Around this time, a new enzymatic class was added, the translocases, consisting of enzymes previously assigned to other classes. Taking this as a starting point, we found that the number of paralogs within each enzymatic class does not always depend on the number of enzymes. Oxidoreductases are the second class with the most entries and contain many paralogous enzymes, most of which are likely to be NAD(P)H dehydrogenases that adopt the Rossmann fold. On the other hand, isomerases and translocases have, on average, the lowest number of entries but show a high ratio of paralogous enzymes. For translocases, we identified that many paralogous enzymes could be involved in ATP biosynthesis or belong to the ABC transporter superfamily. These influx/efflux systems are critical in several environmental conditions, and their diversification could be a way of adapting to new environments.

Isomerases represent a unique case for which it has been quite difficult to explain their high paralogs' ratio. One possibility is that several paralogous sequences are not even isomerases at all but belong to other enzymatic classes (such as chemically different enzymes that are part of the SDR subfamily), as has been identified elsewhere (Martínez Cuesta et al., 2014, 2015). Additional analyses beyond the subclass level could shed more light on why isomerases have a high ratio of paralogs.

The lifestyle of organisms also seems to be related to the content of paralogous enzymes. Free-living organisms have the highest ratio of paralogs for all enzymatic classes, whereas extremophiles and pathogens have similar ratios, and for certain classes, they do not differ significantly. On the other hand, intracellular organisms show the lowest ratios. However, this trend could be due to other variables like genome size or the number of proteins. Further statistical analysis could help to identify the most important factors determining the prevalence of a high ratio of paralogous enzymes in different organisms.

By considering the ratios of paralogous enzymes and other aspects of the genome, we found a clustering of several phyla not only in a phylogenetic but also in a similar-lifestyle context. The most striking example was a group of different phyla whose members share a hyperthermophilic lifestyle. Thus, it seems that a high ratio of certain paralogous enzymes could be useful to cope with this extreme environment. Whether it is due to the same enzymes, or different enzymes belonging to the same class, it is something that our current analysis did not reveal. However, evidence suggests that parts of the biochemical repertoire, like several amino acid biosynthetic pathways, could have evolved independently in different lineages (Hernández-Montes et al., 2008).

To our concern, this study is the first to analyze the content and ratio of paralogous enzymes both in terms of the EC number (considering its recent major update) and taking into account the lifestyle of organisms. Our results support the idea that gene duplication in prokaryotes is a fundamental process to cope with

new environmental conditions (Gevers et al., 2004; Bratlie et al., 2010; Copley, 2020), regardless of organisms' lifestyles.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/ **Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

AÁ-L thanks the Posgrado en Ciencias Biológicas at the Universidad Nacional Autónoma de México, as well as Consejo Nacional de Ciencia y Tecnología (CONACYT) for their support with fellowship No. 747513. Financial support by PAPIIT-UNAM (BV100218) is gratefully acknowledged. Thanks are given to José Alberto Campillo-Balderas, and Ricardo Hernández-Morales for helpful comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.641817/full#supplementary-material

Supplementary Figure 1 | Relation between the enzyme and protein content, and the genome size in free-living organisms. For each pair of variables, a power-law equation is the one that best explains the distribution of the data. The equations and *R*-squared values are as follows: **(A)** $y = 3.02x^{0.7}$; $R^2 = 0.7$; **(B)** $y = 0.07x^{0.63}$; $R^2 = 0.65$; **(C)** $y = 0.005x^{0.89}$; $R^2 = 0.94$.

Supplementary Figure 2 | Relation between the enzyme and protein content, and the genome size in extremophile organisms. For each pair of variables, a linear regression equation is the one that best explains the distribution of the data. The equations and *R*-squared values are as follows: **(A)** $y = 0.18x + 247$; $R^2 = 0.72$, **(B)** $y = 1.46e^{-04}x + 306$; $R^2 = 0.78$; **(C)** $y = 7.65e^{-04}x + 457$; $R^2 = 0.96$.

Supplementary Figure 3 | Relation between the enzyme and protein content, and the genome size in pathogen organisms. For each pair of variables, a power-law equation is the one that best explains the distribution of the data. The equations and *R*-squared values are as follows: **(A)** $y = 2.39x^{0.73}$; $R^2 = 0.83$; **(B)** $y = 0.02x^{0.71}$; $R^2 = 0.82$; **(C)** $y = 0.002x^{0.95}$; $R^2 = 0.97$.

Supplementary Figure 4 | Relation between the enzyme and protein content, and the genome size in intracellular organisms. For panels **(A,B)** (number of enzymes vs. number of proteins, and number of enzymes vs. genome size), a linear equation is the one that best explains the distribution of the data. This is not the case for panel **(C)**, in which the data fits best to a power-law equation. The equations and *R*-squared values are as follows: **(A)** $y = 0.25x + 95$; $R^2 = 0.84$; **(B)** $y = 1.54e^{-04}x + 180$; $R^2 = 0.71$; **(C)** $y = 0.02x^{0.79}$; $R^2 = 0.88$.

Supplementary Figure 5 | Comparison of the ratio of paralogous enzymes across the different lifestyles. **(A)** The ratio for each organism is plotted together with its number of proteins and enzymes. Each of the four colors represents organisms from the same lifestyle. The diameter of each point of the plot is proportional to the ratio of paralogous enzymes, as indicated in the right part of the figure. **(B)** Notched box plots for the average ratio of paralogous enzymes for the organisms grouped by its lifestyle. Graphically, the ratio value differs significantly in all cases because the notches never overlap each other.

Supplementary Figure 6 | Number of paralogous enzymes found within prokaryotic oxidoreductases **(A)**, isomerases **(B)**, and translocases **(C)** subclasses. Each cell of the heatmaps represents the mean value of the phylum for that specific subclass. The values were scaled for each column using the formula $z = (x - u)/s$, where x is the unscaled value, u is the mean of each column, and s is the column's standard deviation.

REFERENCES

Albers, S. V., van de Vossenberg, J. L. C. M., Driessen, A. J. M., and Konings, W. N. (2001). Bioenergetics and solute uptake under extreme conditions. *Extremophiles* 5, 285–294. doi: 10.1007/s007920100214

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3404. doi: 10.1093/nar/25.17.3389

Angert, E. R. (2012). DNA replication and genomic architecture of very large bacteria. *Annu. Rev. Microbiol.* 66, 197–212. doi: 10.1146/annurev-micro-090110-102827

Bartlett, G. J., Borkakoti, N., and Thornton, J. M. (2003). Catalysing new reactions during evolution: economy of residues and mechanism. *J. Mol. Biol.* 331, 829–860. doi: 10.1016/S0022-2836(03)00734-4

Bashton, M., and Chothia, C. (2007). The generation of new protein functions by the combination of domains. *Structure* 15, 85–99. doi: 10.1016/j.str.2006.11.009

Becerra, A. (2021). The semi-enzymatic origin of metabolic pathways: inferring a very early stage of the evolution of life. *J. Mol. Evol.* 89, 183–188. doi: 10.1007/s00239-021-09994-0

Belitsky, B. R. (2004). *Bacillus subtilis* GabR, a protein with DNA-binding and aminotransferase domains, is a PLP-dependent transcriptional regulator. *J. Mol. Biol.* 340, 655–664. doi: 10.1016/j.jmb.2004.05.020

Biegasiewicz, K. F., Cooper, S. J., Emmanuel, M. A., Miller, D. C., and Hyster, T. K. (2018). Catalytic promiscuity enabled by photoredox catalysis in nicotinamide-dependent oxidoreductases. *Nat. Chem.* 10, 770–775. doi: 10.1038/s41557-018-0059-y

Bratlie, M. S., Johansen, J., Sherman, B. T., Huang, D. W., Lempicki, R. A., and Drabløs, F. (2010). Gene duplications in prokaryotes can be associated with environmental adaptation. *BMC Genomics* 11:588. doi: 10.1186/1471-2164-11-588

Breuert, S., Allers, T., Spohn, G., and Soppa, J. (2006). Regulated polyploidy in halophilic archaea. *PLoS One* 1:92. doi: 10.1371/journal.pone.0000092

Bukhari, S. A., and Caetano-Anollés, G. (2013). Origin and evolution of protein fold designs inferred from phylogenomic analysis of CATH domain structures in proteomes. *PLoS Comput. Biol.* 9:e100309. doi: 10.1371/journal.pcbi.1003009

Busta, L., Serra, O., Kim, O. T., Molinas, M., Peré-Fossoul, I., Figueras, M., et al. (2020). Oxidosqualene cyclases involved in the biosynthesis of triterpenoids in *Quercus suber* cork. *Sci. Rep.* 10:8011. doi: 10.1038/s41598-020-64913-5

Caetano-Anollés, G., Yafremava, L. S., Gee, H., Caetano-Anollés, D., Kim, H. S., and Mittenthal, J. E. (2009). The origin and evolution of modern metabolism. *Int. J. Biochem. Cell Biol.* 41, 285–297. doi: 10.1016/j.biocel.2008.08.022

Chen, L. L., Wang, G. Z., and Zhang, H. Y. (2007). Sterol biosynthesis and prokaryotes-to-eukaryotes evolution. *Biochem. Biophys. Res. Commun.* 363, 885–888. doi: 10.1016/j.bbrc.2007.09.093

Conant, G. C., and Wagner, A. (2002). GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res.* 30, 3378–3386. doi: 10.1093/nar/gkf449

Concu, R., and Cordeiro, M. N. D. S. (2019). Alignment-free method to predict enzyme classes and subclasses. *Int. J. Mol. Sci.* 20:5389. doi: 10.3390/ijms20215389

Copley, S. D. (2003). Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol.* 7, 265–272. doi: 10.1016/S1367-5931(03)00032-2

Copley, S. D. (2017). Shining a light on enzyme promiscuity. *Curr. Opin. Struct. Biol.* 47, 167–175. doi: 10.1016/j.sbi.2017.11.001

Copley, S. D. (2020). Evolution of new enzymes by gene duplication and divergence. *FEBS J.* 287, 1262–1283. doi: 10.1111/febs.15299

Cross, R. L., and Müller, V. (2004). The evolution of A-, F-, and V-type ATP synthases and ATPases: reversals in function and changes in the H+/ATP coupling ratio. *FEBS Lett.* 576, 1–4. doi: 10.1016/j.febslet.2004.08.065

Cross, R. L., and Taiz, L. (1990). Gene duplication as a means for altering H+/ATP ratios during the evolution of Fo F1 ATPases and synthases. *FEBS Lett.* 259, 227–229. doi: 10.1016/0014-5793(90)80014-A

Dahlin, P., Srivastava, V., Bulone, V., and McKee, L. S. (2016). The oxidosqualene cyclase from the oomycete *Saprolegnia parasitica* synthesizes lanosterol as a single product. *Front. Microbiol.* 7:1802. doi: 10.3389/fmicb.2016.01802

David, L. A., and Alm, E. J. (2011). Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* 469, 93–96. doi: 10.1038/nature09649

Davidson, A. L., Dassa, E., Orelle, C., and Chen, J. (2008). Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiol. Mol. Biol. Rev.* 72, 317–364. doi: 10.1128/mmbr.00031-07

Díaz-Mejía, J. J., Pérez-Rueda, E., and Segovia, L. (2007). A network perspective on the evolution of metabolism by gene duplication. *Genome Biol.* 8:R26. doi: 10.1186/gb-2007-8-2-r26

Doolittle, W. F. (2020). Evolution: two domains of life or three? *Curr. Biol.* 30, R177–R179. doi: 10.1016/j.cub.2020.01.010

Edwards, H., Abeln, S., and Deane, C. M. (2013). Exploring fold space preferences of new-born and ancient protein superfamilies. *PLoS Comput. Biol.* 9:e1003325. doi: 10.1371/journal.pcbi.1003325

El-Awady, R., Saleh, E., Hashim, A., Soliman, N., Dallah, A., Elrasheed, A., et al. (2017). The role of eukaryotic and prokaryotic ABC transporter family in failure of chemotherapy. *Front. Pharmacol.* 7:535. doi: 10.3389/fphar.2016.00535

Fani, R., and Fondi, M. (2009). Origin and evolution of metabolic pathways. *Phys. Life Rev.* 6, 23–52. doi: 10.1016/j.plrev.2008.12.003

Fischer, J. D., Holliday, G. L., and Thornton, J. M. (2010). The CoFactor database: organic cofactors in enzyme catalysis. *Bioinformatics* 26, 2496–2497. doi: 10.1093/bioinformatics/btq442

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545. doi: 10.1093/genetics/151.4.1531

Furnham, N., Dawson, N. L., Rahman, S. A., Thornton, J. M., and Orengo, C. A. (2016). Large-scale analysis exploring evolution of catalytic machineries and mechanisms in enzyme superfamilies. *J. Mol. Biol.* 428, 253–267. doi: 10.1016/j.jmb.2015.11.01

Furnham, N., Sillitoe, I., Holliday, G. L., Cuff, A. L., Laskowski, R. A., Orengo, C. A., et al. (2012). Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. *PLoS Computat. Biol.* 8:e1002403. doi: 10.1371/journal.pcbi.1002403

Gevers, D., Vandepoele, K., Simillion, C., and van de Peer, Y. (2004). Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol.* 12, 148–154. doi: 10.1016/j.tim.2004.02.007

Goldman, A. D., Baross, J. A., and Samudrala, R. (2012). The enzymatic and metabolic capabilities of early life. *PLoS One* 7:e39912. doi: 10.1371/journal.pone.0039912

Guy, L., and Ettema, T. J. G. (2011). The archaeal "TACK" superphylum and the origin of eukaryotes. *Trends Microbiol.* 19, 580–587. doi: 10.1016/j.tim.2011.09.002

Hernández-Montes, G., Díaz-Mejía, J. J., Pérez-Rueda, E., and Segovia, L. (2008). The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. *Genome Biol.* 9:R95. doi: 10.1186/gb-2008-9-6-r95

Higgins, C. F. (2001). ABC transporters: physiology, structure, and mechanism – an overview. *Res. Microbiol.* 152, 205–210. doi: 10.1016/S0923-2508(01)01193-7

Höllrigl, V., Hollmann, F., Kleeb, A. C., Buehler, K., and Schmid, A. (2008). TADH, the thermostable alcohol dehydrogenase from *Thermus* sp. ATN1: a versatile new biocatalyst for organic synthesis. *Appl. Microbiol. Biotechnol.* 81, 263–273. doi: 10.1007/s00253-008-1606-z

Horowitz, N. H. (1945). On the evolution of biochemical syntheses. *Proc. Natl. Acad. Sci. U.S.A.* 31, 153–157. doi: 10.1073/pnas.31.6.153

Janssen, P. H., Yates, P. S., Grinton, B. E., Taylor, P. M., and Sait, M. (2002). Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions Acidobacteria, Actinobacteria, *Proteobacteria*, and Verrucomicrobia. *Appl. Environ. Microbiol.* 68, 2391–2396. doi: 10.1128/AEM.68.5.2391-2396.2002

Jeffery, C. J. (2020). Enzymes, pseudoenzymes, and moonlighting proteins: diversity of function in protein superfamilies. *FEBS J.* 287, 4141–4149. doi: 10.1111/febs.15446

Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* 30, 409–425. doi: 10.1146/annurev.mi.30.100176.002205

Jörnvall, H., Krook, M., Persson, B., Atrian, S., Gonzàlez-Duarte, R., Jeffery, J., et al. (1995). Short-chain dehydrogenases/reductases (SDR). *Biochemistry* 34, 6003–6013. doi: 10.1021/bi00018a001

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kannan, N., Taylor, S. S., Zhai, Y., Venter, J. C., and Manning, G. (2007). Structural and functional diversity of the microbial kinome. *PLoS Biol.* 5:e17. doi: 10.1371/journal.pbio.0050017

Khersonsky, O., and Tawfik, D. S. (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* 79, 471–505. doi: 10.1146/annurev-biochem-030409-143718

Klenk, H., Clayton, R. A., Tomb, J., Dodson, R. J., Gwinn, M., Hickey, E. K., et al. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390, 364–370. doi: 10.1038/37052

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. R. Soc. B Biol. Sci.* 279, 5048–5057. doi: 10.1098/rspb.2012.1108

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biol.* 3:RESEARCH0008. doi: 10.1186/gb-2002-3-2-research0008

Koonin, E. V., and Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* 36, 6688–6719. doi: 10.1093/nar/gkn668

Kwon, A., Scott, S., Taujale, R., Yeung, W., Kochut, K. J., Eyers, P. A., et al. (2019). Tracing the origin and evolution of pseudokinases across the tree of life. *Sci. Signal.* 12:eaav3810. doi: 10.1126/scisignal.aav3810

Laurino, P., Tóth-Petróczy, Á, Meana-Pañeda, R., Lin, W., Truhlar, D. G., and Tawfik, D. S. (2016). An ancient fingerprint indicates the common ancestry of Rossmann-fold enzymes utilizing different ribose-based cofactors. *PLoS Biol.* 14:e1002396. doi: 10.1371/journal.pbio.1002396

Lazcano, A., and Miller, S. L. (1999). On the origin of metabolic pathways. *J. Mol. Evol.* 49, 424–431. doi: 10.1007/PL00006565

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155. doi: 10.1126/science.290.5494.1151

Maere, S., de Bodt, S., Raes, J., Casneuf, T., van Montagu, M., Kuiper, M., et al. (2005). Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 5454–5459. doi: 10.1073/pnas.0501102102

Manzano-Marín, A., and Latorre, A. (2016). Snapshots of a shrinking partner: genome reduction in *Serratia symbiotica*. *Sci. Rep.* 6:32590. doi: 10.1038/srep32590

Martínez Cuesta, S., Furnham, N., Rahman, S. A., Sillitoe, I., and Thornton, J. M. (2014). The evolution of enzyme function in the isomerases. *Curr. Opin. Struct. Biol.* 26, 121–130. doi: 10.1016/j.sbi.2014.06.002

Martínez Cuesta, S., Rahman, S. A., and Thornton, J. M. (2016). Exploring the chemistry and evolution of the isomerases. *Proc. Natl. Acad. Sci. U.S.A.* 113, 1796–1801. doi: 10.1073/pnas.1509494113

Martínez Cuesta, S., Rahman, S. A., Furnham, N., and Thornton, J. M. (2015). The classification and evolution of enzyme function. *Biophys. J.* 109, 1082–1086. doi: 10.1016/j.bpj.2015.04.020

Martínez-Núñez, M. A., Poot-Hernandez, A. C., Rodríguez-Vázquez, K., and Perez-Rueda, E. (2013). Increments and duplication events of enzymes and transcription factors influence metabolic and regulatory diversity in prokaryotes. *PLoS One* 8:e69707. doi: 10.1371/journal.pone.0069707

Martínez-Núñez, M. A., Rodríguez-Vázquez, K., and Pérez-Rueda, E. (2015). The lifestyle of prokaryotic organisms influences the repertoire of promiscuous enzymes. *Proteins* 83, 1625–1631. doi: 10.1002/prot.24847

McDonald, A. G., and Tipton, K. F. (2014). Fifty-five years of enzyme classification: advances and difficulties. *FEBS J.* 281, 583–592. doi: 10.1111/febs.12530

McDonald, A. G., Boyce, S., and Tipton, K. F. (2009). ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.* 37, D593–D597. doi: 10.1093/nar/gkn582

McDonald, A. G., Boyce, S., and Tipton, K. F. (2015). Enzyme classification and nomenclature. *ELS* x, 1–11. doi: 10.1002/9780470015902.a0000710.pub3

McGrath, C. L., Gout, J. F., Johri, P., Doak, T. G., and Lynch, M. (2014). Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Res.* 24, 1665–1675. doi: 10.1101/gr.173740.114

Moghe, G. D., Leong, B. J., Hurney, S. M., Jones, A. D., and Last, R. L. (2017). Evolutionary routes to biochemical innovation revealed by integrative analysis of a plant-defense related specialized metabolic pathway. *ELife* 6:e28468. doi: 10.7554/eLife.28468

Mudunuri, U., Che, A., Yi, M., and Stephens, R. M. (2009). bioDBnet: the biological database network. *Bioinformatics* 25, 555–556. doi: 10.1093/bioinformatics/btn654

Murphy, T. F., Brauer, A. L., Johnson, A., and Kirkham, C. (2016). ATP-binding cassette (ABC) transporters of the human respiratory tract pathogen, *Moraxella catarrhalis*: role in virulence. *PLoS One* 11:e0158689. doi: 10.1371/journal.pone.0158689

Nicolás, M. F., Barcellos, F. G., Hess, P. N., and Hungria, M. (2007). ABC transporters in *Mycoplasma hyopneumoniae* and *Mycoplasma synoviae*: insights into evolution and pathogenicity. *Genet. Mol. Biol.* 30(Suppl. 1), 202–211. doi: 10.1590/s1415-47572007000200006

O'Brien, P. J., and Herschlag, D. (1999). Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* 6, R91–R105. doi: 10.1016/S1074-5521(99)80033-7

Ohno, S. (1970). *Evolution by Gene Duplication*. New York: Springer.

Panchy, N., Lehti-Shiu, M., and Shiu, S. H. (2016). Evolution of gene duplication in plants. *Plant Physiol.* 171, 2294–2316. doi: 10.1104/pp.16.00523

Pecoraro, V., Zerulla, K., Lange, C., and Soppa, J. (2011). Quantification of ploidy in *proteobacteria* revealed the existence of monoploid, (mero-)oligoploid and polyploid species. *PLoS One* 6:e16392. doi: 10.1371/journal.pone.0016392

Petsko, G. A., Kenyon, G. L., Gerlt, J. A., Ringe, D., and Kozarich, J. W. (1993). On the origin of enzymatic species. *Trends Biochem. Sci.* 18, 372–376. doi: 10.1016/0968-0004(93)90091-Z

Price, D. R. G., and Wilson, A. C. C. (2014). A substrate ambiguous enzyme facilitates genome reduction in an intracellular symbiont. *BMC Biol.* 12:110. doi: 10.1186/s12915-014-0110-4

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reimer, L. C., Vetcininova, A., Carbasse, J. S., Söhngen, C., Gleim, D., Ebeling, C., et al. (2019). BacDive in 2019: bacterial phenotypic data for high-throughput biodiversity analysis. *Nucleic Acids Res.* 47, D631–D636. doi: 10.1093/nar/gky879

Ren, Q., and Paulsen, I. T. (2005). Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. *PLoS Comput. Biol.* 1:e27. doi: 10.1371/journal.pcbi.0010027

Rice, A. M., and McLysaght, A. (2017). Dosage-sensitive genes in evolution and disease. *BMC Biol.* 15:78. doi: 10.1186/s12915-017-0418-y

Rodriguez, G. M., and Smith, I. (2006). Identification of an ABC transporter required for iron acquisition and virulence in *Mycobacterium tuberculosis*. *J. Bacteriol.* 188, 424–430. doi: 10.1128/JB.188.2.424-430.2006

RStudio Team (2020). *RStudio: Integrated Development for R*. Boston, MA: RStudio.

Ruppert, C., Schmid, R., Hedderich, R., and Müller, V. (2001). Selective extraction of subunit D of the Na+-translocating methyltransferase and subunit *c* of the A1A0 ATPase from the cytoplasmic membrane of methanogenic archaea by chloroform/methanol and characterization of subunit *c* of *Methanothermobacter thermoautotrophicus* as a 16-kDa proteolipid. *FEMS Microbiol. Lett.* 195, 47–51. doi: 10.1111/j.1574-6968.2001.tb10496.x

Saier, M. H., and Paulsen, I. T. (1999). Paralogous genes encoding transport proteins in microbial genomes. *Res. Microbiol.* 150, 689–699. doi: 10.1016/S0923-2508(99)00123-0

Sakharkar, K. R., Kumar Dhar, P., and Chow, V. V. T. K. (2004). Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. *Int. J. Syst. Evol. Microbiol.* 54, 1937–1941. doi: 10.1099/ijs.0.63090-0

Saurin, W., Hofnung, M., and Dassa, E. (1999). Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. *J. Mol. Evol.* 48, 22–41. doi: 10.1007/PL00006442

Schmidt, S., Sunyaev, S., Bork, P., and Dandekar, T. (2003). Metabolites: a helping hand for pathway evolution? *Trends Biochem. Sci.* 28, 336–341. doi: 10.1016/S0968-0004(03)00114-2

Sellés-Vidal, L., Kelly, C. L., Mordaka, P. M., and Heap, J. T. (2018). Review of NAD(P)H-dependent oxidoreductases: properties, engineering, and application. *Biochim. Biophys. Acta* 1866, 327–347. doi: 10.1016/j.bbapap.2017.11.005

Serres, M. H., Kerr, A. R. W., McCormack, T. J., and Riley, M. (2009). Evolution by leaps: gene duplication in bacteria. *Biol. Direct* 4, 1–17. doi: 10.1186/1745-6150-4-46

Soppa, J. (2011). Ploidy and gene conversion in archaea. *Biochem. Soc. Trans.* 39, 150–154. doi: 10.1042/BST0390150

Soppa, J. (2017). Polyploidy and community structure. *Nat. Microbiol.* 2:16261. doi: 10.1038/nmicrobiol.2016.261

Spang, A., Caceres, E. F., and Ettema, T. J. G. (2017). Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science* 357:eaaf3883. doi: 10.1126/science.aaf3883

Tanaka, K. J., Song, S., Mason, K., and Pinkett, H. W. (2018). Selective substrate uptake: the role of ATP-binding cassette (ABC) importers in pathogenesis. *Biochim. Biophys. Acta* 1860, 868–877. doi: 10.1016/j.bbamem.2017.08.011

Tian, W., and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* 333, 863–882. doi: 10.1016/j.jmb.2003.08.057

Tipton, K., and Boyce, S. (2000). History of the enzyme nomenclature system. *Bioinformatics* 16, 34–40. doi: 10.1093/bioinformatics/16.1.34

Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* 307, 1113–1143. doi: 10.1006/jmbi.2001.4513

Tóth-Petróczy, Á, and Tawfik, D. S. (2014). The robustness and innovability of protein folds. *Curr. Opin. Struct. Biol.* 26, 131–138. doi: 10.1016/j.sbi.2014.06.007

Tyzack, J. D., Furnham, N., Sillitoe, I., Orengo, C. M., and Thornton, J. M. (2017). Understanding enzyme function evolution from a computational perspective. *Curr. Opin. Struct. Biol.* 47, 131–139. doi: 10.1016/j.sbi.2017.08.003

van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 10, 725–732. doi: 10.1038/nrg2600

van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18, 411–424. doi: 10.1038/nrg.2017.26

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends Genet.* 19, 479–484. doi: 10.1016/S0168-9525(03)00203-8

Waite, D. W., Vanwonterghem, I., Rinke, C., Parks, D. H., Zhang, Y., Takai, K., et al. (2017). Comparative genomic analysis of the class *Epsilonproteobacteria* and proposed reclassification to Epsilonbacteraeota (phyl. nov.). *Front. Microbiol.* 8:682. doi: 10.3389/fmicb.2017.00682

Walsh, B. (2003). Population-genetic models of the fates of duplicate genes. *Genetica* 118, 279–294. doi: 10.1023/A:1024194802441

Wandersman, C., and Delepelaire, P. (2004). Bacterial iron sources: from siderophores to hemophores. *Annu. Rev. Microbiol.* 58, 611–647. doi: 10.1146/annurev.micro.58.030603.123811

Wang, S., and Chen, Y. (2018). Phylogenomic analysis demonstrates a pattern of rare and long-lasting concerted evolution in prokaryotes. *Commun. Biol.* 1:12. doi: 10.1038/s42003-018-0014-x

Wei, J. H., Yin, X., and Welander, P. V. (2016). Sterol synthesis in diverse bacteria. *Front. Microbiol.* 7:990. doi: 10.3389/fmicb.2016.00990

Weng, J. K., Philippe, R. N., and Noel, J. P. (2012). The rise of chemodiversity in plants. *Science* 336, 1667–1670. doi: 10.1126/science.1217411

Wernegreen, J. J. (2015). Endosymbiont evolution: predictions from theory and surprises from genomes. *Ann. N. Y. Acad. Sci.* 1360, 16–35. doi: 10.1111/nyas.12740

Williams, T. A., Foster, P. G., Cox, C. J., and Embley, T. M. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504, 231–236. doi: 10.1038/nature12779

Woese, C. R. (1987). Bacterial evolution. *Microbiol. Rev.* 51, 221–271. doi: 10.1128/mmbr.51.2.221-271.1987

Woods, S., Coghlan, A., Rivers, D., Warnecke, T., Jeffries, S. J., Kwon, T., et al. (2013). Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses. *PLoS Genet.* 9:e1003330. doi: 10.1371/journal.pgen.1003330

Xue, Z., Duan, L., Liu, D., Guo, J., Ge, S., Dicks, J., et al. (2012). Divergent evolution of oxidosqualene cyclases in plants. *New Phytol.* 193, 1022–1038. doi: 10.1111/j.1469-8137.2011.03997.x

Yčas, M. (1974). On earlier states of the biochemical system. *J. Theor. Biol.* 44, 145–160. doi: 10.1016/S0022-5193(74)80035-4

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18, 292–298. doi: 10.1016/S0169-5347(03)00033-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ARTÍCULOS CIENTÍFICOS PUBLICADOS

Journal of Molecular Evolution
https://doi.org/10.1007/s00239-022-10085-x

ORIGINAL ARTICLE

The Fate of Duplicated Enzymes in Prokaryotes: The Case of Isomerases

Alejandro Álvarez-Lugo[1,2] · Arturo Becerra[2]

Received: 11 April 2022 / Accepted: 16 December 2022

Abstract
The isomerases are a unique enzymatic class of enzymes that carry out a great diversity of chemical reactions at the intramolecular level. This class comprises about 300 members, most of which are involved in carbohydrate and terpenoid/polyketide metabolism. Along with oxidoreductases and translocases, isomerases are one of the classes with the highest ratio of paralogous enzymes. Due to its relatively small number of members, it is plausible to explore it in greater detail to identify specific cases of gene duplication. Here, we present an analysis at the level of individual isomerases and identify different members that seem to be involved in duplication events in prokaryotes. As was suggested in a previous study, there is no homogeneous distribution of paralogs, but rather they accumulate into a few subcategories, some of which differ between Archaea and Bacteria. As expected, the metabolic processes with more paralogous isomerases have to do with carbohydrate metabolism but also with RNA modification (a particular case involving an rRNA-modifying isomerase is thoroughly discussed and analyzed in detail). Overall, our findings suggest that the most common fate for paralogous enzymes is the retention of the original enzymatic function, either associated with a dosage effect or with differential expression in response to changing environments, followed by subfunctionalization and, to a much lesser degree, neofunctionalization, which is consistent with what has been reported elsewhere.

Keywords Gene duplication · Isomerases · Paralogous enzymes · Archaea · Bacteria

Introduction

Enzymes are the workforce that allows life's sustenance by means of chemical reactions that constitute metabolism. They are classified into a system established in the early 60 s by the International Commission on Enzymes from the International Union of Biochemistry and Molecular Biology (Tipton & Boyce 2000). Commonly referred to as the Enzyme Commission (EC) System, it groups enzymes in terms of reaction similarity (McDonald et al. 2015) and not by evolutionary-related members. It includes seven classes of enzymes which are further divided into subclasses and sub-subclasses.

Within the EC System it is not uncommon to find evolutionary-related enzymes in different sub-subclasses, subclasses, and classes (which reflect divergent evolution) (Furnham et al. 2012; Martínez-Cuesta et al. 2015). Similarly, enzymes performing remarkably similar biochemical activities without an evolutionary relationship can be found within the same sub-subclass (suggesting evolutionary convergence). This illustrates the complexity within the EC System and highlights the need to address the problem from an evolutionary perspective, which could help to understand certain aspects of enzyme function classification.

De novo enzyme-coding genes can be originated through different processes (Neme & Tautz 2014), but perhaps the most important is gene duplication, i.e., the process through which a preexisting gene can give rise to a new one, leading to two copies known as paralogous genes (Ohno 1970).

It is now widely acknowledged that gene duplication has played a significant role in the evolution of new enzymatic functions (Copley 2020), which has led to the expansion

Handling editor: **David Liberles**.

✉ Arturo Becerra
abb@ciencias.unam.mx

[1] Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México, Mexico City, México

[2] Facultad de Ciencias, Universidad Nacional Autónoma de México, Mexico City, México

Published online: 29 December 2022

of metabolism (Jensen 1976; Díaz-Mejía et al. 2007). In a previous study (Álvarez-Lugo & Becerra 2021), we found that the oxidoreductases, isomerases, and the most recently recognized enzyme class, the translocases, have the highest ratios of duplicated enzymes, and these are concentrated into a few subclasses within each class. This does not exclusively depend on the total number of enzymes identified for each of them: whereas oxidoreductases have many entries, isomerases and translocases are among the enzymatic classes with the fewest number of described enzymes (McDonald et al. 2009).

Together with ligases and translocases, isomerases are one of the enzymatic classes with the lowest number of members. Besides, except for translocases, they are the less abundant type of enzymes in prokaryotes, no matter their lifestyle (Álvarez-Lugo & Becerra 2021), and their functional diversity is relatively low. Three sub-subclasses stand out because they perform reactions that are chemically identical to those of the oxidoreductases (EC 5.3), transferases (EC 5.4), and lyases (EC 5.5). The only difference is that these reactions are performed intramolecularly. One example is the enzyme sterol Δ^{24}-isomerase (24ISO), involved in the biosynthesis of a kind of plant steroid called withanolides. This enzyme catalyzes a reaction similar to that of sterol side chain reductases (SSR) 1 and 2, also involved in the same metabolic pathway. By phylogenetic analysis, it has been shown that the former enzyme arose from a duplication event of SSR1. Although 24ISO still requires the cofactor NADPH, there is no net consumption of it (Knoch et al. 2018).

In the current study we focus on the isomerases, analyze them at the third level of the EC number (the sub-subclass), and further break them down to the individual enzyme level. They comprise about 300 enzymes grouped into seven subclasses and 19 sub-subclasses, catalyze around 4% of the biochemical reactions of central metabolism (Martínez-Cuesta et al. 2014; 2016), and are mainly involved in carbohydrate and terpenoids/polyketides metabolism, but also in lipid and amino acid metabolism to a lesser degree (Fig. 1). Therefore, by analyzing them at the third level of the enzymatic code (which refers to the sub-subclass), we can further identify which individual enzymes have more paralogs and figure out possible scenarios that may have facilitated the fixation of these duplicates.

Materials and Methods

The proteomes of a representative prokaryotic sample comprising 745 organisms (655 bacteria and 90 archaea) were downloaded from the KEGG Database (Kanehisa & Goto 2000). Paralogous sequences were identified with the program BlastP (Altschul et al. 1997) using the same criteria reported elsewhere (Álvarez-Lugo & Becerra 2021). Resulting files were then parsed with ad hoc Perl scripts.

