UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

Análisis de los elementos de regulación transcripcional a través de los *phyla* bacterianos y de archaeas.

TESIS

QUE PARA OPTAR POR EL GRADO DE:

Doctor en Ciencias

PRESENTA:
M.C. Joselyn Cristina Chávez Fuentes

TUTOR PRINCIPAL
Dr. Enrique Merino Pérez
Instituto de Biotecnología

MIEMBROS DEL COMITÉ TUTOR
Dr. José Luis Puente García
Instituto de Biotecnología

Dr. Miguel Ángel Cevallos Gaos
Centro de Ciencias Genómicas

Ciudad de México. Agosto, 2021

ÍNDICE

ÍNDICE DE FIGURAS

Agradecimientos

Al Instituto de Biotecnología, UNAM donde se llevó a cabo este trabajo.

Al Dr. Enrique Merino Pérez por abrirme las puertas de su grupo de trabajo, por su apoyo como tutor y su confianza en mí, por alentarme a crecer como persona y como profesionista, por todas sus recomendaciones que ayudaron a perfeccionar este proyecto y que sirvieron para nutrir mi formación como Doctora en Ciencias. Especialmente por ser no sólo un tutor, sino un amigo.

Al Consejo Nacional de Ciencia y Tecnología, por la beca de Doctorado No. 565669 que hizo posible la realización de este proyecto.

Al Consejo Nacional de Ciencia y Tecnología por el financiamiento del proyecto “Generando nuevos paradigmas de la Biología Sintética aplicados al estudio de estresomas bacterianos” pertenecientes a la convocatoria de “Fronteras de la Ciencia”, con el número de proyecto 2015-02-887 otorgado al Dr. Enrique Merino Pérez para la realización de este trabajo.

A la Dra. Rosa María Gutiérrez Ríos por todas sus recomendaciones y consejos que complementaron el desarrollo de mis habiilidades de programación durante mi proyecto.

Al personal técnico y administrativo del laboratorio de Genómica Computacional, especialmente al M.C. Ricardo Ciria Merce, por manterner nuestro servidor siempre en marcha, así como a la M.B. María Luisa Tabche Barrera y al M.C. José Luis Gama Ferrer por todo su apoyo.

Al personal administrativo del Instituto de Biotecnología, especialmente a Antonio Bolaños y Gloria Villa por su excelente trabajo y completa disposición para facilitar nuestros trámites académicos.

A todos los integrantes del laboratorio del Dr. Enrique Merino y Dra. Guadalupe Espín por su apoyo, compañerismo y amistad, deseo que podamos coincidir en el futuro como colegas.

A la Comunidad de Desarrolladores de Software en Bioinformática y el grupo R-Ladies Cuernavaca, por brindarme un espacio clave para mi desarrollo personal y profesional, además de permitirme conocer grandes personas que se convirtieron en amigos.

A mis padres y hermano, quienes me han apoyado toda la vida para realizar el sueño de convertirme en Doctora en Ciencias y están siempre dispuestos a escuchar mi entusiasmo cuando hablo de este proyecto.

A mi familia y amigos por su apoyo incondicional durante la realización de mi proyecto de Doctorado, en especial a Maritere, Leonel, Julio y Dante cuya amistad es invaluable.

Resumen

En los procariontes, los factores sigma, factores transcripcionales y riboswtiches son los principales reguladores de la expresión génica a nivel transcripcional. Cada uno de estos elementos de regulación tiene características particulares que permiten a los organismos realizar las actividades basales de la célula y responder ante señales ambientales a las cuales se han adaptado de manera evolutiva.

Para comprender su conservación a través de los grupos filogenéticos, los factores sigma y factores transcripcionales se han agrupado de acuerdo con su conservación a nivel de secuencia y función. Estas agrupaciones se pueden encontrar en diversas bases de datos como la base COG (Clusters of Orthologous Groups) y la base KEGG (Kyoto Encyclopedia of Genes and Genomes). En este trabajo se estudió la abundancia y distribución de los grupos COG y KEGG en nueve grupos filogenéticos bacterianos y dos grupos filogenéticos de archaeas a nivel taxonómico de *phylum*.

Este estudio muestra que existe una relación positiva entre la abundancia de factores transcripcionales y factores sigma con el tamaño del genoma, la cual es particular de cada *phylum*. Adicionalmente se encontró que existe una relación entre el número de riboswitches T-box y el tamaño del genoma para uno de los grupos filogenéticos estudiados, los Firmicutes. De manera interesante, se muestra que existe una tendencia de compensación entre la razón de incremento de factores transcripcionales y factores sigma en ciertos *phylum*, de manera que, ante la baja abundancia de uno de estos elementos de regulación el otro presenta una mayor frecuencia.

Adicionalmente, en este trabajo se evaluó el enriquecimiento de grupos COG y KEGG en cada grupo filogenético mostrando que existen grupos de factores transcripcionales o sigma particularmente enriquecidos que responden a las características celulares y ambientales en las cuales viven los organismos de cada *phylum*. Finalmente, en este estudio se analizó el contexto genómico de los genes que codifican para los factores transcripcionales o sigma, dado por su ubicación dentro de un arreglo en operon. Este análisis mostró que los genes que codificas a cada familia de factores sigma presenta una tendencia particular a encontrarse en un arreglo de monocistrón, cabeza de operón o dentro del cuerpo del operón.

Introducción

Los factores sigma como regulador transcripcional en bacterias

Cualquier organismo necesita contar con diversos elementos encargados de mantener la correcta regulación de la expresión de sus genes. En las bacterias se sabe que la RNA polimerasa requiere una subunidad especializada sigma que reconoce a los promotores de los genes y dirige el core catalítico de la RNA Polimerasa al sitio de inicio de la transcripción (Figura 1A); además, los factores sigma permiten el inicio de la transcripción al comenzar la separación de la doble cadena de DNA como primer paso para la formación de la burbuja de transcripción, como se recapitula en el trabajo de (Feklístov et al., 2014; Paget, 2015).

Las bacterias poseen un factor sigma esencial σ^{70} que promueve la transcripción de miles de genes durante la fase de crecimiento al reconocer y unirse a las regiones promotoras -35 y -10 de sus genes regulados, así como factores sigma alternativos que promueven la transcripción de genes específicos que pueden estar involucrados en la respuesta a estrés o en etapas específicas de crecimiento (Feklístov et al., 2014; Paget, 2015).

Entre los factores sigma alternativos se encuentran σ^{24}, σ^{28}, σ^{32}, σ^{38}, σ^{54}, y los factores sigma extracitoplásmicos ECF. Cada uno de estos factores posee funciones y características específicas, por ejemplo el factor σ^{54} conocido también como σ^{N} reconoce las regiones -12 y -24 del promotor, a diferencia del factor esencial σ^{70}. El factor σ^{54} fue originalmente caracterizado por participar en la respuesta al niveles variables de nitrógeno, posteriormente se reportó su participación en la respuesta a metales pesados, en el metabolismo de fuentes alternativas al carbono, en la biosíntesis del pili, así como en la regulación del sistema de secreción tipo III. (Barrios et al., 1999; Ishimoto & Lory, 1989; Köhler et al., 1989; Kustu et al., 1989; Leonhartsberger et al., 2001).

Los factores sigma extracitoplásmicos ECF constituyen el grupo más abundante de factores sigma. Estos factores se encuentran generalmente inactivos debido a la unión de factores anti-sigma, los cuales responden ante diversos estímulos ambientales; una vez activos, los factores ECF promueven la transcripción de genes que permiten responder ante el estímulo y con ello la supervivencia de la célula (Campagne et al., 2015; Lonetto et al., 1994; Mascher, 2013).

En contraste con las bacterias, la RNA polimerasa de las archaeas tiene una mayor similitud con la RNAP de eucariotes que con la enzima de bacterias, por lo que no poseen factores sigma. Las subunidades Rpo1, Rpo2, Rpo3, Rpo11 y Rpo6 de archaeas tienen un homólogo tanto en bacterias (subunidades β', β, α, α y ω) como en eucariotas (RPB1, RPB2, RPB3, RPB11 y RPB6); sin embargo, las subunidades Rpo5, Rpo10, Rpo12, Rpo4 y Rpo7 de archaeas únicamente tienen un homólogo en eucariotes (RPB5, RPB10, RPB12, RPB4 y RPB7). En las archaeas existe variabilidad en la composición de la RNA polimerasa del *phylum* Crenarcheota, que posee una subunidad Rpo8 con un homólogo presente en la RNAPII de eucariotas, mientras que el *phylum* Euryarchaeota no cuenta con esta subunidad (Grohmann & Werner, 2011; Koonin et al., 2007; Werner, 2013).

Los factores transcripcionales en bacterias y archaeas

Los factores transcripcionales modulan la expresión génica al unirse a regiones específicas del promotor cercanas al gene regulado, favoreciendo o impidiendo la unión de la RNA polimerasa al promotor. Estos factores se clasifican respecto a sus dominios, cuya función es detectar señales mediante la unión de un ligando o por interacción proteína-proteína; mientras que otro dominio se encarga de unirse a una secuencia blanco de DNA (Browning & Busby, 2004; Perez-Rueda et al., 2018).

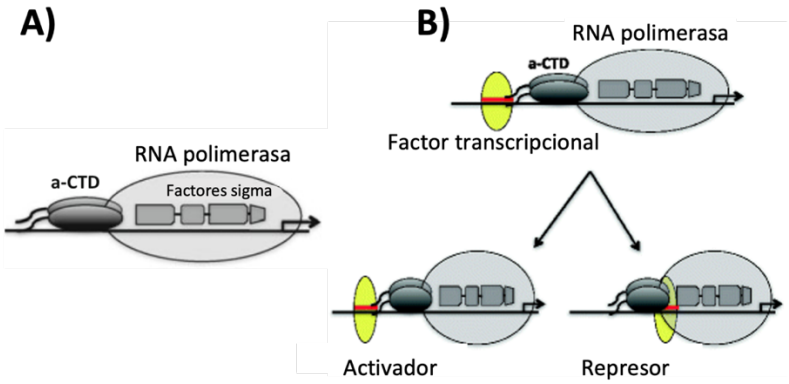

Figura 1. Modo de acción de los factores sigma y factores transcripcionales. A) Los factores sigma reconocen la región promotora y reclutan a la holoenzima de la RNA polimerasa al promotor. B) Los factores transcripcionales se unen a la región promotora actuando como activadores al promover la unión de los factores sigma y la RNA polimerasa, o como represores al impedir la unión de los factores sigma al promotor. Modificado de (Seshasayee et al., 2011).

Los factores transcripcionales pueden funcionar como represores, activadores o tener actividad dual (Figura 1B) dependiendo de la posición del promotor a la cual se unan. Los represores se unen al sitio de reconocimiento del factor sigma, bloqueando o impidiendo la unión de la RNA polimerasa e interfiriendo así con la actividad de esta enzima (Bintu, Buchler, Garcia, Gerland, Hwa, Kondev, & Phillips, 2005; Bintu, Buchler, Garcia, Gerland, Hwa, Kondev, Kuhlman, & Phillips, 2005; Browning et al., 2019). Por otro lado, los factores transcripcionales activadores usualmente se unen a regiones río arriba del promotor de sus genes blanco e interaccionan con la RNA polimerasa para promover su reclutamiento al promotor y el inicio de la transcripción (Balleza et al., 2009; Busby & Ebright, 1994; Lee et al., 2012).

Frecuentemente, la represión o activación dada por los factores transcripcionales se ve potenciada por la formación de oligómeros, es decir, la unión de múltiples factores transcripcionales (Bintu, Buchler, Garcia, Gerland, Hwa, Kondev, & Phillips, 2005). Tal es el caso del represor Lac que se une de forma homotetramérica en regiones llamadas operadores auxiliares, formando un *loop* en el DNA que impide la unión de la RNA polimerasa (Becker et al., 2013). De la misma forma, los activadores pueden inducir cambios conformacionales en el DNA del promotor que permiten el acercamiento de los elementos del promotor para facilitar el reclutamiento de la RNA polimerasa (Brown et al., 2003; Yang et al., 2015). Un ejemplo es el factor transcripcional CRP que genera una curvatura en el DNA que facilita la unión del factor MelR y promueve la activación transcripcional (Bintu, Buchler, Garcia, Gerland, Hwa, Kondev, Kuhlman, & Philips, 2005).

Los factores transcripcionales pueden tener una acción dual, como activadores y represores, dependiendo de la posición de su sitio de unión con respecto al promotor y de su concentración intracelular. La concentración intracelular de los factores transcripcionales favorece una unión diferencial a diversos sitios en el genoma (Weinstein-Fischer & Altuvia, 2007), esta función dual también puede depender de la concentración de ciertas moléculas efectoras que afectan su estructura (Jo et al., 1986). La actividad dual de los factores transcripcionales ha sido descrita principalmente en genes transcritos divergentemente, los cuales poseen regiones promotoras con diferencias en su secuencia que son reconocidas por el factor transcripcional, dando como resultado una mayor afinidad por uno de los

promotores (causando la represión del gene), y a la vez la baja afinidad por el otro promotor (llevando a la activación del gene) (Brödel et al., 2020).

Clasificación de las proteínas en grupos de ortología

Para comprender la relación entre los genes y proteínas de los organismos en los tres dominios de la vida, se han agrupado a los genes en familias de homólogos, que incluyen tanto a ortólogos como parálogos dependiendo del proceso en que se generaron. Los ortólogos se han definido como genes con un ancestro en común a partir del cual sufrieron un evento de especiación, mientras que los parálogos son genes resultado de un evento de duplicación génica. En términos generales, los genes ortólogos tienden a conservar con mayor frecuencia su función a través de la evolución, mientras que los genes parálogos pueden tener divergencia funcional con mayor frecuencia. (Fitch, 1970; Koonin, 2005).

La base de datos COG (Clusters of Orthologous Groups) clasifica a las proteínas homólogas de organismos secuenciados en grupos que consisten en genes ortólogos o grupos ortólogos de parálogos de tres o más grupos filogenéticos, lo que significa que dos proteínas pertenecientes a diferentes grupos filogenéticos asignadas al mismo COG son ortólogos. Esta clasificación asume que cada COG tiene un gene ancestral en común a partir del cual se generaron los genes secuenciados, ya sea por un evento de especiación o duplicación (Tatusov et al., 1997).

La base de datos COG realiza una clasificación bioinformática de cada proteína de un organismo A tomando en cuenta el mejor hit recíproco (Bidirectional Best Hit) con las proteínas de un organismo B. Esta agrupación, basada en criterios de homología, es complementada con la información proveniente de curaciones manuales de las anotaciones de los diversos grupos, lo que permite inferir, con cierta precisión, la función biológica de las proteínas con base al grupo COG al que pertenecen (Galperin et al., 2015; Tatusov et al., 2000).

La base de datos KEGG (Kyoto Encyclopedia of Genes and Genomes), por otro lado, relaciona la información genómica con la anotación funcional de los genes. Para lograr este objetivo, la base KEGG utiliza la información disponible sobre la participación de los genes en los procesos celulares, como las rutas metabólicas, y estandariza la anotación de los

genes de manera computacional. Esta base de datos considera la identificación de ortólogos, la incorporación de evidencia experimental disponible en la literatura acerca de la función de los genes, la asignación de números EC, y la predicción de anotaciones funcionales basadas en la construcción de vías metabólicas (Kanehisa & Goto, 2000).

Tomando en cuenta los criterios de la base COG, los factores transcripcionales de los procariontes se han clasificado en 91 grupos (Galperin et al., 2015; Makarova et al., 2015; Tatusov et al., 1997). Estos grupos de proteínas homólogas pueden ser subdivididos en subgrupos si, además de la similitud de secuencia, se toma en cuenta la caracterización funcional y los procesos celulares en los cuales participan. Así por ejemplo, el número de grupos COG en el que se agrupan los factores transcripcionales es 91, mientras que en la base de datos KEGG este número asciende a 369 grupos (Kanehisa, 2019; Kanehisa et al., 2019; Kanehisa & Goto, 2000). Los factores sigma se clasifican en 4 grupos de acuerdo con la base COG ($\sigma^D/\sigma^S/\sigma^H$, σ^F/SigB, σ^N y los factores ECF). Algunos de ellos comprenden más de un tipo de factor sigma, mientras que en la base KEGG los factores sigma se clasifican de manera que cada grupo KO comprende a un solo tipo de factor sigma, dando como resultado 9 grupos.

La regulación génica efectuada por riboswitches

Además de la regulación por factores sigma y factores transcripcionales, en los procariontes la expresión génica puede ser regulada por atenuación, que es la regulación de la expresión génica dada por la reducción en la transcripción de regiones distales de un operón (Yanofsky, 1981). La atenuación es efectuada por los riboswitches, que son elementos regulatorios transcripcionales localizados generalmente en la región 5' UTR (Untranslated Region) del mRNA donde efectúan su acción de regulación (terminación prematura del mRNA) (Merino & Yanofsky, 2002, 2005; Naville & Gautheret, 2009; Smith et al., 2010).

Los riboswitches están constituidos por dos dominios, el dominio de reconocimiento y el dominio de expresión (Figura 2). El dominio de reconocimiento forma estructuras de tallo y asa que son capaces de unir metabolitos pequeños como iones metálicos, derivados de vitaminas (tiamina pirofosfato, mononucleotido flavina, adenosilcobalamina, etc.), precursores de ácidos nucleicos (guanina y adenina), cofactores enzimáticos, aminoácidos (lisina y glicina), azúcares fosforilados y tRNAs no cargados (Barrick et al., 2004; Grundy et

al., 2003; Mandal et al., 2003; Mandal & Breaker, 2004; Nahvi et al., 2002; Rodionov et al., 2002, 2003; Winkler et al., 2002; Winkler et al., 2002).

En respuesta a la unión del ligando, la secuencia río abajo del aptámero, llamada plataforma de expresión, cambia su estructura secundaria nativa dando lugar a nuevas estructuras secundarias que determinan la respuesta regulatoria. Dicha plataforma de expresión suele estar formada por atenuadores que pueden regular la expresión de los genes río abajo transcripcional o traduccionalmente.

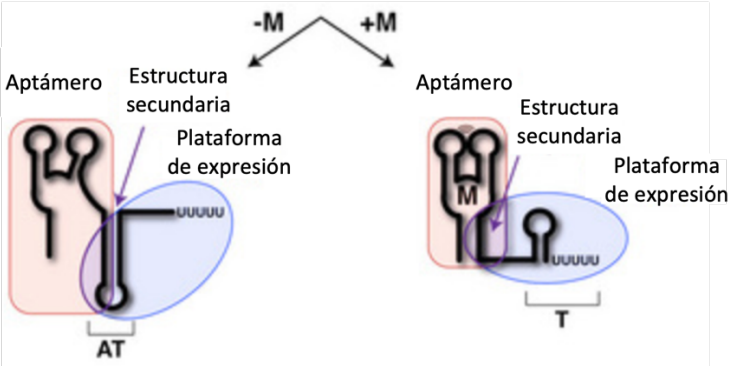

Figura 2. Dominios de los riboswitches. El aptámero reconoce la unión de un metabolito específico, ante la unión del metabolito el dominio de expresión cambia su estructura secundaria adquiriendo una actividad de atenuador transcripcional o traduccional. Modificado de (Edwards & Batey, 2010).

Los riboswitches que actúan a nivel transcripcional poseen una secuencia que dirige la formación de un terminador Rho-independiente. En el caso de las bacterias, este terminador consiste en una estructura corta de tipo tallo-asa seguida de seis o más residuos de uracilo, cuya conformación favorece que la RNA polimerasa termine el proceso de transcripción (Martin et al., 2002; Merino & Yanofsky, 2002, 2005; Naville & Gautheret, 2009). Mientras que en las archaeas, la terminación de la transcripción se da en respuesta a la presencia de secuencias ricas en residuos Timina (T) en la plataforma de expresión, de forma independiente a la formación de estructuras secundarias (French et al., 2007).

Por otro lado, los riboswitches que actúan a nivel traduccional poseen una plataforma de expresión que naturalmente forma una estructura secundaria que mantiene libre la región de unión al ribosoma; ante la unión del ligando, la plataforma de expresión cambia su conformación generando una estructura secundaria que oculta al sitio de unión del ribosoma, impidiendo así el inicio de la traducción (Caron et al., 2012; Rodionov et al., 2003).

Antecedentes

Hace más de 15 años, diversos reportes mostraron la existencia de una relación entre el número de factores transcripcionales y el tamaño del genoma en organismos modelo de bacterias y archaeas. Pérez-Rueda y colaboradores mostraron en un estudio de 90 organismos, entre bacterias y archaeas, que la frecuencia de factores transcripcionales (TFs) aumenta conforme lo hace el número de Open Reading Frames (ORFs) por genoma (Pérez-Rueda et al., 2004). Mientras que Cases y colaboradores, mostraron que la proporción de TFs por gene incrementa conforme lo hace el tamaño de su genoma (Cases et al., 2003).

Al analizar la abundancia de TFs en genomas bacterianos y la relación con su ambiente y estilo de vida, se ha reportado que las bacterias con un ambiente de vida libre suelen tener un mayor número de factores transcripcionales, incluso mayor al esperado de acuerdo con su tamaño de genoma. Mientras que en genomas muy pequeños, generalmente pertenecientes a patógenos intracelulares, se observan muy pocos factores transcripcionales, y la contribución de los factores sigma es tan significativa como la de los TFs en las bacterias de vida libre (Cases et al., 2003; Pérez-Rueda et al., 2004). Se ha observado que en situaciones de una reducción evolutiva del genoma, el número de TFs disminuye, pero los factores con actividad dual suelen mantener su frecuencia en el genoma, como ocurre en los patógenos obligados y endosimbiontes de las γ-proteobacterias, que muestran una reducción promedio de entre 10 y 15% en el número de activadores y represores en comparación con el genoma de *Escherichia coli,* pero mantienen el número esperado de TFs con actividad dual (Galán-Vásquez et al., 2016).

Pese a que los anteriores estudios dan una primera idea de la relación que existe entre el número de TFs con el tamaño del genoma de los organismos, aún no se ha realizado un estudio en el que se contemple, de manera simultánea, la relación que guardan los diferentes efectores de la regulación transcripcional (factores transcripcionales, factores sigma y riboswitches transcripcionales) en relación al tamaño de los genomas y al grupo filogenético al que pertenecen, lo que permitiría comprender si existe una tendencia particular en la regulación dirigida por alguno de estos elementos.

Hipótesis

La frecuencia relativa de las familias de reguladores transcripcionales (factores sigma, factores transcripcionales y riboswitches) en genomas procariontes, varía de acuerdo al *phylum* al que pertenecen dichos organismos.

Objetivos

Objetivo General

Caracterizar la abundancia relativa de los elementos de regulación transcripcional en los *phyla* de bacterias y archaeas, así como su relación con el tamaño de genoma y estilo de vida de cada *phylum*.

Objetivos Particulares

1. Identificar *in silico* los elementos de regulación de la expresión génica; a) factores transcripcionales b) factores sigma y c) regulación por riboswitches.
2. Evaluar la relación entre la abundancia de reguladores transcripcionales y el tamaño de genoma en bacterias y archaeas.
3. Evaluar el enriquecimiento de las familias de reguladores transcripcionales en bacterias y archaeas de acuerdo con el *phylum*.
4. Analizar el contexto genómico de los factores transcripcionales y factores sigma.

Metodología

Conteo de factores sigma y factores de transcripción

A la fecha de la realización de este trabajo, en la base de KEGG se encontraron disponibles 4,852 genomas de bacterias y 277 genomas de archaeas distribuidos en 51 grupos filogenéticos. Con el fin de evitar la redundancia y sobrerepresentación de los datos, se seleccionó el genoma con el mayor número de ORFs por especie. Adicionalmente, se tomaron en cuenta solamente aquellos *phyla* que contuvieran al menos 16 organismos, así como el *phylum* Verrucomicrobia, parte del *superphylum* PVC, lo que resultó en 11 grupos filogenéticos con 2,518 organismos bacterianos y 202 archaeas.

Utilizando el programa de asignación de COGs GeConT desarrollado previamente en nuestro grupo de trabajo con base a Modelos Ocultos de Markov (HMM) (Martinez-Guerrero et al., 2008), se clasificaron todas las proteínas codificadas en los genomas. Adicionalmente, se utilizó la asignación de grupos KO para cada proteína de los genomas disponibles en la base KEGG Orthology. Utilizando la descripción funcional de cada COG/KO se cuantificaron los genes pertenecientes a factores sigma y factores transcripcionales, así como el número de regiones codificantes CDS por organismo.

En el caso de los factores transcripcionales, se tomó en cuenta si el grupo de ortología KEGG se encontraba descrito como represor transcripcional o no y se realizó la comparación entre el número de factores transcripcionales represores y no-represores por organismo.

El conteo de reguladores transcripcionales por genoma se graficó en comparación con el tamaño de genoma. Posteriormente se realizaron modelos de regresión lineal utilizando la función lm() del lenguaje de programación R y se tomaron los valores de la pendiente, que representa el incremento en el número de reguladores transcripcionales (eje y) cuando el genoma incrementa su tamaño en 100 ORFs (eje x). Adicionalmente, se evaluó la correlación entre el tamaño de genoma y el número de reguladores transcripcionales utilizando la función cor() del lenguaje R para encontrar el índice de correlación de Pearson.

Búsqueda de riboswitches

Utilizando la base de datos Rfam (Griffiths-Jones et al., 2003), se obtuvieron los modelos de covarianza de cada riboswitch y se concentraron en una sola matriz; esta matriz se usó como entrada en el programa CMsearch (Cui et al., 2016) para realizar la búsqueda de riboswitches en una región de 400 nt río arriba del inicio de la región codificante de cada CDS. Considerando los genomas de todos los organismos estudiados, la búsqueda se realizó en un total de 9648278 regiones intergénicas. Los resultados se filtraron tomando en cuenta que las secuencias predichas tuvieran un *Score* igual o mayor al valor de confianza *Cutoff* reportado en Rfam. Ya que en algunos genes se detectó más de una secuencia correspondiente a un riboswitch, se seleccionó solamente la secuencia más cercana al extremo 5'.

Predicción de terminadores transcripcionales Rho-independientes.

Para identificar a los riboswitches de tipo terminador transcripcional Rho-independiente se realizó un programa *ad hoc* escrito en lenguaje Perl basado en la metodología reportada en (Merino & Yanofsky, 2005), utilizando el programa RNAfold (Hofacker, 2003) para predecir la formación de estructuras secundarias de RNA.

Siguiendo esta metodología, se analizaron las regiones de 50 nt río abajo de los sitios de predicción del riboswitch, buscando desde el extremo 3' hacia el 5' probables terminadores transcripcionales dados por el cumplimiento de 3 parámetros: 1) que exista una serie de al menos 4 residuos U continuos, 2) la formación de una única estructura secundaria con energía libre menor o igual a -10 kcal/mol en una ventana de 60 nt río arriba de la serie de U's y 3) que el tallo de la estructura secundaria se encuentre a 2 o menos nucleótidos de la serie de U's. Posteriormente, se contó la frecuencia de este tipo de regulador por genoma.

Predicción de terminadores traduccionales.

Para evaluar la presencia de riboswitches de tipo atenuador traduccional se tomaron en cuenta aquellos riboswitches cuya plataforma de expresión no formaba un terminador transcripcional Rho-independiente y que cumplieron los siguientes parámetros: 1) que la secuencia predicha del riboswitch se encontrara a no más de 100 nt del inicio de la traducción, 2) la formación de una estructura secundaria con energía libre menor o igual a -7 kcal/mol en una ventana de 60 nt río arriba del inicio de la traducción y 3) que el tallo de la estructura secundaria se encuentre a no más de 7 nt del inicio de la traducción. Finalmente, se contó la frecuencia de este tipo de reguladores por genoma.

Análisis de enriquecimiento de familias de reguladores

Para determinar si existen grupos de ortología COGs/KOs o familias de riboswitches Rfam enriquecidas de manera específica del *phylum*, se tomó la frecuencia absoluta de cada grupo de ortología o familia Rfam por *phylum* y se realizó una prueba de Fisher, seguida de una corrección FDR utilizando la función enrichment_fisher() del paquete de R *erba* desarrollado durante este proyecto y disponible en GitHub (https://github.com/josschavezf/erba). Posteriormente, se realizó una transformación logarítmica del valor de enriquecimiento *odd ratio* de la prueba de Fisher y los resultados se graficaron en un mapa de calor, utilizando las funciones del paquete de R ComplexHeatmap (Gu et al., 2016).

Identificación de operones

Con la finalidad de identificar el arreglo en operones en los genomas de estudio, se siguió una metodología previamente desarrollada en nuestro grupo de trabajo (Taboada et al., 2018) que toma en cuenta tanto la cercanía en la posición genómica como la relación funcional entre genes vecinos. Para seguir esta metología se tomó como entrada la asignación de grupos de ortología COG descrita previamente y los archivos .gff para cada organismo. Una vez que se identificó el arreglo en operones de cada genoma, se evaluó si los genes que codifican factores sigma o factores transcripcionales correspondían a un arreglo en monocistrón, cabeza de operón o parte del cuerpo de un operón. La frecuencia

relativa en la presencia de estos arreglos se cuantificó para cada *phylum* y se representó gráficamente.

Software usado

Para la búsqueda y conteo de factores sigma, factores transcripcionales y riboswitches se desarrollaron programas *ad hoc* en el lenguaje Perl versión 5.30. Posteriormente, para la generación de gráficas, modelos de regresión, la evaluación del índice de correlación de Pearson y el análisis de enriquecimiento se desarrolló y utilizó la paquetería *erba* en el lenguaje R versión 4.0, disponible en https://github.com/josschavezf/erba.

Resultados

El número de reguladores transcripcionales tiene una relación positiva con el tamaño del genoma.

Como primer paso en el análisis, se seleccionaron grupos filogenéticos pertenecientes a bacterias y archaeas con un número mínimo de especies secuenciadas igual a 16, siendo estos los *phyla* de estudio con su correspondiente número de organismos: Actinobacteria (371), Bacteroidetes (212), Chlamydiae (16), Crenarchaeota (46), Euryarchaeota (156), Firmicutes (487), Planctomycetes (18), Proteobacteria (1271), Spirochaetes (48) y Tenericutes (87). Adicionalmente se incluyó al *phylum* Verrucomicrobia (con 8 organismos) con la finalidad de tener presente al *superphylum* PVC (Wagner & Horn, 2006) en los organismos de estudio. De los 4,852 genomas de bacterias y 277 genomas de archaeas disponibles en la base de KEGG, se tomó un organismo representante por especie, siendo seleccionado aquel con el mayor número de regiones codificantes (CDS) en su genoma.