To identify enzymes for both the proteomes and the paralogous-sequences datasets we employed the online tool *db2db*, which is part of the bioDBnet resource (Mudunuri et al. 2009), and the FTP files provided with the proteomes from the KEGG Database (Kanehisa & Goto 2000). We kept only those sequences for which we knew at least the nature

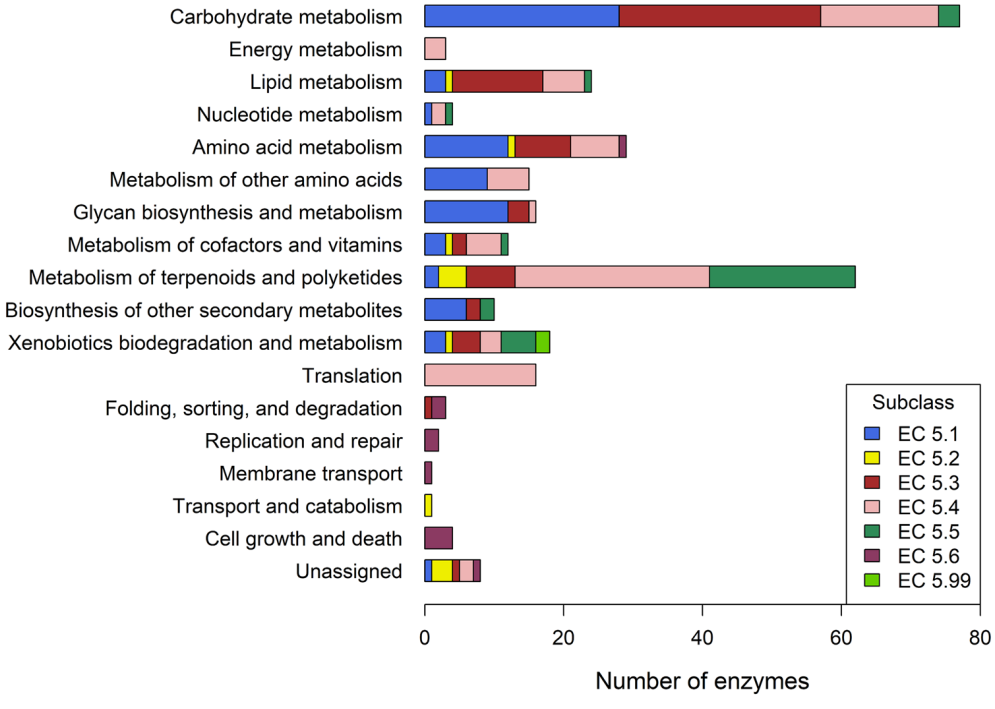


Fig. 1 Distribution of isomerases in metabolism. Here, we only report those metabolic processes within which there is at least one isomerase involved, according to the KEGG database. The "unassigned" category includes isomerases whose function has been described, but that currently are not placed within a specific metabolic pathway

of the substrate involved in the enzymatic reaction, defined by the first three digits of the Enzyme Commission number (EC number). All sequences corresponding to enzymes were then grouped according to their respective subclass and sub-subclass, using the latest version of the ExplorEnz database (McDonald et al. 2009).

KEGG IDs from all bacterial rRNA pseudouridine synthases that had at least one paralog, both those that modify 23S and 16S rRNA (RluE, RluF, RluB, RluD, RluC, RluA, and RsuA, respectively), were searched against the UniProtKB Database to obtain the corresponding sequences. We then grouped the sequences into seven different groups, which correspond to each of the enzymes that we considered. Multiple sequence alignments (MSA) for each group of sequences were performed with the MAFFT software (Katoh & Standley 2013). Output files were given to the trimAl program (Capella-Gutiérrez et al. 2009) to remove gaps and other uninformative sites (gap threshold = 0.9, and conserved at least 60% of positions in the original alignment). We then calculated the best evolutionary model for each pruned alignment with the program IQ-TREE (Nguyen et al. 2015). For RsuA (EC 5.4.99.19) and RluE (EC 5.4.99.20), the best-fitted model was LG+G4. RluB (EC 5.4.99.22), RluD (EC 5.4.99.23), RluC (EC 5.4.99.24), and RluA (EC 5.4.9.29) fit better with the LG+I+G4 model, whereas model JTTDCMut+G4 was the best for RluF (EC 5.4.99.21). IQ-TREE was also used to construct the corresponding maximum likelihood (ML) phylogenetic trees by tree reconstruction+ultrafast bootstrap with 1000 replicates (Hoang et al. 2018) for each set of enzymes. All phylogenetic trees were visualized and annotated with iTOL (Letunic & Bork 2021).

To display the relationships among protein sequences, a sequence similarity network (SSN) and neighborhood connectivity (NC) were then calculated with 1944 isomerases sequences, using the Enzyme Similarity Tool (EFI-EST, Zallot et al. 2019). Transitivity clustering was performed and visualized using Cytoscape (Shannon et al. 2003).

Results

To find out why there is a high ratio of paralogs within the isomerases, we identified the number of paralogs for each subclass and sub-subclass across the whole sample (Fig. 2). However, for a more general picture, we only considered the average values for each phylum. At the subclass level, the category with the highest number of paralogs is "intramolecular transferases" (EC 5.4), followed by "racemases and epimerases" (EC 5.1) (Fig. 2A). Then, at the sub-subclass

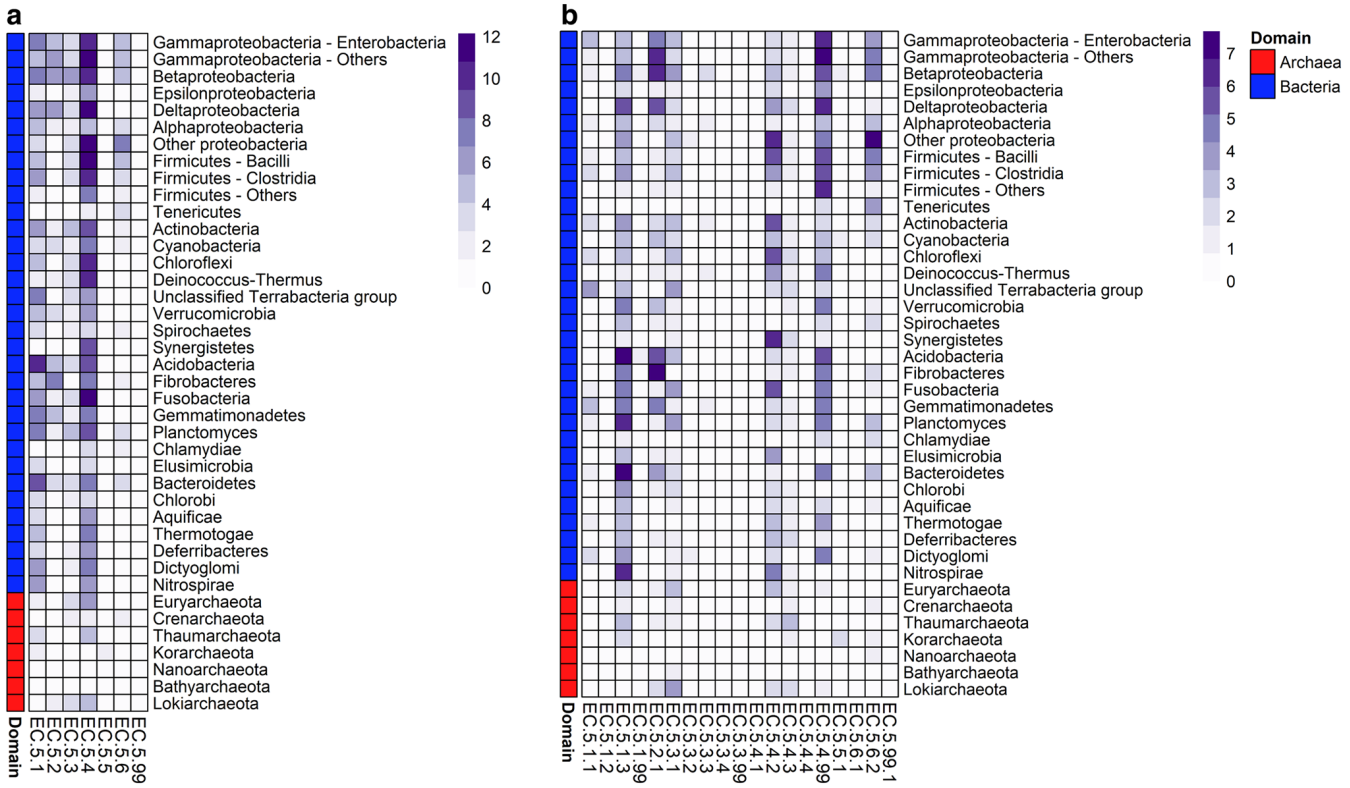


Fig. 2 Ratio of paralogous enzymes within the Isomerases. (**A**) Average ratio of the number of paralogous isomerases, grouped by subclasses. (**B**) Average ratio of the number of paralogous isomerases, grouped by sub-subclasses. The color scale indicates the absolute number of paralogs, which is the average for each phylum (Color figure online)

level, we distinguished six categories with a high ratio of paralogs, which belong to five different subclasses (Fig. 2B and Table 1). From these, "intramolecular transferases transferring other groups" (EC 5.4.99) stands out.

Table 1 shows that sub-subclasses EC 5.4.99 and EC 5.1.3 contain the highest number of both paralogs and entries, according to the ExplorEnz database (McDonald et al. 2009). The former is an interesting case because it comprises a wide diversity of enzymes acting upon different substrates (unlike the other sub-subclasses). However, most paralogs within this category are RNA-modifying enzymes known as pseudouridine synthases, which convert uridine into pseudouridine in different tRNA and rRNA positions.

One could argue that the reason why subclasses EC 5.4.99 and EC 5.1.3 have a greater average number of duplicates is because they both contain many entries. To further investigate this, we performed a correlational analysis between the number of entries identified for each sub-subclass and the number of paralogs that belong to each of them (Supplementary Table 1). Regarding the average number of paralogs of the whole sample, we find a positive but not strong correlation between this variable and the number of described enzymes in each sub-subclass ($r=0.67$). This is also the case when we only consider bacterial paralogs ($r=0.63$). On the other hand, we found a weak correlation when considering archaeal paralogs (0.44).

For bacteria, EC 5.4.99 and EC 5.1.3 are the classes with the highest average number of paralogs per organism (approximately four) (Fig. 3). On the other hand, we found that three sub-subclasses with very similar values within the archaea are the ones with the highest number of paralogs (around two), so we considered all three (EC 5.1.3, EC 5.3.1, and EC 5.4.2) (Fig. 3).

We then identified which enzymes within each of these sub-subclasses were the ones associated with the highest number of paralogs. For this, we also splitted our sample into Bacterial (655; n_B) and archaeal (90; n_B) proteomes and then counted how many times each enzymatic number appeared within each sub-subclass (x).

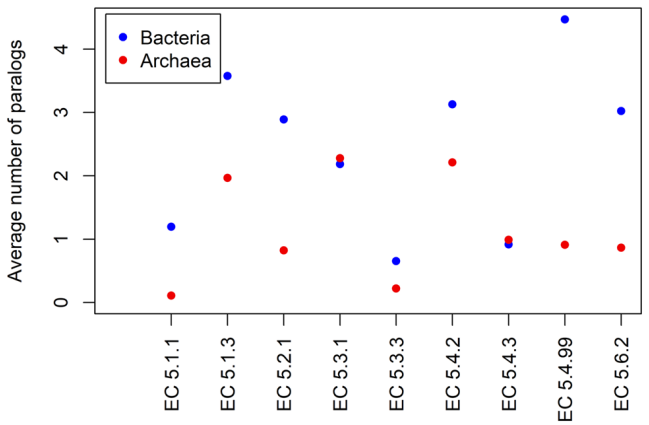


Fig. 3 Average number of paralogous isomerases in Archaea and Bacteria. Each sub-subclass is indicated on the x axis. The dots represent the average number of paralogs found for each sub-subclass in Bacteria (blue dots) and Archaea (red dots). Sub-subclasses in which the average number of paralogs is less than 0.5 for both domains are excluded (Color figure online)

We found several organisms without a single paralog belonging to the specific sub-subclass for each one. This number differs between each of them and between Bacteria and Archaea. To make an appropriate comparison between all sub-subclasses, we made a slight correction for this issue and excluded all those organisms in which there was not a single paralog belonging to each sub-subclass (a). Thus, we calculated the ratio of occurrence (O) with the following formula:

$$O = x/n - a$$

Such a ratio was obtained for every enzymatic number within each sub-subclass.

For simplicity, Fig. 4 depicts the presence of each enzyme with an O value of 0.3 or higher (for a complete depiction of all enzymes for which we identified at least one paralog, please check Supplementary Figure S1). When the corresponding enzyme was present in every organism of a particular phylum, we assigned a ratio of one (indicated by

Table 1 Isomerases' sub-subclasses that contain the highest ratio of paralogs, without distinguishing between Archaea and Bacteria

Subclass	Sub-subclass	EC code	No. paralogs (average)	No. entries
Intramolecular transferases	Transferring other groups	EC 5.4.99	4.04	67
Racemases and epimerases	Acting on carbohydrates and derivatives	EC 5.1.3	3.38	43
Intramolecular transferases	Phosphomutases	EC 5.4.2	3.02	13
Altering macromolecular conformation	Altering nucleic acid conformation	EC 5.6.2	2.76	2
cis–trans isomerases	*cis–trans* isomerases	EC 5.2.1	2.64	14
Intramolecular oxidoreductases	Interconverting aldoses and ketoses, and related compounds	EC 5.3.1	2.2	31

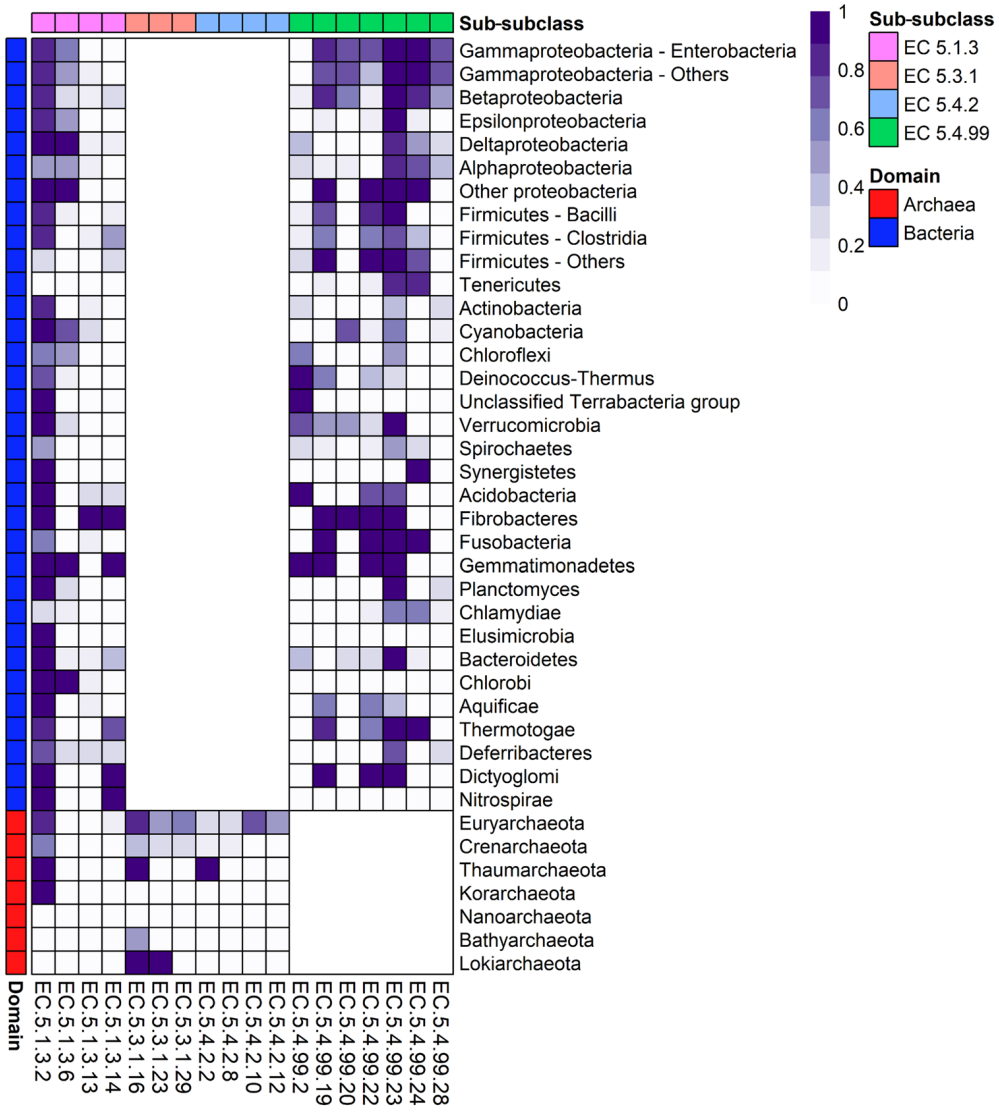


Fig. 4 Ratio of paralogous isomerases with an *O* value of 0.3 or higher in Archaea and Bacteria. The listed EC numbers belong to sub-subclasses with more paralogs, but these differ between Archaea and Bacteria. The greater the intensity of each cell, the higher the number of identified paralogs within each phylum. Subclass EC 5.4.99 is not considered for Archaea, whereas sub-subclasses EC 5.3.1 and EC 5.4.2 are not considered for Bacteria (blank areas at the lower right corner and in the middle of the figure, respectively)

the color scale at the right of the heatmap). On the other hand, its ratio was zero if it was not present in any organism. Large blank areas correspond to the sub-subclasses that were not considered for Bacteria (EC 5.3.1 and EC 5.4.2) and Archaea (EC 5.4.99).

Enzymes with an *O* value of 0.3 or higher (as in Fig. 4) are listed in Table 2. We also indicate the pathway in which each enzyme takes part. Finally, in the last column of the table we suggest the scenario that we consider most likely to explain the retention of each set of duplicates, according to what is reported about such enzymes in the literature (see Discussion section).

One limitation of the above results is that they may not be directly comparable because the sub-subclasses with more paralogs are not the same for Archaea and Bacteria, except for EC 5.1.3. Therefore, to evaluate the content of duplicates within different metabolic processes so that they are comparable to each other, we considered the same subclasses for both cellular domains (EC 5.1.3, EC 5.3.1, EC 5.4.2, and EC 5.4.99). In addition, we identified each enzyme's biochemical function according to the BRENDA database (Schomburg et al. 2004).

To define how representative each process was, we identified the number of times that each EC number appeared within each sample ($x_i \ldots x_n$). Dividing each value by the sum of them all (X) gives us the representativeness percentage (%R) per function. In cases where two or more enzymes had the same assigned role, we grouped them into a single value, so each function was considered only once.

Figure 5 compares the abundance of duplicates within different metabolic and cellular processes in Archaea and Bacteria. Paralogs are not equally distributed throughout all metabolic processes. Most of them concentrate less than 1% of paralogous isomerases, whereas several others accumulate around 20% of them (for example, "23S rRNA modification" in Bacteria and "Degradation of hexoses" in both Archaea and Bacteria). A striking case is that of "23S rRNA modification". This process is also present in

Table 2 Isomerases with the highest number of paralogs

No. paralogs	EC number	Enzyme name	Pathway	Abundance (%)	Probable scenario(s) for paralogs' retention
Bacteria					
903	EC 5.1.3.2	UDP-glucose 4-epimerase	Galactose metabolism	23.44	Subfunctionalization
783	EC 5.4.99.23	23S rRNA pseudouridine1911/1915/1917 synthase	23S rRNA modification	20.32	Molecular backup* and subfunctionalization
310	EC 5.4.99.28; EC 5.4.99.29	tRNA pseudouridine32 synthase	tRNA modification	8.05	–
308	EC 5.4.99.19	16S rRNA pseudouridine516 synthase	16S rRNA modification	7.99	–
298	EC 5.4.99.24	23S rRNA pseudouridine955/2504/2580 synthase	23S rRNA modification	7.73	Subfunctionalization
292	EC 5.4.99.2	methylmalonyl-CoA mutase	Propanoate metabolism	7.58	Conservation of enzyme function**
210	EC 5.4.99.22	23S rRNA pseudouridine2605 synthase	23S rRNA modification	5.45	
210	EC 5.1.3.6	UDP-glucuronate 4-epimerase	Ascorbate metabolism	5.45	Sub/neofunctionalization
185	EC 5.1.3.14	UDP-N-acetylglucosamine 2-epimerase (non-hydrolyzing)	Polysaccharide biosynthesis	4.8	*Conservationof enzyme function** / Subfunctionalization*
181	EC 5.1.3.13	dTDP-4-dehydrorhamnose 3,5-epimerase	Biosynthesis of cell wall	4.7	Conservation of enzyme function**
173	EC 5.4.99.20	23S rRNA pseudouridine2457 synthase	23S rRNA modification	4.49	–
Archaea					
124	EC 5.1.3.2	UDP-glucose 4-epimerase	Galactose metabolism	27.02	Subfunctionalization/ conservation of enzyme function**
81	EC 5.3.1.16	1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino]imidazole-4-carboxamide isomerase	Histidine metabolism	17.65	Neofunctionalization
63	EC 5.4.2.12	phosphoglycerate mutase (2,3-diphosphoglycerate-independent)	Glycolysis	13.73	–
45	EC 5.4.2.10	phosphoglucosamine mutase	Protein N-glycosylation	9.8	*Conservation of enzyme function**
40	EC 5.3.1.29	ribose 1,5-bisphosphate isomerase	AMP metabolism	8.71	Conservation of enzyme function**
38	EC 5.3.1.23	S-methyl-5-thioribose-1-phosphate isomerase	Cysteine and methionine metabolism	8.28	–
24	EC 5.1.3.14	UDP-N-acetylglucosamine 2-epimerase (non-hydrolyzing)	Polysaccharide biosyntheses	5.23	*Conservation of enzyme function**
22	EC 5.4.2.8	phosphomannomutase	D-mannose degradation	4.79	–
22	EC 5.4.2.2; EC 5.4.2.8	phosphoglucomutase (alpha-D-glucose-1,6-bisphosphate-dependent)	Glycogenolysis	4.79	–

The first column indicates the number of copies with the corresponding EC number that we identified, which is also reflected in the Abundance column, in terms of percentages. In the last column, based on our results and research reported in the literature, we indicate the probable scenarios that could be involved in paralogs' retention

Those in italics indicate that this scenario needs to be further investigated, whereas a dash (—) indicates that the corresponding enzyme was not analyzed. Information from the "Pathway" column was obtained from the KEGG Database (Kanehisa & Goto 2000)

(*)Refers to RluD (EC 5.4.99.23) copies with the same function

(**)In this work, conservation of enzyme function indicates that duplicates perform the same enzymatic reaction although the cellular context may be different

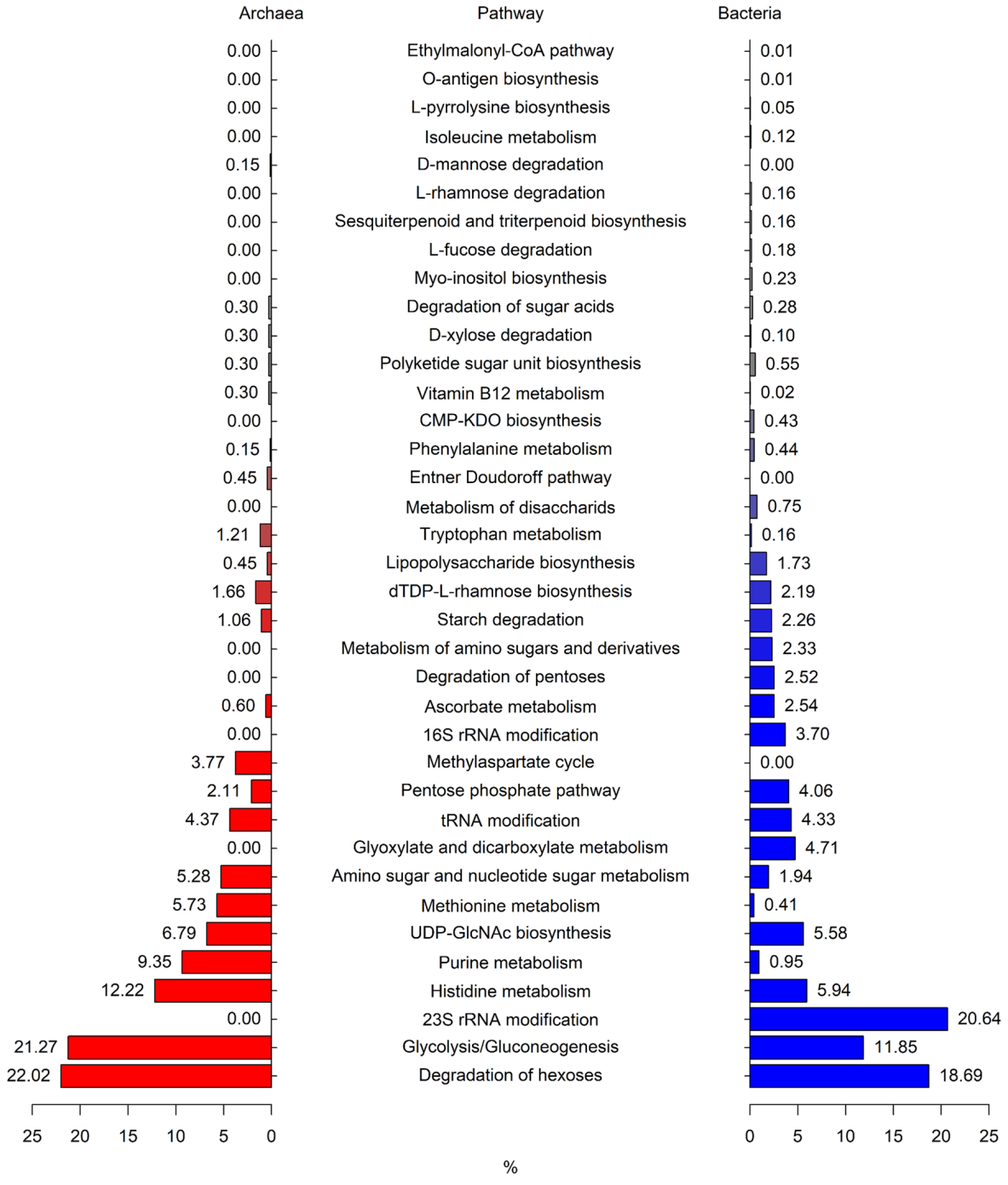


Fig. 5 Abundance of paralogous isomerases within different metabolic and cellular processes. Functions comprising higher numbers of paralogs appear at the bottom of the figure, and their %R is indicated next to each horizontal bar

Archaea, and enzymes associated with it are homologous to bacterial ones. However, according to our analyses, paralogs could not be identified for any archaeal enzymes involved in this post-transcriptional modification process. Similar situations, although to a much lesser degree, occur for the "Methylaspartate cycle" (paralogs are only found within the archaeal genomes) and for "Glyoxylate and dicarboxylate metabolism", "Degradation of pentoses", and "Metabolism of amino sugars and derivatives", processes in which we only find paralogs within Bacteria.

Several processes identified in both cellular domains have a similar parlog's abundance, like "Glycolysis/Gluconeogenesis", "UDP-GlcNAc biosynthesis", and "tRNA modification" whereas others, like "Degradation of hexoses" and "Purine metabolism", show more remarkable

differences between domains (in these cases, a higher ratio of paralogs is found within Archaea) (Fig. 5).

We identified many paralogs related to rRNA modification, all belonging to the superfamily of pseudouridine synthases. Phylogenetic analyses reveal large differences in the number of paralogs for each analyzed enzyme (Fig. 6 and Supplementary Figure S2). For some, most identified paralogs correspond to an enzyme with a different EC number, whereas others possess one or more paralogs that share the same one. The enzyme 23S rRNA pseudouridine(1911/1915/1917) synthase (RluD, EC 5.4.99.23) (Fig. 6A) is the one with the highest number of copies in many organisms. In contrast, its closest paralog, RluC (EC 5.4.99.24), which modifies uridines at positions 955/2504/2580 of the 23S rRNA, is almost always found once, and in no case did we find a single organism

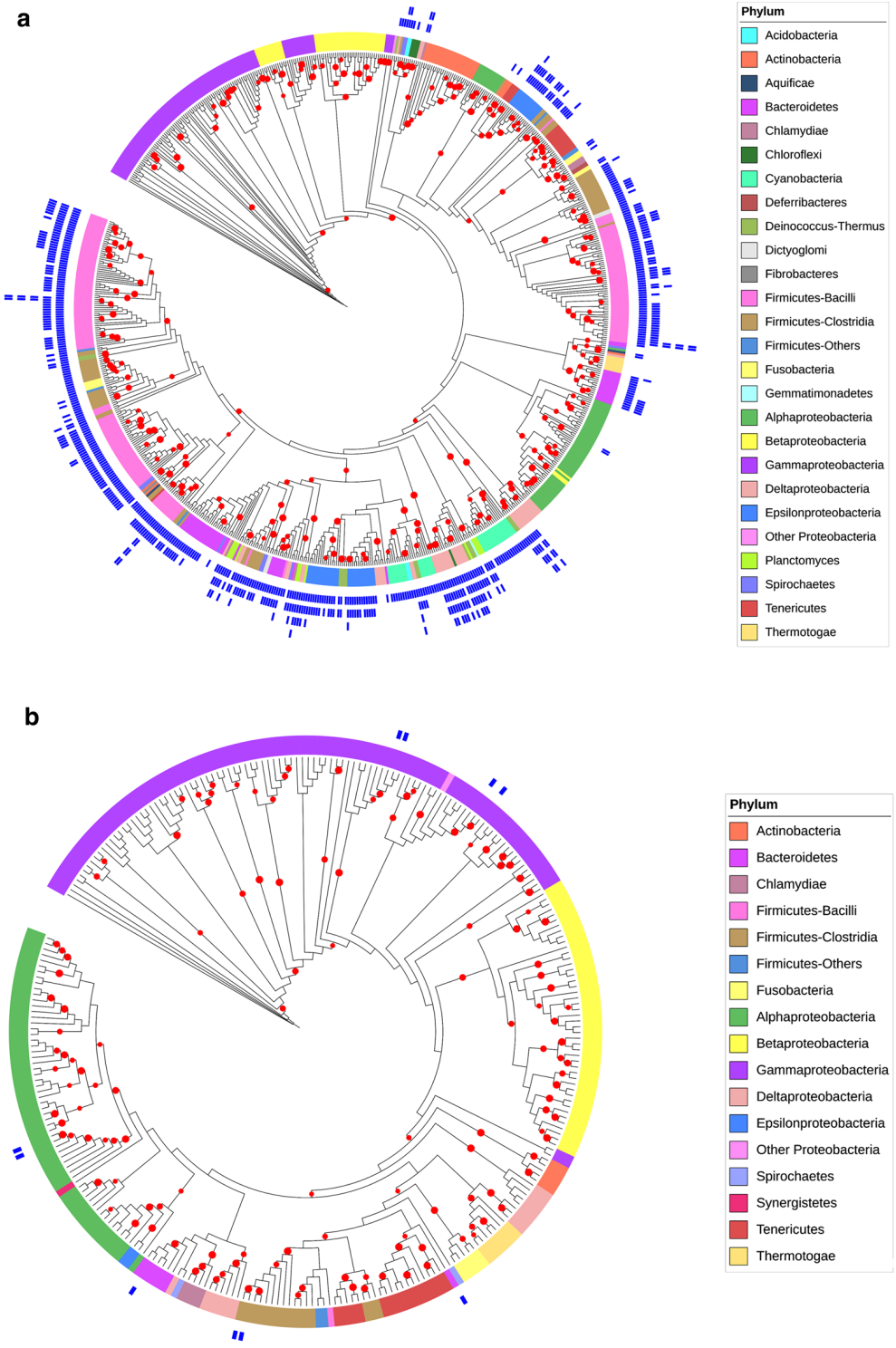


Fig. 6 Phylogenetic distribution of rRNA pseudouridine synthases RluD (**A**) and RluC (**B**) in our bacterial sample. All sequences in our sample for which at least one paralog was detected (regardless of its function) were included in the phylogenetic analysis. Blue bars from the outer circles indicate the number of paralogous enzymes with the same assigned function for each of the sequences that were considered to build the phylogenies. The absence of a blue bar indicates that paralogs have different functions. Red dots indicate bootstrap values of 90 or higher, and their size is proportional to their value (Color figure online)

that had more than two copies of this enzyme (Fig. 6B). The other pseudouridine synthases have a pattern similar to that of RluC, except for RluB (EC 5.4.99.22) and RluA (EC 5.4.99.29) (Supplementary Figure S2).

The SSN was calculated for the 1944 isomerase sequences using the Enzyme Similarity Tool (see Material and Methods section) networks shown in Fig. 7. The network statistics displayed 1944 nodes, and 181,265 edges, with a clustering coefficient of 0.973. The transitivity clustering was performed by Cytoscape and is displayed in Fig. 7A, where all sequences included in each cluster have a high similarity value and the same assigned function (i.e., 5.4.99.23). Furthermore, Fig. 7B shows the neighborhood connectivity distribution, where RluD and RluC have the largest connectivity values. Each of the two major clusters corresponds to two of the four families belonging to the Pseudouridine Synthase Superfamily: RluA and RsuA. The former includes subfamilies RluA, RluC, and RluD (left cluster), whereas RluB, RluE, RluF and RsuA belong to the later (right cluster). This is in agreement with what has been reported elsewhere.