En este trabajo se utilizaron dos bases de datos que toman en cuenta la clasificación de proteínas por grupos de ortología, aunque con algunas diferencias. La base de datos COG agrupa las proteínas basándose en la ortología como resultado de la identificación del *Bidirectional Best Hit* (Tatusov et al., 2000), mientras que la base de datos KEGG complementa la agrupación por ortología con la anotación funcional de los genes basada en evidencia experimental (Kanehisa et al., 2016), lo que hace posible obtener su implicación en rutas de relevancia en el metabolismo. Tomando en cuenta la clasificación de las proteínas en ambas bases de datos, se contó la abundancia de los genes cuya función corresponde a factores sigma y factores transcripcionales.

Por otro lado, para evaluar la presencia de riboswitches en los genomas de estudio, se utilizaron los modelos de covarianza de los 39 riboswitches reportados en la base de datos Rfam y se siguió la metodología descrita por (Merino & Yanofsky, 2002), así como programas desarrollados durante este trabajo para la predicción de riboswitches de tipo atenuador transcripcional Rho-independiente. Finalmente se contó la abundancia de riboswitches encontrados en cada organismo y se evaluó su relación con el número de genes del genoma.

Al contar el número de factores sigma y factores transcripcionales encontrados de acuerdo con la clasificación de COG y KEGG, como se muestra en la Figura 3, se observó que el número de factores transcripcionales con asignación COG es significativamente mayor a los encontrados en KEGG. El número de factores sigma con asignación COG también fue mayor, aunque con un número más cercano al encontrado en KEGG. Esto puede deberse a que la composición y función de los factores sigma se encuentra mucho más conservada, por lo que la clasificación funcional no tendrá variaciones importantes respecto a la identidad de las proteínas basadas en su secuencia.

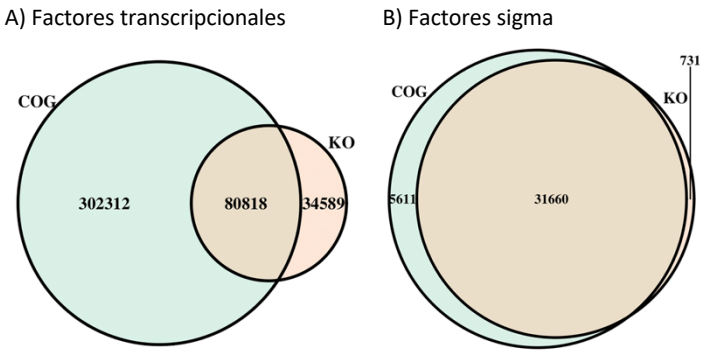

Figura 3. Número de reguladores encontrados en las bases de datos COG y KEGG. Comparación del número total de A) factores transcripcionales y B) factores sigma encontrados en los genomas de bacterias y archaeas, de acuerdo con su clasificación de grupos de ortología COG (círculo verde) y KEGG (círculo amarillo).

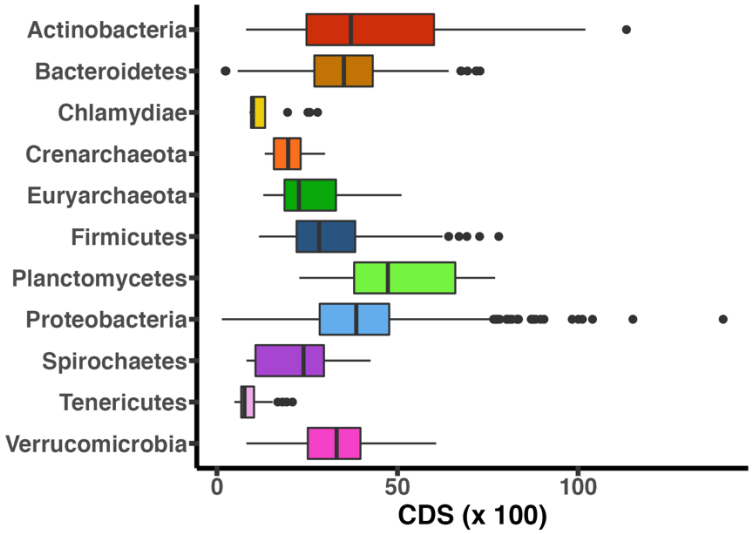

Figura 4. La distribución en el tamaño de genoma es específica de *phylum*. En el eje x se muestra el tamaño de genoma que poseen las bacterias y archaeas en unidades de 100 CDS. Cada caja muestra la distribución del tamaño de genoma de los organismos dentro de cada *phylum*.

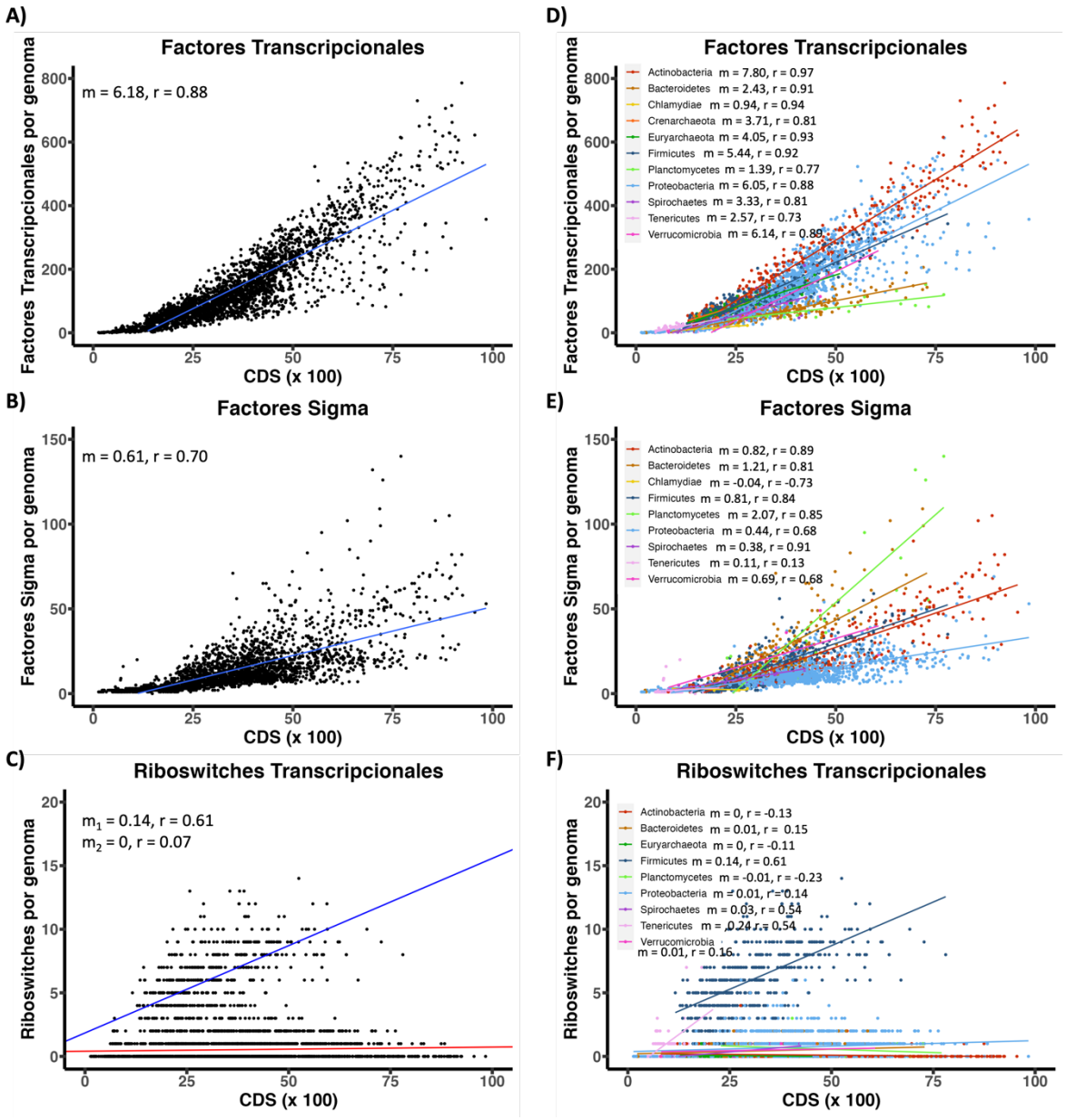

Figura 5. El número de reguladores transcripcionales incrementa conforme lo hace el tamaño del genoma, de acuerdo con la base COG. Se muestra el número de A) factores transcripcionales, B) factores sigma y C) riboswitches de tipo terminador transcripcional Rho-independiente por genoma contra el número de CDS. D-F) El número de reguladores se separó por color de acuerdo con el *phylum* de los organismos. Las pendientes del modelo de regresión de lineal (m) y el coeficiente de correlación de Pearson (r) se muestran sobre la gráfica.

Clasificación de KEGG

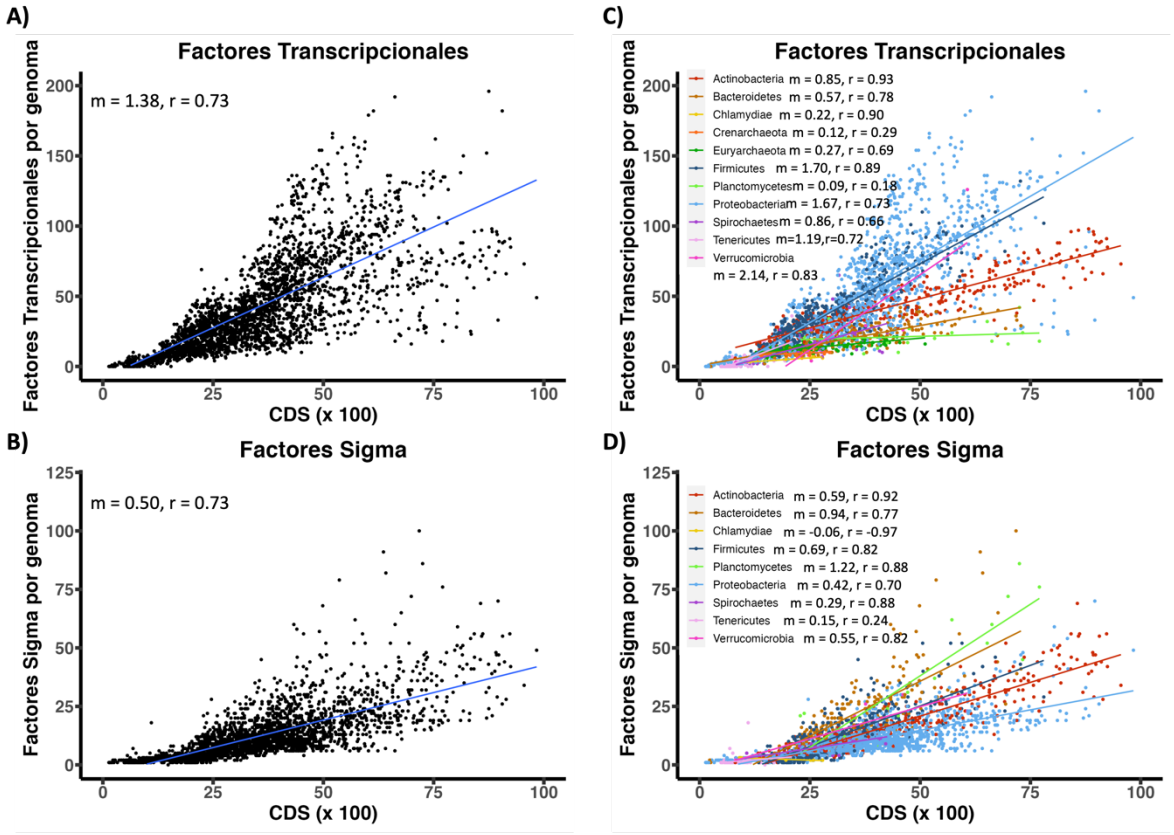

Figura 6. El número de reguladores transcripcionales incrementa conforme lo hace el tamaño del genoma, de acuerdo con la base KEGG. Se muestra el número de A) factores transcripcionales y B) factores sigma por genoma contra el número de CDS. C-D) El número de reguladores se separó por color de acuerdo con el *phylum* de los organismos. Las pendientes del modelo de regresión de lineal (m) y el coeficiente de correlación de Pearson (r) se muestran sobre la gráfica.

Adicionalmente, se evaluó la distribución de cada grupo filogenético en cuanto al tamaño de genoma, dado por el número de regiones codificantes (CDS) proporcionadas en la anotación de los genomas de la base de datos KEGG. Como se puede observar en la Figura 4, algunos *phyla* como los Tenericutes y las Chlamydias únicamente poseen genomas pequeños, de hasta 2,000 o 3,000 CDS, mientras que las Actinobacterias y Proteobacterias comprenden organismos de muy diversos tamaños de genoma, desde los más pequeños (~1,000 CDS) hasta los más grandes (10,000 o más CDS).

Al comparar el número de reguladores transcripcionales contra el número de CDS por genoma, como se muestra en las Figuras 5 y 6, se observó que el número de factores transcripcionales y factores sigma incrementaba conforme lo hacia el tamaño del genoma. Esta relación se vio reflejada en su pendiente (m) ya que, al incrementar el tamaño de genoma en una unidad de 100 CDS, el número de factores transcripcionales aumentó en 6.18 y 1.38 unidades, mientras el número de factores sigma incrementó 0.61 y 0.50 unidades de acuerdo con la clasificación de COGs y KOs, respectivamente. Para corroborar si existe una correlación entre el número de reguladores y el número de CDS, se realizó un análisis de correlación representado por el coeficiente de Pearson (r) para cada grupo de organismos.

Por otro lado, los riboswitches de tipo terminador transcripcional Rho-independiente mostrados en Figura 5C presentaron dos comportamientos diferentes, el grupo de los Firmicutes mostró una relación positiva reflejada en su pendiente (m=0.14), mientras que el resto de los grupos filogenéticos tuvieron una pendiente cercana a cero, lo que significa que en estos grupos no existen variaciones importantes en el número de riboswitches aún con el incremento en el tamaño del genoma.

La relación entre el número de reguladores transcripcionales y el tamaño del genoma es *phylum*-dependiente y muestra un efecto de compensación.

Con el fin de evaluar si la relación entre el número de reguladores transcripcionales y el tamaño del genoma es dependiente del grupo filogenético al que pertenece el organismo analizado, se separaron los conteos de reguladores transcripcionales por organismo de acuerdo con el *phylum* correspondiente.

Cada grupo filogenético mostró una tasa de incremento específica en los reguladores transcripcionales respecto al número de CDS por genoma, como se muestra en las Figuras 5D-F y 6C-D. Adicionalmente, se observó un efecto de compensación en la abundancia de los elementos de regulación cuando uno de ellos está poco representado en el *phylum*. Por ejemplo, los grupos Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria, Spirochaetes y Verrucomicrobia presentaron un mayor número de factores transcripcionales en comparación al número de factores sigma por organismo, como se muestra en la Figura 7, mientras que en el grupo Planctomycetes se observó un mayor número de factores sigma en comparación con el número de factores transcripcionales, y la tasa de incremento de

factores sigma con el tamaño de genoma también fue mayor que los TF. Por otro lado, en las Chlamydias se observó una tasa de incremento poco significativa de factores sigma con respecto al genoma y, al igual que en los Tenericutes, la proporción de factores transcripcionales y factores sigma por genoma fue poco variable.

En cuanto al conteo de riboswitches, los Firmicutes fueron por mucho, el grupo con mayor número de riboswitches por genoma, dado principalmente por la presencia del riboswitch T-box; mientras que el número de riboswitches en otros grupos filogenéticos se mantuvo relativamente constante e independiente del tamaño de genoma (Figura 5F).