Discussion

A plausible explanation for the high ratio of duplicates within the isomerases may lie in the fate of such enzymes. It has been reported that, throughout its evolutionary history, only around 20% of the changes in activity remain within the isomerization category, whereas the other 80% involves a huge functional change from the isomerases to another enzymatic class (Furnham et al. 2012; Martínez-Cuesta et al. 2014; 2015). One of the reasons behind this could be the catalytic capabilities of certain folds. For example, in the archaea, *Pyrococcus furiosus*, it has been shown that the Triosephosphate isomerase (EC 5.3.1.1), which belongs to the $(\beta/\alpha)_8$-barrel (TIM barrel) fold superfamily, also exhibits an endoglucanase/cellulase activity (EC 3.2.1.4) at different temperature maximums (Sharma & Guptasarma 2017). This is further supported by the tremendous functional diversity found within TIM barrel proteins, among which many isomerases and many other enzymes encompass all general enzymatic functions (Goldman et al. 2016). This vast functional diversity could provide the raw material for the evolution of new enzymes and functions, and gene duplication seems to play an essential role in this process (Lang

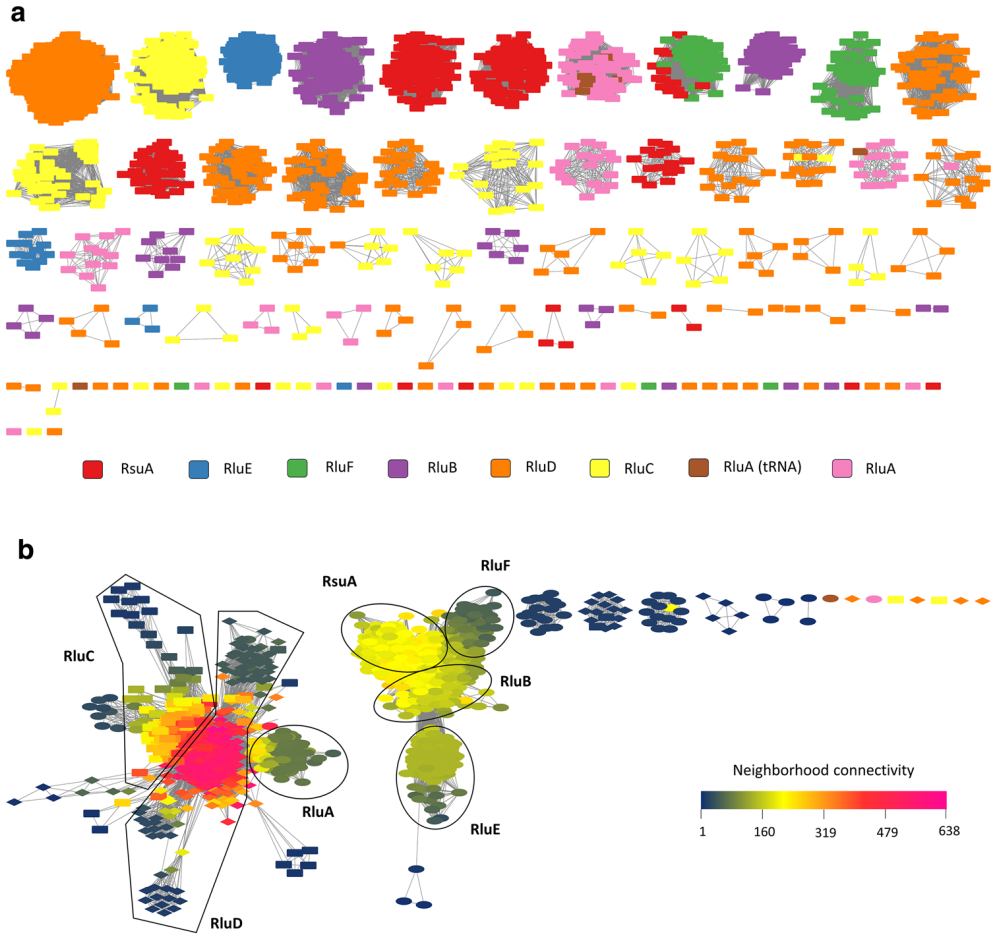


Fig. 7 Sequence similarity network (SSN) for members of the Pseudouridine Synthase superfamily that are found in our sample. Part ***a*** shows a similarity network built with the *transitivity algorithm*. Color code indicates that despite the presence of multiple clusters, most of these are exclusively made up of a single enzymatic subfamily. Part ***b*** depicts a network built with the *neighborhood algorithm*. Here, most clusters of the same color in part *a* group together into a single one, which in turn is well separated from the others (except only for RluB, RluF and RsuA). Nodes with a diamond and a rectangle correspond to RluD and RluC, respectively. All other nodes, representing the rest of rRNA-modifying pseudouridine synthases, are depicted with an ellipse. Nodes are colored according to the number of connections in the cluster to which they belong (Color figure online)

et al. 2000; Höcker et al. 2001; Sterner & Höcker 2006; Goldman et al. 2016).

Before we go on, it is important to clarify what we mean by "retention of the original function" in the context of this study. Here, the concept covers only those paralogous enzymes that after a duplication event still catalyze the same enzymatic reaction on the same positions of the same substrate in contrast to, for example, some tyrosine phosphatases (which share the same EC number) that modify different protein parts despite being homologous and performing the same overall reaction (Sweeney et al. 2005).

Bacterial Pseudouridine Synthases

By analyzing the ratio of paralogs within the isomerases notable differences between Bacteria and Archaea can be found. This is illustrated by the fact that sub-subclasses with the highest number of paralogs differ between both cellular domains (Fig. 3).

As it was noted previously, most paralogs belonging to the intramolecular transferases transferring *other* groups are pseudouridine synthases involved in tRNA or rRNA modification (sometimes both, but under very specific circumstances). All of these are stand-alone enzymes that perform the same overall reaction: the conversion of uridine to pseudouridine (Ψ) by means of the same chemical mechanism (Hamma & Ferré D'Amaré, 2006).

In terms of the number of paralogs and phylogenetic distribution, RluD is quite different from the other pseudouridine synthases: it is the enzyme with more duplicates and the broadest phylogenetic distribution (Fig. 6a). A plausible explanation for the presence of multiple copies in many organisms may have to do with the importance of the ribosomal region that contains the uridine residues modified by this enzyme. Such uridines are in positions 1911, 1915, and 1917 of 23S rRNA, in a region known as helix-loop 69 (HL69) (Leppik et al. 2007). This structure constitutes a bridge between ribosomal subunits and ensures translation fidelity (Jiang et al. 2014), which makes it very important for translation to take place. It has been shown that, at least in vitro, mutations at Ψ1917 inhibit translation (Liiv et al. 2005), and this could represent a strongly deleterious effect in vivo (Gutgsell et al. 2001; Ofengand et al. 2001). This might explain why many bacteria have this enzyme and why it is found in more than one copy in multiple organisms. The extra copies could provide robustness (i.e.,the ability for the phenotype to remain unchanged despite genotypic changes) by means of genetic redundancy, as some sort of *molecular backup* for this vital function, so that if one of the copies is mutated or gets compromised another can rescue the function and thus ensure translation.

One possible caveat from our study has to do with the sequences that we considered for each of the phylogenetic analyzes. Functional annotation was the inclusion criteria, so we grouped each sequence into one of the seven sub-families of rRNA-modifying pseudouridine synthases. This raises the possibility of including wrongly annotated sequences even after checking the MSA. To corroborate that this was not the case, we performed a sequence similarity network analysis (Fig. 7) in which we included all 1944 sequences considered for tree construction. Color homogeneity in the different clusters in Fig. 7a indicates that all sequences included in each one have a high similarity value and the same assigned function, which suggests that annotations in the KEGG database are correct. If they were not, we would expect to see a mixture of enzymes with different annotations in more than one of the clusters generated in the transitivity clustering analysis, which is not the case (although we find a few exceptions). Besides, network topology in Fig. 7b shows that the neighborhood connectivity distribution correlates with the enzyme function (EC code). Here, most enzymes that are scattered in different clusters in Fig. 7a are grouped into a single one. This indicates that sequences annotated with the same EC number are more similar to each other than enzymes with different annotations.

Furthermore, the fact that all RluD similarity clusters are made up exclusively (with only a few exceptions) of correctly annotated enzymes further suggests that many organisms from different phylogenetic groups have multiple copies of this enzyme and that all of them still perform the same overall reaction. It is important to note that the concept of "molecular backup" is substantially different from the classical scenario of the requirement of an additional gene-dosage, which can be the primary selective pressure for the duplicated gene to conserve the original function after being fixed. Therefore, we consider Kafri et al. (2005) definition in which molecular backup refers to those scenarios in which a requirement for additional dosage is not the reason for retaining the extra copies.

It was previously thought that divergence after gene duplication events would be so fast that most paralogs would rarely provide backup to the other copy due to the acquisition of a new function (Wagner 2005). However, most recent evidence suggests that if a gene is going to rescue the function of its paralog, it may not be necessary that both share high sequence or regulation similarity (they could be very divergent) (Ihmels et al. 2007). This could occur with RluD and its paralogs with the same EC number: most sequences show high similarity scores (red and pink diamonds in Fig. 7b) but there are others that lie outside the main cluster (blue and gray diamonds), which indicates a lower similarity value. In many cases, these belong to organisms that have three or more RluD copies.

Most genes associated with translation almost always occur as singletons (Jordan et al. 2004; Bratlie et al. 2010a; 2010b), but RluD seems to be a notable exception. Given

the cellular context in which this enzyme takes part, it is likely that the extra copies are only expressed under situations in which the original one is compromised, so as not to disrupt the balance of the other components of the translation machinery.

Despite many organisms having several RluD copies, it is unlikely that most of them, if not all, survive in the long term. Krakauer and Plotkin (2002) have suggested that natural selection in small populations would favor robustness mechanisms but this may not happen in large ones, such as those of many bacteria. This prediction is supported by in silico experiments where the mutational robustness tends to increase with mutation rate and decrease with population size (Elena et al. 2007). Likewise, an interplay between genetic drift and a mutational bias toward the deletion of redundant copies usually shapes bacterial genomes (Bobay & Ochman 2017). For these copies to be preserved after many generations, they would have to acquire some level of distinct or independent activity on which selection could act and not serve merely as a backup copy (Ghosh & O'Connor 2017; Putty et al. 2013). Consequently, more analysis and experiments are needed to explain why we detected several RluD copies in many organisms.

Other Bacterial Isomerases

Besides the high amount of paralogs associated with RNA modification we also found a high number related to propionate metabolism (Fig. 5 & Table 2). These correspond to a single enzyme, methylmalonyl-coenzyme A (methylmalonyl-CoA) mutase (MCM) (EC 5.4.99.2), which catalyzes the reversible isomerization of (R)-methylmalonyl-CoA into succinyl-coenzyme A (succinyl-CoA). This enzyme has several essential roles in bacterial cells. For example, it is crucial for the biosynthesis of complex polyketides, a vast category of diverse natural products that comprise pigments, antibiotics, and immunosuppressants, among others (Donadio et al. 1991). Perhaps even more important is its role in the glyoxylate cycle (GC). This is an essential mechanism for converting acetyl-CoA to succinate which can be used for carbohydrate biosynthesis (Kondrashov et al. 2006). This pathway is mainly found in methylotrophic organisms but is also suggested to be more widely distributed in other kinds of bacteria (Korotkova et al. 2002), specifically those with aerobic metabolisms (Ahn et al. 2016). Consistent with this point, we found MCM in different phyla from our sample, but we could not identify it in most phyla that include anaerobic organisms; we also detected that there is more than one copy of this enzyme for many organisms (data not shown). Since GC has a significant role as a carbon assimilation strategy in the absence of typical carbon sources such as glucose (Chew et al. 2019) and given the importance of MCM in polyketide biosynthesis, we suggest that a gene-dosage effect could be responsible for maintaining paralogs with the same function in many different organisms. The extra copies could enable the constant production of succinyl-CoA, which could go to polyketide-synthetizing pathways or the glyoxylate cycle, especially when typical carbon sources are scarce. Recently, it has been demonstrated that activating this pathway is a crucial acclimation strategy in marine bacteria subjected to iron limitation (Koedooder et al. 2018), which further highlights its importance in coping with changing environments.

In different pathogenic bacteria, the nucleotide sugar precursor dTDP-Rhamnose is a crucial component of the cell wall polysaccharides (Tsukioka et al. 1997). One of the enzymes involved in its biosynthesis, dTDP-4-dehydrorhamnose 3,5-epimerase (EC 5.1.3.13), was found among the isomerases with the highest number of paralogs. It catalyzes the conversion of dTDP-4-dehydro-6-deoxy-α-D-glucopyranose to dTDP-4-dehydro-β-L-rhamnose and is typically associated with another enzyme, dTDP-4-keto-L-rhamnose reductase (EC 1.1.1.133), which catalyzes the last step of the pathway leading to dTDP-L-rhamnose. It has been shown that this molecule is essential for the growth of bacteria from the genus *Mycobacterium* (Ma et al. 2002) and that disruption of this pathway leads to pleiotropic effects in *Azospirillum brasilense* (Jofré et al. 2004). We found that most organisms with paralogous dTDP-4-dehydrorhamnose 3,5-epimerase, have two or more enzymes with the same EC number. This suggests that a gene-dosage effect could be the reason for having two or more redundant copies of this enzyme, perhaps to satisfy the right dTDP-rhamnose amount to ensure pathogenic viability and virulence given the pivotal role of this molecule. Because of this pathway's importance, all its enzymes have been proposed as targets of a new class of antibiotics that can inhibit dTDP-Rhamnose biosynthesis (van der Beek et al. 2019).

UDP-glucuronate 4-epimerase (EC 5.1.3.6), like UDP-glucose 4-epimerase, is an isomerase that plays an important role in nucleotide diphosphate (NDP) sugars biosynthesis. It converts UDP-glucuronate to UDP-D-galacturonate, which is a pivotal molecule for moenomycin biosynthesis, a class of natural antibiotics that disrupts cell wall formation through the inhibition of peptidoglycan glycosyltransferases (PGTs) (Ostash & Walker 2010). This enzyme, along with other eight that also act upon NDP-sugars, belongs to the short-chain dehydrogenases/reductases (SDR) superfamily, shares common features in terms of bond change and reaction center similarity, and has the same domain composition (Martínez-Cuesta et al. 2014), which suggests a possible origin by gene duplication and later divergence of some of its members. Four of this superfamily members are also isomerases, and two of them, UDP-glucose 4-epimerase (EC 5.1.3.2) and ADP-glyceromanno-heptose 6-epimerase (EC 5.1.3.20) have also an important number

of paralogs in our bacterial sample. This suggests that these three enzymes could share a common origin and may have originated via gene duplication. The fact that UDP-glucuronate 4-epimerase is highly specific (Broach et al. 2012; Sun et al. 2020) further suggests a relatively recent origin through duplication and further specialization of another member of the superfamily, representing a possible case of sub/neofunctionalization.

Archaeal Isomerases

Unlike bacteria, archaea use a ribonucleoprotein complex to convert uridine residues into pseudouridine ones (Liang et al. 2009), which includes a pseudouridine synthase, a guide RNA, and several auxiliary proteins. Despite archaeal pseudouridine synthases being homologous to bacterial ones (Fitzek et al. 2018) and performing the same reaction in equivalent positions, we found no duplication signals for any of these enzymes.

We identified different duplicated enzymes involved in protein glycosylation (Fig. 5 and Table 2). One of these, UDP-glucose-4-epimerase (EC 5.1.3.2) has the largest number of paralogs (Table 2). This enzyme takes part in galactose metabolism, specifically in the Leloir pathway, which goes from β-D-galactose to glucose 1-phosphate (Holden et al. 2003). The main role of this enzyme seems to be the biosynthesis of different kinds of carbohydrates in glycoproteins, glycolipids, and cell walls through the production of galactosyl units (Frey 1996). Analyses in bacteria show that mutants for this enzyme produce different glycoforms which are not found in wild-type cells (Lee et al. 1999). As for the archaea, variations in glycosylation patterns, especially N-glycosylation, are thought to be associated with adaptations to different environments (Calo et al. 2010; Jarrell et al. 2014), so it seems likely that UDP-glucose-4-epimerase, as well as phosphoglucosamine mutase (EC 5.4.2.10) and UDP-N-acetylglucosamine 2-epimerase (EC 5.1.3.14), for which we also found many duplicates, contribute in some degree to glycan diversity. Some of their corresponding paralogs might be able to synthesize slightly different glycans, or be expressed under different environmental conditions.

The enzyme 1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino) methylideneamino] imidazole-4-carboxamide isomerase, better known as HisA (EC 5.3.1.16) is the second one with the highest number of paralogs identified for Archaea (Table 2). It is involved in histidine biosynthesis and produces the substrate PRFAR, which in turn is taken by another enzyme, HisF (EC 4.3.2.10), to synthesize the molecule AICAR, a pivotal molecule that connects histidine and purine biosynthesis besides having other important roles in cell's biochemistry (Vázquez-Salazar et al. 2018). HisA and HisF are paralogous enzymes (members of the $(\beta\alpha)_8$ superfamily of proteins). It has been proposed that they arose by a combination of gene duplication followed by gene elongation of an ancestral enzyme that was half the size ($(\beta\alpha)_4$) of extant $(\beta\alpha)_8$ barrels (Fani et al. 1994; Lang et al. 2000). Despite being universally distributed, there is a heterogeneous organization and distribution of *his* genes within the archaea, which suggests that the arrangement typically found in bacteria could have been absent in the Archaeal ancestor and also in the Last Universal Common Ancestor (LUCA) (Fondi et al. 2009). Of all enzymes that participate in histidine biosynthesis, HisA and HisF are the only ones whose ancient gene fusion is universally conserved (Fani et al. 2007). Given this, one would expect that when finding one of them within an organism, the other would almost certainly be there too. Additionally, given the functional diversity throughout $(\beta\alpha)_8$ proteins (Nagano et al. 2002), enzymes with this fold seem to be candidates in which a different or novel function could arise. This has been shown for $(\beta\alpha)_8$ barrels evolved from the combination of two different halves (or $(\beta\alpha)_4$ barrels) performing different functions (Höcker et al. 2004), and also for the *hisA* gene of *Salmonella enterica* which, after being amplified to a high copy number, acquired several advantageous mutations (Näsvall et al. 2012). This study was performed in vitro but could resemble a mechanism that can also occur in vivo, leading to the evolution of new enzymes and functions through duplication and divergence (Francino 2005).

Regarding central metabolism there are notable differences between Archaea and Bacteria. One of them is found in glycolysis, which is thought to be an ancient metabolic pathway (Fothergill-Gilmore 1986). Although its outcome is the same in both groups of organisms, there are striking variations specific to each cellular domain (Verhees et al. 2003). Additionally, the oxidative pentose phosphate pathway (OPPP), which operates parallel with glycolysis to synthesize ribulose-5-phosphate and is needed for nucleotide biosynthesis, was thought to be absent in Archaea, though it was recently reported in the archaeon *Haloferax volcanii* (Pickl & Schönheit 2015). An alternative pathway involved in AMP metabolism, known as pentose bisphosphate pathway (PBP), generates the intermediates needed for nucleotide biosynthesis (ribose 1,5-bisphosphate (R15P) and ribulose 1,5-bisphosphate). It has been identified in *Thermococcus kodakarensis* (Aono et al. 2015) and seems to be quite familiar to many archaea (Finn & Tabita 2004). The intermediate ribulose 1,5-bisphosphate is the substrate of ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO), an enzyme considered exclusive of the Calvin-Benson cycle of photosynthetic bacteria, but that is also part of the archaeal pathway previously mentioned (Kono et al. 2017). Our analysis identified a relatively high number of paralogs for another enzyme of this pathway: ribose 1,5-bisphosphate isomerase (EC 5.3.1.29), which catalyzes the conversion of α-D-ribose 1,5-bisphosphate to D-ribulose 1,5-bisphosphate,

and its activity is greatly enhanced in the presence of AMP (Aono et al. 2012). It is striking that another enzyme with a high ratio of duplicates, phosphoglycerate mutase (PGM) (EC 5.4.2.12), is related to the activity of R15P isomerase. PGM is a glycolytic enzyme that catalyzes the conversion of 3-phosphoglycerate (3PGA) to 2-phosphoglycerate (2PGA).

The archaeal AMP degradation pathway, which comprises R15P isomerase, RuBisCo and AMP phosphorylase, generates 3PGA as the product, which will ultimately be directed to glycolysis and converted into 2PGA by PGM (Aono et al. 2012). However, it is unclear if there is a dependence between the high ratio of paralogs of R15P and PGM. We suggest that for the case of R15P isomerase such a ratio might be associated with energetic metabolism. An ADP-dependent phosphofructokinase that produces AMP has been found in many archaea (Kengen et al. 2001). This generates large concentrations of intracellular AMP, leading to a higher amount of 3PGA via PBP. Subsequently, 3PGA can fuel ATP production, which would be very useful for the cell's bioenergetics if there are low energy levels (Sato et al. 2007). Thus, paralogs of R15P isomerase could be associated with dosage effects that could enable some archaea to thrive in environments where ATP production is limited, as in anaerobic environments.

In this study, we identify several cases in which the retention of duplicates is likely to be related to a dosage effect (i.e., a need to increase the metabolic flux of a specific reaction). This is perhaps the simplest conservation-of-function scenario because it only requires an additional gene copy that is expressed under similar conditions as the original one (instead of upregulating a single copy and enhancing the production of its product, a similar effect could be obtained if its paralog is retained and expressed simultaneously).

However, another scenario could be equally likely. Paralogs may still perform the same enzymatic reaction although their expression may not be simultaneous but dependent on environmental fluctuations, which is why they are known as ecoparalogs (Sanchez-Perez et al. 2008). For example, in *Lactobacillus johnsonii*, three enolase copies have been identified but only two of them are expressed under the same culture conditions (Antikainen et al. 2007). The halophilic bacterium *Salinibacter ruber* has several pairs of paralogs whose members differ in their electrostatic potential, both on the surface and on the active site, resulting in different halophillicity (Sanchez-Perez et al. 2008). And there are cases in which the expression of one member of a pair of paralogous genes associated with pathogenesis (and with the same function and biological role) depends on the extracellular concentration of certain ions (Mouammine et al. 2014; Fortuna et al. 2022).

Examples of ecoparalogs are also found in Archaea. *Haloferax volcanii* has three copies of group II chaperonin whose differential expression depends on salt concentrations and temperature (Kapatai et al. 2006). In *Haloarcula marismortui* paralogs FlaA2 and FlaB, involved in the biosynthesis of the archaellum (the archaeal flagellum), are well adapted to changing environments and not expressed simultaneously. As in *H. volcanii*, their expression is dependent on temperature and salt concentration (Syutkin et al. 2014; 2019). *H. marismortui* has also three copies of rRNA operons of which two are almost identical at the sequence level (operons A and C). The other, identified as operon B, has more than 130 polymorphisms compared to operons A and C, even in the promoter region, and is overexpressed at higher temperatures and underexpressed at lower ones (López-López et al. 2007).

It could be argued that the examples mentioned above represent cases of neo or subfunctionalization because such paralogs are not expressed under the same conditions. Besides, mutational analyses reveal that some genes can not provide a backup function if one or more paralogs are inactivated, which leads to phenotypic differences (Kapatai et al. 2006; López-López et al. 2007; Tripepi et al. 2013). As previously mentioned, when we say that two paralogous genes have the same function we are referring only to the enzymatic reaction as defined by the EC number. Whether those copies fulfill or not the same cellular role is something that can not be identified with our current analyses.

Concluding Remarks

The relatively small number of isomerases allowed us to analyze their paralogs' content in deeper detail at the levels of sub and sub-subclasses. We found a great heterogeneity at both levels, that is, only a few subclasses and sub-subclasses have a high ratio of paralogs, whereas the majority are underrepresented.

Interestingly, many enzymes and functions associated with a high number of duplicates are involved in the biosynthesis of several antibiotics, molecules required for pathogenicity, and processes involving RNA and/or modified ribonucleotides. An important example was the enzyme RluD, which modifies the 23S rRNA at positions crucial for proper ribosome assembly, and in many organisms, more than one paralog performs the same function. This might represent a case of molecular backup that could rescue RluD's essential function if the original enzyme gets compromised. However, additional studies are needed to identify if such copies serve a redundant role or if they are in the pseudogenization process for their subsequent removal from the genome.

Although Bacteria and Archaea share many enzymes, as indicated by the presence of the same EC numbers in different bacterial and archaeal proteomes, there are notable differences in retention of duplicates. Sub-subclasses with more paralogs differ between both domains, except

for the racemases and epimerases acting on carbohydrates and derivatives (EC 5.1.3). Moreover, even when the same enzymes are identified for both domains, different reasons are likely to explain the high number of duplicates. The most striking example corresponds to the UDP-glucose 4-epimerase, for which the highest number of copies was found in bacteria and archaea. However, the reasons behind this may be different for each domain. Whereas in Archaea the paralogs of this enzyme could be related to the production of a wider glycan diversity, as well as with a differential dosage requirement according to specific environments, for Bacteria it seems more likely that subfunctionalization was involved. An example is UDP-glucuronate 4-epimerase, an enzyme that belongs to the same superfamily as UDP-glucose 4-epimerase but is involved in the biosynthesis of the antibiotic moenomycin.

Our findings suggest that conservation of the same enzymatic reaction, either associated with fixation due to additional dose requirement or differential expression in response to environmental fluctuations, could be the most common fate for duplicated genes. It is likely that some of these extra copies would be able to accumulate mutations that in the long term could bestow them the ability to perform low levels of secondary activities (i.e., making them *promiscuous*), which could be the starting point for the evolution of new functions or the optimization of the ancestral ones (Copley 2020). Further phylogenetic and functional analyses of the isomerases may help to explain why there is a high ratio of paralogs within this enzymatic class.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00239-022-10085-x.

Acknowledgements AÁ-L thanks the Posgrado en Ciencias Biológicas at the Universidad Nacional Autónoma de México, as well as Consejo Nacional de Ciencia y Tecnología (CONACYT) for their support with fellowship No. 747513. Financial support by PAPIIT-UNAM (IN214421) is gratefully acknowledged. Thanks are given to José Alberto Campillo-Balderas and Ricardo Hernández-Morales for helpful comments on the manuscript.

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Ahn S, Jung J, Jang IA, Madsen EL, Park W (2016) Role of glyoxylate shunt in oxidative stress response. J Biol Chem 291(22):11928–11938. https://doi.org/10.1074/jbc.M115.708149

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3404. https://doi.org/10.1093/nar/25.17.3389

Álvarez-Lugo A, Becerra A (2021) The role of gene duplication in the divergence of enzyme function: a comparative approach. Front Genet 12:1–16. https://doi.org/10.3389/fgene.2021.641817

Antikainen J, Kuparinen V, Lähteenmäki K, Korhonen TK (2007) Enolases from Gram-positive bacterial pathogens and commensal lactobacilli share functional similarity in virulence-associated traits. FEMS Immunol Med Microbiol 51(3):526–534. https://doi.org/10.1111/j.1574-695X.2007.00330.x

Aono R, Sato T, Imanaka T, Atomi H (2015) A pentose bisphosphate pathway for nucleoside degradation in Archaea. Nat Chem Biol 11(5):355–360. https://doi.org/10.1038/nchembio.1786

Aono R, Sato T, Yano A, Yoshida S, Nishitani Y, Miki K, Imanaka T, Atomi H (2012) Enzymatic characterization of amp phosphorylase and ribose-1,5-bisphosphate isomerase functioning in an archaeal amp metabolic pathway. J Bacteriol 194(24):6847–6855. https://doi.org/10.1128/JB.01335-12

van der Beek SL, Zorzoli A, Çanak E, Chapman RN, Lucas K, Meyer BH, Evangelopoulos D, de Carvalho LPS, Boons GJ, Dorfmueller HC, van Sorge NM (2019) Streptococcal dTDP-L-rhamnose biosynthesis enzymes: functional characterization and lead compound identification. Mol Microbiol 111(4):951–964. https://doi.org/10.1111/mmi.14197

Bobay LM, Ochman H (2017) The evolution of bacterial genome architecture. Front Genet 8:1–6. https://doi.org/10.3389/fgene.2017.00072

Bratlie MS, Johansen J, Drabløs F (2010) Relationship between operon preference and functional properties of persistent genes in bacterial genomes. BMC Genomics 11:1. https://doi.org/10.1186/1471-2164-11-71

Bratlie MS, Johansen J, Sherman BT, Huang DW, Lempicki RA, Drabløs F (2010) Gene duplications in prokaryotes can be associated with environmental adaptation. BMC Genomics 11:1. https://doi.org/10.1186/1471-2164-11-588

Broach B, Gu X, Bar-Peled M (2012) Biosynthesis of UDP-glucuronic acid and UDP-galacturonic acid in Bacillus cereus subsp. cytotoxis NVH 391–98. FEBS J 279(1):100–112. https://doi.org/10.1111/j.1742-4658.2011.08402.x

Calo D, Kaminski L, Eichler J (2010) Protein glycosylation in Archaea: sweet and extreme. Glycobiology 20(9):1065–1076. https://doi.org/10.1093/glycob/cwq055

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Chew SY, Chee WJ, Than LTL (2019) The glyoxylate cycle and alternative carbon metabolism as metabolic adaptation strategies of *Candida glabrata*: Perspectives from *Candida albicans* and *Saccharomyces cerevisiae*. J Biomed Sci 26(1):1–10. https://doi.org/10.1186/s12929-019-0546-5

Copley SD (2020) Evolution of new enzymes by gene duplication and divergence. FEBS J 287(7):1262–1283. https://doi.org/10.1111/febs.15299

Díaz-Mejía JJ, Pérez-Rueda E, Segovia L (2007) A network perspective on the evolution of metabolism by gene duplication. Genome Biol 8(2):1–10. https://doi.org/10.1186/gb-2007-8-2-r26

Donadio S, Staver MJ, McAlpine JB, Swanson SJ, Katz L (1991) Modular organization of genes required for complex polyketide biosynthesis. Science 252:675–679

Elena SF, Wilke CO, Ofria C, Lenski RE (2007) Effects of population size and mutation rate on the evolution of mutational robustness. Evolution 61(3):666–674. https://doi.org/10.1111/j.1558-5646.2007.00064.x

Fani R, Brilli M, Fondi M, Lió P (2007) The role of gene fusions in the evolution of metabolic pathways: The histidine biosynthesis case. BMC Evolut Biol 7:2. https://doi.org/10.1186/1471-2148-7-S2-S4

Fani R, Liò P, Chiarelli I, Bazzicalupo M (1994) The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the hisA and hisF genes. J Mol Evol 38(5):489–495. https://doi.org/10.1007/BF00178849

Finn MW, Tabita RR (2004) Modified pathway to synthesize ribulose 1,5-bisphosphate in methanogenic archaea. J Bacteriol 186(19):6360–6366. https://doi.org/10.1128/JB.186.19.6360-6366.2004

Fitzek E, Joardar A, Gupta R, Geisler M (2018) Evolution of eukaryal and archaeal pseudouridine synthase pus10. J Mol Evol 86(1):77–89. https://doi.org/10.1007/s00239-018-9827-y

Fondi M, Emiliani G, Liò P, Gribaldo S, Fani R (2009) The evolution of histidine biosynthesis in archaea: insights into the his genes structure and organization in luca. J Mol Evol 69(5):512–526. https://doi.org/10.1007/s00239-009-9286-6

Fortuna A, Collalto D, Schiaffi V, Pastore V, Visca P, Ascenzioni F, Rampioni G, Leoni L (2022) The *Pseudomonas aeruginosa* DksA1 protein is involved in H2O2 tolerance and within-macrophages survival and can be replaced by DksA2. Sci Rep 12(1):1–11. https://doi.org/10.1038/s41598-022-14635-7

Fothergill-Gilmore LA (1986) The evolution of the glycolytic pathway. Trends Biochem Sci 11(1):47–51. https://doi.org/10.1016/0968-0004(86)90233-1

Francino MP (2005) An adaptive radiation model for the origin of new gene functions. Nat Genet 37(6):573–577. https://doi.org/10.1038/ng1579

Frey PA (1996) The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose. FASEB J 10(4):461–470

Furnham N, Sillitoe I, Holliday GL, Cuff AL, Laskowski RA, Orengo CA, Thornton JM (2012) Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. PLoS Comput Biol 8:3. https://doi.org/10.1371/journal.pcbi.1002403

Ghosh S, O'Connor TJ (2017) Beyond paralogs: The multiple layers of redundancy in bacterial pathogenesis. Front Cell Infect Microbiol 7:1–14. https://doi.org/10.3389/fcimb.2017.00467

Goldman AD, Beatty JT, Landweber LF (2016) The TIM Barrel architecture facilitated the early evolution of protein-mediated metabolism. J Mol Evol 82(1):17–26. https://doi.org/10.1007/s00239-015-9722-8

Gutgsell NS, Del Campo M, Raychaudhuri S, Ofengand J (2001) A second function for pseudouridine synthases: a point mutant of RluD unable to form pseudouridines 1911, 1915, and 1917 in Escherichia coli 23S ribosomal RNA restores normal growth to an RluD-minus strain. RNA 7(7):990–998. https://doi.org/10.1017/S1355838201000243

Hamma T, Ferré-D'Amaré AR (2006) Pseudo uridine synthases. Chem Biol 13(11):1125–1135. https://doi.org/10.1016/j.chembiol.2006.09.009

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 35(2):518–522. https://doi.org/10.1093/molbev/msx281.PMID:29077904;PMCID:PMC5850222

Höcker B, Beismann-Driemeyer S, Hettwer S, Lustig A, Sterner R (2001) Dissection of a (βα)8-barrel enzyme into two folded halves. Nat Struct Biol 8(1):32–36. https://doi.org/10.1038/83021

Höcker B, Claren J, Sterner R (2004) Mimicking enzyme evolution by generating new (βα) 8-barrels from (βα)4-half-barrels. Proc Natl Acad Sci USA 101(47):16448–16453. https://doi.org/10.1073/pnas.0405832101

Holden HM, Rayment I, Thoden JB (2003) Structure and function of enzymes of the leloir pathway for galactose metabolism. J Biol Chem 278(45):43885–43888. https://doi.org/10.1074/jbc.R300025200

Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS (2007) Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. Mol Syst Biol 3:86. https://doi.org/10.1038/msb4100127

Jarrell KF, Ding Y, Meyer BH, Albers S-V, Kaminski L, Eichler J (2014) N-Linked glycosylation in Archaea: a structural, functional, and genetic analysis. Microbiol Mol Biol Rev 78(2):304–341. https://doi.org/10.1128/mmbr.00052-13

Jensen RA (1976) Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30(1):409–425. https://doi.org/10.1146/annurev.mi.30.100176.002205

Jiang J, Aduri R, Chow CS, Santa Lucia J (2014) Structure modulation of helix 69 from *Escherichia coli* 23S ribosomal RNA by pseudouridylations. Nucleic Acids Res 42(6):3971–3981. https://doi.org/10.1093/nar/gkt1329

Jofré E, Lagares A, Mori G (2004) Disruption of dTDP-rhamnose biosynthesis modifies lipopolysaccharide core, exopolysaccharide production, and root colonization in *Azospirillum brasilense*. FEMS Microbiol Lett 231(2):267–275. https://doi.org/10.1016/S0378-1097(04)00003-5

Jordan IK, Wolf YI, Koonin EV (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol Biol 4:1–11. https://doi.org/10.1186/1471-2148-4-22

Kafri R, Bar-Even A, Pilpel Y (2005) Transcription control reprogramming in genetic backup circuits. Nat Genet 37(3):295–299. https://doi.org/10.1038/ng1523

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30. https://doi.org/10.1093/nar/28.1.27

Kapatai G, Large A, Benesch JLP, Robinson CV, Carrascosa JL, Valpuesta JM, Gowrinathan P, Lund PA (2006) All three chaperonin genes in the archaeon *Haloferax volcanii* are individually dispensable. Mol Microbiol 61(6):1583–1597. https://doi.org/10.1111/j.1365-2958.2006.05324.x

Kengen SW, Tuninga J, Verhees CH, van der Oost J, Stams AJM, de Vos WM (2001) ADP-dependent glucokinase and phosphofructokinase from pyrococcus furiosus. Methods Enzymol 331:41–53

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30(4):772–780. https://doi.org/10.1093/molbev/mst010

Knoch E, Sugawara S, Mori T, Poulsen C, Fukushima A, Harholt J, Fujimoto Y, Umemoto N, Saito K (2018) Third DWF1 paralog in Solanaceae, sterol Δ24-isomerase, branches withanolide biosynthesis from the general phytosterol pathway. Proc Natl Acad Sci USA 115(34):E8096–E8103. https://doi.org/10.1073/pnas.1807482115

Koedooder C, Guéneuguès A, van Geersdaële R, Vergé V, Bouget FY, Labreuche Y, Obernosterer I, Blain S (2018) The role of the glyoxylate shunt in the acclimation to iron limitation in marine heterotrophic bacteria. Front Marine Sci 5:1–12. https://doi.org/10.3389/fmars.2018.00435

Kondrashov FA, Koonin E, Morgunov IG, Tv F, Kondrashova MN (2006) Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. Biol Direct 1:1–14. https://doi.org/10.1186/1745-6150-1-31

Kono T, Mehrotra S, Endo C, Kizu N, Matusda M, Kimura H, Mizohata E, Inoue T, Hasunuma T, Yokota A, Matsumura H, Ashida H (2017) A RuBisCO-mediated carbon metabolic pathway in methanogenic archaea. Nat Commun 8:14007. https://doi.org/10.1038/ncomms14007

Korotkova N, Chistoserdova L, Kuksa V, Lidstrom ME (2002) Glyoxylate regeneration pathway in the methylotroph. Microbiology 184(6):1750–1758

Krakauer DC, Plotkin JB (2002) Redundancy, antiredundancy, and the robustness of genomes. Proc Natl Acad Sci USA 99(3):1405–1409. https://doi.org/10.1073/pnas.032668599

Lang D, Thoma R, Henn-Sax M, Sterner R, Wilmanns M (2000) Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. Science 289(5484):1546–1550. https://doi.org/10.1126/science.289.5484.1546

Lee FKN, Gibson BW, Melaugh W, Zaleski A, Apicella MA (1999) Relationship between UDP-glucose 4-epimerase activity and oligoglucose glycoforms in two strains of *Neisseria meningitidis*. Infect Immun 67(3):1405–1414. https://doi.org/10.1128/iai.67.3.1405-1414.1999

Leppik M, Peil L, Kipper K, Liiv A, Remme J (2007) Substrate specificity of the pseudouridine synthase RluD in *Escherichia coli*. FEBS J 274(21):5759–5766. https://doi.org/10.1111/j.1742-4658.2007.06101.x

Letunic I, Bork P (2021) Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49(W1):W293–W296. https://doi.org/10.1093/nar/gkab301

Liang B, Zhou J, Kahen E, Terns RM, Terns MP, Li H (2009) Structure of a functional ribonucleoprotein pseudouridine synthase bound to a substrate RNA. Nat Struct Mol Biol 16(7):740–746. https://doi.org/10.1038/nsmb.1624

Liiv A, Karitkina D, Maiväli Ü, Remme J (2005) Analysis of the function of E coli 23S rRNA helix-loop 69 by mutagenesis. BMC Mol Biol 6:1–9. https://doi.org/10.1186/1471-2199-6-18

López-López A, Benlloch S, Bonfá M, Rodríguez-Valera F, Mira A (2007) Intragenomic 16s rDNA divergence in *Haloarcula marismortui* is an adaptation to different temperatures. J Mol Evol 65(6):687–696. https://doi.org/10.1007/s00239-007-9047-3

Ma Y, Pan F, McNeil M (2002) Formation of dTDP-rhamnose is essential for growth of mycobacteria. J Bacteriol 184(12):3392–3395. https://doi.org/10.1128/JB.184.12.3392-3395.2002

Martínez Cuesta S, Rahman SA, Furnham N, Thornton JM (2015) The classification and evolution of enzyme function. Biophys J 109(6):1082–1086. https://doi.org/10.1016/j.bpj.2015.04.020

Martínez-Cuesta S, Furnham N, Rahman SA, Sillitoe I, Thornton JM (2014) The evolution of enzyme function in the isomerases. Curr Opin Struct Biol 26(1):121–130. https://doi.org/10.1016/j.sbi.2014.06.002

Martínez-Cuesta S, Rahman SA, Thornton JM (2016) Exploring the chemistry and evolution of the isomerases. Proc Natl Acad Sci USA 113(7):1796–1801. https://doi.org/10.1073/pnas.1509494113

McDonald AG, Boyce S, Tipton KF (2009) ExplorEnz: The primary source of the IUBMB enzyme list. Nucleic Acids Res 37(SUPPL):1. https://doi.org/10.1093/nar/gkn582

McDonald AG, Boyce S, Tipton KF (2015) Enzyme classification and nomenclature. ELS. https://doi.org/10.1002/9780470015902.a0000710

Mouammine A, Lanois A, Pagès S, Lafay B, Molle V, Canova M, Girard PA, Duvic B, Givaudan A, Gaudriault S (2014) Ail and PagC-related proteins in the entomopathogenic bacteria of photorhabdus genus. PLoS ONE 9:10. https://doi.org/10.1371/journal.pone.0110060

Mudunuri U, Che A, Yi M, Stephens RM (2009) bioDBnet: the biological database network. Bioinformatics 25(4):555–556. https://doi.org/10.1093/bioinformatics/btn654

Nagano N, Orengo CA, Thornton JM (2002) One-fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 321(5):741–765. https://doi.org/10.1016/S0022-2836(02)00649-6

Näsvall J, Sun L, Roth JR, Andersson DI (2012) Real-time evolution of new genes by innovation, amplification, and divergence. Science 338(6105):384–387. https://doi.org/10.1126/science.1226521

Neme R, Tautz D (2014) Evolution: dynamics of de novo gene emergence. Curr Biol 24(6):R238–R240. https://doi.org/10.1016/j.cub.2014.02.016

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32(1):268–274. https://doi.org/10.1093/molbev/msu300

Ofengand J, Malhotra A, Remme J, Gutgsell NS, Del Campo M, Jean-Charles S, Peil L, Kaya Y (2001) Pseudouridines and pseudouridine synthases of the ribosome. Cold Spring Harb Symp Quant Biol 66:147–159. https://doi.org/10.1101/sqb.2001.66.147

Ohno S (1970) Evolution by gene duplication. Springer, New York

Ostash B, Walker S (2010) Moenomycin family antibiotics: chemical synthesis, biosynthesis, and biological activity. Nat Prod Rep 27(11):1594–1617. https://doi.org/10.1039/c001461n

Pickl A, Schönheit P (2015) The oxidative pentose phosphate pathway in the haloarchaeon *Haloferax volcanii* involves a novel type of glucose-6-phosphate dehydrogenase—the archaeal Zwischenferment. FEBS Lett 589(10):1105–1111. https://doi.org/10.1016/j.febslet.2015.03.026

Putty K, Marcus SA, Mittl PRE, Bogadi LE, Hunter AM, Arur S, Berg DE, Sethu P, Kalia A (2013) Robustness of *helicobacter pylori* infection conferred by context-variable redundancy among cysteine-rich paralogs. PLoS ONE 8:3. https://doi.org/10.1371/journal.pone.0059560

Sanchez-Perez G, Mira A, Nyiro G, Pasić L, Rodriguez-Valera F (2008) Adapting to environmental changes using specialized paralogs. Trends Genet 24(4):154–158. https://doi.org/10.1016/j.tig.2008.01.005

Sato T, Atomi H, Imanaka T (2007) Archaeal type III RuBisCOs function in a pathway for AMP metabolism. Science 315(5814):1003–1006. https://doi.org/10.1126/science.1135999

Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D (2004) BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 32:431–433. https://doi.org/10.1093/nar/gkh081

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504

Sharma P, Guptasarma P (2017) Endoglucanase activity at a second site in *Pyrococcus furiosus* triosephosphate isomerase—promiscuity or compensation for a metabolic handicap? FEBS Open Bio 7(8):1126–1143. https://doi.org/10.1002/2211-5463.12249

Sterner R, Höcker B (2006) Catalytic versatility, stability, and evolution of the (βα)8-barrel enzyme fold. Chem Rev 105(11):4038–4055. https://doi.org/10.1021/cr030191z

Sun H, Ko TP, Liu W, Liu W, Zheng Y, Chen CC, Guo RT (2020) Structure of an antibiotic-synthesizing UDP-glucuronate 4-epimerase MoeE5 in complex with substrate. Biochem Biophys Res Commun 521(1):31–36. https://doi.org/10.1016/j.bbrc.2019.10.035

Sweeney MC, Wavreille AS, Park J, Butchar JP, Tridandapani S, Pei D (2005) Decoding protein-protein interactions through combinatorial chemistry: Sequence specificity of SHP-1, SHP-2, and SHIP SH2 domains. Biochemistry 44(45):14932–14947. https://doi.org/10.1021/bi051408h

Syutkin AS, Pyatibratov MG, Galzitskaya OV, Rodríguez-Valera F, Fedorov OV (2014) *Haloarcula marismortui* archaellin genes as ecoparalogs. Extremophiles 18(2):341–349. https://doi.org/10.1007/s00792-013-0619-4

Syutkin AS, van Wolferen M, Surin AK, Albers SV, Pyatibratov MG, Fedorov OV, Quax TEF (2019) Salt-dependent regulation of archaellins in *Haloarcula marismortui*. MicrobiologyOpen 8(5):1–8. https://doi.org/10.1002/mbo3.718

Tipton K, Boyce S (2000) History of the enzyme nomenclature system. Bioinformatics 16(1):34–40. https://doi.org/10.1093/bioinformatics/16.1.34

Tripepi M, Esquivel RN, Wirth R, Pohlschröder M (2013) *Haloferax volcanii* cells lacking the flagellin FlgA2 are hypermotile. Microbiology (united Kingdom) 159(PART11):2249–2258. https://doi.org/10.1099/mic.0.069617-0

Tsukioka Y, Yamashita Y, Oho T, Nakano Y, Koga T (1997) Biological function of the dTDP-rhamnose synthesis pathway in *Streptococcus mutans*. J Bacteriol 179(4):1126–1134. https://doi.org/10.1128/jb.179.4.1126-1134.1997

Vázquez-Salazar A, Becerra A, Lazcano A (2018) Evolutionary convergence in the biosyntheses of the imidazole moieties of histidine and purines. PLoS ONE 13(4):1–22. https://doi.org/10.1371/journal.pone.0196349

Verhees CH, Kengen SW, Tuninga JE, Schut GJ, Adams MWW, de Vos WM, van der Oost J (2003) The unique features of glycolytic pathways in Archaea. Biochem J 375(2):231–246

Wagner A (2005) Distributed robustness versus redundancy as causes of mutational robustness. BioEssays 27(2):176–188. https://doi.org/10.1002/bies.20170

Zallot R, Oberg N (2019) Gerlt JA (2019) the EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. Biochemistry 58(41):4169–4182

DISCUSIÓN GENERAL

Consideraciones acerca del genoma y el nivel de ploidía en organismos procariontes

A diferencia de lo que ha ocurrido a lo largo de la evolución de distintos linajes eucariontes, la duplicación a nivel de genomas completos no parece haber jugado un papel tan importante en procariontes. Hace algunas décadas, antes de que se secuenciara el primer genoma procarionte, se observó que había cierta relación, en términos de la posición en el genoma, de algunos genes funcionalmente similares y que eran producto de una duplicación, tanto en *Escherichia coli* (Zikas & Riley 1975) como en *Streptomyces coelicolor* (Hopwood 1967). Esto sugería que la duplicación del genoma completo en los respectivos ancestros podría haber estado relacionada con la evolución de estos organismos. Años más tarde, y basándose en datos genómicos de más de 50 organismos, Herdman (1985) propuso que incluso varias rondas de WGD habrían ocurrido a lo largo de la evolución de los procariontes, por lo que los organismos más recientes tendrían los genomas más grandes. Sin embargo, estudios más recientes con una muestra más grande y de mejor calidad no apoyan esta hipótesis. Muchos de los organismos con los genomas más pequeños no son antiguos sino que han pasado por procesos de reducción de su genoma, lo cual se observa a lo largo de diferentes clados (Islas *et al*. 2000).

Existe una diferencia conceptual muy importante con respecto a la duplicación del genoma completo en eucariontes y procariontes. En los primeros, la linealidad de los cromosomas permite que estos puedan incorporar cantidades importantes de material genético sin que haya grandes restricciones de por medio, lo cual queda de manifiesto en muchos organismos diferentes cuyos genomas presentan evidencia de haber pasado por una o más rondas de duplicación del genoma completo en el pasado (van de Peer 2004; Wolfe 2001). Por el contrario, los procariontes poseen cromosomas circulares sobre los cuales intervienen diversos factores que modulan y mantienen relativamente constante el tamaño y organización de su genoma, tales como su posición filogenética (Maistrenko *et al*. 2020; Martínez-Gutiérrez & Aylward 2022), la tasa de mutación (Marais *et al*. 2020), el ambiente en el que habitan (Maistrenko *et al*. 2020), su potencial metabólico (Rodríguez-Gijón *et al*. 2023) e incluso el número de virus, plásmidos y sistemas CRISPR con los que estén asociados (los dos primeros se asocian con la expansión del genoma mientras que el último favorece la reducción) (Gao *et*

al. 2019). Una excepción importante son las bacterias del género *Streptomyces*, las cuales poseen tanto cromosomas (Leblond & Decaris 1999) como plásmidos (Chater & Kinashi 2007) lineales. Estas bacterias poseen un porcentaje importante de genes duplicados (alrededor del 10%), mismos que se originaron a través de duplicaciones de uno o unos cuantos genes y no mediante la duplicación de todo el genoma (Zhou *et al*. 2012).

Estrictamente hablando sí existe la poliploidía en procariontes, aunque no en el mismo sentido que en los eucariontes. Por ejemplo, en la bacteria *Achromatium oxaliferum* se ha observado que existen múltiples copias del genoma, separadas físicamente una de otra, las cuales no son idénticas. Una vez que ocurre la división celular, cada una de las células hijas adquirirá copias diferentes, algunas de las cuales podrían estar ausentes en la otra (Soppa 2022). En ocasiones, la diversidad genética entre dos células hermanas es tal que podría pensarse que se trata de especies diferentes que pertenecen al mismo género (Ionescu *et al*. 2017).

De manera general, la presencia de muchas copias del genoma en procariontes parece ser un fenómeno más común de lo que originalmente se pensaba y puede representar una ventaja en ciertos casos. Por ejemplo, para ayudar a reducir la tasa de mutaciones espontáneas (muta un gen o genes determinados en una de las copias solamente) (Pecoraro *et al*. 2011), como reservorios de fosfato o un mecanismo de protección contra el rompimiento del DNA (sería improbable que todas las copias sufrieran daño simultáneamente) (van de Peer *et al*. 2017), para optimizar la expresión de proteínas específicas en alguna región celular (como proteínas asociadas al transporte celular en regiones cercanas a la membrana plasmática) sobre todo en bacterias de gran tamaño (Angert 2012) o como un mecanismo de resistencia a condiciones adversas que pudieran provocar el rompimiento de las hebras de DNA como la desecación o los rayos X (Kottemann *et al*. 2005).

Además de los puntos mencionados previamente, se ha observado que distintas especies de procariontes pueden modificar su nivel de ploidía a lo largo de su ciclo de vida. En las arqueas halófilas *Halobacterium salinarum* y *Haloferax volcanii* el número de copias cromosómicas aumenta durante la fase de crecimiento exponencial y disminuye cuando regresan a la fase estacionaria (Breuert *et al*. 2006). Y en cultivos de *Escherichia coli* que se reproducen a velocidades diferentes se ha observado que esta bacteria es monoploide si su crecimiento es lento y poliploide cuando este es rápido (Pecoraro *et al*. 2011).

Todo lo anterior sugiere que la duplicación de genomas en procariontes tiene que ver más con una serie de estrategias fisiológicas que parecen estar muy relacionadas con la variación de distintas condiciones en el medio que rodea a estos organismos, y no tanto con la diversificación de linajes como ocurre en eucariontes (Crow & Wagner 2005).

Prevalencia de la duplicación de genes en los genomas procariontes

Desde hace varios años se sabe que la evolución de los genomas procariontes es un proceso dinámico que consta básicamente de dos etapas: la adquisición de nuevos genes y nuevas funciones y la pérdida de estos. El incremento en el contenido génico está determinado por dos procesos: la duplicación y el transporte horizontal de genes, mientras que la pérdida de genes puede darse de manera directa o pasar por un proceso más largo que involucra la formación de pseudogenes (Mira *et al*. 2001). Ahora bien, con respecto a la ganancia de genes, aún existe un amplio debate acerca de si lo que predomina en procariontes es la duplicación o la transferencia horizontal. Inicialmente se pensaba que la duplicación génica en pequeña escala era el mecanismo principal para explicar la enorme diversidad funcional que observamos en procariontes (Serres *et al*. 2009). Por otra parte, trabajos más recientes sugieren lo contrario, es decir, que el transporte horizontal es bastante más común en arqueas y bacterias (Treangen & Rocha 2011) y que las duplicaciones podrían ser hasta 100 veces menos comunes (Tria & Martin 2021).

El transporte horizontal de genes muchas veces es mediado por vectores de material genético extracromosómico como los plásmidos (Harrison & Brockhusrst 2012; Rodríguez-Beltrán *et al*. 2021). Por ello, con el fin de evitar que nuestros resultados se vieran afectados por los genes originados por este mecanismo, decidimos excluir del análisis todos aquellos genes que estuvieran presentes en plásmidos. Esto no garantiza que no incluyamos genes xenólogos (aquellos originados por transferencia horizontal), pero sí reduce bastante la probabilidad de que esto ocurra. Se ha demostrado que en muchas especies de bacterias de diferentes phyla únicamente alrededor del 1% de los genes localizados en el cromosoma principal se originaron por transferencia horizontal (Oliveira *et al*. 2017).

Para evitar sesgos asociados con la sobrerrepresentación de algún grupo particular de organismos y con la redundancia en el caso de aquellos cuya posición filogenética fuera muy cercana, decidimos utilizar una muestra representativa reportada previamente por Martínez-Núñez *et al*. (2013). Esta se construyó a partir de la concatenación de 23 proteínas que se encuentran conservadas en los tres dominios celulares, las cuales tienen que ver, en su mayoría, con el proceso de traducción, y unas pocas con transcripción, replicación del DNA y metabolismo celular (Brown *et al*. 2001). Se construyó una filogenia basada en máxima verosimilitud, utilizando todos los genomas completamente secuenciados disponibles hasta ese momento y, posteriormente, se conservaron únicamente aquellos que fueran filogenéticamente distantes. De los 745 proteomas resultantes, solo en 221 identificamos la presencia de plásmidos (datos no mostrados). De estos, la mayoría posee dos plásmidos y solo un pequeño porcentaje de organismos posee más de cuatro (Figura 1). Asimismo, en la mayoría de estos organismos el porcentaje de genes localizados en plásmidos es menor al 5% (Figura 1).

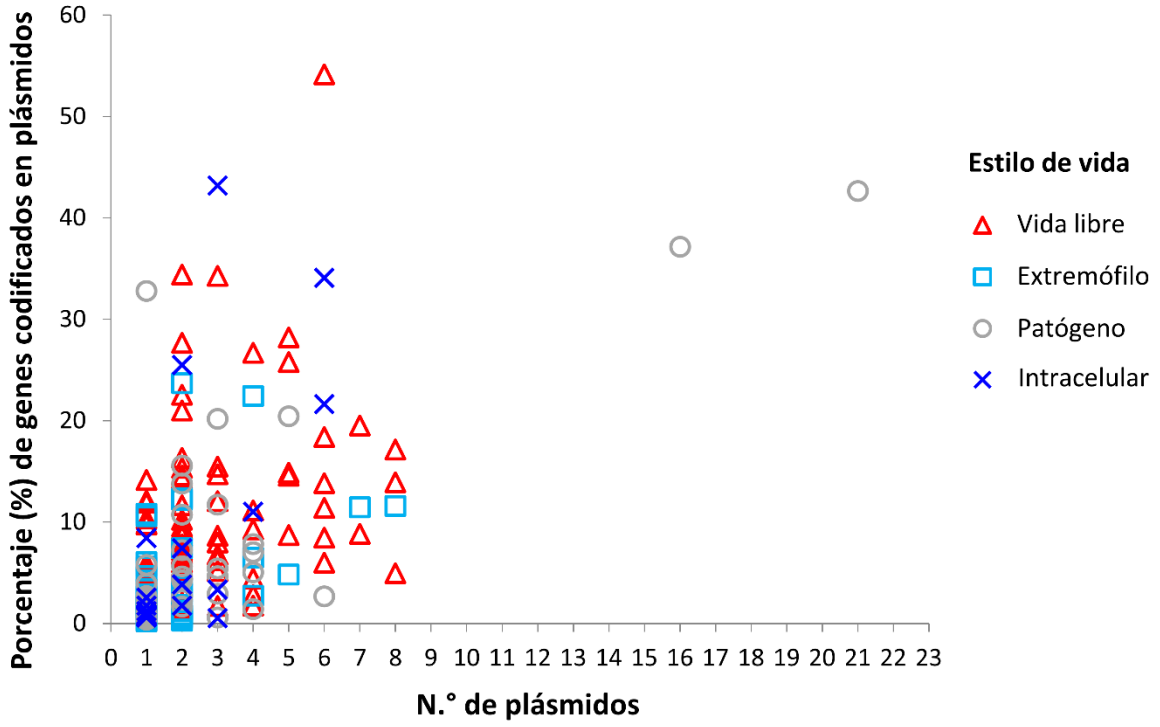


Figura 1. Relación entre el número de plásmidos y el porcentaje de genes que se encuentran en estos. A pesar de que el porcentaje de genes en plásmidos va de poco más de 0 hasta más de 50%, la mediana corresponde a menos del 5%.

Uno de los puntos fundamentales de este trabajo fue el establecimiento de los criterios a utilizar para que un par de secuencias se consideraran parálogas. Para ello nos basamos en diferentes estudios previos que cumplieran con las dos siguientes condiciones: i) que hicieran una búsqueda de secuencias duplicadas en genomas procariontes y ii) que dicha búsqueda fuera intragenómica. En la Tabla 3 se enlistan los estudios en cuestión. Después de contrastar dicha información, elegimos los parámetros siguientes: un valor *e* de 10^{-07}, un porcentaje de identidad de 20% o superior y una cobertura de la secuencia *query* en la que, por lo menos, el 75% de los residuos estuvieran alineados. En el caso del porcentaje de identidad, elegimos dicho valor porque, como se ha propuesto previamente (Rost 1997; 1999) por debajo de este valor es muy probable hallar un gran número de secuencias cuya identidad se deba al azar y no a homología. Para la búsqueda de homólogos, cada una de las secuencias de un genoma se consideró como un *query* y se buscó contra el resto de las secuencias de dicho genoma.

Tabla 3. Parámetros de búsqueda identificados en estudios previos en los que se llevó a cabo una búsqueda de parálogos intragenómica en procariontes.

Valor *e*	Porcentaje de identidad	Cobertura de la secuencia query	Referencia
10^{-05}	>30%	>150 residuos alineados	Gevers *et al*. 2004
10^{-05}	>15%	—	Tekaia & Dujon 1999
10^{-07}	>40%	>70 residuos alineados	Conant & Wagner 2002
10^{-05}	>75%	—	Bratlie *et al*. 2010

Se ha sugerido previamente que podría existir una relación entre la frecuencia de los eventos de duplicación de genes y el estilo de vida procarionte. Por ejemplo, la prevalencia de enzimas promiscuas duplicadas es relativamente alta en organismos de vida libre tanto mesófilos como extremófilos, los cuales poseen alrededor de tres veces más enzimas de este tipo que los organismos intracelulares (Martínez-Núñez *et al*. 2015; Martínez-Núñez & Pérez-Rueda 2016). Basándonos en ambos trabajos, clasificamos a los organismos de nuestra muestra en uno de cuatro estilos de vida diferentes: vida libre (que habiten en condiciones mesófilas), extremófilo, patógeno e intracelular (que incluye tanto endosimbiontes como patógenos intracelulares). De estos cuatro grupos de organismos, el único que podría resultar un tanto problemático es el de los patógenos. A diferencia de lo que ocurre en los otros tres grupos, existe cierta ambigüedad en lo que abarca el término “patógeno” debido a que, en

muchas ocasiones, estos organismos no se comportan de la misma manera en todos y cada uno de sus hospederos naturales. Por ejemplo, organismos como *Aspergillus fumigatus* que solo provoca enfermedad en personas inmunodeprimidas o *Staphylococcus aureus* que en una tercera parte de los hospederos no provoca daño alguno (Casadevall & Pirofski 2014). Por ello, es importante aclarar que en esta submuestra incluimos a todos aquellos organismos no intracelulares que se sabe que, al menos bajo ciertas condiciones, infectan a uno o más hospederos.

Una función de poder ($y = 0.004x^{0.9}$; $R^2 = 0.95$) es la que mejor describe la relación entre el número de proteínas y el tamaño de los genomas procariontes (Figura 2A). Esto nos dice que, a diferencia de los eucariontes, quienes poseen grandes cantidades de DNA, conforme aumenta el tamaño del genoma procarionte, también lo hará el número de proteínas de manera proporcional (Figura 2A). Indirectamente, esto confirma que los genomas de arqueas y bacterias poseen, en su mayoría, DNA codificante (Koonin & Wolf 2008). Pero esto no significa que el incremento en el número de enzimas siga la misma tendencia. De hecho, como se aprecia en la Figura 2B, ocurre algo bastante diferente. Si bien en los genomas más grandes aumenta el número de proteínas cuasilinealmente, en la mayoría de ellos el número de enzimas suele ser de alrededor de 1,000. Evidentemente, una posibilidad es que haya bastantes más enzimas que no han sido identificadas por métodos bioinformáticos. Pero lo que parece más probable es que sí exista un incremento significativo en el número de proteínas no enzimáticas. Por ejemplo, se ha encontrado que existe un incremento exponencial en el número de factores de transcripción a medida que aumenta el tamaño de los genomas procariontes (van Nimwegen 2003). Esto parece tener que ver con el hecho de que un pequeño incremento en el número de enzimas metabólicas parece requerir de un gran incremento en el número de proteínas reguladoras (Martínez-Núñez *et al*. 2013; Sanchez *et al*. 2020), probablemente para mantener en equilibrio los flujos metabólicos centrales.

La proporción de enzimas con por lo menos un parálogo también se ajusta a una función de poder ($y = 0.003x^{0.67}$; $R^2 = 0.68$) y solo en 7 organismos de nuestra muestra detectamos una proporción superior a 0.6. Como se muestra en la Figura 2C, la mayoría de los organismos posee entre un 20 y 40 % de enzimas duplicadas. Dicha proporción no parece depender del estilo de vida de los organismos, salvo en la mayoría de los casos en que esta es menor a 0.2, en los que observamos que esos organismos tienen un estilo de vida intracelular.

La relación entre el tamaño del genoma y el número de enzimas y proteínas varía considerablemente cuando separamos a los organismos de nuestra muestra de acuerdo con su estilo de vida. Solamente en los organismos de vida libre (Figura S1A-S1C) y en los patógenos (Figura S1G-S1I) la relación entre estas variables se ajusta a una función de poder. Por su parte, para los organismos extremófilos, la relación entre el número de enzimas y proteínas (Figura S1D), número de enzimas y tamaño de genoma (Figura S1E) y número de enzimas y tamaño de genoma (Figura S1F) es de tipo lineal. Finalmente, en los organismos intracelulares también hallamos que una regresión lineal es la que mejor explica la relación entre el número de enzimas y el número de proteínas (y=0.25x + 95; R^2=0.84; Figura S1J) y entre el número de enzimas y el tamaño del genoma (y=*1.54e*$^{-04}$$x$ + *180;* R^2=*0.71*; Figura S1K). En cambio, la relación entre el número de proteínas y el tamaño del genoma se ajusta mejor a una función de poder (y=0.02$x^{0.79}$; R^2=0.88; Figura S1L). Cabe aclarar que, en algunos casos, los coeficientes de determinación correspondientes a un modelo lineal y a una función de poder son muy similares, por lo que no siempre es posible identificar visualmente la tendencia. En la Tabla S1 se indican dichos coeficientes para cada una de las correlaciones en la Figura S1.

En general, consideramos tres escenarios posibles que podrían explicar que la proporción de enzimas parálogas no aumente linealmente con respecto al número total de enzimas en el genoma:

a) Sí existen más enzimas duplicadas pero, dados los filtros que consideramos para definir las secuencias parálogas, no nos es posible detectarlas todas.
b) No todos los genes poseen la misma probabilidad de ser retenidos en el genoma después de un evento de duplicación; esto podría tener que ver con qué tan esencial o no es cada gen (Woods *et al*. 2013).
c) Es probable que haya varios casos en los que sí se identifican parálogos pero estos no llevan a cabo función enzimática alguna a pesar de poseer uno o más dominios típicamente asociados a enzimas. A estos se les conoce comúnmente como *pseudoenzimas* (Jeffrey 2020).

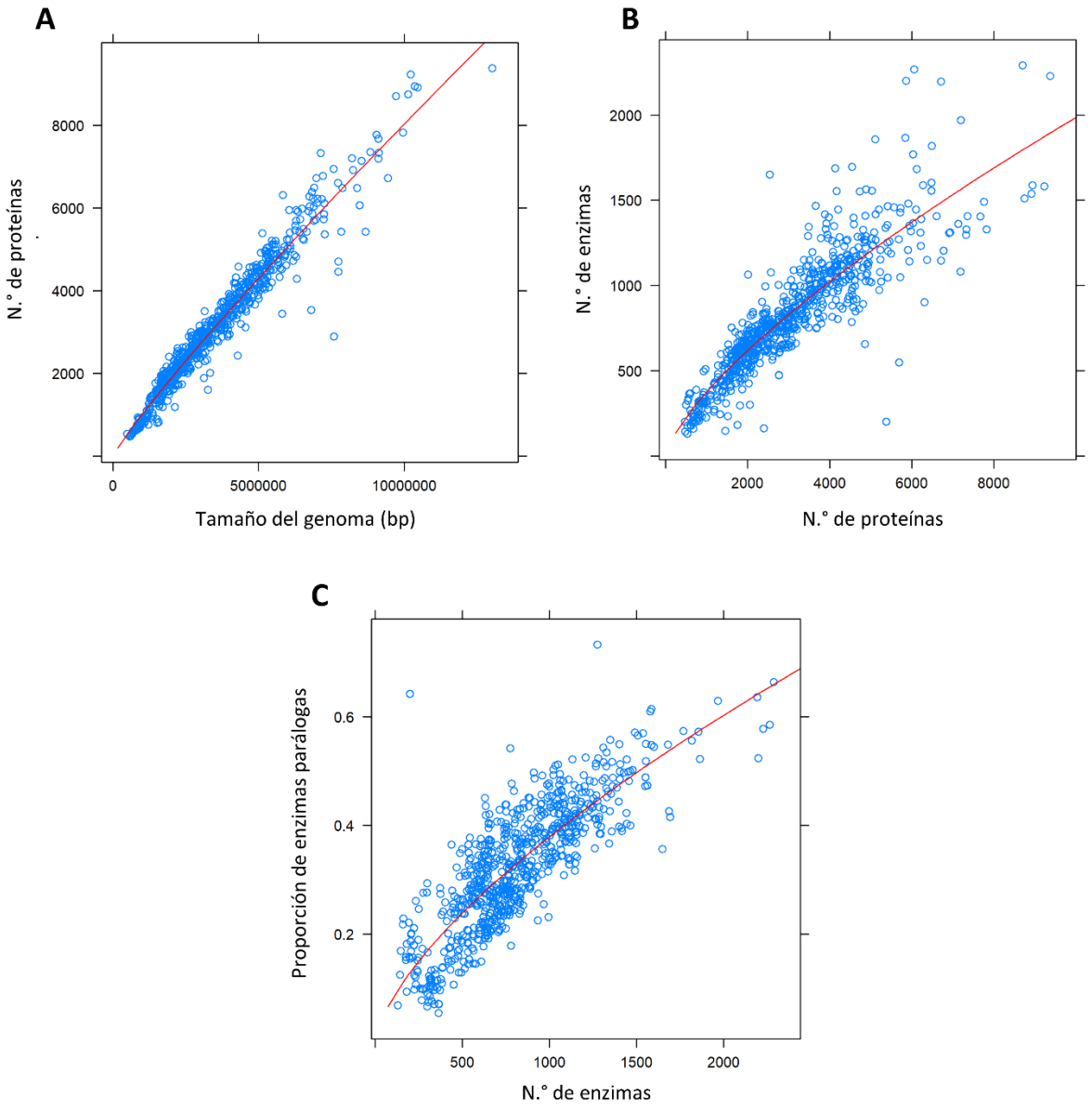


Figura 2. Relación entre distintos aspectos del genoma y proteoma procariontes. (A) Existe una relación casi lineal, aunque definida por una función de poder, entre el número de proteínas y el tamaño de los genomas procariontes ($y = 0.004x^{0.9}$; $R^2 = 0.95$). (B) Conforme crecen los genomas, a pesar de que el número de proteínas aumenta a una razón cercana a 1Mb : 1,000 proteínas, no ocurre un incremento similar en cuanto al número de enzimas ($y = 2.51x^{0.72}$; $R^2 = 0.79$). (C) Para la enorme mayoría de los proteomas analizados, la proporción de enzimas parálogas es menor que 0.6 ($y = 0.003x^{0.67}$; $R^2 = 0.68$) (Figura modificada a partir de Álvarez-Lugo & Becerra 2021).