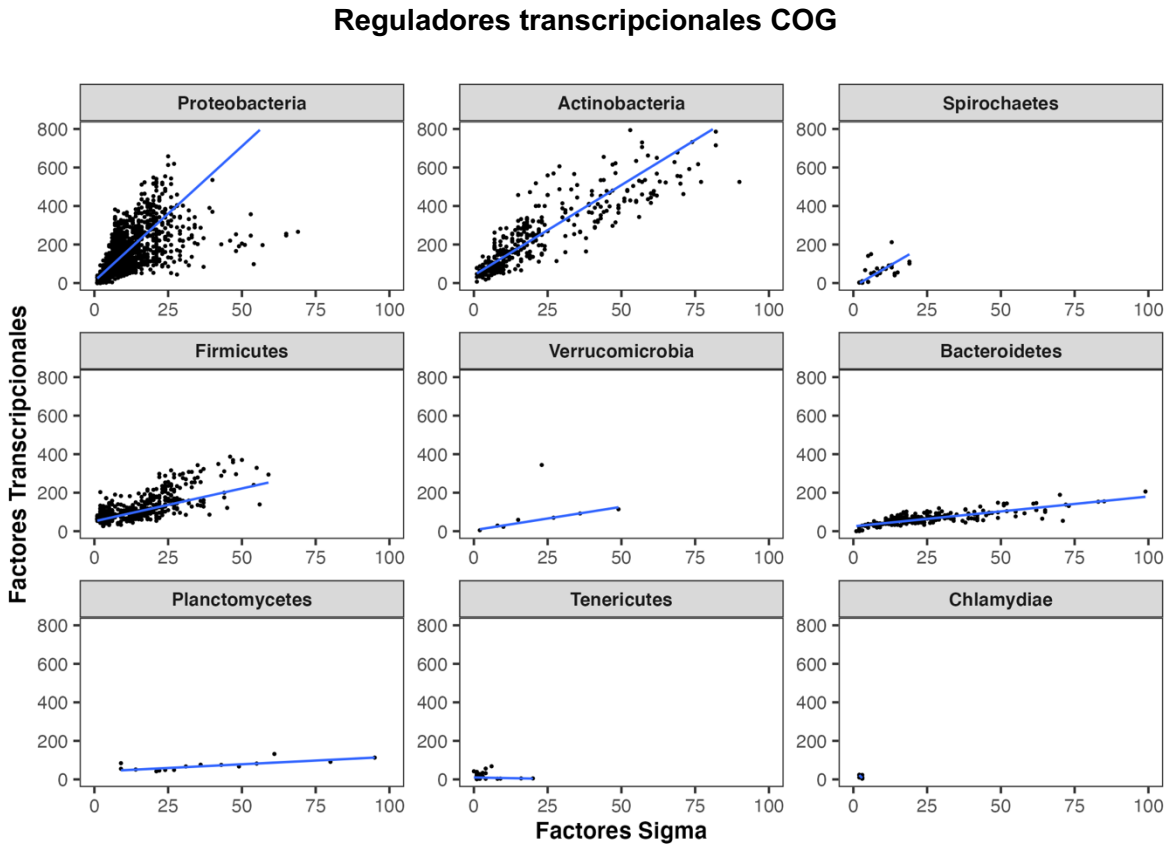

Figura 7. La tasa de factores transcripcionales por factor sigma es específica de *phylum*. Se muestra el número de factores transcripcionales y factores sigma por organismo para los diferentes *phyla*. La mediana de la proporción de TFs por factor sigma se muestra como el valor de la pendiente en la regresión lineal.

El enriquecimiento de familias de reguladores es característica para cada *phylum*.

Para evaluar si existen familias de reguladores particularmente enriquecidas en ciertos *phyla*, se realizó una prueba de Fisher utilizando las frecuencias de cada familia de regulador (factores sigma, factores transcripcionales y riboswitches) por *phylum* y se tomó el $\log_{10}$(odd.ratio) como valor de enriquecimiento. Esta prueba evalúa si el número de reguladores de una familia, clasificada con su clave COG, KO o Rfam, muestra un enriquecimiento al ser comparada contra tres valores: 1) El total de reguladores dentro del *phylum*, 2) El número de reguladores de la misma familia en los otros *phyla* y 3) El número de reguladores de las familias restantes en los otros *phyla*. Estos valores de enriquecimiento se colocaron en un mapa de calor, donde los valores mayores a cero representan que la familia se encuentra enriquecida en el *phylum* y los valores por debajo de cero representan que la familia tiene una baja representación en comparación con los otros *phyla*.

Al analizar los valores de enriquecimiento (Figura 8A), se observó que la mayoría de las familias de TFs de la base COG se encontraron presentes a través de los *phyla*; entre los más ampliamente distribuidos se encuentran DtxR (COG1321), miembros de la familia de TFs con dominios helix-turn-helix (HTH)-like (COG2865) y la familia GntR (COG1725). Por otro lado, algunas familias se encontraron mayormente representadas en ciertos *phyla,* como el caso de las archaeas (Euryarchaeota y Crenarchaeota), donde la familia de TFs con dominio HTH (COG3373), miembros de la familia de TFs involucrados en biosíntesis de tiamina (COG1992) y la familia de TFs con función desconocida (COG1709) se encontraron altamente representados. En los Firmicutes, algunas familias enriquecidas fueron el regulador CodY (COG4465), el factor de transcripción ComK (COG4903), y algunos TFs predichos que contienen dominios CBS (COG4109). Mientras que, en Proteobacterias las familias más enriquecidas fueron MltR (COG3722) y los TFs dependientes de σ^{54} (COG4650).

Posteriormente, se realizó el mismo análisis tomando en cuenta la clasificación funcional de los factores transcripcionales de los grupos de ortología KOs (Figura 8B) y se observó que un número importante de familias de factores transcripcionales se encontraron únicamente en ciertos *phyla* (observados como bloques rojos en un solo *phylum),* y completamente ausentes en otros *phyla* (observado como bloques azul obscuro); este fue el caso de las Proteobacterias, Firmicutes, Euryarchaeota y Actinobacteria.

El *phylum* Proteobacteria presentó el mayor número de familias de TFs con clasificación KO, ya que se encontraron miembros de 301 diferentes grupos, de los cuales 114 fueron exclusivos de este *phylum*. Entre los TFs únicamente presentes en Proteobacterias se encontraron 25 grupos KO pertenecientes a la familia LysR, 21 KOs de la familia LuxR, 18 de la familia AraC y 12 de la familia TetR/AcrR.

En Firmicutes, se encontraron de manera exclusiva 28 grupos KOs de factores transcripcionales pertenecientes a las familias MarR con dominio helix-turn-helix y MerR. Por otro lado, los 4 grupos KO encontrados de manera exclusiva en Actinobacterias corresponden a la familia WhiB-like de factores transcripcionales, previamente reportados como exclusivos de este *phylum*.

En el *phylum* Bacteroidetes se encontraron solamente 2 grupos KO de forma exclusiva, correspondientes a las familias CRP/FNR y HTH-type. De forma interesante en Euryarchaeota y Crenarchaeota, ambos *phyla* de las archaeas, se encontraron 4 grupos KO exclusivos cuya función aún es desconocida.

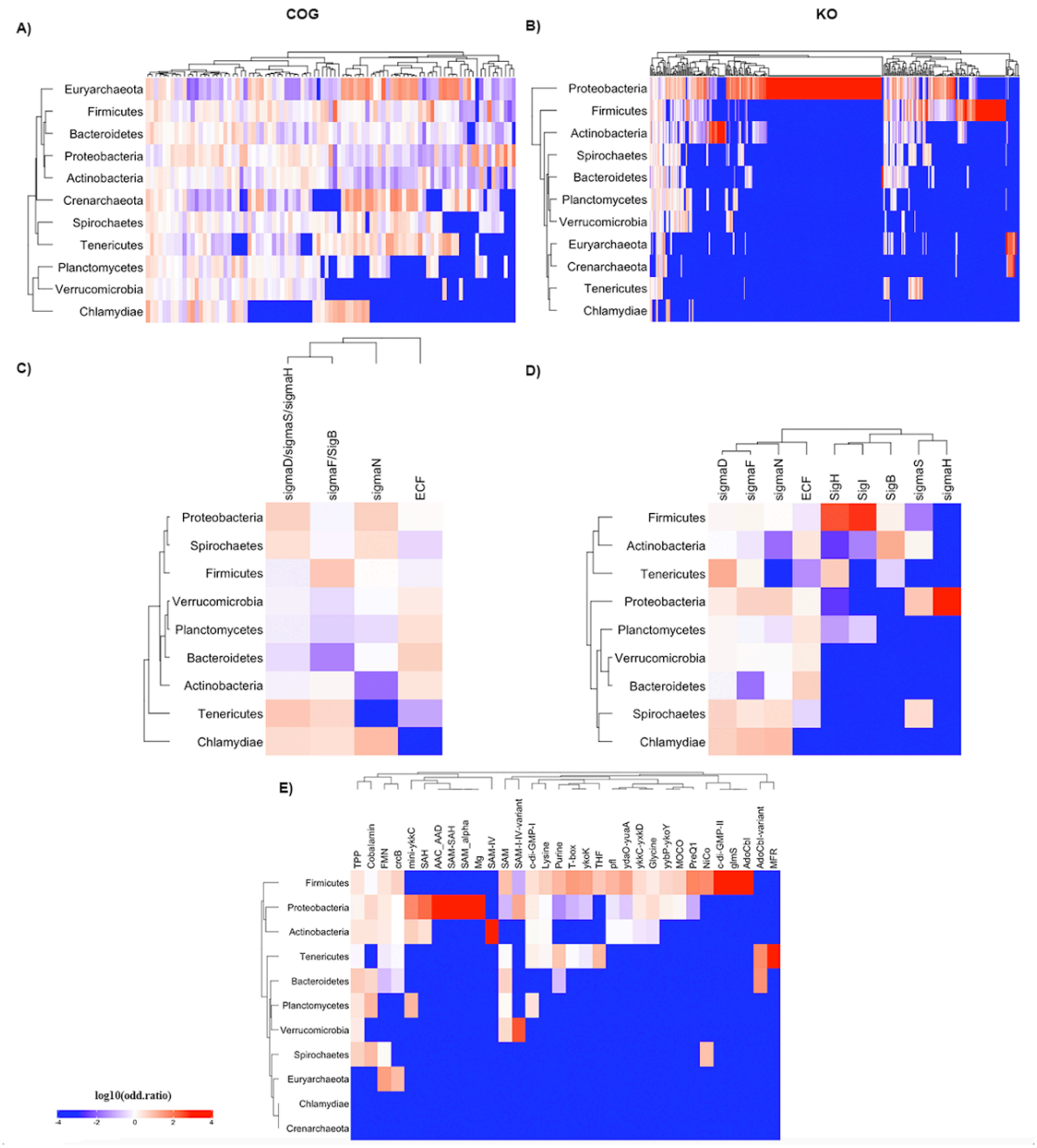

Figura 8. El enriquecimiento de las familias de reguladores transcripcionales es característico del *phylum*. Se muestra el enriquecimiento dado por el $\log_{10}$(odd.ratio) de cada familia de factores transcripciones clasificados mediante A) COGs y B) KOs, así como el enriquecimiento de los factores sigma clasificados por C) COGs y D) KOs. E) Se muestra el enriquecimiento de riboswitches de tipo terminador transcripcional Rho-independiente de acuerdo con su clasificación en Rfam.

Al analizar el enriquecimiento de las familias de factores sigma (Figura 8C-D) se observó que, con excepción de los Tenericutes y Chlamydias, los *phyla* contienen al menos un elemento de cada familia COG de factores sigma, mientras que solamente los factores σ^D y σ^F se encontraron en todos los *phyla* cuando se tomó en cuenta la clasificación por KOs, donde los Firmicutes y Actinobacterias presentaron la mayor diversidad de familias. Es sabido que los factores extracitoplásmicos tienen una alta abundancia y diversidad en las bacterias; sin embargo, estos factores no se encontraron presentes en las Chlamydias, *phylum* en el cual se encontraron enriquecidos los factores σ^D, σ^F y σ^N. Por otro lado, los factores σ^H, SigH y SigI se encontraron particularmente enriquecidos en los *phyla* Proteobacteria y Firmicutes, respectivamente (Figura 8D).

En cuanto a los riboswitches, cabe mencionar que, de los 2,674 organismos estudiados, solamente se identificaron riboswitches de tipo terminador transcripcional Rho-independiente en 1,103 de ellos. Como se puede observar en la Figura 8E, algunos riboswitches se encontraron de forma exclusiva en ciertos *phyla*, tal es el caso de las Proteobacterias, donde se encontraron dos riboswitches con unión a SAM y el sensor de Mg. En los Tenericutes se encontraron de forma exclusiva riboswitches con respuesta a Purina y Cobalamina; en las Actinobacterias se encontraron riboswitches particulares de respuesta a SAM y Guanidina, los Euryarchaeota mostraron de forma específica al riboswitch creB con respuesta a Flúor y los Verrucomicrobia tuvieron específicamente al riboswitch variante SAM I-IV.

Por otro lado, se observó un particular enriquecimiento de los riboswitches con unión a c-di-GMP-II, Cobalamina, GlcN6p y PreQ1 en el *phylum* Firmicutes; además, en este *phylum* se encontró particularmente enriquecido el riboswitch T-box, que fue el riboswitch con mayor abundancia e incremento relacionado con el tamaño de genoma, como se mostró en la Figura 4F. De manera interesante, ningún riboswitch de tipo atenuador transcripcional fue detectado en los *phyla* Chlamydia y Crenarchaeota.

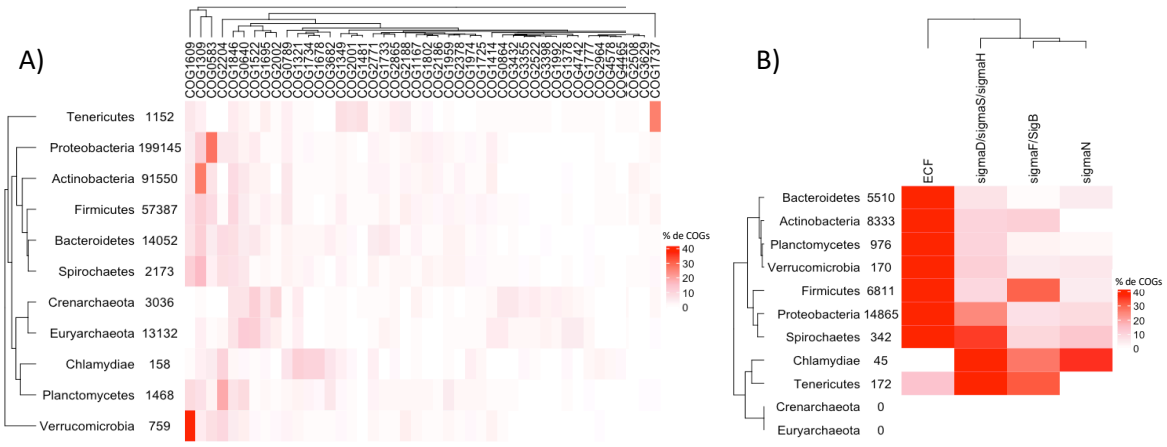

Figura 9. Existen familias de reguladores enriquecidas dentro de cada *phylum*. Se muestra el porcentaje de abundancia de cada familia de A) Factores transcripcionales y B) Factores sigma. Para cada *phylum* se tomó como referencia el número total de TFs o factores sigma. Para los TFs se muestran solamente las familias con al menos una abundancia de 3%.