Proporción de enzimas parálogas y su relación con el estilo de vida de los organismos

Al agrupar las enzimas de acuerdo con su clase enzimática (Tipton & Boyce 2000; McDonald & Tipton 2014) y compararlas entre sí, podemos observar que su distribución es muy similar a lo que se ha reportado en las bases de datos. Es decir, para aquellas clases que presentan una alta representación en cuanto al número de enzimas según lo que se ha reportado en la literatura desde hace varias décadas (oxidorreductasas, transferasas e hidrolasas) (ver, por ejemplo, McDonald *et al*. 2009), ocurre lo mismo a nivel proteoma, y algo similar sucede con las menos representadas (liasas, isomerasas, ligasas y translocasas) (Figura 3A). Este mismo patrón lo observamos también en la distribución de enzimas parálogas (Figura 3B). Sin embargo, esto no se observa al analizar la proporción de dichas enzimas. En este caso, las clases enzimáticas con una mayor proporción de duplicados son las translocasas, seguidas de las oxidorreductasas y las isomerasas. Además, la proporción de parálogos en estas tres clases es significativamente diferente a las de las otras cuatro. Esto se corroboró inicialmente con la prueba de Kruskall–Wallis ($P \leq 2.2e^{-16}$) y, posteriormente, por medio de la prueba de Dunn con un ajuste de Bonferroni.

El siguiente paso fue comparar la proporción de enzimas duplicadas pero tomando en cuenta el estilo de vida de los organismos de la muestra (vida libre, extremófilos, patógenos e intracelulares) (Figura 4). De manera general, hallamos que los organismos de vida libre son los que poseen una mayor proporción de parálogos pertenecientes a todas las clases enzimáticas. Solo para las oxidorreductasas (P=1; α=0.25) y las ligasas (P=1; α=0.25) presentan valores que no difieren significativamente de los organismos extremófilos (Tabla S2). Por su parte, los extremófilos y los patógenos presentan proporciones de enzimas parálogas que no difieren significativamente en tres clases: hidrolasas (P=1; α=0.25), isomerasas (P=0.074; α=0.25) y translocasas (P=0.645; α=0.25) (Tabla S2). Finalmente, el grupo de los intracelulares, el cual incluye tanto a endosimbiontes como a parásitos obligadamente intracelulares, posee la menor proporción de parálogos para todas las clases enzimáticas, la cual es significativamente diferente en todos los casos. En la Tabla S2 se presentan los valores P de la prueba de Dunn de las comparaciones pareadas entre proporciones de enzimas parálogas. Estas se hicieron por cada clase enzimática, comparando los distintos estilos de vida. Sin embargo, en la mayoría de estas la diferencia fue significativa (los únicos valores no significativos son los que se mencionaron previamente).

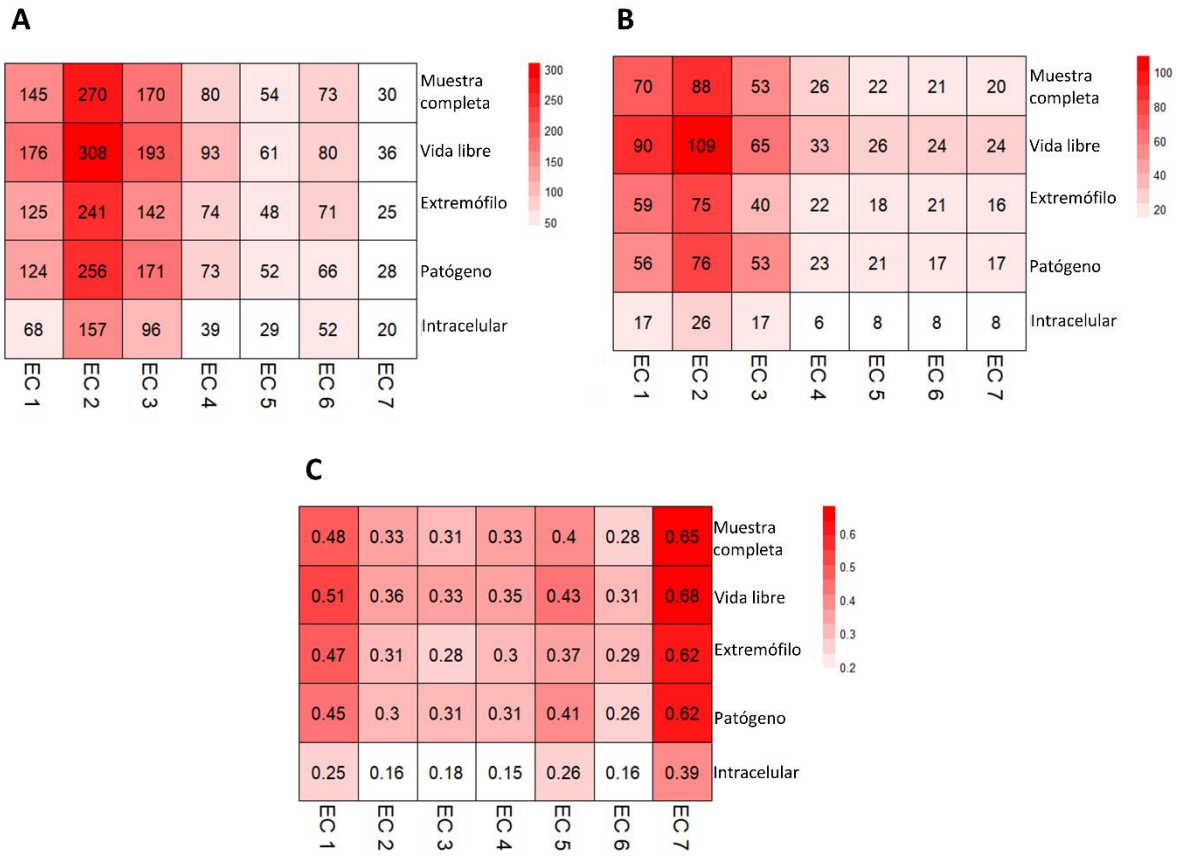


Figura 3. Número y proporción de enzimas parálogas en organismos procariontes. El valor que se encuentra al interior de cada celda corresponde al promedio de cierta clase enzimática en cada uno de los cuatro estilos de vida considerados y en la muestra total (fila superior en cada heatmap). (A) Número promedio de enzimas totales para cada clase enzimática. (B) Número promedio de enzimas parálogas para cada clase enzimática. (C) Proporción promedio de enzimas parálogas para cada clase enzimática.

Una de las preguntas centrales de este trabajo fue si existía una relación entre el contenido de enzimas parálogas y el estilo de vida y/o posición filogenética de los organismos. Para abordarla, realizamos un análisis de componentes principales (PCA) en el que consideramos una serie de aspectos que tienen que ver con características genómicas y enzimáticas (tamaño del genoma, número de proteínas, número de enzimas totales y parálogas y la proporción de enzimas parálogas para cada una de las clases enzimáticas). A partir de esto, pudimos identificar que distintos phyla de organismos con un estilo de vida similar, o que son filogenéticamente cercanos, aparecen muy cercanos uno de otro, lo cual indica similitud a nivel de las diferentes variables que consideramos.

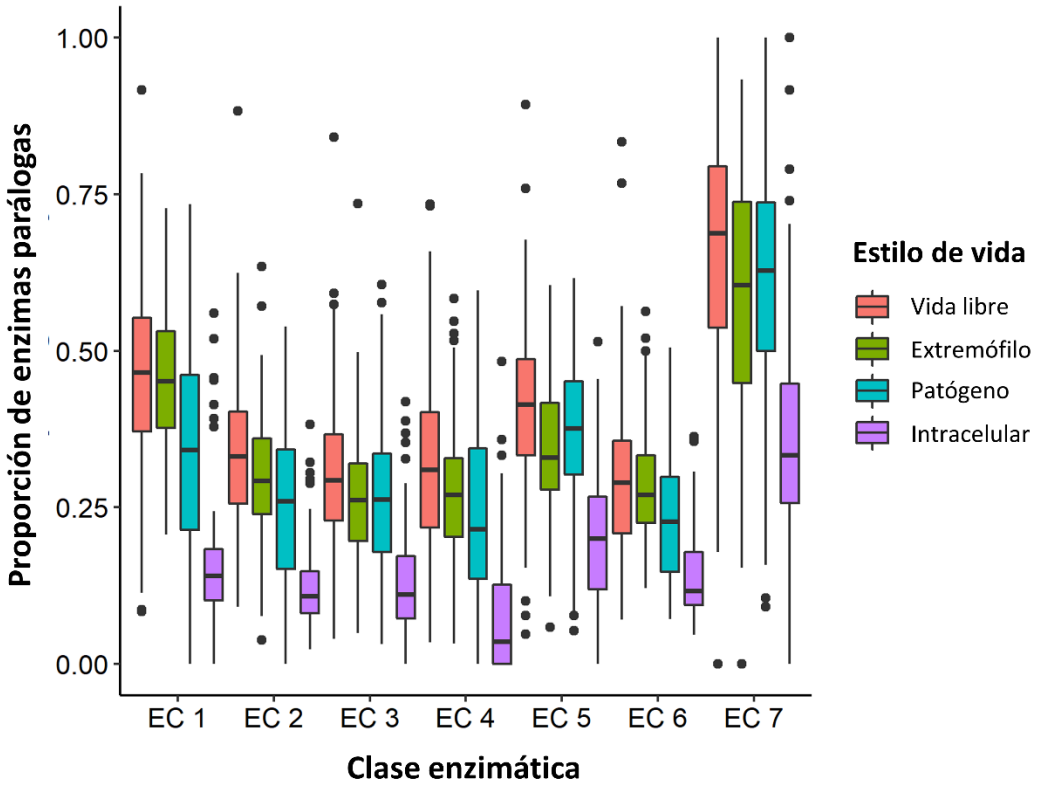


Figura 4. Comparación de la proporción de duplicados en cada clase enzimática, entre los cuatro estilos de vida diferentes que consideramos en el presente análisis. Las barras arriba y debajo de cada caja corresponden a la desviación estándar. Los puntos negros arriba y debajo de las cajas corresponden a los outliers (Figura modificada de Álvarez-Lugo & Becerra 2021).

Los phyla Deinococcus-Thermus, Chlorobi, Aquifecae, Thermotogae, Dictyoglomi y Euryarchaeota aparecen muy cercanos en la representación gráfica del PCA (elipse color verde), como puede observarse en la parte central de la Figura 5. En este caso, la mayoría de ellos son filogenéticamente distantes, salvo Thermotogae y Aquifecae, los cuales, además, son dos phyla considerados como tempranamente divergentes (Ciccarelli *et al*. 2006). Pero a nivel de estilo de vida, todos ellos poseen un gran número de extremófilos, particularmente termófilos (datos no mostrados). Esto sugiere que, en este tipo de organismos, existen estrategias similares para prosperar en ambientes con temperaturas altas.

Un caso adicional de agrupamiento de phyla en los que muchos de sus miembros comparten estilo de vida lo vemos en la parte inferior derecha de la Figura 5. Dentro de los phyla Tenericutes, Chlamydia y Elusimicrobia (elipse color anaranjado) hay un gran número de

especies intracelulares, las cuales comparten ciertas características como genomas reducidos, pérdida de genes, rutas metabólicas incompletas, etc. (Merhej *et al*. 2009). Además, se ha observado que la retención de proteínas específicas, como aquellas que tienen que ver con la ingesta de nutrientes (Saier & Paulsen 1999; Wandersman & Delepelaire 2004) es algo común en distintos grupos de este tipo de organismos. Por su parte, otros phyla que poseen un gran número de representantes intracelulares, como es el caso de las Alphproteoba (parte central de la Figura 5), quedan relativamente alejados del clúster Tenericutes-Chlamydia-Elusimicrobia. La explicación más probable podría tener que ver con el hecho de que, dentro de este grupo de proteobacterias, también consideramos a una gran cantidad de organismos con otros estilos de vida. Es probable que, de haber considerado exclusivamente a los organismos intracelulares de este grupo y otros con una situación similar, habrían quedado muy cercanos a otros que en su mayoría están conformados por intracelulares.

En la parte superior derecha de la Figura 5 (elipse color azul) se encuentran Crenarchaeota y Thaumarchaeota, además de otros dos phyla arqueanos que fueron descubiertos recientemente y para los que hay muy pocos miembros con su genoma completamente secuenciado: Korarchaeota y Bathyarchaeota. A pesar de que habitan en ambientes muy diferentes (Spang *et al*. 2017), todos ellos son filogenéticamente cercanos y pertenecen al mismo grupo, denominado TACK o Proteoarchaeota (Guy & Ettema 2011; Spang *et al*. 2017). Sin embargo, este resultado es poco informativo y debería considerarse preliminar dado el reducido número de genomas disponibles y el hecho de que algunos de ellos poseen un gran número de proteínas pobremente anotadas.

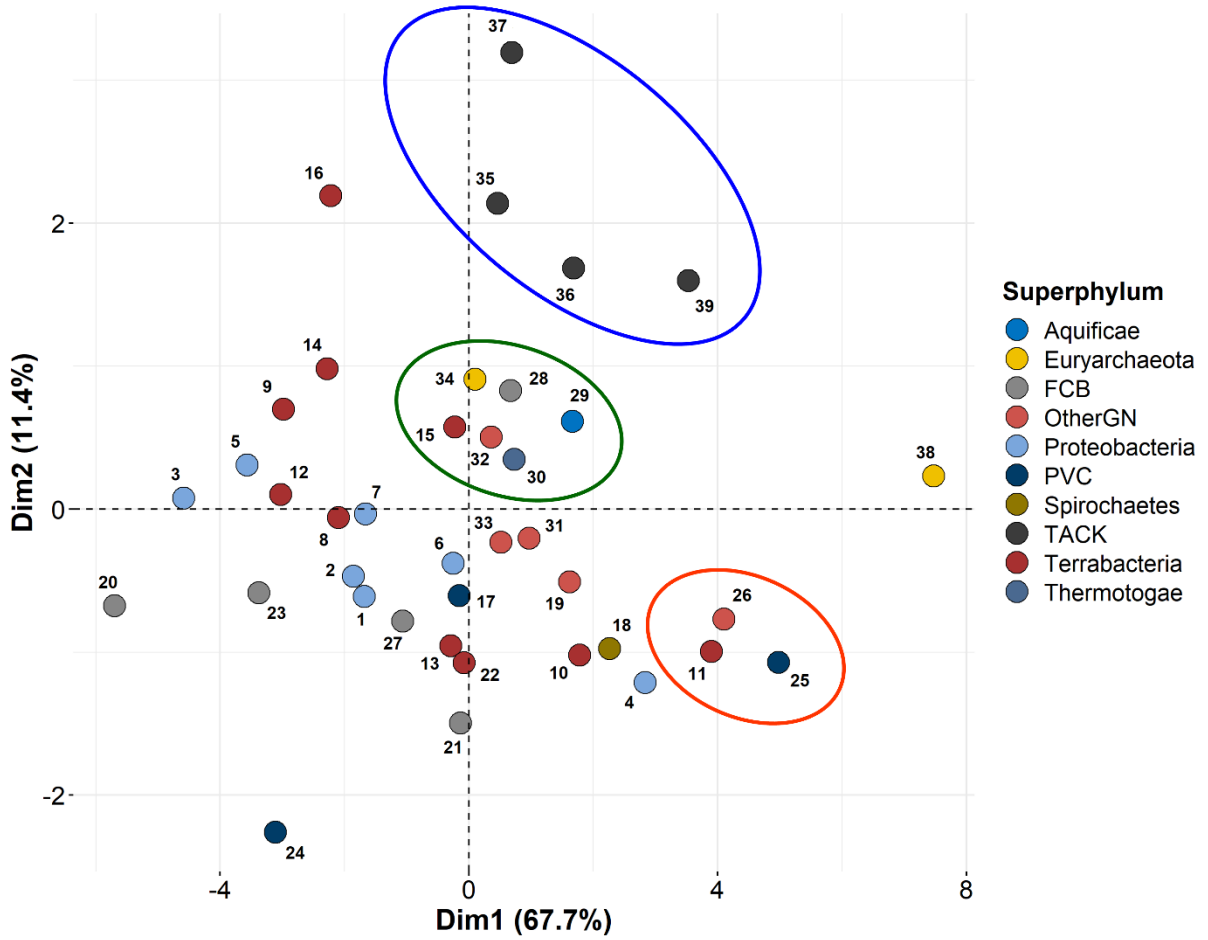


Figura 5. Análisis de componentes principales en el que se evalúa la similitud genómica y enzimática de los phyla procariontes. Por cada phylum, se consideró el valor promedio de 11 variables: tamaño del genoma, número de proteínas, número de enzimas, número de proteínas parálogas y la proporción de parálogos para cada clase enzimática. Cada círculo representa a un phylum y aquellos que pertenecen al mismo superphylum se indican con el mismo color. Las excepciones son Aquifecae, Thermotogae y Spirochaetes, los cuales no pertenecen a un superphylum definido. (1) Gammaproteobacteria-Enterobacteria, (2) Gammaproteobacteria-Others, (3) Betaproteobacteria, (4) Epsilonproteobacteria, (5) Deltaproteobacteria, (6) Alphaproteobacteria, (7) Other proteobacteria, (8) Firmicutes-Bacilli, (9) Firmicutes-Clostridia, (10) Firmicutes-Others, (11) Tenericutes, (12) Actinobacteria, (13) Cyanobacteria, (14) Chloroflexi, (15) Deinococcus-Thermus, (16) Unclassified Terrabacteria Group, (17)Verrucomicrobia, (18) Spirochaetes, (19) Synergistetes, (20) Acidobacteria, (21) Fibrobacteres, (22) Fusobacteria, (23) Gemmatimonadetes, (24) Planctomyces, (25) Chlamydia, (26) Elusimicrobia, (27) Bacteroidetes, (28) Chlorobi, (29) Aquificae, (30) Thermotogae, (31) Deferribacteres, (32) Dictyoglomi, (33) Nitrospirae, (34) Euryarchaeota, (35) Crenarchaeota, (36) Thaumarchaeota, (37) Korarchaeota, (38) Nanoarchaeota, (39) Bathyarchaeota. Las elipses engloban a cuatro phyla de arqueas del frupo TACK (color azul), seis phyla en donde predominan organismos termófilos (color verde) y tres phyla con un gran número de organismos intracelulares (color anaranjado). (Figura modificada de Álvarez-Lugo & Becerra 2021).

Finalmente está el cluster de las proteobacterias (parte izquierda del PCA). Para este análisis, dividimos a las γ-proteobacterias en los mismos dos grupos que aparecen en la base de datos KEGG (Kanehisa & Goto 2000): enterobacterias y otras; el resto de los phyla se consideraron sin subdivisiones. Como era de esperarse, las γ-proteobacterias aparecen juntas en el PCA (n.^{os} 1 y 2 en la Figura 5). Por el contrario, las β-proteobacteria, mismas que se consideran el grupo hermano de las γ-proteobacteria debido a su posición en distintas filogenias (Gupta 2000; Jun *et al*. 2010), aparecen bastante alejadas (n.^{o} 3 en la Figura 5). La ubicación de las ε-protebacteria en el PCA (n.^{o} 6 en la Figura 5; parte derecha) también refleja su posición filogenética con respecto a las otras proteobacterias. Este grupo se considera como el más filogenéticamente distante entre los distintos grupos de proteobacterias (Gupta 2000; Jun *et al*. 2010), y recientemente se ha propuesto que, dadas las grandes diferencias en ciertos aspectos genéticos y metabólicos (Waite *et al*. 2017), así como el hecho de que no siempre conforma un grupo monofilético con el resto de las proteobacterias (Hug *et al*. 2016; Rinke *et al*. 2013; Zhang & Sievert 2014) debería ser reclasificado como un grupo independiente.

Oxidorreductasas y promiscuidad enzimática

Las oxidorreductasas son enzimas que, como su nombre lo indica, llevan a cabo reacciones de oxidación-reducción, en las cuales una molécula donadora cede un par de electrones a una aceptora.

Esta clase enzimática es la segunda con un mayor número de enzimas descritas. Sin embargo, es aquella dentro de la cual existe la mayor diversidad en términos de funcionales; prueba de ello es que no hay otra clase enzimática con un número mayor de subclases y sub-subclases.

Las oxidorreductasas parálogas no se distribuyen de manera homogénea en las diferentes subclases sino que predominan en solo unas cuantas de ellas (Figura 6). Estas se distinguen por el grupo de donadores sobre los que actúan, los cuales son: grupo Ch-OH (EC 1.1), grupo aldehído u oxo (EC 1.2), grupo CH-CH (EC 1.3) y grupo sulfuro (EC 1.8). A pesar de que el Sistema de Clasificación Enzimática agrupe a las enzimas con base en similitud de reacciones y no bajo una óptica evolutiva, resulta interesante que, dentro de las cuatro

subclases previamente mencionadas, se encuentren muchas enzimas con el mismo plegamiento y que, por lo tanto, estarían relacionadas evolutivamente: el plegamiento Rossmann, mismo que podría haber estado presente desde etapas tempranas de la vida, incluso antes del Último Ancestro Común (Laurino *et al*. 2016). Además, prácticamente todas las enzimas con este plegamiento utilizan al NAD(P)H como cofactor (Fischer *et al*. 2010).

Un aspecto peculiar de las oxidorreductasas, particularmente de aquellas que adoptan el plegamiento Rossmann, es su habilidad de llevar a cabo reacciones adicionales (tanto similares como diferentes) a la reacción o actividad nativa. Cuando la reacción es básicamente la misma pero cambia la molécula que modifica la enzima, hablamos de promiscuidad a nivel de sustrato. Un ejemplo es la alcohol deshidrogenasa II (ADHII) presente en la bacteria *Zymomonas mobilis*, la cual puede catalizar la misma reacción sobre tres alcoholes diferentes (Kinoshita *et al*. 1985). Por otra parte, existen enzimas que pueden llevar a cabo dos o más reacciones químicas diferentes, ya sea sobre el mismo sustrato o, en ciertos casos, sobre sustratos diferentes, lo cual se conoce como promiscuidad catalítica. Ciertas alcohol deshidrogenasas, tanto de bacterias como de algunos eucariontes, no solo oxidan alcoholes a su respectivo aldehído sino que, además, pueden catalizar la oxidación de oximas (compuestos con una estructura química similar a la de los alcoholes pero con un átomo de nitrógeno unido al grupo hidroxilo) a su respectivo alcohol (Arora *et al*. 2014). Y finalmente, existen otras oxidorreductasas que presentan ambos tipos de promiscuidad enzimática, como es el caso de la alcohol deshidrogenasa de *Thermus* sp. ATN1, la cual puede oxidar alcoholes diferentes pero también sintetizar ácidos carboxílicos a partir de la dismutación de aldehídos (Höllrigl *et al*. 2008).

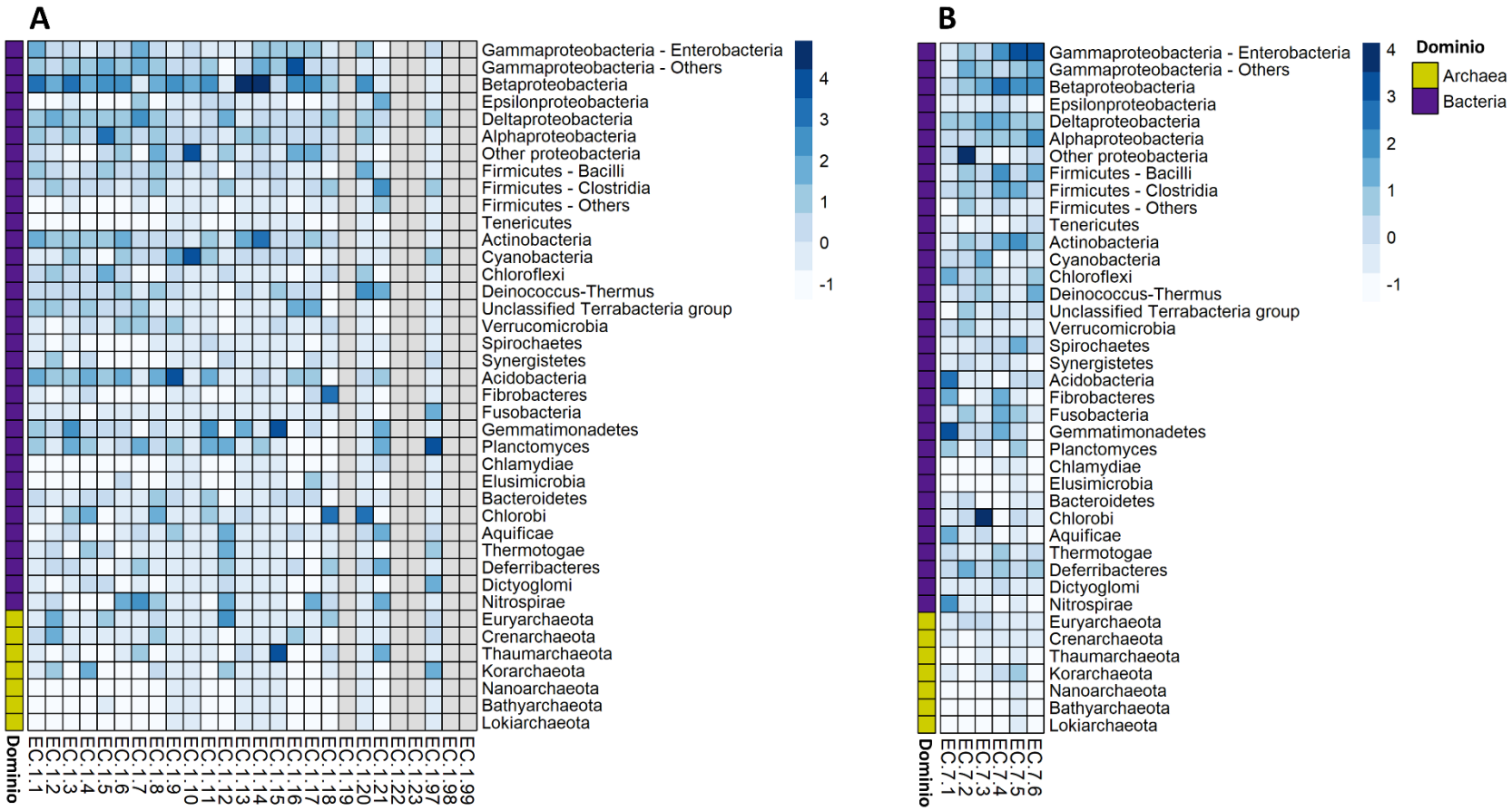


Figura 6. Número de enzimas parálogas que pertenecen a las subclases de (A) oxidorreductasas y (B) translocasas en genomas procariontes. Dado que los valores absolutos varían mucho entre las distintas subclases y entre los distintos phyla, cada uno de estos se normalizó por medio de la fórmula $z = (x - u) / s$, en la que x corresponde al número total de enzimas parálogas, u es el valor promedio para cada columna y s es la desviación estándar para esa misma columna. Las columnas cuyas celdas están en color gris corresponden a subclases para las que no identificamos ni una sola enzima. (Figura modificada de Álvarez-Lugo & Becerra 2021).

Los altos niveles de promiscuidad en muchas de las oxidorreductasas con plegamiento Rossmann podrían ser ventajosos para algunos organismos bajo ciertas condiciones. Pero es probable que este no sea el único factor que propicie la gran diversidad de funciones que existe en estas oxidorreductasas. La *evolucionabilidad*, definida como la capacidad de cambio a lo largo del tiempo, es otro aspecto muy importante. Esta se encuentra determinada por la capacidad de un plegamiento para acumular mutaciones que prácticamente no afecten la función y la estructura (robustez), así como por la habilidad para adquirir nuevas funciones enzimáticas en términos de reacción y/o sustrato (innovabilidad) (Tóth-Petróczy & Tawfik 2014). En conjunto, la promiscuidad enzimática, de la mano con una propensión a ser más evolucionables, podrían ser la razón por la que vemos una gran proporción de duplicados que presentan el plegamiento Rossmann. Análisis adicionales en los que se contraste el número de parálogos con el número de sustratos y reacciones diferentes en las enzimas con este plegamiento podrían ser útiles para extender nuestras conclusiones.

Un punto importante a tener en cuenta es que las subclases para las que hallamos un mayor número de parálogos se encuentran entre aquellas para las que se ha descrito un mayor número de enzimas. Debido a esto, existe la posibilidad de que el gran número de parálogos dentro de ellas se deba meramente a este factor, por lo que resulta un poco engañoso trabajar con los valores absolutos del número de parálogos. Para corroborar este punto, en un análisis subsecuente normalizamos dichos valores, lo cual se puede observar en la Figura 6A. De manera general, podemos considerar que nuestra interpretación previa se mantiene para las oxidorreductasas que actúan en los grupos Ch-OH (EC 1.1), aldehído u oxo (EC 1.2), CH-CH (EC 1.3) y sulfuro (EC 1.8) de los donadores de electrones, aunque otras como las que actúan sobre grupos CN-NH (EC 1.5), NAD/P (EC 1.6), otros compuestos nitrogenados (EC 1.7) e hidrógeno (EC 1.12) también presentan valores altos. Este análisis también nos permitió identificar que grupos como la mayoría de las proteobacterias (principalmente Betaproteobacteria), Actinobacteria, Acidobacteria y Planctomyces suelen tener una mayor proporción de oxidorreductasas parálogas en comparación de los otros phyla bacterianos.

Translocasas y su relación con el ATP

Las translocasas constituyen la clase enzimática más reciente; fue propuesta y definida hace poco más de cuatro años, después de más de 60 años en que el sistema de clasificación enzimática había permanecido sin cambios, en términos de las clases que lo conformaban.

La clasificación de estas enzimas resulta un tanto problemática debido a que, a diferencia de las otras clases, la reacción química que llevan a cabo no es el criterio de agrupación. De hecho, prácticamente todas las enzimas que integran este nuevo grupo (poco menos de 100), pertenecían originalmente a otra clase enzimática, principalmente a las hidrolasas. En este caso, la característica que comparten todas las enzimas que fueron reubicadas a esta nueva clase es su capacidad para transportar sustratos (desde iones hasta moléculas grandes) de un lado de la membrana al otro. Así, este se vuelve el factor principal, mientras que el tipo de reacción química necesaria para lograr el transporte pasa a segundo plano.

Tomando en cuenta lo anterior, no sorprende que un gran número parálogos dentro de esta clase pertenezcan a la superfamilia de los transportadores dependientes de ATP, también llamados transportadores ABC. Esta superfamilia es una de las más antiguas, presenta una distribución universal en procariontes y eucariontes y se sabe que muchos de sus miembros han pasado por varios eventos de duplicación (Higgins 2001; Saier & Paulsen 1999; Saurin *et al*. 1999). Que hayamos encontrado un gran número de transportadores ABC parálogos a lo largo de distintas subclases en la mayoría de los organismos, sin importar su estilo de vida, nos habla de la importancia que tienen estas proteínas tanto en la internalización de nutrientes como en la expulsión de sustancias tóxicas o de desecho.

Hallamos bastantes parálogos relacionados con la enzima ATP sintasa y movimiento de protones a través de la membrana (la mayoría de los cuales pertenecen a la subclase EC 7.1), la cual es crucial en el metabolismo energético. Esta enzima (o mejor dicho, complejo o máquina molecular), al igual que los transportadores ABC, también está ampliamente distribuida en procariontes. Además, distintos eventos de duplicación génica parecen formar parte de su historia evolutiva, y es común hallar organismos que presenten más de una copia (Cross & Taiz 1990; Klenk *et al*. 1997; Ruppert *et al*. 2001). Sin embargo, al ser un complejo molecular conformado por varias subunidades, es muy probable que las secuencias parálogas

anotadas con el mismo número enzimático no sean copias de la enzima completa sino que correspondan a varias de las subunidades que la conforman. Esto mismo podría ocurrir con otras máquinas moleculares para las que se halle una gran cantidad de parálogos.

Al igual que con las oxidorreductasas, normalizamos los valores absolutos del número promedio de translocasas por phylum para evitar sesgos relacionados con el número total de enzimas por cada subclase (Figura 6B). Es interesante observar que la subclase relacionada con el movimiento de protones (EC 7.1) ya no es la única que posee un número alto de duplicados, salvo por algunos phyla. Adicionalmente, las translocasas involucradas en el movimiento de aminoácidos y péptidos (EC 7.4), subclase a la que pertenece un gran número de transportadores ABC, también tienen un gran número de parálogos. En cambio, la subclase que engloba a las translocasas que catalizan el transporte de iones inorgánicos (EC 7.2), que a su vez es la que posee el mayor número de enzimas descritas, presenta un número de parálogos similar al de otras subclases. Esto podría deberse a que en esta categoría hay relativamente pocos transportadores ABC que, como se ha mencionado previamente, constituyen una de las superfamilias con más duplicados, por lo que casi no aportarían al número de parálogos de este grupo de enzimas. Y en términos de grupos de organismos, es en las Gamma, Beta, Delta y Alphaproteobacteria en donde observamos, de manera general, el mayor número de translocasas duplicadas.

El destino de las enzimas parálogas en procariontes: las isomerasas como una clase modelo

El enfoque que se ha presentado hasta el momento presenta ciertas limitaciones. A pesar de que nos revela un panorama general acerca de grupos específicos de enzimas que podrían tener bastantes parálogos entre sus representantes, nada nos dice acerca de los casos puntuales. Tomando en cuenta esto, decidimos expandir el alcance de nuestra metodología hasta llegar al nivel de enzimas individuales en una de las clases enzimáticas para la que identificamos una alta proporción de parálogos: las isomerasas.

Evidentemente, es necesario desglosar poco a poco los diferentes niveles antes de llegar al de enzimas individuales. Así como ocurre con las oxidorreductasas y las transferasas, la mayoría de las isomerasas parálogas se distribuyen en unas cuantas subclases,

específicamente las transferasas intramoleculares (EC 5.4) y las racemasas y epimerasas (EC 5.1) (Álvarez-Lugo & Becerra 2023).

Ahora bien, si seguimos al nivel inferior del número enzimático, el de la sub-subclase, observamos algo similar a lo que ocurre con las subclases: la distribución de enzimas duplicadas no es homogénea sino que se restringe a unas pocas categorías (Álvarez-Lugo & Becerra 2023). En este caso, la que más resalta es la de las transferasas intramoleculares que transfieren “otros” grupos.