Debido a que el análisis de enriquecimiento mostrado en el Figura 8 permite comparar únicamente a las familias de reguladores entre los *phyla* y con la finalidad de comparar el enriquecimiento de las familias de reguladores dentro del propio *phylum*, se tomó en cuenta el porcentaje de abundancia de cada familia COG de factores sigma y factores transcripcionales, tomando como 100% el número de TFs o SFs totales encontrados dentro del *phylum*.

Como se puede observar en la Figura 9A, se encontraron familias de TFs particularmente abundantes en ciertos phyla, como la familia AcrR en Actinobacterias, la familia LysR en Proteobacteras, la familia MurR/RpiR con dominios HTH y SIS en Tenericutes, y la familia LacI/PurR en Verrucomicrobia. Por otro lado, en la Figura 9B se puede observar que los factores sigma de tipo ECF se encontraron altamente representados en la mayoría de los *phyla*, mientras que el COG que comprende a los factores σ^D, σ^S y σ^H se encontró preponderantemente abundante en Chlamydias y Tenericutes. Adicionalmente el factor σ^N se encontró muy abundante en Chlamydias mientras que en otros *phyla* no representó la mayor contribución a los factores sigma de los organismos. Este análisis también se realizó a un nivel taxonómico más profundo, tomando las Clases dentro del *phylum* Proteobacteria como un grupo de prueba, donde se observaron diferencias en la abundancia de ciertos reguladores como los factores sigma ECF y los factores transcripcionales NtrC y LysR.

Existe una tendencia en el uso de riboswitches de tipo atenuador transcripcional y traduccional dependiente del *phylum*.

Durante el proceso de identificación de riboswitches en las regiones intergénicas, se detectaron posibles secuencias de riboswitches en algunos genomas de Chlamydiae y Crenarchaeota, pero no cumplieron con los criterios establecidos para ser considerados riboswitches con terminador transcripcional, por lo que se realizó la búsqueda de riboswitches de tipo atenuador traduccional para complementar el análisis.

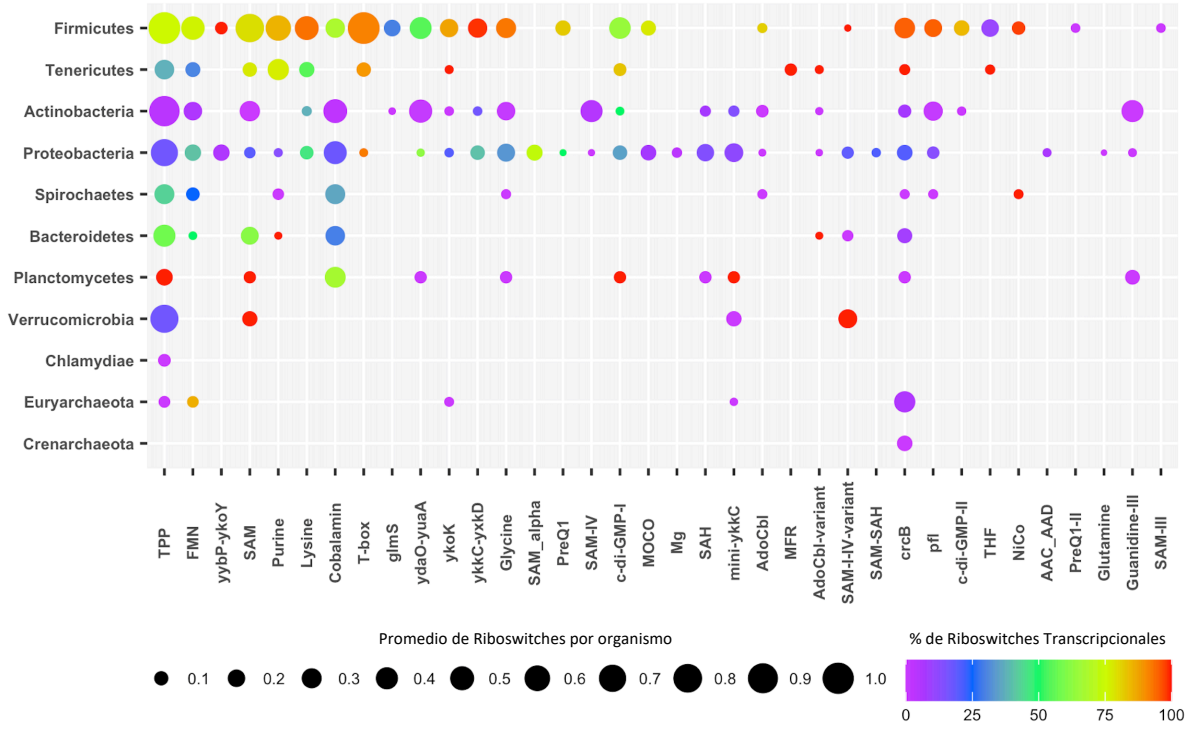

Figura 10. Comparación de la abundancia de riboswitches con acción de atenuador transcripcional o traduccional. Para cada *phylum* se muestra el número de riboswitches promedio de cada familia Rfam, dado por el tamaño del círculo. La escala de color muestra el porcentaje de riboswitches que actúan como atenuadores transcripcionales (rojo) o traduccionales (morado) en cada familia de riboswitches y para cada *phylum*.

Con la finalidad de comparar el porcentaje de riboswitches que actúan a nivel transcripcional o traduccional en cada *phylum*, se tomó el promedio de riboswitches encontrados por *phylum* y la proporción de ellos que corresponde a cada tipo de riboswitch para cada familia de Rfam. Como se muestra en la Figura 10, este análisis permitió

determinar que los riboswitches encontrados en los *phyla* Chlamydiae y Crenarchaeota corresponden completamente a atenuadores traduccionales. Adicionalmente, se observó que ciertos *phyla* tienen una tendencia a poseer riboswitches de tipo terminador traduccional, como es el caso de las Actinobacterias, donde la mayoría de los riboswitches predichos es de tipo atenuador traduccional y solamente el riboswitch c-di-GMP-I posee una proporción igualitaria de riboswitches con actividad de terminador transcripcional y terminador traduccional.

El contexto genómico de los reguladores transcripcionales es específico de los grupos filogenéticos.

Con el fin de determinar si la preponderancia de ciertas familias de factores transcripcionales y factores sigma va acompañada de una composición específica en el contexto genómico, se evaluó la presencia de operones en los genomas, siguiendo la metodología descrita por (Taboada et al., 2018). Una vez que se determinó la presencia de operones, se evaluó para cada gen, que codifica un factor transcripcional o un factor sigma, si su posición correspondía a ser cabeza de operón, si formaba parte del cuerpo del operón o si se encontraba como monocistrón.

Como se puede observar en la Figura 11A, una importante proporción de TFs se encontró en arreglo de monocistrón y la proporción de TFs que se encontraron como parte del cuerpo del operón o como cabeza de operón fue variable entre los diferentes *phyla*. Por otro lado, los factores sigma (Figura 11B) se encontraron mayormente como cabeza de operón en algunos *phyla*, como el caso de los Bacteroidetes y Planctomycetes, mientras que en las Chlamydias y Tenericutes se encontraron mayoritariamente como parte del cuerpo del operón. Considerando que las Chlamydias y Tenericutes presentaron pocos o ningún factor ECF en los análisis previos, se separaron los conteos de los factores sigma por familia COG (Figura 11C) encontrando de manera interesante que la mayoría de los factores ECF fueron cabeza de operón, mientras el grupo de los factores sigmaD, sigmaS y sigmaH se encontraron en la mayoría de los casos como parte del cuerpo del operón. De forma interesante, ninguno de los factores sigmaN o sigmaF/B de las Chlamydias se encontró como cabeza de operón.

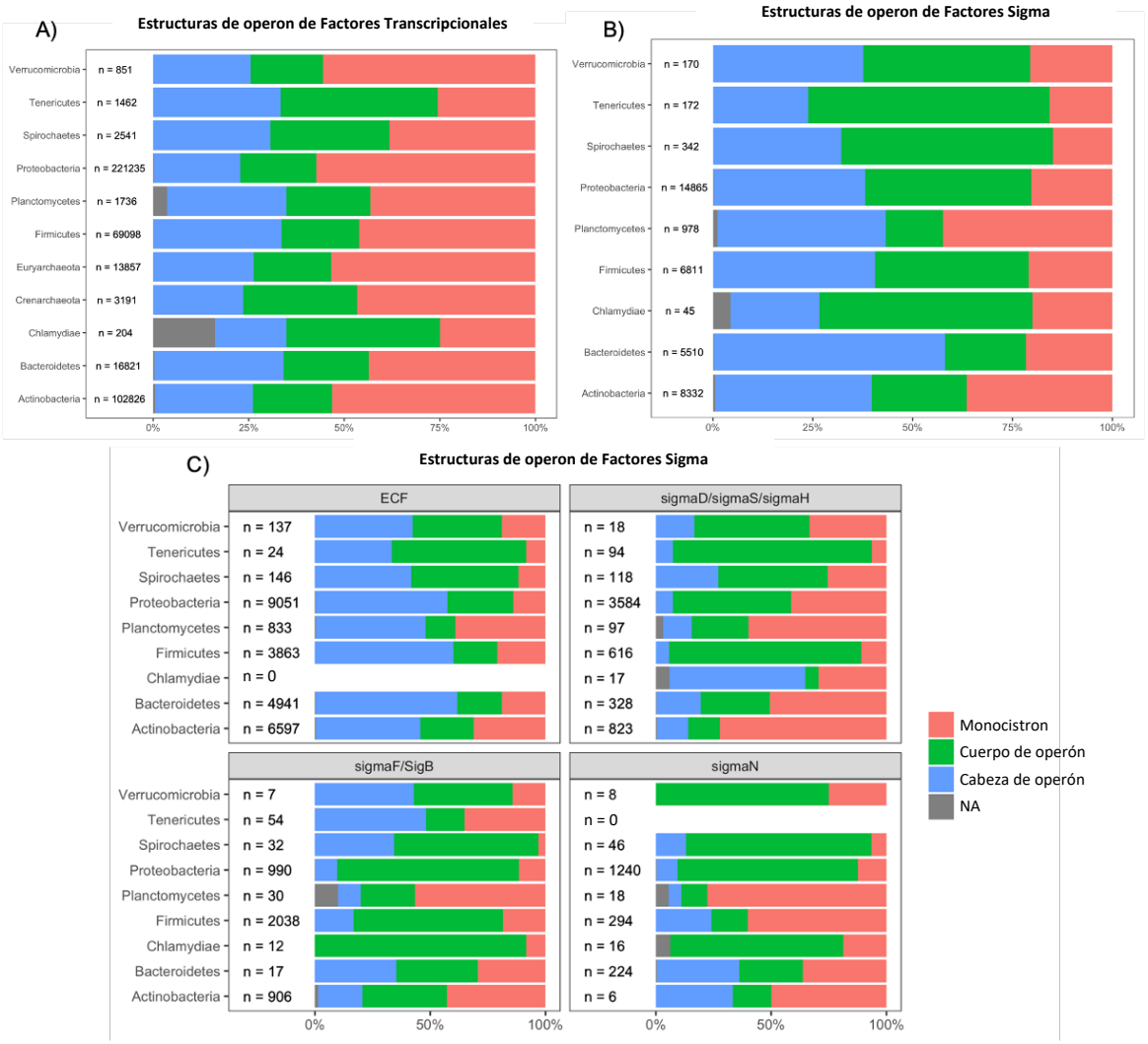

Figura 11. Conformación de operones de los reguladores transcripcionales. Se muestra el número de A) Factores transcripcionales, B) Factores sigma y C) Factores sigma separados por familia COG que corresponden a una estructura de monocistrón, cabeza de operón o cuerpo de operón. Los porcentajes se encuentran normalizados conforme al total de factores sigma o factores transcripcionales de cada familia de factores para cada *phylum*.

Discusión

En este trabajo se analizaron de manera conjunta a los tres elementos de regulación transcripcional de bacterias y archaeas dada por factores transcripcionales, factores sigma y riboswitches, así como su relación con el tamaño de genoma, distribución y enriquecimiento a través de los diversos grupos filogenéticos. En este trabajo se muestra que la frecuencia tanto de los factores transcripcionales como de los factores sigma guarda una relación positiva con el tamaño del genoma de los organismos. A pesar de que la regulación por TFs suele ser preponderante, encontramos grupos filogenéticos con una relación inversa en la frecuencia de TFs y factores sigma, describiendo por primera vez la existencia de un efecto de compensación entre la abundancia de estos dos reguladores, de manera que, ante la baja abundancia de factores transcripcionales existe un mayor número de factores sigma codificados en el genoma. Adicionalmente, este trabajo mostró que los riboswitches suelen tener una frecuencia invariante respecto al tamaño de genoma, con la excepción de los Firmicutes que parecen tener una regulación más amplia dada por el riboswitch T-box.

La comparación en la frecuencia relativa de los reguladores transcripcionales, considerando dos bases de datos ampliamente utilizadas para la clasificación de proteínas, nos permitió determinar la importancia de la asignación funcional en la búsqueda de reguladores particulares de cada *phylum* o clase. Por ejemplo, al tomar en cuenta la clasificación KEGG de los factores transcripcionales, se pudo verificar que el grupo WhiB-like se encuentra de manera específica dentro del *phylum* Actinobacteria, lo que coincide con los reportes previos; mientras que al considerar su asignación COG, estas proteínas se encuentran en una clasificación más general que no permite su identificación dentro de un *phylum* en particular. Esta familia de factores transcripcionales se ha descrito previamente en Mycobacterias por su participación en la respuesta a óxido nítrico y cAMP como parte de la regulación del anabolismo de lípidos durante la infección o virulencia (Singh et al., 2009; L. J. Smith et al., 2010). También se ha reportado que en Streptomyces, uno de los géneros más extensos de las Actinobacterias, esta familia de TFs está involucrada en patogenicidad, resistencia a antibióticos, y es esencial en el control de la diferenciación celular hacia la esporulación (Bush, 2018; Bush et al., 2016).

En este trabajo se identificaron grupos de ortología exclusivos de los *phyla* Euryarchaeota y Crenarchaeota, los cuales resultan de gran interés para futuros estudios que podrían llevar al descubrimiento de vías de regulación específicas de las archaeas.

Al analizar el enriquecimiento de los factores sigma, se observó que los factores extra citoplásmicos (ECF) no se encontraron particularmente enriquecidos a pesar de su gran frecuencia en los genomas, esto se debe a que en general estos factores se encuentran ampliamente distribuidos en las bacterias, por lo que no representan un tipo de factor específico de *phylum*. En *Escherichia coli* estos factores son importantes en la respuesta a estrés osmótico, choque de calor, estrés oxidativo o altos niveles de proteínas no plegadas (Alba et al., 2002; Rowley et al., 2006). Por lo tanto, su amplia distribución entre los grupos filogenéticos y su alta frecuencia dentro del genoma reflejan la capacidad de respuesta de los organismos a las condiciones ambientales.

Por el contrario, resultó notorio que las Chlamydias no poseen factores ECF, y muestran un enriquecimiento en los factores sigmaN y sigmaF. En *Salmonella* el factor sigmaF regula el operón flagelar y componentes del sistema de secreción necesarios para la infección celular (Eichelberg & Galán, 2000; Ohnishi et al., 1990); sin embargo en Chlamydia, donde no existe un flagelo, sigmaF se expresa durante el ciclo de infección y en respuesta a choque térmico (Shen et al., 2004), lo que parece suplir las funciones de los factores sigma ECF y puede explicar la ausencia de tales factores en este *phylum*. El factor sigmaN ha sido estudiado por su función como regulador de genes de motilidad, resistencia a estrés, patogenicidad, transporte y biosíntesis de componentes extracelulares (Riordan & Mitra, 2017). En el *phylum* Chlamydia no se ha comprendido completamente cuál es su función ya que la mayoría de los componentes nitrogenados son tomados de la célula huésped, sin embargo algunos reportes muestran que sigmaN se expresa durante las fases activa y persistente de la infección (Douglas & Hatch, 2000; Gérard et al., 2002) y controla la expresión de genes necesarios para el paso de la célula desde cuerpos replicativos reticulados hacia células infecciosas, entre ellos algunos componentes de remodelación de la membrana celular, que son críticos para determinar la capacidad infecciosa de las Chlamydias (Soules et al., 2020).

Es importante señalar que debido a que la clasificación por COGs contempla únicamente la ortología basada en similitud de secuencia, una familia COG puede contener proteínas de varios grupos KO. Esto se observó claramente en los factores sigma, donde el COG0568

contiene proteínas correspondientes a los factores sigmaD, sigmaS y sigmaH. Esto también explica por qué existen KOs únicamente en ciertos *phyla*, denotado por los cuadros rojos del mapa de calor, mientras que no se encontraron grupos COG con este nivel de enriquecimiento. Adicionalmente, la mayor caracterización de proteínas en ciertos *phyla,* y la falta de ella en otros, puede generar una mayor diferencia en el enriquecimiento de las familias de reguladores estudiados utilizando la base KEGG.

En cuanto al análisis de enriquecimiento de los diferentes tipos de riboswitches de tipo terminador transcripcional, llamó la atención encontrar al riboswitch T-box en el *phylum* Proteobacteria. En las Proteobacterias no es común encontrar este riboswitch, de hecho, se conoce que en *E. coli* no existe este elemento; sin embargo, un pequeño grupo (12 organismos) de Deltaproteobacterias presentó un riboswitch tipo T-box con terminador transcripcional. Considerando que este elemento se encontró siempre en la región 5' UTR de genes de la familia LeuA, involucrada en la biosíntesis de leucinas, su presencia podría deberse a un evento de transferencia horizontal.

Por otro lado, se encontró que ciertos *phyla* tienen una tendencia a poseer un mayor número de riboswitches de tipo terminador traduccional y un bajo o nulo número de riboswitches de tipo atenuador transcripcional, como fue el caso de las Chlamydias y Crenarchaeotas. Se tiene la hipótesis que esto podría deberse a variaciones en la velocidad de transcripción, ya que mientras más rápida sea la transcripción sería más difícil que un riboswitch pueda actuar como atenuador transcripcional y tendría que haber sufrido una adaptación evolutiva para actuar a nivel traduccional.

Por último, al analizar el contexto genómico de los factores transcripcionales y factores sigma, dado por el arreglo de los genes en operones, se encontraron comportamientos particulares (en ocasiones inesperados) dentro de cada *phylum*. Esto abre nuevas líneas de estudio que permitirán comprender mejor cómo las vías de regulación han sufrido adaptaciones tanto en la presencia o ausencia de reguladores, como en el arreglo de estos en el genoma. El estudio de los genes que se encuentran localizados dentro del mismo operón de una familia de reguladores transcripcionales y su posición dentro de este arreglo genómico nos permitiría comprender mejor las adaptaciones evolutivas de los componentes de la regulación génica en respuesta al ambiente en el cual vive cada especie.

Conclusiones

1. La abundancia de factores transcripcionales y factores sigma tiene una relación positiva con el tamaño del genoma.
2. La relación en el incremento de los factores transcripcionales y factores sigma con el tamaño del genoma es específica de cada *phylum*.
3. Existe un efecto de compensación en la relación de incremento de los factores transcripcionales y factores sigma con el tamaño del genoma.
4. Los abundancia de riboswitches no tiene variaciones significativas con el tamaño del genoma, con excepción de los Firmicutes.
5. El enriquecimiento de las familias de reguladores transcripcionales responde al estilo de vida de cada *phylum*.
6. Existen tendencias específicas de *phylum* en el uso de riboswitches de tipo atenuador transcripcional y traduccional.
7. El arreglo genómico de los factores sigma es específico de cada familia de regulador.

Perspectivas

8. Evaluar el enriquecimiento de reguladores transcripcionales considerando el pangenoma de cada especie de bacterias y archaeas.
9. Realizar la búsqueda de motivos de secuencia conservados en las regiones de regulación de factores transcripcionales y factores sigma para cada familia de regulador.
10. Comparar la simlitud de motivos de secuencia conservados de los reguladores transcripcionales a diversos niveles taxonómicos.
11. Analizar la abundancia y enriquecimiento de reguladores traduccionales y evaluar su relación con el tamaño de genoma y el grupo filogenético.
12. Complementar el análisis del contexto genómico de los factores transcripcionales y sigma en la regulación génica específica de *phylum*.

Anexo 1: Artículo publicado como resultado de este trabajo

Los resultados de este trabajo dieron lugar a la publicación titulada "*Complementary Tendencies in the Use of Regulatory Elements (Transcription Factors, Sigma Factors, and Riboswitches) in Bacteria and Archaea*" publicada en Enero de 2021 en la revista Journal of Bacteriology, donde fue reconocida como parte de los artículos Spothligth del número en el cual fue publicada.

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Bacteriology

Home Articles For Authors About the Journal Subscribe

Spotlight

Articles of Significant Interest in This Issue

DOI: 10.1128/JB.00613-20 Check for updates

Article Info & Metrics PDF

Complementary Tendencies in the Use of Regulatory Elements (Transcription Factors, Sigma Factors, and Riboswitches) in Bacteria and Archaea

Gene expression in prokaryotes is regulated mainly at the level of transcription by transcription factors, sigma factors, and riboswitches. Chávez et al. (e00413-20) provided a comprehensive description of the most abundant COG, KEGG, and Rfam families of transcriptional regulators present in prokaryotic genomes according to their genome size and phylogenetic origin. Furthermore, they found a clear tendency for organisms to compensate for the low frequencies of a particular type of regulatory element (transcription factors) with a high frequency of other types of regulatory elements (sigma factors).

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Bacteriology®

RESEARCH ARTICLE

Complementary Tendencies in the Use of Regulatory Elements (Transcription Factors, Sigma Factors, and Riboswitches) in Bacteria and Archaea

Joselyn Chávez,[a] Damien P. Devos,[b] Enrique Merino[a]

[a]Department of Molecular Microbiology, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México
[b]Centro Andaluz de Biología del Desarrollo, Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide/Junta de Andalucía, Seville, Spain

ABSTRACT In prokaryotes, the key players in transcription initiation are sigma factors and transcription factors that bind to DNA to modulate the process, while premature transcription termination at the 5′ end of the genes is regulated by attenuation and, in particular, by attenuation associated with riboswitches. In this study, we describe the distribution of these regulators across phylogenetic groups of bacteria and archaea and find that their abundance not only depends on the genome size, as previously described, but also varies according to the phylogeny of the organism. Furthermore, we observed a tendency for organisms to compensate for the low frequencies of a particular type of regulatory element (i.e., transcription factors) with a high frequency of other types of regulatory elements (i.e., sigma factors). This study provides a comprehensive description of the more abundant COG, KEGG, and Rfam families of transcriptional regulators present in prokaryotic genomes.

IMPORTANCE In this study, we analyzed the relationship between the relative frequencies of the primary regulatory elements in bacteria and archaea, namely, transcription factors, sigma factors, and riboswitches. In bacteria, we reveal a compensatory behavior for transcription factors and sigma factors, meaning that in phylogenetic groups in which the relative number of transcription factors was low, we found a tendency for the number of sigma factors to be high and vice versa. For most of the phylogenetic groups analyzed here, except for *Firmicutes* and *Tenericutes*, a clear relationship with other mechanisms was not detected for transcriptional riboswitches, suggesting that their low frequency in most genomes does not constitute a significant impact on the global variety of transcriptional regulatory elements in prokaryotic organisms.

KEYWORDS genome size, phylum-specific trends, riboswitches, sigma factors, transcription factors

Gene expression regulation is a common mechanism in all living organisms in response to intracellular and environmental changes. In general terms, the regulation of gene expression in prokaryotic organisms is performed at the transcriptional and translational levels during the initiation, elongation, or termination stages (1, 2). Additionally, regulation can take place posttranscriptionally (i.e., mRNA processing or degradation [reviewed in references 3 and 4]) or posttranslationally (i.e., degradation or modification by phosphorylation, acetylation, hydroxylation, methylation, or glycosylation, among others [reviewed in reference 5]).

The three primary key players in prokaryotic transcriptional regulation are transcription factors (TFs), sigma factors, and riboswitches. TFs are proteins for which DNA binding results in the activation or repression of gene transcription. In some cases, TFs may have dual activity as activators or repressors, depending on the position in which they bind to the DNA with respect to the promoter position. Considering the conser-

Citation Chávez J, Devos DP, Merino E. 2021. Complementary tendencies in the use of regulatory elements (transcription factors, sigma factors, and riboswitches) in bacteria and archaea. J Bacteriol 203:e00413-20. https://doi.org/10.1128/JB.00413-20.

Editor Michael Y. Galperin, NCBI, NLM, National Institutes of Health

Address correspondence to Enrique Merino, merino@ibt.unam.mx.

Received 15 July 2020
Accepted 10 October 2020
Accepted manuscript posted online 19 October 2020
Published 18 December 2020

vation of protein sequences, prokaryotic TFs are found in 91 groups in the COG (Clusters of Orthologous Groups) database (6–8) (see Table S1 in the supplemental material). These TF groups can be subdivided into smaller groups if, in addition to the sequence similarity criterion, the functions of the proteins and the cellular processes in which they participate are also considered. Thus, prokaryotic TFs are grouped into 369 KEGG orthology (KO) groups in the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (9–11) (Table S1).

In addition to transcriptional regulation based on TFs, gene expression in bacteria is also transcriptionally modulated by sigma factors, which are polypeptides that are required by the bacterial RNA polymerase holoenzyme to initiate transcription by conferring promoter recognition selectivity and participating in the initial steps of RNA synthesis. In *Escherichia coli* and other bacterial organisms, the transcription of most genes, including all of the essential or "housekeeping" genes, depends on sigma70 (12). Nevertheless, under certain specific stresses or developmental pathways, such as sporulation, the participation of alternative sigma factors might change the RNA polymerase preferences for a set of particular promoters, resulting in the coordinated transcription of large numbers of specific genes (13). According to the COG database, bacterial sigma factors can be clustered into four groups (Table S1). These groups can be further subdivided when the processes of the genes that they regulate are considered, as in the KEGG database, where bacterial sigma factors are clustered into nine groups (Table S1). In contrast to bacteria, no sigma factors have been found in archaea. Instead, recruitment of the unique RNA polymerase depends on the TATA-binding protein (TBP) and transcription factor B (TFB), homologous to the eukaryotic transcription factor TFIIB (2, 14).

The regulation of gene expression in bacteria and archaea can also be modulated by a regulatory mechanism known as transcription attenuation, in which the cells sense a specific metabolic signal that originates a response, causing RNA polymerase to terminate transcription prematurely or to continue transcribing the subsequent genes of an operon (15–18). In bacteria, this response is based on the formation, in the nascent transcript, of two alternative RNA hairpin structures that are mutually exclusive, the terminator and a competing transcription antiterminator (15–18). In contrast to bacteria, transcription termination in archaea takes place in response to oligo(T)-rich sequences and seems to be independent of any RNA secondary structures (19). There are different molecular mechanisms used for sensing intracellular signals during the attenuation process. One is ribosome-mediated transcription attenuation, in which the intracellular level of charged tRNAs determines whether a ribosome translates a small leader peptide with or without pausing. Pausing during the translation of the leader peptide favors the formation of the antiterminator secondary structure, while translation without pausing favors the formation of the transcription terminator structure. An example of this kind of attenuation regulates the transcription of the tryptophan biosynthetic operon in *E. coli* (20). Another is protein-mediated transcription attenuation. In this case, an RNA-binding protein can either interact with the nascent transcript and prevent the formation of an antiterminator structure, as occurs in the *Bacillus subtilis* pyrimidine nucleotide operon (21), or stabilize this antiterminator structure, as in the *E. coli bgl* operon (22). The last is riboswitch-mediated transcription attenuation, in which the untranslated RNA leader can be folded into three-dimensional structures capable of sensing intracellular signals with high specificity and sensitivity in the total absence of other factors, including proteins (23–25).

Riboswitches are composed of two platforms or domains, the recognition or sensor domain and the expression domain. The recognition domain is folded into three-dimensional structures that include highly selective binding pockets whose structure is complementary to the shape of their corresponding binding targets. The sizes of the target molecules sensed by riboswitches are commonly small and include vitamin derivatives (thiamine pyrophosphate, flavin mononucleotide, adenosylcobalamin, etc.) (24–27), purines and their derivatives (guanine and adenine [28, 29]), amino acids (lysine and glycine [23, 30, 31]), and a phosphorylated sugar (27, 28). In addition to

these compounds, a unique kind of riboswitch called the T-box can sense different types of uncharged tRNAs. This property of recognizing tRNAs as signal molecules implies that the T-box was not originally recognized as a riboswitch. However, given the similarity in its mechanism of action and because it fulfills all of the other characteristics of riboswitches, the T-box is currently described as a riboswitch in the reference Rfam database as well as in many other articles. The T-box regulates the expression of aminoacyl-tRNA synthetases and amino acid biosynthesis and transport (32–34).

The specific recognition of the ligand by the riboswitch is performed by ligand-RNA interactions through specific hydrogen bonds, electrostatic interactions, or stacking interactions. The sensing of the metabolite by the recognition platform stabilizes it. This action commonly produces a conformational change in the expression platform, which is placed contiguously on the sensor platform and is the active regulatory element of the riboswitch. The expression platform is commonly formed by transcriptional or translational attenuators; however, on some occasions, this platform can be folded, forming a ribozyme or defining alternative splicing patterns in eukaryotes (35, 36).

A general relationship between the number of TFs and the genome size of model organisms was described 15 years ago by Cases et al. and Pérez-Rueda et al., who demonstrated an overrepresentation of TFs in larger genomes, usually from free-living organisms, while fewer TFs were found in the smallest genomes, commonly from intracellular organisms (37, 38). In this regard, these authors proposed that transcriptional regulation in intracellular and extremophile bacteria relies almost exclusively on TFs. By contrast, in pathogenic bacteria, the contribution of sigma factors is more significant and nearly as significant as that of TFs in free-living bacteria (39). Our analysis aims to obtain an updated view of the trends in transcriptional regulation of organisms by considering not only a more representative number of organisms but also other key factors in transcription regulation, such as sigma factors and riboswitches. We identified compensatory behaviors for transcription factors and sigma factors in bacteria. A similar compensatory behavior for riboswitches was also exclusively found in the bacterial phyla *Firmicutes* and *Tenericutes*. Our study recognizes that the differential use of these regulatory elements might vary depending on the phylogenetic group to which the organisms belong. Our survey on the distribution of regulatory elements in prokaryotic organisms is discussed in light of possible events during the evolution of these organisms.