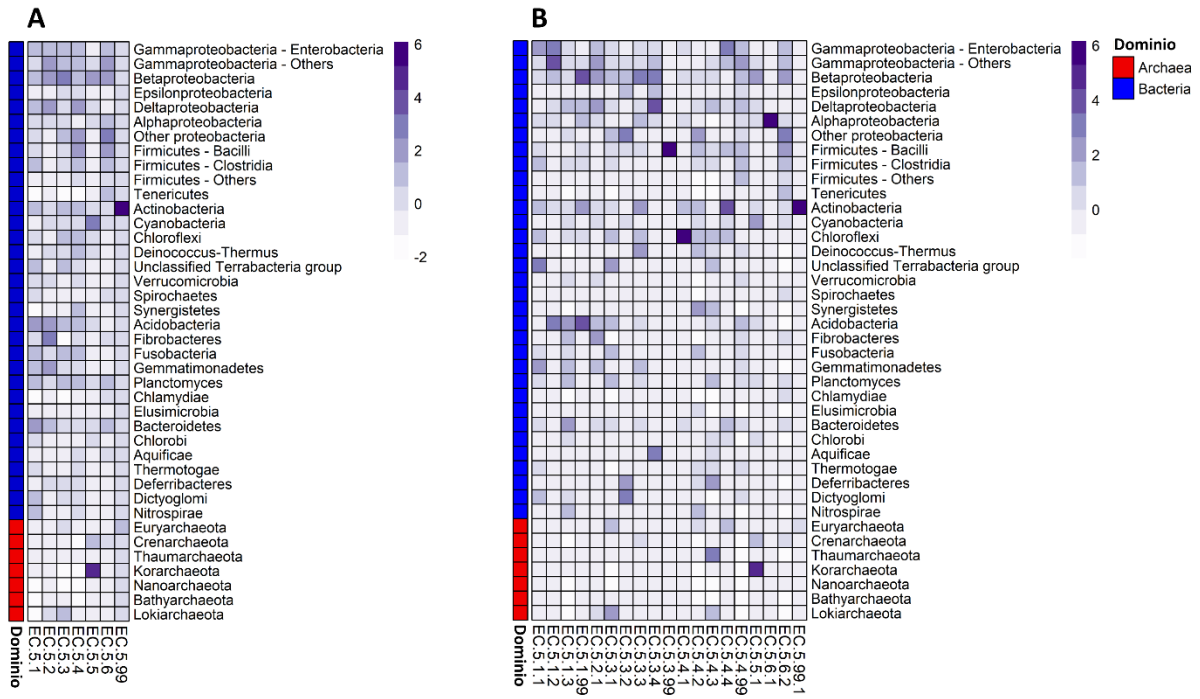


Figura 7. Número de enzimas parálogas en las diferentes (A) subclases y (B) sub-subclases de isomerasas. El valor absoluto correspondiente a cada se normalizó por medio de la fórmula $z = (x - u) / s$, en la que x corresponde al número total de enzimas parálogas, u es el valor promedio para cada columna y s es la desviación estándar para esa misma columna.

Lo anterior podría sugerir que el gran número de enzimas duplicadas que hallamos en algunas subcategorías se debe a que en el proteoma encontramos, a su vez, muchas enzimas descritas para dicha subcategoría (mientras mayor sea el número de enzimas totales, más parálogos). Sin embargo, a pesar de que existe una correlación positiva entre el número total de enzimas para cada subcategoría y el número promedio de parálogos, esta no es muy fuerte (r = 0.67), y disminuye aún más al considerar por separado bacterias (r = 0.63) y arqueas (r =

0.44) (Tabla S3). Adicionalmente, al normalizar el número de enzimas parálogas en las distintas subclases (Figura 7A) y sub-subclases de isomerasas (Figura 7B) obtenemos un panorama muy diferente al que resulta de considerar únicamente el número total de parálogos en cada una de ellas. De esta forma, ya no se observa una o más subclases y sub-subclases dominantes sino que, de hecho, en varias de ellas vemos una distribución similar en cuanto al número de duplicados. En algunos grupos de organismos encontramos un número alto de parálogos de ciertas subcategorías mientras que, en la mayoría, el número es bajo. Y al analizarlo en términos phyla vemos que, de manera general, son las Gamma, Beta y Deltaproteobacteria las que poseen el mayor número de isomerasas duplicadas, así como las Actinobacteria y Chloroflexi, aunque en menor medida.

Antes de continuar, me parece necesario aclarar un punto en particular, a fin de evitar una futura confusión. De las siete subclases de isomerasas, hay tres que destacan debido a que su nombre corresponde al de tres clases enzimáticas: oxidorreductasas, transferasas y liasas intramoleculares (EC 5.3, EC 5.4 y EC 5.5, respectivamente). Como su nombre lo sugiere, estos grupos de enzimas llevan a cabo reacciones que son químicamente idénticas a las que catalizan sus contrapartes de otras clases enzimáticas, con la salvedad de que todos los pasos ocurren intramolecularmente. En algunos casos, ambas enzimas pueden ser homólogas y requerir del mismo cofactor; la diferencia es que en la isomerasa no habrá consumo neto de este (ver, por ejemplo, Knoch *et al*. 2018).

Ahora bien, lo que se observa en la Figura 7 resulta poco informativo en términos de distribución en el metabolismo. Debido a ello, en un análisis posterior consideramos a todas las isomerasas con por lo menos un parálogo que logramos identificar al interior de las clases enzimáticas con una mayor representación absoluta en bacterias (EC 5.4.99 (Transferasas intramoleculares que transfieren otros grupos) y EC 5.1.3 (Racemasas y epimerasas que actúan sobre carbohidratos y derivados)) y arqueas (EC 5.1.3 (Racemasas y epimerasas que actúan sobre carbohidratos y derivados), EC 5.3.1 (Oxidorreductasas intramoleculares que interconvierten aldosas, cetosas y compuestos afines) y EC 5.4.2 (Fosfomutasas)). Al analizar la distribución de estos parálogos en el metabolismo, identificamos que la mayoría de ellos están involucrados en el metabolismo de carbohidratos, modificación de ácidos nucleicos y modificación de aminoácidos específicos (Figura 8). Esto coincide, en gran parte, con la abundancia de isomerasas en los distintos procesos metabólicos (Álvarez-Lugo & Becerra 2023; Martínez Cuesta *et al*. 2014). Una diferencia importante es que, a pesar de que la

categoría “metabolismo de terpenos y policétidos” es la segunda en la que están involucradas un mayor número de isomerasas, a nivel de parálogos casi no identificamos rutas metabólicas específicas comprendidas dentro de dicha categoría.

A partir del análisis anterior pudimos identificar que la mayoría de las isomerasas parálogas en el dominio Bacteria tienen que ver con modificaciones al 23S rRNA (parte inferior de la Figura 8). Estas enzimas, conocidas como rRNA pseudouridina sintasas, se encuentran presentes en los tres dominios de la vida (Spenkuch *et al*. 2014). El hecho de que conviertan residuos de uridina en pseudouridina por medio del mismo mecanismo (Ge & Yu 2013; Hamma & Ferré-D’Amaré 2006), así como su similitud a nivel de secuencia (Xie *et al*. 2022) y estructural (Mueller 2002) indica que todas ellas poseen un origen común y que pudieron haber evolucionado a partir de divergencia funcional después de eventos de duplicación génica.

De las siete 23S rRNA pseudouridina sintasas que identificamos, destacan dos en particular: RluD (EC 5.4.99.23) y RluC (EC 5.4.99.24). La primera modifica las uridinas que se encuentran en las posiciones 1911, 1915 y 1917 (Leppik *et al*. 2007) mientras que la segunda actúa sobre las posiciones 955, 2504 y 2580 (Conrad *et al*. 1998). A pesar de que la modificación de uridina a pseudouridina se considera un paso fundamental para la maduración del ribosoma, la inactivación de la enzima RluC no parece tener ningún efecto *in vivo* (Conrad *et al*. 1998; Huang *et al*. 1998). En cambio, si se inactiva la enzima RluD, se observa un efecto deletéreo que se refleja en una drástica reducción en la tasa de crecimiento (Huang *et al*. 1998). Análisis más específicos revelan que la falta de isomerización en U1917 es la responsable de los efectos deletéreos (Liiv *et al*. 2005). En cambio, que si la posición 1911 y/o 1915 no son isomerizadas, la viabilidad celular no se afecta.

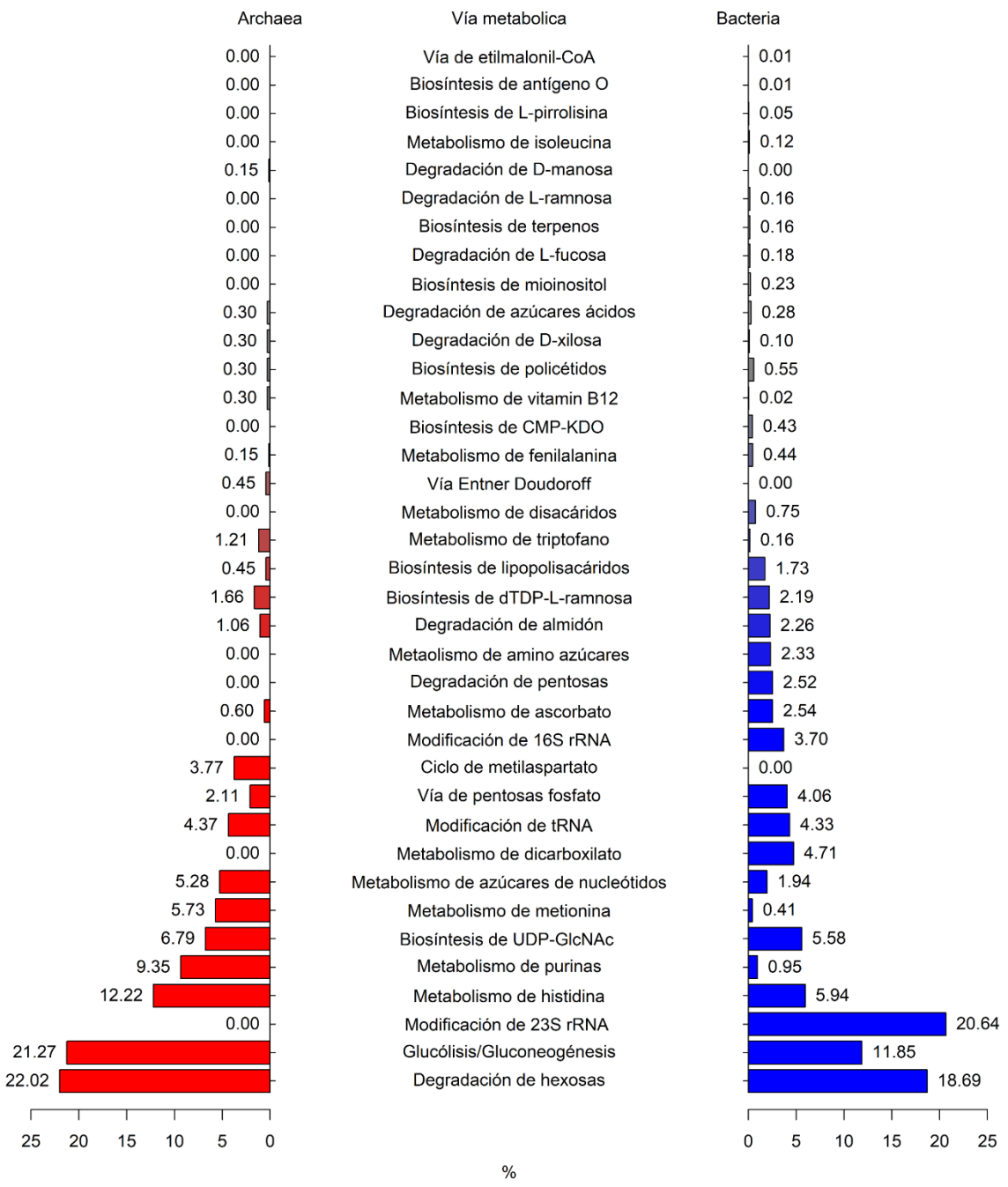


Figura 8. Distribución metabólica de las isomerasas con por lo menos un parálogo. En cada grupo de organismos, la suma de los diferentes valores equivale al 100% para ese dominio en particular (Figura modificada de Álvarez-Lugo & Becerra 2023).

En nuestro análisis pudimos identificar que en muchos organismos las enzimas RluD y RluC se encuentran mutuamente al momento de hacer la búsqueda de parálogos, y solo en muy pocos casos aparece como 'hit' alguna otra pseudouridina sintasa. Esto sugiere que RluD y RluC podrían representar un caso de parálogos que divergieron hace poco, probablemente a partir de un ancestro más generalista en términos del número de residuos en posiciones diferentes que pudiera modificar. Adicionalmente, identificamos bastantes organismos que presentaban dos o más copias de la enzima RluD con la misma función asignada, pero solo unos cuantos con más de una copia de RluC y nunca con más de dos.

Los análisis filogenéticos revelan que RluD se encuentra distribuida no solo en un mayor número de organismos que RluC sino también en un mayor número de phyla (Figura 9). Esto es un punto a favor del grado de importancia de cada una de estas enzimas. En la Figura 9 se presentan las filogenias de ambas enzimas y, además, se indica el número de copias con la misma función asignada (barras azules en la parte externa de cada filogenia). Como puede observarse, no en todos los casos se hallan parálogos con estas características. Por ejemplo, en el caso de las Gammaproteobacteria no identificamos a un solo organismo con más de una enzima anotada como RluD, pero sí con un parálogo cuya anotación correspondía a RluC.

Esta enorme disparidad en cuanto al número de parálogos que identificamos para cada enzima nos hizo suponer inicialmente que, probablemente, se trataba de un artefacto relacionado con problemas en la anotación de las secuencias, la cual se hace automáticamente en la mayoría de los casos. Un problema adicional fue que no se hizo una nueva búsqueda de ortólogos cuyo resultado sirviera como input para la construcción de las filogenias, sino que nos basamos en la presencia de determinada enzima en los proteomas que conformaban la muestra.

Para tratar de subsanar los puntos anteriores, decidimos hacer un análisis basado en redes de similitud de secuencias en el cual incluimos a las siete rRNA pseudouridina sintasas que identificamos en nuestra muestra de proteomas (seis modifican 23S rRNA y una 16S rRNA) (Figura 10). En total se consideraron 1944 secuencias, mismas que se utilizaron como input en el software Enzyme Similarity Tool (Zallot *et al*. 2019) para construir la red de similitud de secuencias (SSN). Este archivo de salida se le dio al programa Cytoscape (Shannon *et al*. 2003) para poder visualizar las relaciones entre los distintos grupos de secuencias.

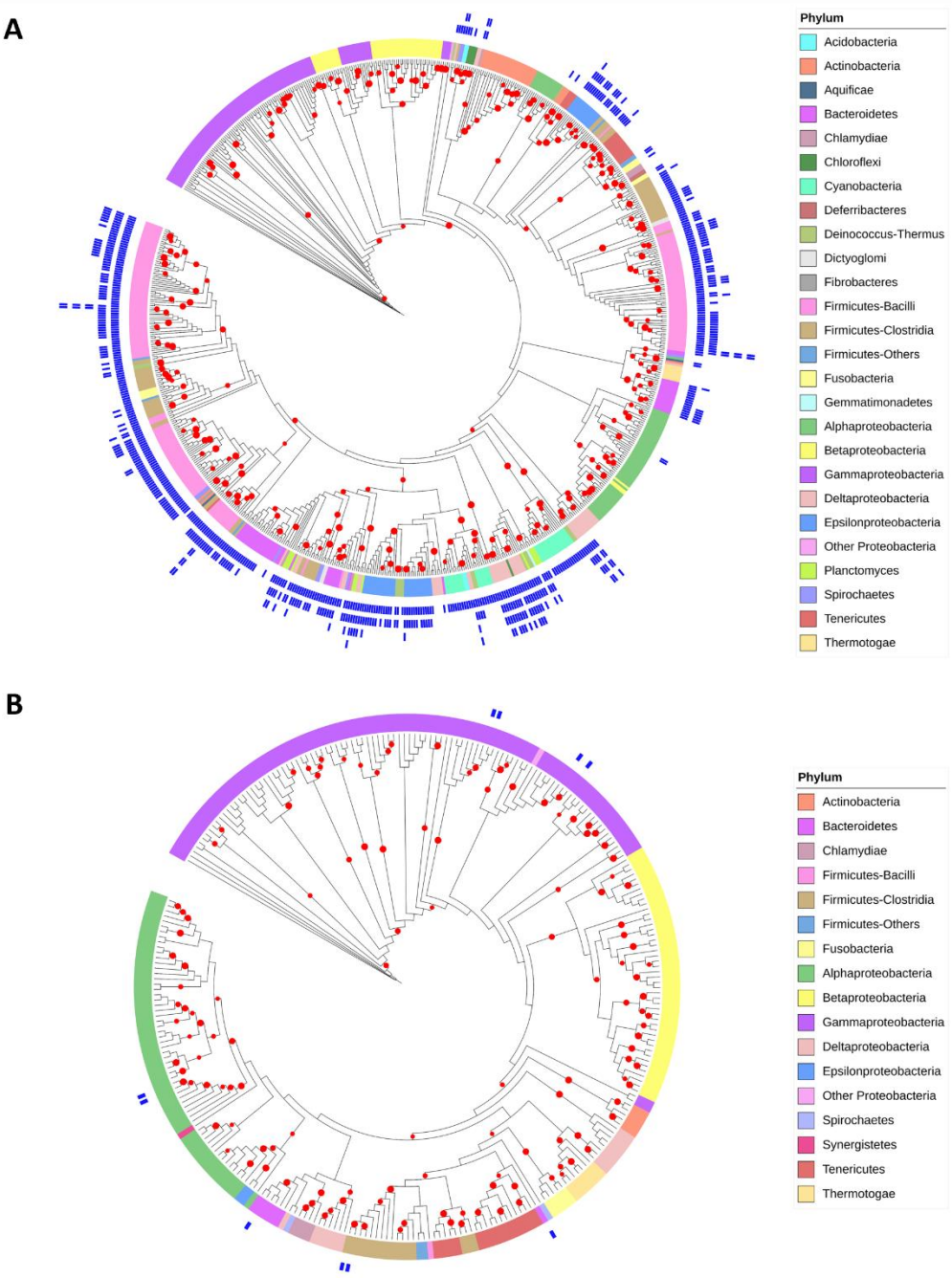


Figura 9. Análisis de la distribución filogenética de las rRNA pseudouridina sintasas bacterianas RluD (A) y RluC (B). Para cada una de las filogenias consideramos aquellas secuencias que tuvieran por lo menos un parálogo en el genoma del que provienen, sin importar que tuviera la misma u otra función. Las barras azules externas indican el número de parálogos que poseen la misma función asignada. Por el contrario, la ausencia de estas barras indica que los parálogos de la secuencia en cuestión llevan a cabo una función diferente. Los puntos rojos en las diferentes ramas representan valores de Bootstrap de 90 o superiores y su tamaño es proporcional a dicho valor (Figura modificada de Álvarez-Lugo & Becerra 2023).

En la Figura 10A se observa un gran número de cúmulos en los que cada nodo corresponde a una secuencia. Esta SSN fue construida por medio del algoritmo de *transitividad*. Es claramente notoria la homogeneidad en cuanto al color de cada cúmulo: la gran mayoría de ellos están conformados por secuencias que están anotadas de la misma manera. De cierta forma, esto nos ayuda a descartar que estuviéramos mezclando secuencias de dos o más enzimas diferentes en nuestro análisis filogenético. Sin embargo, surge una pregunta adicional. ¿Qué relación existe entre los diferentes cúmulos del mismo color? Esto es algo que el algoritmo de transitividad no nos permite identificar, y fue la razón por la cual hicimos un segundo análisis pero esta vez utilizamos el algoritmo de *vecindad* (Figura 10B). De esta forma, pudimos comprobar que los distintos cúmulos pertenecientes a la misma enzima que se observan en la Figura 10A se agrupan en un solo cúmulo principal en la Figura 10B. Existen algunas excepciones que quedan fuera de los dos grandes grupos de nodos (parte derecha de la figura), pero esto no necesariamente querría decir que esas secuencias están mal anotadas sino que, probablemente, hayan divergido bastante. Además, los dos grandes cúmulos corresponden a dos de las familias de pseudouridina sintasas, según se ha reportado en otros lados (Tabla 4).

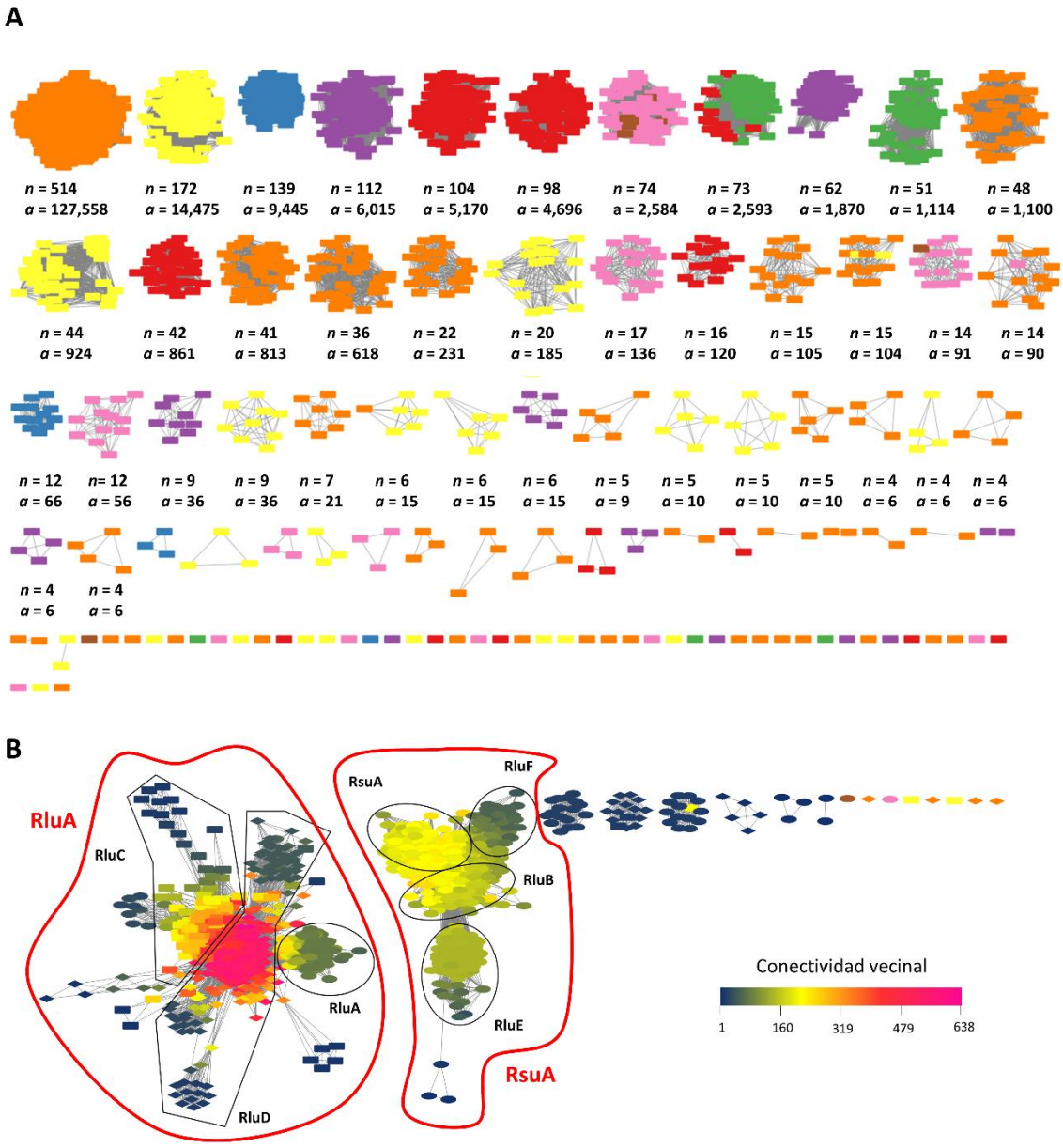


Figura 10. Redes de similitud de secuencias (SSN) para las pseudouridina sintasas presentes en nuestra muestra. (A) Red construida mediante el algoritmo de *transitividad*. La enorme mayoría de los clusters están conformados por secuencias que corresponden a la misma pseudouridina sintasa, lo cual está indicado por la homogeneidad a nivel de colores. Debajo de cada cluster de cuatro o más secuencias se indica el número de nodos (*n*) y el número de aristas (*a*) que lo componen. (B) En este caso, la red se construyó por medio del algoritmo de *vecindad*. La tonalidad de cada nodo depende del número de conexiones con otros nodos. Casi todos los nodos de la misma enzima que estaban dispersos en distintos clusters en (A) quedan unidos entre sí en (B), y estos a su vez están bien separados de los clusters de enzimas diferentes. Los nodos representados por diamantes y rectángulos corresponden a las enzimas RluD y RluC respectivamente (Figura modificada de Álvarez-Lugo & Becerra 2023).

Tabla 4. Clasificación de las diferentes familias de RNA pseuduridina sintasa. En itálicas y negritas se indican las enzimas de esta familia que consideramos para nuestro análisis.

Familia	Enzimas	Sustrato
Pseudouridina sintasa I	TruA	tRNA
Pseudouridina sintasa II	TruB	tRNA
Pseudouridina sintasa RsuA	***Rsua***, ***RluB***, ***RluE***, ***RluF***	rRNA
Pseudouridina sintasa RluA	***RluA***, TruC, ***RluC***, ***RluD***	rRNA (RluC y RluD)
		tRNA (solo TruC)
		tRNA y rRNA (solo RluA)
Pseudouridina sintasa TruD	TruD	tRNA

El posible papel de la duplicación de genes en el origen y evolución del metabolismo aerobio

Como hemos visto hasta el momento, la retención de uno o más genes duplicados es un mecanismo que puede favorecer la adaptación de los organismos a ambientes con condiciones cambiantes (Bratlie *et al*. 2010; Gevers *et al*. 2004; Kondrashov, 2012). Tomando esto en cuenta, decidimos ahondar un poco más acerca del posible papel que pudo haber jugado la duplicación génica en respuesta a uno de los cambios ambientales más drásticos de los que se tiene registro: la transición de una atmósfera anoxigénica a una oxidante.

La rápida oxigenación de la atmósfera primitiva, conocida como “el Gran Evento Oxidativo”, fue un proceso que tuvo lugar hace aproximadamente 2.5 Ga, aunque parece ser que comenzó lentamente unos 500 ma antes, durante el eón Arqueano, y culminó hace casi 2 Ga en el eón Proterozoico (Canfield 2005; Holland 2006; Lyons *et al*. 2014). Este importantísimo cambio ambiental se debió, en gran parte, a la liberación de enormes cantidades de oxígeno molecular a la atmósfera como resultado de la fotosíntesis oxigénica, la cual tuvo su origen en el linaje de las cianobacterias (Schirrmeister *et al*. 2013; 2015).

La evidencia sugiere que la fotosíntesis oxigénica en cianobacterias evolucionó a partir de componentes más antiguos que muy posiblemente estaban involucrados en un tipo de fotosíntesis anoxigénica basada en ácido sulfhídrico (H_2S) (Martin *et al*. 2018; Hamilton 2019). Eventualmente se agotaría el azufre como fuente de electrones, lo cual pudo haber

representado una presión selectiva que llevara al uso del agua como un sustituto para la obtención de electrones (Saito 2012). Análisis filogenéticos sugieren que algunos de los componentes fotosintéticos parecen haber evolucionado a partir de más de un evento de duplicación génica (Cardona 2015), antes de que las moléculas de agua se usaran como fuente de electrones (Ben-Shem *et al*. 2004).

David y Alm (2011) identificaron que, entre hace 3.3 y 2.8 Ga, durante el eón Arqueano, hubo un breve periodo de rápida innovación a nivel metabólico durante el cual se originaron poco más del 25% de las familias de genes actuales, lo cual también coincidió con una gran diversificación de linajes bacterianos. Y aunque la duplicación génica y la transferencia horizontal de genes no parecen haber sido tan relevantes en esta pequeña ventana de tiempo, la tasa de ambos eventos, en especial la duplicación, se fue incrementando poco a poco. Ahora bien, a partir de este trabajo podemos identificar dos puntos relevantes: 1) ese periodo de rápida innovación genética y metabólica concluyó entre 200 y 300 Ma antes del GEO y 2) fue posteriormente a este cuando la tasa de duplicación génica comenzó a aumentar considerablemente. El hecho de que muchos de los genes que surgieron después de la expansión durante el Arqueano requieran de oxígeno molecular (David & Alm 2011) sugiere que la duplicación génica pudo haber jugado un papel fundamental en la diversificación metabólica y en la adaptación de los organismos a condiciones cada vez más oxidantes.

El uso del oxígeno como una herramienta de datación relativa

Utilizar al oxígeno como un marcador molecular nos permite datar relativamente ciertas enzimas y procesos metabólicos. De manera general, podemos considerar que aquellas enzimas y rutas que dependen del oxígeno debieron surgir una vez que la concentración de este gas en la atmósfera alcanzó niveles bastante altos, hace poco más de 2 Ga. Aun así, existen ciertas enzimas oxígeno-dependientes cuyo origen parece remontarse a por lo menos 400 Ma antes del GEO, cuando los niveles de oxígeno atmosférico comenzaban a aumentar muy lentamente (Wang *et al*. 2011).

La oxigenación de la atmósfera dio pie a una enorme expansión y diversificación de las capacidades metabólicas de los seres vivos (Jiang *et al*. 2012; Raymond & Segrè 2006). No solo permitió la aparición de reacciones que serían termodinámicamente desfavorables en

ausencia de oxígeno molecular (Jabłońska & Tawfik 2022) sino que, además, dio lugar a que muchas enzimas O_2-independientes fueran reemplazadas por una versión O_2-dependiente, mismas que poseen la ventaja de ser más eficientes y llevar a cabo reacciones irreversibles (Raymond & Blankenship 2006).

Recientemente, Jabłońska & Tawfik (2022) identificaron tres posibles orígenes para las enzimas O_2-dependientes: 1) a partir de una familia de oxidorreductasas O_2-independientes; 2) a partir de enzimas pertenecientes a otra clase enzimática y 3) a partir de precursores no enzimáticos, como ciertas proteínas que se unen a un ligando específico. Para ello consideraron 54 pares de reacciones análogas (es decir, versiones aerobias y anaerobias de la misma reacción), basándose en la clasificación de la base de datos Pfam. El resultado más revelador de este trabajo es que, de 136 familias de proteínas que contienen por lo menos una enzima O_2-dependiente, en 81 de ellas (60%) el ancestro parece haber sido una enzima de la misma naturaleza, mientras que las 55 restantes (40%) podrían haber divergido a partir de enzimas que no tenían que ver con el oxígeno, las cuales probablemente precedieron al GEO.

A partir del punto anterior surge la siguiente pregunta: ¿cómo se originaron las enzimas O_2-dependientes para las cuales aún es posible identificar una versión anaerobia de la reacción que catalizan? Es probable que algunas de estas sean producto de un evento de duplicación ancestral, el cual pudo haberse dado a partir de la enzima análoga o de una cuya actividad no tuviera relación alguna. Precisamente, este fue el punto que decidimos investigar.

Comparación entre reacciones dependientes e independientes de oxígeno

Es innegable que el trabajo de Jabłońska & Tawfik, publicado hace menos de un año, abarca ciertos puntos de esta segunda parte del presente proyecto. A pesar de ello, existen resultados y diferencias notables que se desarrollarán a continuación.

En primer lugar, Jabłońska & Tawfik (2022) consideran únicamente a las enzimas que parecen haber surgido a partir de un ancestro cuya actividad no dependiera del oxígeno, mientras que nosotros incluimos también al otro grupo de enzimas, es decir, aquellas cuyo ancestro también fuera una enzima O_2-dependiente. En segundo lugar, mientras que Jabłońska & Tawfik establecen 54 pares de reacciones análogas, nosotros logramos identificar 99. Para

ello, revisamos manualmente todas las reacciones de oxidorreducción (EC 1.-.-.-) presentes en la base de datos MetaCyc (Caspi *et al*. 2020). Como puede observarse en la Tabla 4, este número no coincide con el total de códigos enzimáticos para cada tipo de enzima (O_2-dependiente y O_2-independiente). Esto se debe a que puede haber más de una enzima que lleve a cabo la misma reacción, lo cual ocurre con mayor frecuencia para las O_2-independientes.

En términos de diversidad funcional, las enzimas O_2-dependientes se distribuyen a lo largo de 11 subclases enzimáticas diferentes, mientras que las O_2-independientes abarcan 12 subclases (Tabla 5 y Figura 11A-B). Sin embargo, la razón a nivel de sub-subclases es casi 2:1. En este caso, 38 sub-subclases agrupan a las enzimas que no requieren oxígeno, en comparación de las 20 que comprenden al otro grupo de enzimas (Tabla 5 y Figura 11C-D).

Tabla 5. Distribución funcional y estructural de los pares de reacciones análogas que identificamos en este trabajo. La información contenida fue tomada de las bases de datos MetaCyc (Caspi *et al*. 2020), ExplorEnz (McDonald *et al*. 2009); CATH (Sillitoe *et al*. 2021) y Mechanism and Catalytic Site Atlas (M-CSA) (Ribeiro *et al*. 2018).