RESULTS AND DISCUSSION

The abundance of transcriptional regulators correlates with the genome size. To elucidate how transcriptional regulators are distributed and whether their frequencies vary depending on the genome size of the organisms, we evaluated the total number of TFs and sigma factors in the genomic sequence of a set of 2,720 representative bacterial and archaeal organisms available in the KEGG database (see Table S2 in the supplemental material).

In the first instance, a protein was considered a TF or a sigma factor if its description corresponded to a transcription factor or a sigma factor according to the COG or KEGG databases (see Table S1 in the supplemental material). Note that while proteins in the COG database are grouped based on their sequence similarity, proteins in the KEGG database are grouped not only based on similarity but also considering up-to-date annotations of gene functions. Although this functional subdivision adds a degree of precision in identifying functional orthologs, the number of TFs with KO assignments is much lower than the number with COG classification. Thus, KEGG offers a more precise but limited view of their frequencies in the sequenced genomes. This limitation is less significant for sigma factors since most of their functions can be inferred based on amino acid sequence similarity (see Fig. S1 in the supplemental material).

We plotted the total number of each type of regulatory element versus the number of protein coding sequences (CDS) for each genome. The variation rate between these two variables (the number of regulators versus the number of CDS) was expressed as the slope (*m*) of the regression line of the plotted values (Fig. 1A to C; see also Fig. S2

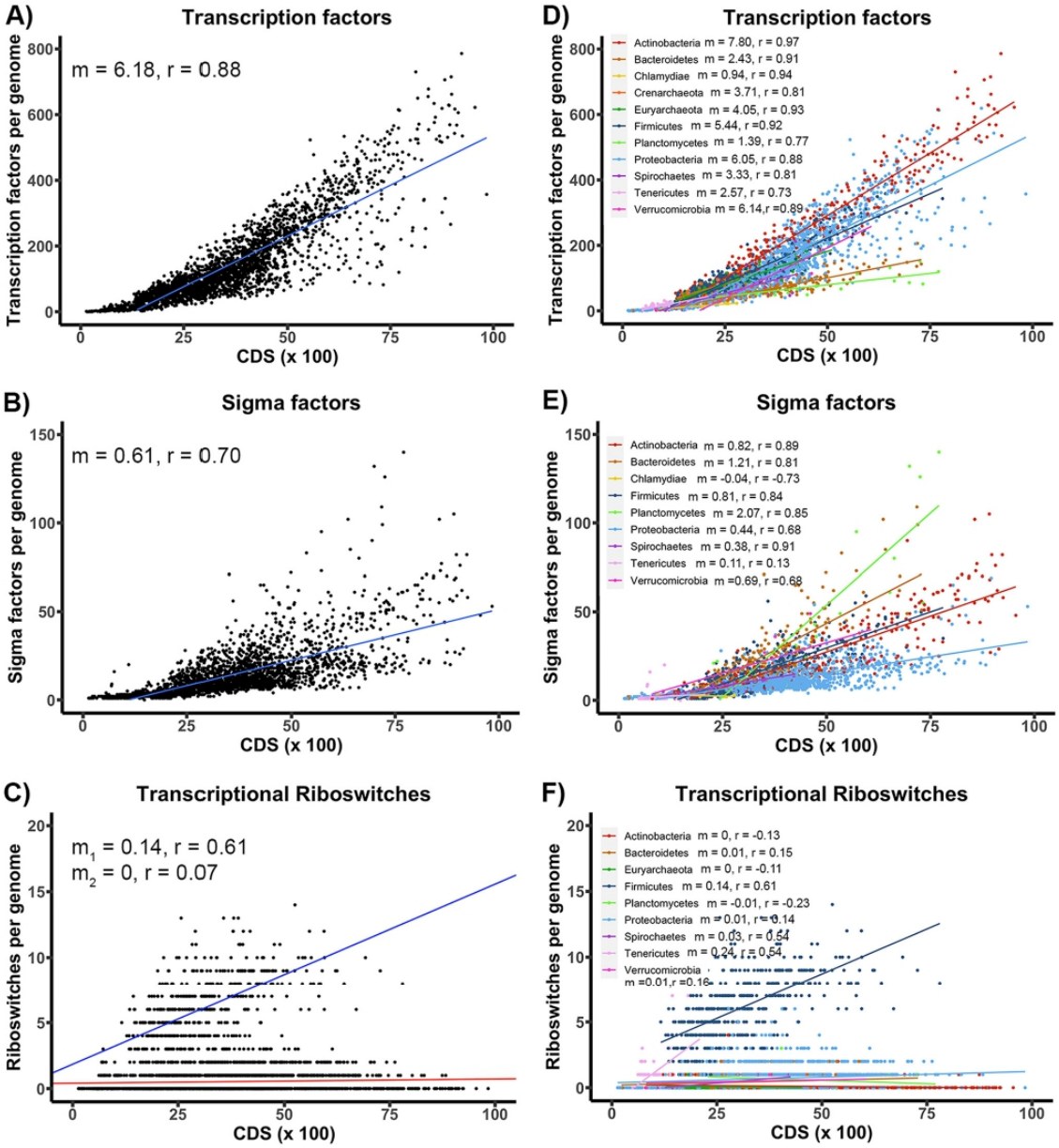

FIG 1 The relationship between the abundance of transcriptional regulatory elements and genome size is phylum dependent. (A and B) The frequency of transcriptional factors (A) and sigma factors (B) in bacteria and archaea based on the COG database versus the number of CDS (grouped into hundreds as a unit) per genome. (C) The abundance of rho-independent transcriptional terminator riboswitches in bacteria and archaea versus the number of CDS per genome. (D to F) The abundance of each regulator is split by phylum. The slope (*m*) from the linear regression model and the Pearson correlation coefficient (*r*) are shown for each phylum. The blue line in panel C corresponds to a linear regression model in which the number of riboswitches and the genome sizes are correlated. The red line in panel C corresponds to a linear regression model in which the number of riboswitches seemed to be invariant with respect to the genome size.

in the supplemental material), such that a value of $m = 1$ would correspond to the group of genomes in which there is one regulatory element for every 100 CDS. Although the exponential regression model has been used previously to compare the number of TFs in relation to the size of the genomes (37, 38), in our case, we used a

linear model as it facilitates our comparison of the trends between different phylogenetic groups without a significant decrease of fit value.

Consistent with previous results (37, 38), we observed that the number of genes that code for TFs vary proportionally in relation to the size of their corresponding genomes (Fig. 1A; see also Fig. S2A). Regarding sigma factors, the numbers of genes that code for these regulatory elements in the genomes of our study were 37,271 and 32,391 according to the annotations in the COG or KEGG databases, respectively (see Tables S3 and S4 in the supplemental material). Our results indicate that they have a similar trend to that described for TFs, that is, the number of genes coding for sigma factors tends to vary proportionally to the total number of genes in the organism (Fig. 1B; see also Fig. S2B).

To search for riboswitches in the 5′ untranslated regions (UTRs) of the genes, we used the covariance models defined in the Rfam database (40–42) and the CMsearch program (43). We searched for these regulatory elements in the 5′ UTR of all of the genes in the 2,720 representative genomes of bacteria and archaea and found 9,565 riboswitches (see Materials and Methods for further detail). To identify the level of regulation by the riboswitches, we evaluated the sequence and secondary structure of their expression platforms. We considered that a riboswitch acts at the transcriptional level if its expression platform could fold into rho-independent transcriptional terminators in the case of bacteria or if its expression platform contains an oligo(T)-rich sequence in the case of archaea. Conversely, we considered that a riboswitch acts at the translational level if its expression platform could fold into secondary structures capable of sequestering the Shine-Dalgarno region (see Materials and Methods). Of the set of 9,565 riboswitches identified in our analysis, 4,099 were classified as transcriptional acting riboswitches. These riboswitches belong to 32 different riboswitch families according to their Rfam classification (see Table S5 in the supplemental material). We quantified the number of different riboswitch families per genome and plotted this number in relation to the number of CDS in the genomes. Unlike TFs and sigma factors, the number of riboswitches in most prokaryotic genomes is low and does not increase as the size of their genomes increases (Fig. 1C). From this figure, two different behaviors were observed; the first is one in which the genome sizes and the numbers of riboswitches were correlated (Fig. 1C, blue line), and the second is one in which the number of riboswitches seemed to be invariant with respect to the genome size with a low number of occurrences (Fig. 1C, red line).

The ratio of transcriptional regulators versus genome size depends on the phylogenetic origin of the organisms. The fact that the number of transcriptional regulatory elements increases in proportion to the size of the genomes does not necessarily imply that the relationship between these two variables is the same for all organisms. Our study was intended to explain the different trends in the frequencies of transcriptional regulator use in prokaryotes, and it was performed by considering an evolutionary perspective and accounting for the phylogenetic relationships among these organisms. We analyzed the genomic sequences of 2,518 bacteria and 202 archaea representative of their species and clustered them based on their phylum. The resulting graphs are shown in Fig. 1D to F. From these figures, it can be observed that the points fit their respective regression models better and that the slope (m) varies significantly depending on the phylogenetic groups in question. For example, for TFs, *Actinobacteria* ($m = 7.80$) followed by *Verrucomicrobia* ($m = 6.14$), *Proteobacteria* ($m = 6.05$), and *Firmicutes* ($m = 5.44$) were the phylogenetic groups that contained the highest ratio of genes that code for TFs in relation to the number of CDS in their genomes. Additionally, *Chlamydiae* and *Planctomycetes*, both members of the same superphylum, were the groups with the smallest such values at $m = 0.94$ and 1.39, respectively (Fig. 1D). Note that although the slope values of the regression lines were low, the Pearson correlation coefficients (r) were significant. Regarding the sigma factors (Fig. 1E), *Planctomycetes* were the phylogenetic group with the highest ratio of this kind of regulatory element per number of CDS ($m = 2.07$). By contrast, in *Chla-*

mydiae, the number of genes encoding sigma factors does not seem to have a significant relationship with the number of CDS per genome ($m = -0.04$).

Similar trends for riboswitches were only observed in association with *Firmicutes* and *Tenericutes*, which were by far the phylogenetic groups that encoded a more significant number of transcriptional riboswitches in relation to the number of CDS in their genomes, as can be inferred from their corresponding slope values of 0.14 and 0.24, respectively (Fig. 1F). For the remaining phyla, the number of transcriptional riboswitches per genome was notoriously lower (*Spirochaetes*, *Proteobacteria*, and *Bacteroidetes*) or nonexistent (*Crenarchaeota* and *Chlamydiae*) and appeared to be unrelated to the genome size as can be inferred from the regression line slopes (m), with values close to or equal to zero, and the low values of the corresponding Pearson correlation coefficients (r) (Fig. 1F). Similar behavior is observed when accounting for the number of copies of riboswitches per genome in which *Firmicutes* and *Tenericutes* show both the highest copy numbers and increasing ratios of transcriptional riboswitches versus genome size, mostly due to T-box overrepresentation in their genomes. Here, the slope values of their regression lines are 0.60 and 0.57, respectively (see Fig. S3 in the supplemental material).

Trends in genomic frequency compensation for transcription factors, sigma factors, and riboswitches. From the slope values of the regression lines mentioned above (Fig. 1; see also Fig. S2), a compensatory tendency can be observed between the genomic abundances of the different types of transcriptional regulators. To visualize these compensatory tendencies, we directly compared the number of TFs versus the number of sigma factors per genome and adjusted the regression lines by using the Theil-Sen estimator, which takes the median of all of the slopes in the data (see Materials and Methods for further details) (Fig. 2; see also Fig. S4 in the supplemental material). The slope values (m) of the regression lines for these figures represent the median number of TFs per sigma factor in their genomes.

When TFs represent the primary number of regulators in a specific genome, sigma factors tend to have lower representation in that genome. This is true of *Proteobacteria*, with a median ratio of 14 TFs per sigma factor (Fig. 2A); *Actinobacteria*, with a median ratio of 9.27 TFs per sigma factor (Fig. 2B); and *Spirochaetes*, with a median ratio of 8.85 TFs per sigma factor (Fig. 2C). On the contrary, when the number of genes encoding TFs in a specific phylum is lower, the number of genes that encode sigma factors is higher. Some of the clearest examples of this trend are presented in *Bacteroidetes* and *Planctomycetes*, with median ratios of 1.53 and 0.74 TFs per sigma factor, respectively (Fig. 2F and G). Furthermore, there are specific phylogenetic groups that possess almost invariant low numbers of some kinds of regulators, regardless of their genome sizes. The first example is the case of sigma factors in *Chlamydiae*, the members of which have only two or three sigma factors but variable frequencies of TFs, resulting in the negative slopes shown in Fig. 2I and Fig. S4I. The second case of almost invariant low numbers of regulatory elements is transcriptional riboswitches. With the exceptions of *Firmicutes* and *Tenericutes*, transcriptional riboswitches in bacteria and archaea tend to exist in small numbers. In many organisms, the regulation of riboswitches occurs at the translational level; consequently, its effect on the transcriptional regulatory compensation behavior, as observed for TFs and sigma factors, is not significant.

To analyze whether the phylogenetic dependence on the relationship values between the genome size and the number of genes coding for transcription factors observed at the phylum level could be observed at deeper levels, we repeated our analysis at the taxonomic class level. As shown in Fig. S5 in the supplemental material, class-specific rates and compensatory effects were observed. As an example, organisms from *Betaproteobacteria* and *Gammaproteobacteria* had the highest increasing ratio of genes coding for transcription factors in comparison with the observed sigma factors. By contrast, *Deltaproteobacteria* tended to have the highest ratio of genes coding for sigma factors in comparison with TFs (Fig. S5).

We repeated this kind of analysis in different phyla but did not observe any trends similar to those found in *Proteobacteria* (data not shown). The class-specific rates

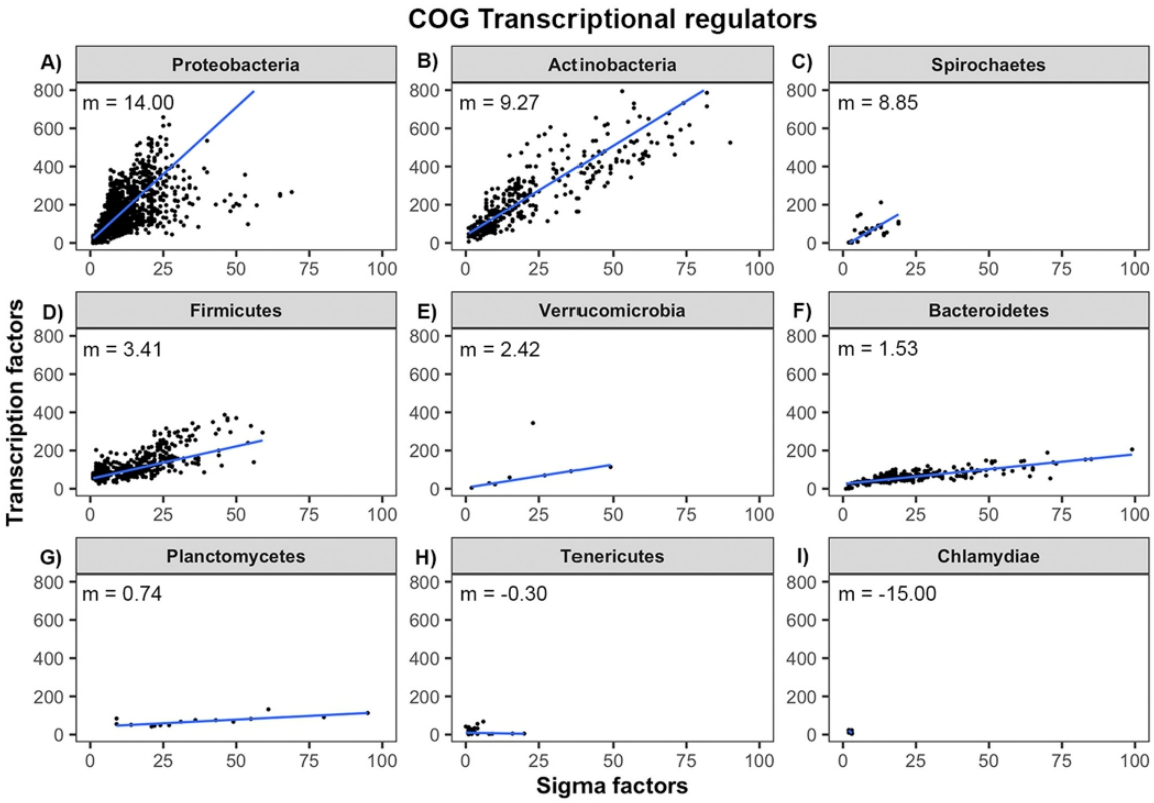

FIG 2 Compensatory trends in the frequencies of transcriptional factors and sigma factors are phylum specific. The total numbers of sigma factors (*x* axis) and transcription factors (*y* axis) per genome are shown for each phylum. The identification of TFs or sigma factors was performed based on the descriptions from the COG database. The linear regression models were adjusted using the Theil-Sen estimator. The slope (*m*) representing the median of the data slopes is shown for each phylum.

observed in *Proteobacteria* could be explained by the fact that *Proteobacteria* encompass an enormous amount of morphological, physiological, and metabolic diversity (44).