	O_2-dependientes	O_2-independientes
Pares de reacciones análogas	99	
No. Códigos EC	101	161
No. Reacciones	103	169
No. Códigos incompletos	11	23
Diversidad de subclases	11	12
Diversidad de sub-subclases	20	38
Pares de reacciones con estructura terciaria disponible	26	
Superfamilias exclusivas	9	13
Superfamilias compartidas	5	
Mecanismos de reacción para ambos miembros del par	3	

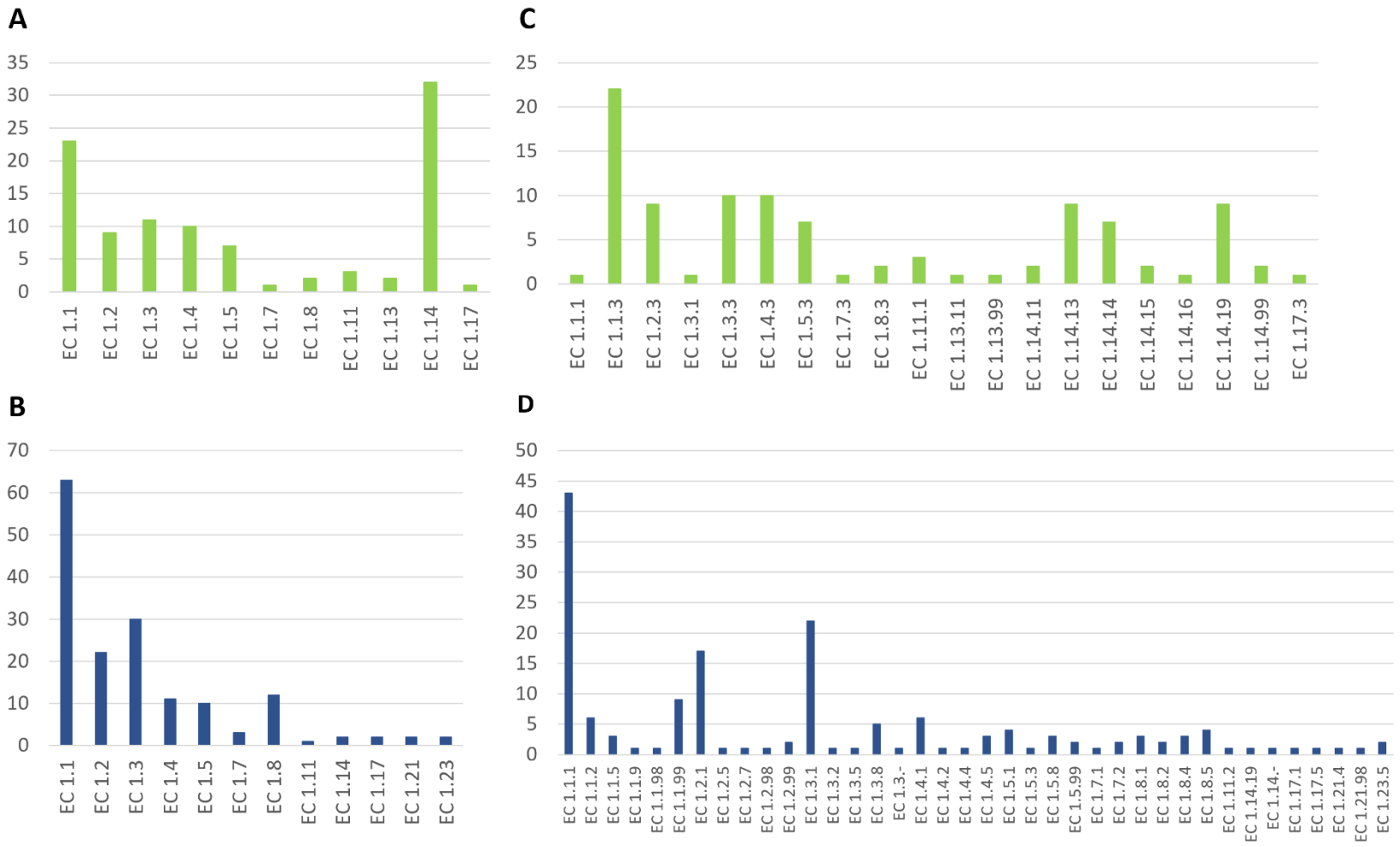
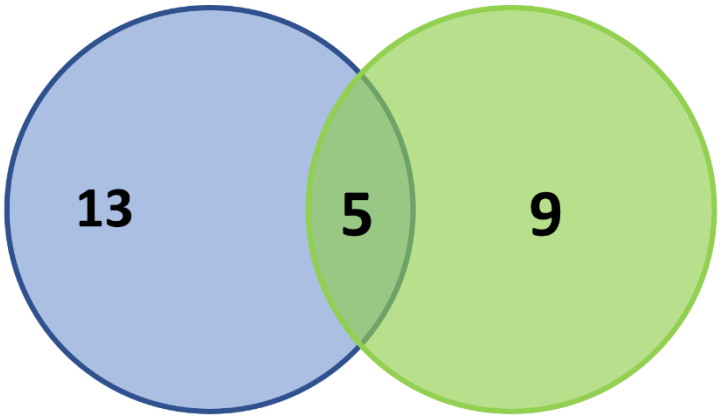


Figura 11. Distribución a nivel de subclases y sub-subclases de las enzimas dependientes e independientes de oxígeno. Aquí se incluye el total de números EC (Tabla 4) que identificamos para las enzimas de aquellas reacciones que poseen versiones aerobias y anaerobias, independientemente de si poseen o no estructura terciaria resuelta. El eje *y* de cada gráfica indica el número de enzimas (frecuencia) dentro de cada una de las categorías en el eje *x*. En (A) y (B) se muestra lo referente a las subclases, mientras que (C) y (D) indican la frecuencia a nivel de sub-subclase. Las gráficas con columnas verdes corresponden a las enzimas dependientes de oxígeno.



O_2-independent

Alanine racemase
Aldehyde Dehydrogenase; Chain A, domain 1
Aldehyde Dehydrogenase; Chain A, domain 2
Aldehyde Ferredoxin Oxidoreductase Protein, subunit A, domain 2
Cytochrome c
Deoxyhypusine Synthase
Glutamate Dehydrogenase, chain A, domain 3
Methanol Dehydrogenase, chain B
N-(1-d-carboxylethyl)-l-norvaline Dehydrogenase; domain 2
NAD(P)-binding Rossmann-like Domain
NADP-dependent oxidoreductase
Quinoprotein alcohol dehydrogenase-like superfamily

[2Fe-2S]-binding domain
Aldolase class I
Alpha-glycerophosphate oxidase, cap domain
Butyryl-Coa Dehydrogenase, subunit A, domain 1
Glutaredoxin

O_2-dependent

Butyryl-Coa Dehydrogenase, subunit A, domain 1
Catalase HpII, Chain A, domain 1
Cholesterol Oxidase; domain 2
D-Amino Acid Oxidase, subunit A, domain 2
FAD/NAD(P)-binding domain
Glucose Oxidase, domain 3
Guanine Nucleotide Dissociation Inhibitor, domain 1
Metallo-beta-lactamase, chain A
Sulfite Oxidase; Chain A, domain 2
Vanillyl-alcohol Oxidase; Chain A, domain 4

Figura 12. Superfamilias a las que pertenecen las enzimas dependientes e independientes de oxígeno molecular para las que existe estructura terciaria disponible. Existen 13 superfamilias exclusivas para las enzimas O_2-independientes de este grupo, 9 para las O_2-dependientes y 5 que son comunes a ambos grupos.

Con respecto a la estructura terciaria, ésta solamente se encuentra reportada para el 26% de los pares de reacciones. En este caso, la diversidad a nivel de superfamilias resulta similar en ambos grupos de enzimas: hay 10 exclusivas para las enzimas O_2-dependientes, 12 para las O_2-independientes y cinco más que incluyen miembros de ambos grupos de enzimas, de acuerdo con la base de datos CATH (Sillitoe *et al*. 2021). (Tabla 4 y Figura 12). Y en cuanto al mecanismo de reacción, únicamente hallamos tres casos en los que está descrito tanto para la reacción aerobia como para la anaerobia, de acuerdo con la base de datos M-CSA (Ribeiro *et al*. 2018).

La superfamilia NAD/P-Rossman fold es la que concentra un mayor número de las enzimas O_2-independientes, seguida de "Aldehído deshidrogenasa" y "Dominio de unión [2Fe-2S]" (Figura 13A). El resto de las superfamilias solamente poseen 1 ó 2 miembros. Por su parte, para las enzimas O_2-dependientes, las superfamilias "dominio de unión [2Fe-2S]", "Butiril-CoA deshidrogenasa" y "Vanilil-alcohol oxidasa" son las que contienen más miembros, mientras que las demás solo incluyen 1 ó 2 enzimas (Figura 13B).

En las reacciones aerobias (O_2-dependientes) de oxidorreducción, el oxígeno molecular actúa como aceptor de electrones. En cambio, la transferencia de electrones en las oxidorreductasas O_2-independientes es posible gracias a una gran diversidad de cofactores, en su mayoría orgánicos (Fischer *et al*. 2010). Esta misma tendencia la observamos también en nuestra muestra de 161 reacciones redox que poseen una versión análoga dependiente de oxígeno: el 90% de ellas requiere de un cofactor orgánico (Figura 14A). Entre los cofactores orgánicos involucrados destacan aquellos derivados de ribonucleótidos (72% del total), seguidos de las quinonas y de los derivados de tetrapirroles (Figura 14B). Llama la atención que, de las 145 enzimas que utilizan cofactores orgánicos, la gran mayoría de ellas utiliza al NAD^+ o a su forma fosforilada (NADP) (Figura 14C). Esto podría representar una evidencia indirecta de que este tipo de enzimas son bastante antiguas y, casi seguramente, mucho más viejas que sus contrapartes O_2-dependientes. Se han identificado distintas enzimas putativamente ancestrales cuyos residuos de aminoácidos (neutros en su mayoría) en los sitios de unión a ligandos están mucho más asociados a moléculas orgánicas que a metales de transición y pequeñas moléculas inorgánicas (Ji *et al*. 2008). Al igual que las reacciones en nuestra muestra, la mayoría de las que Ji et al. consideran como "ancestrales" también requieren del cofactor NAD^+/NADP.

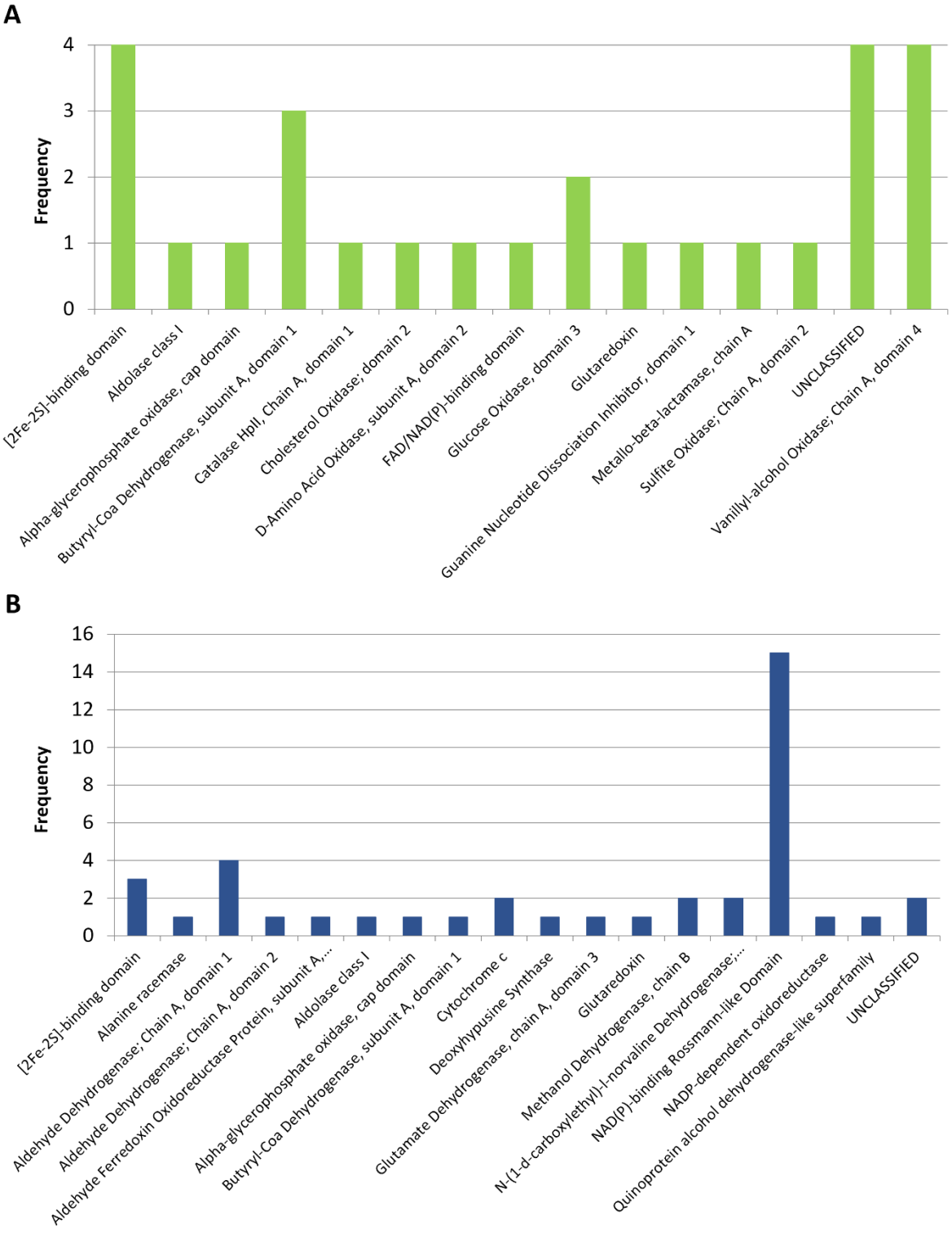


Figura 13. Distribución a nivel de superfamilias de las enzimas dependientes (A) e independientes (B) de oxígeno para las cuales existe estructura terciaria resuelta.

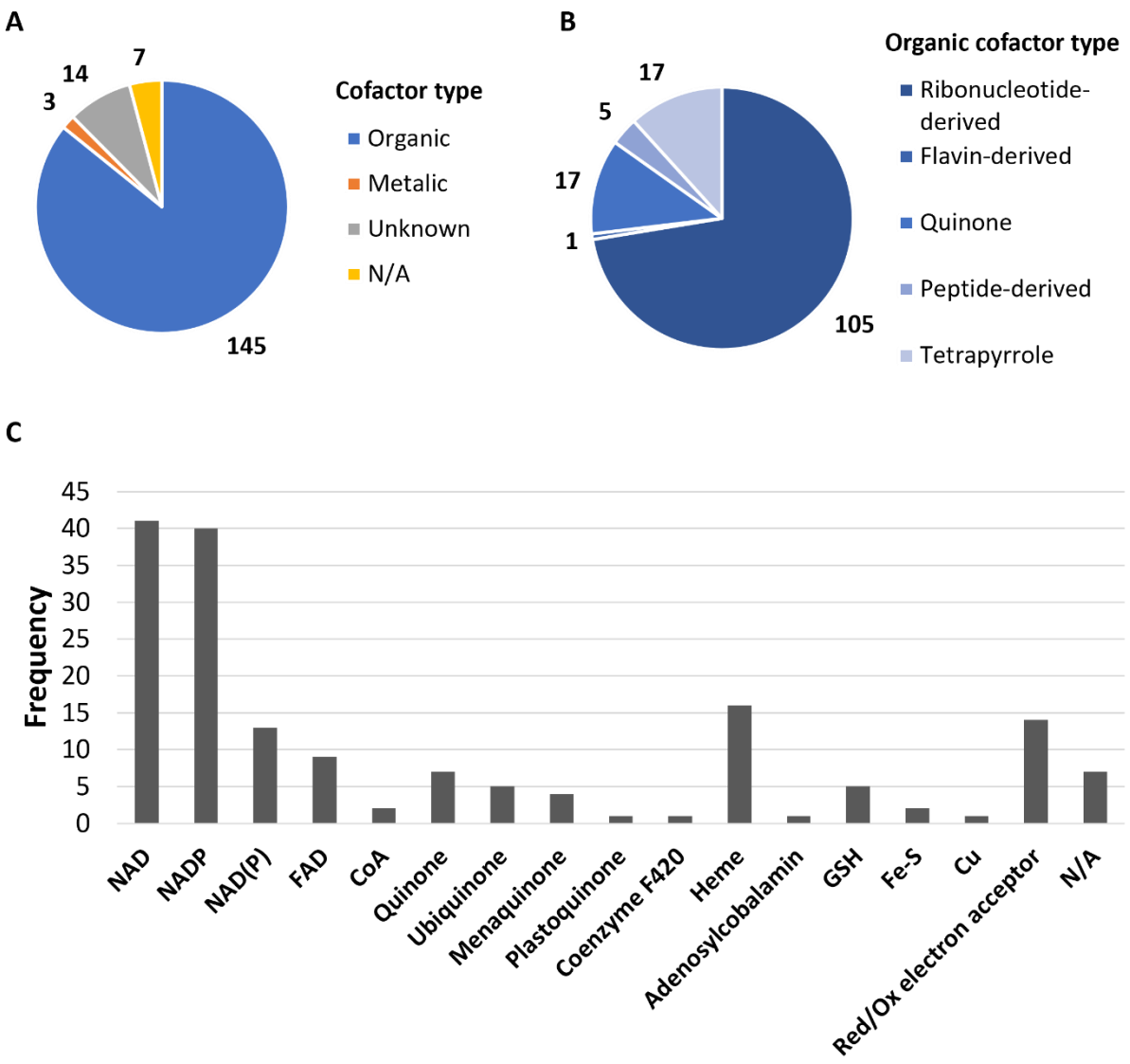


Figura 14. Uso de cofactores en nuestra muestra de enzimas dependientes e independientes de oxígeno molecular. (A) Tipo de cofactor que participa en cada reacción diferente. (B) Tipo de cofactor orgánico utilizado por las 145 enzimas identificadas en (A). (C) Frecuencia de los cofactores orgánicos involucrados en las reacciones tipo redox de nuestra muestra. La información con respecto al uso de cofactores la obtuvimos a partir de la base de datos CoFactor (Fischer *et al*. 2010)

A partir de nuestra muestra de 26 pares de reacciones análogas en los que para cada miembro del par existe una estructura terciara resuelta, pudimos identificar tres casos en los que ambos miembros del par son homólogos y pertenecen a la misma superfamilia. Estos corresponden a las enzimas L-lactato oxidasa (EC 1.1.3.2) - L-lactato deshidrogenasa (EC

1.1.2.3), acil-CoA oxidasa (EC 1.3.3.6) - acil-CoA deshidrogenasa de cadena media (EC 1.3.8.7) y glicerol 3-fosfato oxidasa (EC 1.1.3.21) – glicerol 3-fosfato deshidrogenasa (EC 1.1.5.3) (Figura 15). En este último caso, identificamos otras dos enzimas que llevan a cabo la misma reacción (sn-glicerol 3-fosfato → glicerona fosfato) pero que poseen una estructura terciaria no-homóloga.

La xantina oxidasa/deshidrogenasa representa un caso único debido a que se trata de una misma enzima que es capaz de llevar a cabo tanto la versión aerobia (Figura 16A) como la anaerobia (la cual es dependiente de NAD^+; Hellsten 2000; Figura 16B) de la misma reacción: la oxidación de hipoxantina a ácido úrico. Dicha reacción ocurre en dos pasos; primeramente se oxida la hipoxantina a xantina y, posteriormente, la xantina se oxida para dar lugar al ácido úrico. Para que esto pueda ocurrir, es necesario que la enzima pase por un proceso de transformación (Nishino *et al*. 1998) que puede ser reversible o irreversible. La transformación reversible se logra de distintas maneras como el someter a la enzima a una temperatura de -20°C (Della Corte & Stirpe 1968), ponerla en una disolución junto con la enzima sulfhidrilo oxidasa (Clare *et al*. 1981) o en presencia de reactivos con grupos tiol como el sulfato de cobre, 5,5'-Ditiobis(ácido 2-nitrobenzoico), la *N*-etilmaleimida, entre otros (Della Corte & Stirpe 1972). Por su parte, la transformación irreversible solo parece obtenerse en presencia de ciertas enzimas proteolíticas presentes en el hígado (Saksela *et al*. 1999) y en el intestino de mamíferos (Battelli *et al*. 1972).

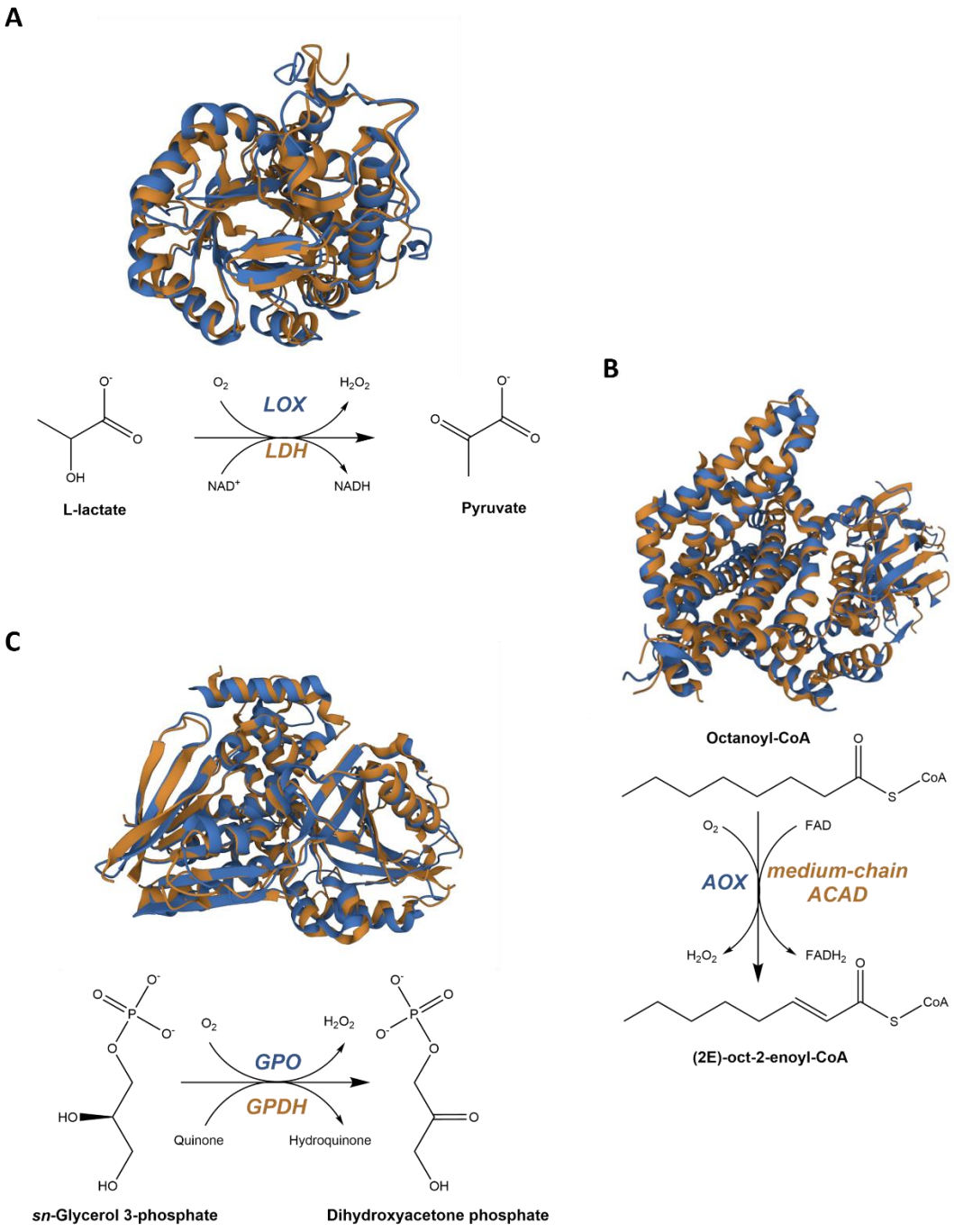


Figura 15. Superposición estructural de las enzimas homólogas que llevan a cabo versiones análogas de la misma reacción. (A) L-lactato oxidasa (LOX), L-lactato deshidrogenasa (LDH) y oxidación de lactato a piruvato (B) Acil-CoA oxidasa (AOX), acil-CoA deshidrogenasa de cadena media (ACAD) y oxidación de octanoil-CoA a (2E)-oct-e-enoil-CoA. (C) Glicerol-3-fosfato oxidasa (GPO), glicerol-3-fosfato deshidrogenasa (GPDH) y oxidación de glicerol 3-fosfato a dihidroxiacetona fosfato. Las enzimas O_2-dependientes se representan en color azul. Códigos PDB: 2J6X (LOX), 1FCB (LDH), 1W07 (AOX), 6KPT (ACAD), 2RGH (GPO) y 2QCU (GPDH).

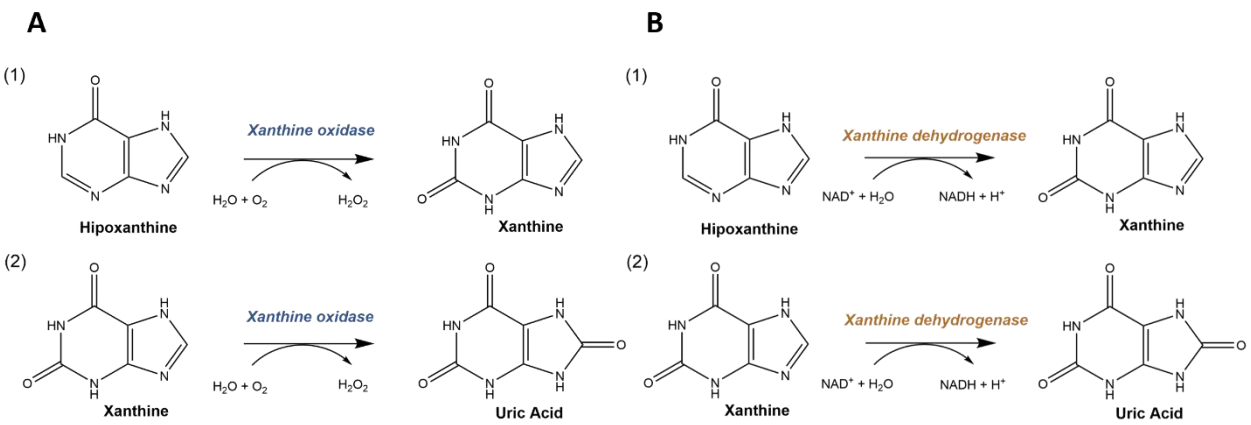


Figura 16. Oxidación de la xantina, la cual ocurre en dos pasos: primero a partir de la oxidación de la hipoxantina a xantina para que, posteriormente, esta se oxide y de lugar a ácido úrico. (A) Versión aerobia; (B) versión anaerobia, dependiente del cofactor NAD^+.

Perspectivas finales

Con los resultados obtenidos hasta el momento no nos es posible establecer el papel que la duplicación génica pudo haber jugado en la expansión del metabolismo, particularmente el aerobio, como consecuencia del GEO. Aun así, existen varios puntos interesantes que pueden desarrollarse más y/o dar lugar a futuros análisis que nos permitan responder, aunque quizá de manera parcial, nuestra pregunta inicial. Dichos puntos son:

1) Para la mayoría de los pares de reacciones análogas no existe por lo menos una estructura terciaria disponible para ambos miembros del par. En la mayoría de los casos, sin embargo, al menos uno de los miembros posee una estructura asociada. Tomando esto en cuenta, análisis filogenéticos a nivel de estructura primaria parecen ser la opción más viable para tratar de dilucidar parte de la historia evolutiva de cada una de estas enzimas.
2) De los 26 pares de reacciones en las que existe estructura terciaria para ambos miembros de cada par, en más de 20 no parece haber homología dado que cada enzima pertenece a una superfamilia diferente. Esto sugiere que las enzimas aerobias tuvieron un origen independiente (y probablemente posterior) al de las anaerobias, representando así casos de evolución convergente.
3) Aunado al punto anterior, existen ejemplos en los cuales una enzima que no requiere oxígeno (un miembro del par A) es parte de una superfamilia a la cual también pertenece

una enzima O_2-dependiente pero que cataliza una reacción completamente diferente (un miembro del par B). Este tipo de ejemplos son de los más relevantes en nuestra opinión dado que nos hablan acerca de la diversidad funcional que existe dentro de esas superfamilias.

4) Otro tipo de ejemplos de superfamilias con gran versatilidad funcional son los tres que involucran a dos enzimas que catalizan la versión aerobia y anaerobia de la misma reacción y que, además, pertenecen a la misma superfamilia. En estos casos, sería posible llevar a cabo distintos análisis filogenéticos a partir de los cuáles podríamos identificar con relativa facilidad (a) cuál de las dos enzimas surgió primero y (b) si la enzima dependiente de oxígeno se originó a partir de un evento de duplicación, el cual podría involucrar a su contraparte anaerobia o, quizá, a alguna otra de las enzimas de esa superfamilia.
5) Finalmente, valdría la pena llevar a cabo análisis filogenéticos basados en estructura terciaria para los casos en que dicha estructura esté resuelta para ambos miembros de un par de reacciones. Particularmente, varias de las enzimas anaerobias podrían haber surgido hace más de 2Ga, por lo que es probable que en algunos casos no sea posible detectar homología entre estas y otros miembros de sus respectivas superfamilias si nos limitamos únicamente a análisis basados en estructura primaria.

CONCLUSIÓN GENERAL

El estilo de vida de los organismos parece tener cierta relación con la proporción de genes duplicados que hallamos en ellos, siendo los de vida libre quienes, en general, poseen los valores más altos, seguidos de los patógenos y extremófilos y, finalmente, de los intracelulares. Análisis estadísticos más finos podrían ayudarnos a identificar las variables más importantes para esta tendencia que observamos.

Asimismo, hallamos que distintos grupos de organismos filogenéticamente distantes pero con estilos de vida similares (como algunos termófilos e intracelulares) poseen aspectos genómicos y metabólicos en común. Esto sugiere que podrían compartir ciertas estrategias relacionadas con la adaptación a condiciones ambientales comunes, mismas que podrían recaer en funciones relacionadas evolutivamente (genes homólogos) o que sean producto de evolución convergente (genes análogos).

A nivel del Sistema de Clasificación Enzimática, las Translocasas (EC 7), Oxidorreductasas (EC 1) e Isomerasas (EC 5) son, en ese orden, las clases de enzimas que presentan las proporciones de parálogos más altas. Tales proporciones difieren significativamente entre sí y con respecto al resto de las clases enzimáticas.

El número relativamente bajo de isomerasas nos permitió analizar esta clase a un detalle mucho mayor. Así, pudimos identificar que, en procariontes, hay un gran número de isomerasas duplicadas que participan en los procesos de patogenicidad, biosíntesis de antibióticos y otros que involucran directamente al RNA y/o a ribonucleótidos modificados, como es el caso de la pseudouridina sintasa RluD. Además, casos como el de esta y otras enzimas nos indican que, a pesar de que arqueas y bacterias comparten muchas isomerasas, esto no significa que en ambos grupos de organismos hallemos proporciones similares de duplicados y, más aún, aunque dos enzimas presenten un número similar de parálogos tanto en arqueas como en bacterias, la razón detrás de su retención podría ser muy diferente en cada uno.

REFERENCIAS BIBLIOGRÁFICAS

Angert, E. R. (2012). DNA replication and genomic architecture of very large bacteria. *Annual Review of Microbiology*, *66*, 197–212. https://doi.org/10.1146/annurev-micro-090110-102827

Arora, B., Mukherjee, J., & Gupta, M. N. (2014). Enzyme promiscuity: using the dark side of enzyme specificity in white biotechnology. *Sustainable Chemical Processes*, *2*(1), 1–9. https://doi.org/10.1186/s40508-014-0025-y

Battelli, M. G., Della Corte, E., & Stirpe, F. (1972). Xanthine Oxidase Type D (Dehydrogenase) in the Intestine and other Organs of the Rat By. *Biochemical Journal*, *126*, 747–749. 10.1042/bj1260747

Ben-Shem, A., Frolow, F., & Nelson, N. (2004). Evolution of photosystem I - From symmetry through pseudosymmetry to asymmetry. *FEBS Letters*, *564*(3), 274–280. https://doi.org/10.1016/S0014-5793(04)00360-6

Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. A., Boudreau, M. E. R., Nesbø, C. L., Case, R. J., & Doolittle, W. F. (2003). Lateral Gene Transfer and the Origins of Prokaryotic Groups. *Annual Review of Genetics*, *37*, 283–328. https://doi.org/10.1146/annurev.genet.37.050503.084247

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., & Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nature Genetics*, *28*(3), 281–285. https://doi.org/10.1038/90129

Canfield, D. E. (2005). The early history of atmospheric oxygen: Homage to Robert M. Garrels. *Annual Review of Earth and Planetary Sciences*, *33*, 1–36. https://doi.org/10.1146/annurev.earth.33.092203.122711

Cardona, T. (2015). A fresh look at the evolution and diversification of photochemical reaction centers. *Photosynthesis Research*, *126*(1), 111–134. https://doi.org/10.1007/s11120-014-0065-x

Carrington, Y., Guo, J., Le, C. H., Fillo, A., Kwon, J., Tran, L. T., & Ehlting, J. (2018). Evolution of a secondary metabolic pathway from primary metabolism: shikimate and quinate biosynthesis in plants. *Plant Journal*, *95*(5), 823–833. https://doi.org/10.1111/tpj.13990

Casadevall, A., & Pirofski, L. A. (2014). Ditch the term pathogen. *Nature*, *516*(7530), 165–166. https://doi.org/10.1038/516165a

Caspi, R., Billington, R., Keseler, I. M., Kothari, A., Krummenacker, M., Midford, P. E., Ong, W. K., Paley, S., Subhraveti, P., & Karp, P. D. (2020). The MetaCyc database of metabolic pathways and enzymes-a 2019 update. *Nucleic Acids Research*, *48*(D1), D455–D453. https://doi.org/10.1093/nar/gkz862

Chater, K.F., Kinashi, H. (2007). *Streptomyces* Linear Plasmids: Their Discovery, Functions, Interactions with Other Replicons, and Evolutionary Significance. In: Meinhardt, F., Klassen, R. (eds) *Microbial Linear Plasmids. Microbiology Monographs, vol 7* (pp. 1–32). Springer, Berlin, Heidelberg. https://doi.org/10.1007/7171_2007_097

Ciccarelli, F. D., Doerks, T., Von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science*, *311*(5765), 1283–1287. https://doi.org/10.1126/science.1123061

Clare, D. A., Blakistone, B. A., Swaisgood, H. E., & Horton, H. R. (1981). Sulfhydryl Oxidase-Catalyzed Conversion of Xanthine Dehydrogenase to Xanthine Oxidase. *Archives of Biochemistry and Biophysics*, *211*(1), 44–47.

Conant, G. C., & Wagner, A. (2002). GenomeHistory: A software tool and its application to fully sequenced genomes. *Nucleic Acids Research*, *30*(15), 3378–3386. https://doi.org/10.1093/nar/gkf449

Conant, G. C., & Wolfe, K. H. (2007). Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Molecular Systems Biology*, *3*(129). https://doi.org/10.1038/msb4100170

Conrad, J., Sun, D., Englund, N., & Ofengand, J. (1998). The rluC gene of Escherichia coli codes for a pseudouridine synthase that is solely responsible for synthesis of pseudouridine at positions 955, 2504, and 2580 in 23 S ribosomal RNA. *Journal of Biological Chemistry*, *273*(29), 18562–18566. https://doi.org/10.1074/jbc.273.29.18562

Copley, S. D. (2003). Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. *Current Opinion in Chemical Biology*, *7*(2), 265–272. https://doi.org/10.1016/S1367-5931(03)00032-2

Crow, K. D., & Wagner, G. P. (2006). What is the role of genome duplication in the evolution of complexity and diversity? *Molecular Biology and Evolution*, *23*(5), 887–892. https://doi.org/10.1093/molbev/msj083

Della Corte, E., & Stirpe, F. (1968). The regulation of rat-liver xanthine oxidase: Activation by proteolytic enzymes. *FEBS Letters*, *2*(2), 83–84. https://doi.org/10.1016/0014-5793(68)80107-3

Della Corte, E., & Stirpe, F. (1972). The regulation of rat liver xanthine oxidase. Involvement of thiol groups in the conversion of the enzyme activity from dehydrogenase (type D) into oxidase (type O) and purification of the enzyme. *The Biochemical Journal*, *126*(3), 739–745. https://doi.org/10.1042/bj1260739

Des Marais, D. L., & Rausher, M. D. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, *454*(7205), 762–765. https://doi.org/10.1038/nature07092

Díaz-Mejía, J. J., Pérez-Rueda, E., & Segovia, L. (2007). A network perspective on the evolution of metabolism by gene duplication. *Genome Biology*, *8*(2), 1–10. https://doi.org/10.1186/gb-2007-8-2-r26

Durbin, M. L., McCaig, B., & Clegg, M. T. (2000). Molecular evolution of the chalcone synthase multigene family in the morning glory genome. *Plant Molecular Biology*, *42*, 79–92. https://doi.org/10.1023/A

Fani, R. (2012). The Origin and Evolution of Metabolic Pathways: Why and How did Primordial Cells Construct Metabolic Routes? *Evolution: Education and Outreach*, *5*(3), 367–381. https://doi.org/10.1007/s12052-012-0439-5

Fani, R., & Fondi, M. (2009). Origin and evolution of metabolic pathways. *Physics of Life Reviews*, *6*(1), 23–52. https://doi.org/10.1016/j.plrev.2008.12.003

Fani, R., Gallo, R., & Liò, P. (2000). Molecular evolution of nitrogen fixation: The evolutionary history of the nifD, nifK, nifE, and nifN genes. *Journal of Molecular Evolution*, *51*(1), 1–11. https://doi.org/10.1007/s002390010061

Fani, R., Liò, P., Chiarelli, I., & Bazzicalupo, M. (1994). The evolution of the histidine biosynthetic genes in prokaryotes: A common ancestor for the hisA and hisF genes. *Journal of Molecular Evolution*, *38*(5), 489–495. https://doi.org/10.1007/BF00178849

Feschotte, C., & Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics*, *41*, 331–368. https://doi.org/10.1146/annurev.genet.40.110405.090448

Fischer, J. D., Holliday, G. L., Rahman, S. A., & Thornton, J. M. (2010). The structures and physicochemical properties of organic cofactors in biocatalysis. *Journal of Molecular Biology*, *403*(5), 803–824. https://doi.org/10.1016/j.jmb.2010.09.018

Fischer, J. D., Holliday, G. L., & Thornton, J. M. (2010). The CoFactor database: Organic cofactors in enzyme catalysis. *Bioinformatics*, *26*(19), 2496–2497. https://doi.org/10.1093/bioinformatics/btq442

Fondi, M., Brilli, M., Emiliani, G., Paffetti, D., & Fani, R. (2007). The primordial metabolism: An ancestral interconnection between leucine, arginine, and lysine biosynthesis. *BMC Evolutionary Biology*, *7*(SUPPL. 2), 1–14. https://doi.org/10.1186/1471-2148-7-S2-S3

Freeburg, S. H., Engelbrecht, E., & Powell, W. H. (2017). Subfunctionalization of paralogous aryl hydrocarbon receptors from the frog Xenopus laevis: Distinct target genes and differential responses to specific agonists in a single cell type. *Toxicological Sciences*, *155*(2), 337–347. https://doi.org/10.1093/toxsci/kfw212

Gao, N. L., Chen, J., Wang, T., Lercher, M. J., & Chen, W. H. (2019). Prokaryotic Genome Expansion Is Facilitated by Phages and Plasmids but Impaired by CRISPR. *Frontiers in Microbiology*, *10*(October), 1–8. https://doi.org/10.3389/fmicb.2019.02254

Ge, J., & Yu, Y. T. (2013). RNA pseudouridylation: New insights into an old modification. *Trends in Biochemical Sciences*, *38*(4), 210–218. https://doi.org/10.1016/j.tibs.2013.01.002

Gout, J. F., & Lynch, M. (2015). Maintenance and loss of duplicated genes by dosage subfunctionalization. *Molecular Biology and Evolution*, *32*(8), 2141–2148. https://doi.org/10.1093/molbev/msv095

Gupta, R. S. (2000). The phylogeny of proteobacteria: Relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiology Reviews*, *24*(4), 367–402. https://doi.org/10.1016/S0168-6445(00)00031-0

Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity*, *100*(5), 605–617. https://doi.org/10.1093/jhered/esp047

Hamilton, T. L. (2019). The trouble with oxygen: The ecophysiology of extant phototrophs and implications for the evolution of oxygenic photosynthesis. *Free Radical Biology and Medicine*,

140(September 2018), 233–249. https://doi.org/10.1016/j.freeradbiomed.2019.05.003

Harrison, E., & Brockhurst, M. A. (2012). Plasmid-mediated horizontal gene transfer is a coevolutionary process. *Trends in Microbiology*, *20*(6), 262–267. https://doi.org/10.1016/j.tim.2012.04.003

Hellsten, Y. (2000). The role of xanthine oxidase in exercise. In C. K. Sen, L. Packer, & O. Hänninen (Eds.), *Handbook of Oxidants and Antioxidants in Exercise* (pp. 153–176). Elsevier Science B.V. https://doi.org/10.1016/b978-044482650-3/50007-9

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K., & Hood, L. (1997). Gene families: The taxonomy of protein paralogs and chimeras. *Science*, *278*(5338), 609–614. https://doi.org/10.1126/science.278.5338.609

Herdman M (1985) The evolution of bacterial genomes . In: Cavalier Smith T (ed) *The Evolution of genome size*. John Wiley, London

Holland, H. D. (2006). The oxygenation of the atmosphere and oceans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1470), 903–915. https://doi.org/10.1098/rstb.2006.1838

Hooper, S. D., & Berg, O. G. (2003). On the nature of gene innovation: Duplication patterns in microbial genomes. *Molecular Biology and Evolution*, *20*(6), 945–954. https://doi.org/10.1093/molbev/msg101

Hopwood, D. A. (1967). Genetic analysis and genome structure in Streptomyces coelicolor. *Bacteriological Reviews*, *31*(4), 373–403. https://doi.org/10.1128/mmbr.31.4.373-403.1967

Horowitz, N. H. (1945). On the Evolution of Biochemical Syntheses. *Proceedings of the National Academy of Sciences of the United States of America*, *31*(6), 153–157.

Hu, J. Y., Zhang, Y. P., & Yu, L. (2012). Summary of Laurasiatheria (mammalia) phylogeny. *Dong Wu Xue Yan Jiu = Zoological Research / "Dong Wu Xue Yan Jiu" Bian Ji Wei Yuan Hui Bian Ji*, *33*(E5-6), 65–74. https://doi.org/10.3724/sp.j.1141.2012.e05-06e65

Huang, L., Ku, J., Pookanjanatavip, M., Gu, X., Wang, D., Greene, P. J., & Santi, D. V. (1998). Identification of two Escherichia coli pseudouridine synthases that show multisite specificity for 23S RNA. *Biochemistry*, *37*(45), 15951–15957. https://doi.org/10.1021/bi981002n

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(5), 1–6. https://doi.org/10.1038/nmicrobiol.2016.48

Ionescu, D., Bizic-Ionescu, M., De Maio, N., Cypionka, H., & Grossart, H. P. (2017). Community-like genome in single cells of the sulfur bacterium Achromatium oxaliferum. *Nature Communications*, *8*(1), 1–12. https://doi.org/10.1038/s41467-017-00342-9

Islas, S., Castillo, A., Vázquez, H.G., Lazcano, A. (2000). On the Role of Genome Duplications in the Evolution of Prokaryotic Chromosomes. In: Chela-Flores, J., Lemarchand, G.A., Oró, J. (eds) *Astrobiology*. Springer, Dordrecht. https://doi.org/10.1007/978-94-011-4313-4_30

Jabłońska, J., & Tawfik, D. S. (2022). Innovation and tinkering in the evolution of oxidases. *Protein Science*, *31*(5), 1–11. https://doi.org/10.1002/pro.4310

Jangam, D., Feschotte, C., & Betrán, E. (2017). Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends in Genetics*, *33*(11), 817–831. https://doi.org/10.1016/j.tig.2017.07.011

Jensen, R. A. (1976). Enzyme Recruitment in Evolution of New Function. *Annual Review of Microbiology*, *30*(1), 409–425. https://doi.org/10.1146/annurev.mi.30.100176.002205

Ji, H. F., Chen, L., & Zhang, H. Y. (2008). Organic cofactors participated more frequently than transition metals in redox reactions of primitive proteins. *BioEssays*, *30*(8), 766–771. https://doi.org/10.1002/bies.20788

Jiang, Y. Y., Kong, D. X., Qin, T., Li, X., Caetano-Anollés, G., & Zhang, H. Y. (2012). The impact of oxygen on metabolic evolution: A chemoinformatic investigation. *PLoS Computational Biology*, *8*(3), 1–8. https://doi.org/10.1371/journal.pcbi.1002426

Jun, S. R., Sims, G. E., Wu, G. A., & Kim, S. H. (2010). Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(1), 133–138. https://doi.org/10.1073/pnas.0913033107

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Research*, *20*(10), 1313–1326. https://doi.org/10.1101/gr.101386.109

Kafri, R., Bar-Even, A., & Pilpel, Y. (2005). Transcription control reprogramming in genetic backup circuits. *Nature Genetics*, *37*(3), 295–299. https://doi.org/10.1038/ng1523

Kawashima, T., Kawashima, S., Tanaka, C., Murai, M., Yoneda, M., Putnam, N. H., Rokhsar, D. S., Kanehisa, M., Satoh, N., & Wada, H. (2009). Domain shuffling and the evolution of vertebrates. *Genome Research*, *19*(8), 1393–1403. https://doi.org/10.1101/gr.087072.108

Kinoshita, S., Kakizono, T., Kadota, K., Das, K., & Taguchi, H. (1985). Purification of two alcohol dehydrogenases from Zymomonas mobilis and their properties. *Applied Microbiology and Biotechnology*, *22*(4), 249–254. https://doi.org/10.1007/BF00252025

Kottemann, M., Kish, A., Iloanusi, C., Bjork, S., & DiRuggiero, J. (2005). Physiological responses of the halophilic archaeon Halobacterium sp. strain NRC1 to desiccation and gamma irradiation. *Extremophiles*, *9*(3), 219–227. https://doi.org/10.1007/s00792-005-0437-4

Kuzmin, E., Taylor, J. S., & Boone, C. (2022). Retention of duplicated genes in evolution. *Trends in Genetics*, *38*(1), 59–72. https://doi.org/10.1016/j.tig.2021.06.016

Leblond, P., & Decaris, B. (1999). Unstable Linear Chromosomes: the Case of *Streptomyces*. In: Charlebois, R. L. (ed) *Organization of the Prokaryotic Genome* (pp. 235–261). American Society for Microbiology, Washington, D.C.

Lindner, C., Nijland, R., Van Hartskamp, M., Bron, S., Hamoen, L. W., & Kuipers, O. P. (2004). Differential Expression of Two Paralogous Genes of Bacillus subtilis Encoding Single-Stranded DNA Binding Protein. *Journal of Bacteriology*, *186*(4), 1097–1105. https://doi.org/10.1128/JB.186.4.1097-1105.2004

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, *290*(5494), 1151–1155. https://doi.org/10.1126/science.290.5494.1151

Lynch, Michael, & Conery, J. S. (2003). The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics*, *3*, 35–44. https://doi.org/10.1023/A

Lyons, T. W., Reinhard, C. T., & Planavsky, N. J. (2014). The rise of oxygen in Earth's early ocean and atmosphere. *Nature*, *506*(7488), 307–315. https://doi.org/10.1038/nature13068

Maistrenko, O. M., Mende, D. R., Luetge, M., Hildebrand, F., Schmidt, T. S. B., Li, S. S., Rodrigues, J. F. M., von Mering, C., Pedro Coelho, L., Huerta-Cepas, J., Sunagawa, S., & Bork, P. (2020). Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME Journal*, *14*(5), 1247–1259. https://doi.org/10.1038/s41396-020-0600-z

Marais, G. A. B., Batut, B., & Daubin, V. (2020). Genome Evolution: Mutation Is the Main Driver of Genome Size in Prokaryotes. *Current Biology*, *30*(19), R1083–R1085. https://doi.org/10.1016/j.cub.2020.07.093

Martin, W. F., Bryant, D. A., & Beatty, J. T. (2018). A physiological perspective on the origin and evolution of photosynthesis. *FEMS Microbiology Reviews*, *42*(2), 205–231. https://doi.org/10.1093/FEMSRE/FUX056

Martinez-Gutierrez, C. A., & Aylward, F. O. (2022). Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales. *PLoS Genetics*, *18*(5), 1–17. https://doi.org/10.1371/journal.pgen.1010220

Martínez-Núñez, M. A., & Pérez-Rueda, E. (2016). Do lifestyles influence the presence of promiscuous enzymes in bacteria and Archaea metabolism? *Sustainable Chemical Processes*, *4*(1), 3–7. https://doi.org/10.1186/s40508-016-0047-8

Martínez-Núñez, M. A., Rodríguez-Vázquez, K., & Pérez-Rueda, E. (2015). The lifestyle of prokaryotic organisms influences the repertoire of promiscuous enzymes. *Proteins: Structure, Function and Bioinformatics*, *83*(9), 1625–1631. https://doi.org/10.1002/prot.24847

McClintock, J. M., Carlson, R., Mann, D. M., & Prince, V. E. (2001). Consequences of Hox gene duplication in the vertebrates: An investigation of the zebrafish Hox paralogue group 1 genes. *Development*, *128*(13), 2471–2484. https://doi.org/10.1242/dev.128.13.2471

Meléndez-Hevia, E., Waddell, T. G., & Cascante, M. (1996). The puzzle of the Krebs citric acid cycle: Assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. *Journal of Molecular Evolution*, *43*(3), 293–303. https://doi.org/10.1007/BF02338838

Merhej, V., Royer-Carenzi, M., Pontarotti, P., & Raoult, D. (2009). Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biology Direct*, *4*, 1–25. https://doi.org/10.1186/1745-6150-4-13

Mira, A., Ochman, H., & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*, *17*(10), 589–596. https://doi.org/10.1016/S0168-9525(01)02447-7

Mueller, E. G. (2002). Chips off the old block. *Nature Structural Biology*, *9*(5), 320–322.

Nishino, T., Okamoto, K., Nakanishi, S., Hori, H., & Nishino, T. (1998). The Mechanism of Conversion of Xanthine Dehydrogenase to Xanthine Oxidase. In Y. Ishimura, H. Shimada, & M. Suematsu (Eds.), *Oxygen Homeostasis and Its Dynamics. Keio University Symposia for Life Science and Medicine, vol 1* (pp. 333–339). Springer, Tokyo. https://doi.org/10.1007/978-4-431-68476-3_42

Oliveira, P. H., Touchon, M., Cury, J., & Rocha, E. P. C. (2017). The chromosomal organization of horizontal gene transfer in bacteria. *Nature Communications*, *8*(1), 25–28. https://doi.org/10.1038/s41467-017-00808-w

Pecoraro, V., Zerulla, K., Lange, C., & Soppa, J. (2011). Quantification of ploidy in proteobacteria revealed the existence of monoploid, (mero-)oligoploid and polyploid species. *PLoS ONE*, *6*(1). https://doi.org/10.1371/journal.pone.0016392

Pérez-Pantoja, D., González, B., & Pieper, D. H. (2016). Aerobic Degradation of Aromatic Compounds. In F. Rojo (Ed.), *Aerobic Utilization of Hydrocarbons, Oils and Lipids* (pp. 1–44). Springer International Publishing. https://doi.org/10.1007/978-3-319-39782-5

Perrin, E., Fondi, M., Bosi, E., Mengoni, A., Buroni, S., Scoffone, V. C., Valvano, M., & Fani, R. (2017). Subfunctionalization influences the expansion of bacterial multidrug antibiotic resistance. *BMC Genomics*, *18*(1), 1–14. https://doi.org/10.1186/s12864-017-4222-4

Piontkivska, H., Rooney, A. P., & Nei, M. (2002). Purifying selection and birth-and-death evolution in the histone H4 gene family. *Molecular Biology and Evolution*, *19*(5), 689–697. https://doi.org/10.1093/oxfordjournals.molbev.a004127

Raymond, J., & Blankenship, R. E. (2004). Biosynthetic pathways, gene replacement and the antiquity of life. *Geobiology*, *2*(4), 199–203. https://doi.org/10.1111/j.1472-4677.2004.00037.x

Raymond, J., & Segrè, D. (2006). The Effect of Oxygen on Biochemical Networks and the Evolution of Complex Life. *Science*, *311*(5768), 1764–1767.

Ribeiro, A. J. M., Holliday, G. L., Furnham, N., Tyzack, J. D., Ferris, K., & Thornton, J. M. (2018). Mechanism and Catalytic Site Atlas (M-CSA): A database of enzyme reaction mechanisms and active sites. *Nucleic Acids Research*, *46*(D1), D618–D623. https://doi.org/10.1093/nar/gkx1012

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W. T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437. https://doi.org/10.1038/nature12352

Rison, S. C. G., & Thornton, J. M. (2002). Pathway evolution, structurally speaking. *Current Opinion in Structural Biology*, *12*(3), 374–382. https://doi.org/10.1016/S0959-440X(02)00331-7

Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R. C., & San Millán, Á. (2021). Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology*, *19*(6), 347–359. https://doi.org/10.1038/s41579-020-00497-1

Rodríguez-Gijón, A., Buck, M., Andersson, A. F., Izabel-Shen, D., Nascimento, F. J. A., & Garcia, S. L. (2023). Linking prokaryotic genome size variation to metabolic potential and environment. *ISME Communications*, *3*(1), 1–12. https://doi.org/10.1038/s43705-023-00231-x

Rost, B. (1997). Protein structures sustain evolutionary drift. *Folding and Design*, *2*(3), 19–24. https://doi.org/10.1016/S1359-0278(97)00059-X

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, *12*(2), 85–94. https://doi.org/10.1093/protein/12.2.85

Saito, M. A. (2012). The rise of oxygen and aerobic biochemistry. *Structure*, *20*(1), 1–2. https://doi.org/10.1016/j.str.2011.12.006

Sanchez, I., Hernandez-Guerrero, R., Mendez-Monroy, P. E., Martinez-Nuñez, M. A., Ibarra, J. A., & Pérez-Rueda, E. (2020). Evaluation of the abundance of DNA-binding transcription factors in Prokaryotes. *Genes*, *11*(1). https://doi.org/10.3390/genes11010052

Schirrmeister, B. E., De Vos, J. M., Antonelli, A., & Bagheri, H. C. (2013). Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(5), 1791–1796. https://doi.org/10.1073/pnas.1209927110

Schirrmeister, B. E., Gugger, M., & Donoghue, P. C. J. (2015). Cyanobacteria and the Great Oxidation Event: Evidence from genes and fossils. *Palaeontology*, *58*(5), 769–785. https://doi.org/10.1111/pala.12178

Scossa, F., & Fernie, A. R. (2020). The evolution of metabolism: How to test evolutionary hypotheses at the genomic level. *Computational and Structural Biotechnology Journal*, *18*, 482–500. https://doi.org/10.1016/j.csbj.2020.02.009

Sikosek, T., Chan, H. S., & Bornberg-Bauer, E. (2012). Escape from adaptive conflict follows from weak functional trade-offs and mutational robustness. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(37), 14888–14893. https://doi.org/10.1073/pnas.1115620109

Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S. D., Berka, K., Varekova, I. H., Svobodova, R., Lees, J., & Orengo, C. A. (2021). CATH: Increased structural coverage of functional space. *Nucleic Acids Research*, *49*(D1), D266–D273. https://doi.org/10.1093/nar/gkaa1079

Soppa, J. (2022). Non-equivalent genomes in polyploid prokaryotes. *Nature Microbiology*, *7*(2), 186–188. https://doi.org/10.1038/s41564-021-01034-3

Spenkuch, F., Motorin, Y., & Helm, M. (2014). Pseudouridine: Still mysterious, but never a fake (uridine)! *RNA Biology*, *11*(12), 1540–1554. https://doi.org/10.4161/15476286.2014.992278

Sturtevant, A. H. (1925). The effects of unequal crossing over At the Bar locus in Drosophila. *Genetics*, *10*(2), 117–147. https://doi.org/10.1093/genetics/10.2.117

Tekaia, F., & Dujon, B. (1999). Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. *Journal of Molecular Evolution*, *49*(5), 591–600.

https://doi.org/10.1007/PL00006580

Treangen, T. J., & Rocha, E. P. C. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genetics*, *7*(1). https://doi.org/10.1371/journal.pgen.1001284

Tria, F. D. K., & Martin, W. F. (2021). Gene Duplications Are At Least 50 Times Less Frequent than Gene Transfers in Prokaryotic Genomes. *Genome Biology and Evolution*, *13*(10), 1–14. https://doi.org/10.1093/gbe/evab224

Van De Peer, Y. (2004). Computational approaches to unveiling ancient genome duplications. *Nature Reviews Genetics*, *5*(10), 752–763. https://doi.org/10.1038/nrg1449

van de Peer, Y., Ashman, T. L., Soltis, P. S., & Soltis, D. E. (2021). Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell*, *33*(1), 11–26. https://doi.org/10.1093/plcell/koaa015

Van De Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, *18*(7), 411–424. https://doi.org/10.1038/nrg.2017.26

Velasco, A. M., Leguina, J. I., & Lazcano, A. (2002). Molecular evolution of the lysine biosynthetic pathways. *Journal of Molecular Evolution*, *55*(4), 445–459. https://doi.org/10.1007/s00239-002-2340-2

Voordeckers, K., Brown, C. A., Vanneste, K., van der Zande, E., Voet, A., Maere, S., & Verstrepen, K. J. (2012). Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication. *PLoS Biology*, *10*(12). https://doi.org/10.1371/journal.pbio.1001446

Wang, M., Jiang, Y. Y., Kim, K. M., Qu, G., Ji, H. F., Mittenthal, J. E., Zhang, H. Y., & Caetano-Anollés, G. (2011). A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. *Molecular Biology and Evolution*, *28*(1), 567–582. https://doi.org/10.1093/molbev/msq232

Weisman, C. M. (2022). The Origins and Functions of De Novo Genes: Against All Odds? *Journal of Molecular Evolution*, *90*(3–4), 244–257. https://doi.org/10.1007/s00239-022-10055-3

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics*, *2*(5), 333–341. https://doi.org/10.1038/35072009

Xie, G., Keyhani, N. O., Bonner, C. A., & Jensen, R. A. (2003). Ancient origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiology and Molecular Biology Reviews*, *67*(3), 303–342. https://doi.org/10.1128/mmbr.67.3.303-342.2003

Xie, Y., Gu, Y., Shi, G., He, J., Hu, W., & Zhang, Z. (2022). Genome-Wide Identification and Expression Analysis of Pseudouridine Synthase Family in Arabidopsis and Maize. *International Journal of Molecular Sciences*, *23*(5). https://doi.org/10.3390/ijms23052680

Ycas, M. (1974). On earlier states of the biochemical system. *Journal of Theoretical Biology*, *44*(1), 145–160.

Zhang, J. (2003). Evolution by gene duplication: An update. *Trends in Ecology and Evolution*, *18*(6), 292–298. https://doi.org/10.1016/S0169-5347(03)00033-8

Zhang, Y., & Sievert, S. M. (2014). Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in Epsilonproteobacteria. *Frontiers in Microbiology*, *5*(MAR), 1–13. https://doi.org/10.3389/fmicb.2014.00110

Zhou, Z., Gu, J., Li, Y. Q., & Wang, Y. (2012). Genome plasticity and systems evolution in Streptomyces. *BMC bioinformatics*, *13 Suppl 10*(Suppl 10), S8. https://doi.org/10.1186/1471-2105-13-S10-S8

Zhou, Y., Zhang, C., Zhang, L., Ye, Q., Liu, N., Wang, M., Long, G., Fan, W., Long, M., & Wing, R. A. (2022). Gene fusion as an important mechanism to generate new genes in the genus Oryza. *Genome Biology*, *23*(1), 1–23. https://doi.org/10.1186/s13059-022-02696-w

Zipkas, D., & Riley, M. (1975). Proposal concerning mechanism of evolution of the genome of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, *72*(4), 1354–1358. https://doi.org/10.1073/pnas.72.4.1354

Zou, C., Lehti-Shiu, M. D., Thomashow, M., & Shiu, S. H. (2009). Evolution of stress-regulated gene expression in duplicate genes of Arabidopsis thalianae. *PLoS Genetics*, *5*(7). https://doi.org/10.1371/journal.pgen.1000581

ANEXO

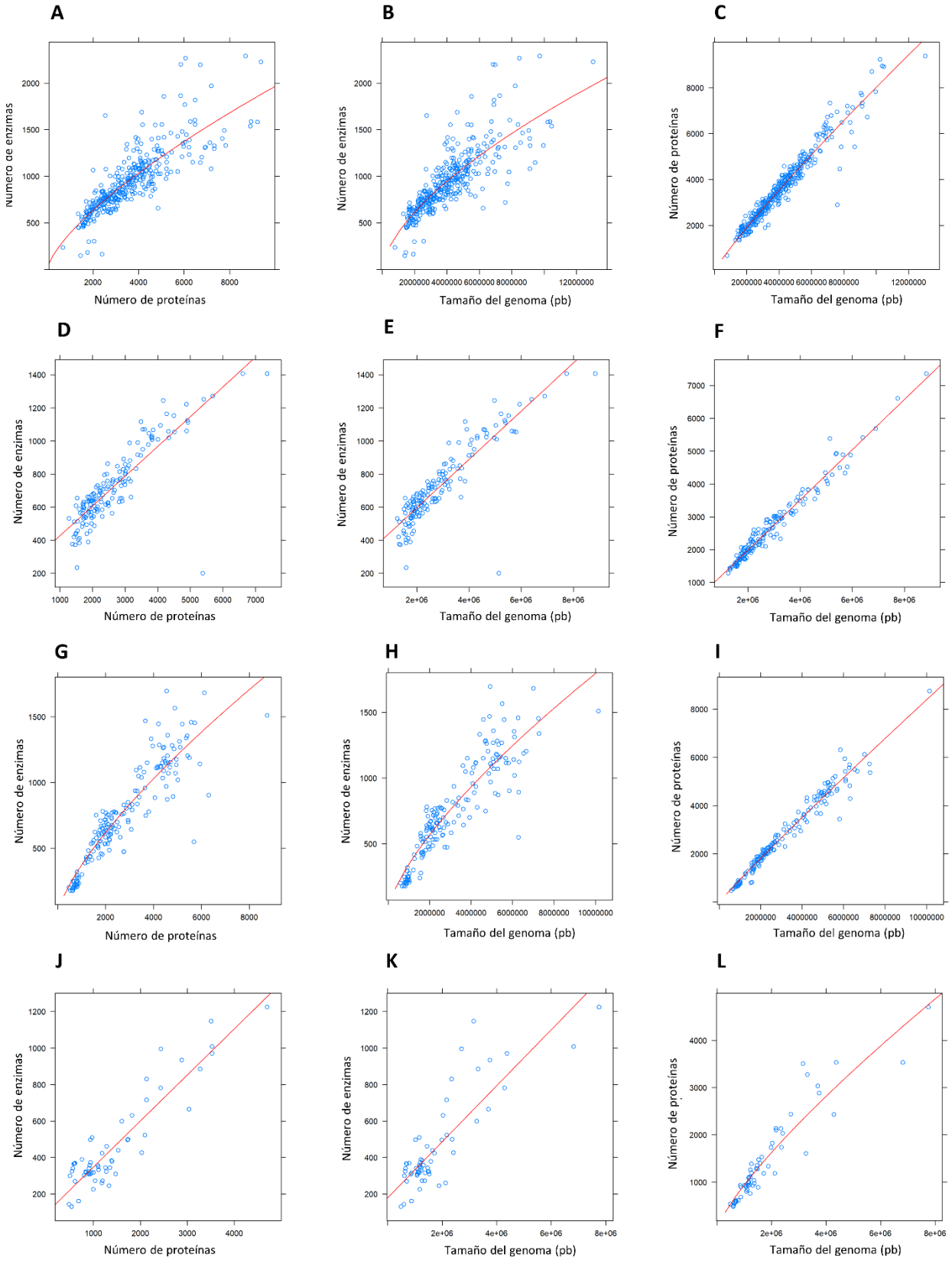

A
numero de enzimas
Número de proteínas
B
Número de enzimas
Tamaño del genoma (pb)
C
Número de proteínas
Tamaño del genoma (pb)
D
Número de enzimas
Número de proteínas
E
Número de enzimas
Tamaño del genoma (pb)
F
Número de proteínas
Tamaño del genoma (pb)
G
Número de enzimas
Número de proteínas
H
Número de enzimas
Tamaño del genoma (pb)
I
Número de proteínas
Tamaño del genoma (pb)
J
Número de enzimas
Número de proteínas
K
Número de enzimas
Tamaño del genoma (pb)
L
Número de proteínas
Tamaño del genoma (pb)

Figura S1. Relación entre el número de enzimas, número de proteínas y tamaño del genoma en los diferentes grupos de organismos que consideramos en nuestro estudio. En organismos de vida libre, una función de poder explica la distribución de los datos (**A**) $y=3.02x^{0.7}$; $R^2=0.7$; **B**) $y=0.07x^{0.63}$; $R^2=0.65$; **C**) $y=0.005x^{0.89}$; $R^2=0.94$). Los extremófilos son el único grupo de organismos en los que la relación entre las tres variables que consideramos es de tipo lineal (**D**) $y=0.18x + 247$; $R^2=0.72$, **E**) $y=1.46e^{-04}x + 306$; $R^2=0.78$; **F**) $y=7.65e^{-04}x + 457$; $R^2=0.96$). En los patógenos, al igual que en los de vida libre, la relación entre cada par de variables se ajusta mejor a una función de poder (**G**) $y=2.39x^{0.73}$; $R^2=0.83$; **H)** $y=0.02x^{0.71}$; $R^2=0.82$; **I**) $y=0.002x^{0.95}$; $R^2=0.97$). Y en los organismos intracelulares, una función lineal es la que explica mejor la relación entre el número de enzimas y proteínas (**J**) $y=0.25x + 95$; $R^2=0.84$) y entre el número de enzimas y el tamaño del genoma (**K**) $y=1.54e^{-04}\ x + 180$; $R^2 =0.71$). Por el contrario, la relación entre el número de proteínas y el tamaño del genoma se ajusta a una función de poder (**L**) $y=0.02x^{0.79}$; $R^2 =0.88$). (Figura modificada de Álvarez-Lugo & Becerra 2021).

Tabla S1. Coeficientes de determinación (R^2) para la comparación entre el número de enzimas, número de proteínas y tamaño del genoma procarionte. Se indican los coeficientes que corresponden tanto a una función de poder (parte izquierda) como a una de tipo lineal (parte derecha). En cada caso, el coeficiente con el valor más alto se indica en negritas e itálicas.

	Función de poder			**Función lineal**		
	N.° enzimas	N.° proteínas	Tamaño del genoma	N.° enzimas	N.° proteínas	Tamaño del genoma
Vida libre						
N.° enzimas	—	***0.695***	***0.654***	—	0.686	0.638
N.° proteínas	—	—	***0.938***	—	—	0.936
Extremófilos						
N.° enzimas	—	0.714	0.767	—	***0.723***	***0.782***
N.° proteínas	—	—	0.962	—	—	***0.965***
Patógenos						
N.° enzimas	—	***0.833***	***0.816***	—	0.811	0.792
N.° proteínas	—	—	***0.965***	—	—	0.961
Intracelulares						
N.° enzimas	—	0.831	0.705	—	***0.839***	***0.714***
N.° proteínas	—	—	***0.878***	—	—	0.857

Tabla S2. Valores de *P* resultantes de la prueba de Dunn para las comparaciones pareadas entre proporciones de enzimas parálogas. Las comparaciones se hicieron por cada clase enzimática. En este caso, el valor *P* fue definido como *α*/2, lo cual equivale a 0.025. Aquellos valores que no fueron significativos están sombreados en gris.

Clase enzimática		**Estilo de vida**		
		Extremófilos	Vida libre	Intracelulares
Óxidorreductadas	Vida libre	1.0000		
	Intracelulares	0.0000	0.0000	
	Patógenos	0.0000	0.0000	0.0000
Transferasas	Vida libre	0.0052		
	Intracelulares	0.0000	0.0000	
	Patógenos	0.0074	0.0000	0.0000
Hidrolasas	Vida libre	0.0005		
	Intracelulares	0.0000	0.0000	
	Patógenos	1.0000	0.0007	0.0000
Liasas	Vida libre	0.0164		
	Intracelulares	0.0000	0.0000	
	Patógenos	0.0193	0.0000	0.0000
Isomerasas	Vida libre	0.0000		
	Intracelulares	0.0000	0.0000	
	Patógenos	0.0738	0.0036	0.0000
Ligasas	Vida libre	1.0000		
	Intracelulares	0.0000	0.0000	
	Patógenos	0.0000	0.0000	0.0000
Translocasas	Vida libre	0.0002		
	Intracelulares	0.0000	0.0000	
	Patógenos	0.6449	0.0000	0.0000

Tabla S3. Análisis de correlación entre el número promedio deisomerasas duplicadas y el número de enzimas descritas para cada sub-subclase. Las columnas, de izquierda a derecha en la parte superior de la tabla, indican: sub-subclases de isomerasas, número de enzimas para cada una de ellas y número promedio de duplicados en la muestra completa, en bacterias y en arqueas. En la parte inferior se muestra la matriz de correlación, la cual se coloreó automáticamente con el Complemento de Análisis de Datos de Microsoft Excel 365.

		Número promedio de duplicados		
Sub-subclase	**No. Enzimas**	**Muestra completa**	**Bacteria**	**Archaea**
EC 5.1.1	24	0.7964	1.1954	0.1111
EC 5.1.2	7	0.0332	0.0916	0.0000
EC 5.1.3	43	3.2445	3.5756	1.9667
EC 5.1.99	8	0.1028	0.2794	0.0111
EC 5.2.1	12	1.8057	2.8885	0.8222
EC 5.3.1	32	1.8152	2.1850	2.2778
EC 5.3.2	8	0.1082	0.1147	0.0556
EC 5.3.3	19	0.2840	0.6529	0.2222
EC 5.3.4	1	0.0348	0.0736	0.0000
EC 5.3.99	10	0.0035	0.0137	0.0000
EC 5.4.1	3	0.0200	0.0366	0.0000
EC 5.4.2	12	2.5852	3.1282	2.2111
EC 5.4.3	9	1.0397	0.9160	0.9889
EC 5.4.4	8	0.0418	0.1101	0.1000
EC 5.4.99	67	3.1797	4.4687	0.9111
EC 5.5.1	34	0.1791	0.2508	0.2111
EC 5.6.1	9	0.0009	0.0046	0.0000
EC 5.6.2	2	1.5916	3.0214	0.8667
EC 5.99.1	2	0.0030	0.0138	0.0111

	No. Enzimas	*Muestra completa*	*Bacteria*	*Archaea*
No. Enzimas	1			
Muestra completa	0.671011129	1		
Bacteria	0.634349048	0.973838414	1	
Archaea	0.436054724	0.846041252	0.76777397	1