Trends in the use of TF families in bacterial and archaeal genomes. To assess whether the use of regulatory elements exhibits differences among phylogenetic groups, we calculated the frequency of each COG, KO, or Rfam family per phylum (see Tables S4 to S6 in the supplemental material) and performed a Fisher test (see Materials and Methods). Fisher's test compares the frequency of a TF family within a particular phylum versus its frequencies in other phyla, such that an enrichment $\log_{10}$ (odd value) greater than zero represents an overrepresentation of a certain TF family on a particular phylum in comparison with other phyla, while a value less than zero reflects an underrepresentation of that family. Note that a TF family with high global frequency but homogeneously distributed in all phyla might have an enrichment value close to zero, while a family with low abundance but present only in one phylum will have a high enrichment value. These Fisher test values were used to construct heat maps to compare the trends in the different regulator families among phylogenetic groups. The frequencies of the TFs or sigma factors were evaluated using either the COG groups (Fig. 3A and C) or the KO groups (Fig. 3B and D) as a reference.

Although the results on the abundances of transcriptional regulatory elements may be biased by significant differences in the number of organisms sequenced and their degree of characterization in the different phyla, our analysis of the TF distribution

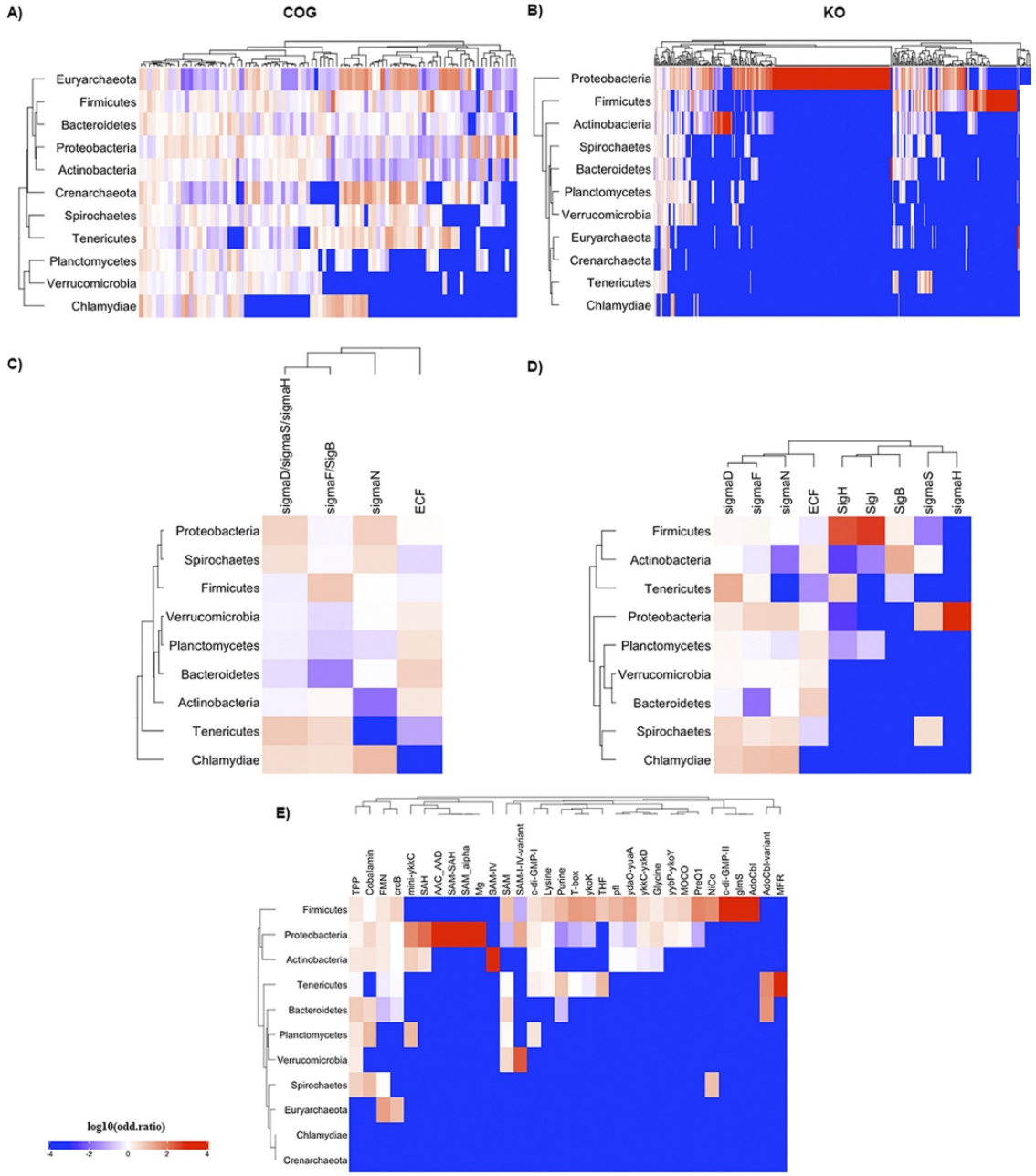

FIG 3 The enrichment of transcriptional regulator families is phylum dependent. (A and B) The enrichment of transcription factor families per phylum according to the COG (A) and KEGG (B) databases. (C and D) The enrichment of sigma factor families per phylum according to the COG (C) and KEGG (D) databases. (E) Enrichment of transcriptional riboswitches classified by their Rfam family. The color scale shows the $\log_{10}$ (odds ratio) value from the Fisher test.

according to phylogenetic origin indicates that, to date, 91 families of TFs are known according to the orthologous relationships defined in the COG database. A vast majority of these TF families are present in all of the phylogenetic groups studied in our

work, although their proportions can vary significantly (Fig. 3A). Their ability to recognize different intracellular or extracellular stimuli depends on the great diversity of their ligand-binding domains.

When TFs are classified according to COG groups, the most widely distributed TFs belong to the DtxR family of regulators (COG1321), members of the TF family with predicted helix-turn-helix (HTH)-like domains (COG2865), and members of the GntR family (COG1725). By contrast, some TF families are highly specific to particular phylogenetic clades, such as the members of the TF family with an HTH domain (COG3373), members of the TF family involved in thiamine biosynthesis (COG1992), or TFs of a family of regulators with unknown function (COG1709) from the archaeal domain as represented by the phyla *Euryarchaeota* and *Crenarchaeota*. In *Firmicutes*, some of the most enriched TF families belong to transcriptional regulator CodY (COG4465), competence transcription factor ComK (COG4903), and predicted transcriptional regulator containing CBS domains (COG4109). Moreover, in *Proteobacteria*, the most enriched TF families were those of a predicted transcription regulator (COG4957), the MltR family (COG3722), and sigma54-dependent TF (COG4650). The phyla with fewer identified TFs were *Verrucomicrobia*, *Planctomycetes*, and *Chlamydiae*, which were grouped into a branch of the dendrogram shown in Fig. 3A, consistent with the fact that they constitute the *Planctomycetes-Verrucomicrobia-Chlamydiae* (PVC) superphylum (45).

A similar TF enrichment analysis was performed using the KO groups as references. In this case, the TF grouping represents not only the orthologous relationships of the proteins but also a common function defined by their regulatory target genes due to the classification itself. In accordance with this criterion, *Proteobacteria* is the phylum with the greatest diversity of TFs, with 301 different KO groups; by contrast, the phylum *Chlamydiae* only has 11 different KO groups of TFs. In addition, *Proteobacteria*, *Firmicutes*, *Actinobacteria*, and, to a lesser extent, *Bacteroidetes* and *Euryarchaeota* present a clear tendency to have phylum-specific KO groups of TFs (Fig. 3B; see also Table S7 in the supplemental material). We found that 25 of the 114 KO groups exclusively present in *Proteobacteria* belong to the LysR family, making it the most enriched class of TFs in the phylum, followed by the LuxR, AraC, and TetR/AcrR TF families, which were present in 21, 18, and 12 KO groups, respectively. In *Firmicutes*, 28 KOs were exclusive, primarily containing members of the TF with helix-turn-helix domain MarR and MerR TF families, which were present in 7, 4, and 3 KO groups, respectively. Among the KO groups exclusively found in *Actinobacteria*, 4 belong to the WhiB TF family. In addition, there are only two KO groups exclusive to the *Bacteroidetes* phylum, the CRP/FNR family and the HTH-type of TFs. Notably, *Euryarchaeota* and *Crenarchaeota*, both members of the Archaea domain, share 4 KO groups of TFs that are uniquely present in both phyla; most are defined as putative and nonfunction-associated TFs. These TFs can be considered excellent candidates to identify new models of regulation in these organisms.

Trends in the use of sigma factor families in bacterial genomes. In terms of their frequency in bacterial genomes, sigma factors are the second most common type of transcriptional regulators. They confer promoter selectivity to the RNA polymerase to start transcription. Similar to how we addressed the TFs, we quantified the enrichment of sigma factor families in genomes across the different bacterial phylogenetic groups. Members of the RpoD housekeeping sigmaD factor (sigma70), the RpoS stationary phase sigmaS factor (sigma38), and the RpoH heat shock sigmaH factor (sigma32) families are grouped into COG0568. As expected, all of the bacterial organisms possessed a housekeeping sigma and, therefore, at least one member of this COG in their genome (Fig. 3C). Other widely distributed sigma factors among bacterial organisms are the flagellar synthesis and chemotaxis sigmaF (sigma28/FliA), the nitrogen limitation sigmaN (sigma54/RpoN), and the extracytoplasmic function (ECF) sigma factors that constitute the most ubiquitous and diverse family of alternative sigma factors in bacterial genomes (Fig. 3C and D). ECF sigma factors commonly regulate the expression of genes coding for outer membrane or periplasmic space proteins involved in the

response to unfolding proteins during the stationary phase and in responding to different kinds of cellular stresses, for instance, heat shock and osmotic and oxidative stresses as well as genes associated with bacterial virulence, among others (46–48); thus, the high abundance of this class of sigma factors reflects a wide response to environmental changes. Consistent with a previous study (49), we found a high abundance of ECF sigma factors in organisms from the *Planctomycetes* phylum, most of which are free living in the sea or soil (50), with up to 74 genes coding for proteins of this family in the case of *Singulisphaera acidiphila*; nevertheless, we did not find these sigma factors in the closely related *Chlamydiae* phylum. This lack of ECFs in *Chlamydiae* could be explained by their nature as intracellular pathogens.

Conversely, there are sigma factors that are mostly specific to a phylum, such as SigH, which is involved in the transition to the post-exponential phase at the beginning of the sporulation process, and SigI, which is involved in the regulation of cell wall metabolism in response to heat stress. These two sigma factors are highly enriched in *Firmicutes* (Fig. 3D). In addition, sigmaH (sigmaH32/RpoH) was found exclusively in *Proteobacteria*, and it was transcribing genes involved in the heat shock response. Notably, both *Chlamydia* and *Tenericutes* contain organisms with small genomes and a low abundance of alternative sigma factors (Fig. 3C and D).

In general terms, the amount and type of alternative sigma factors vary substantially between different bacterial organisms and usually reflect their lifestyle. For example, an obligate intracellular organism living in almost undisturbed environments often possesses only the housekeeping sigma factor. By contrast, some free-living bacteria, such as *Streptomyces coelicolor*, which inhabit environments with diverse fluctuating physicochemical conditions, may have several dozen alternative sigma factors (ScoDB at http://strepdb.streptomyces.org.uk) (51).

By themselves, sigma factors have a very limited capacity to modify their activity in response to the recognition of intracellular signals or stimuli, although exceptional examples of transcriptional switching through the direct interaction of small molecules with RNA polymerase have been described; this is the case for the so-called stringent response, in which alarmone guanosine tetraphosphate (ppGpp) is recognized and induces an arrest in RNA synthesis in response to nutrient starvation or other stress conditions (52, 53). Regulations by sigma factors work as global regulators, allowing bacteria to switch between different transcriptional programs based exclusively on the type of promoters that the RNA polymerase recognizes. These transcriptional programs might include the coordinated expression of virulence-associated genes in pathogenic bacteria or genes that respond to specific nutritional conditions, developmental processes, or stress-related signals (54, 55). Commonly, the availability of alternative sigma factors is controlled at the level of their synthesis, by proteolysis, or by the reversible interaction with their corresponding anti-sigma factors (12).

Note that neither eukaryotes nor archaebacteria, as represented by the phyla *Euryarchaeota* and *Crenarchaeota*, have sigma factors. The transcription machinery in archaebacteria has been widely documented to resemble that of eukaryotes in that the archaea RNA polymerase uses transcription factors such as TATA-binding protein (TBP) and TFIIB instead of sigma factors for initiation (2).

Trends in the use of transcriptional acting riboswitch families in bacterial and archaeal genomes. Unlike TFs and sigma factors, riboswitches are RNA elements that act in *cis*, and without counting some exceptional cases, they exclusively affect the genes and operons that are immediately downstream from them. Riboswitches recognize their target molecules with high specificity, binding vitamin derivatives, nucleotides, amino acids, phosphorylated sugars, and metal ions; therefore, the genes directly regulated by riboswitches are commonly those that are involved in the synthesis pathway or the transport of the sensed molecule.

Considering the number of different families of riboswitches per genome, we performed a Fisher test to analyze the enrichment of transcriptional acting riboswitch families per phylum. As shown in Fig. 3E, the riboswitch TPP was the most widespread among bacteria, including the phyla with the lowest (e.g., *Verrucomicrobia*, *Planctomy-*

cetes, and *Spirochaetes*) or the highest (e.g., *Firmicutes* and *Proteobacteria*) number of riboswitch families. For example, 48 of the 97 riboswitches found in *Bacteroidetes*, 6 of the 13 riboswitches found in *Spirochaetes*, and 3 of the 10 riboswitches from *Planctomycetes* belong to the TPP riboswitch.

As noted previously, the *Firmicutes* phylum has by far the highest abundance of transcriptional acting families of riboswitches compared with that of other phyla, with an average of 6 riboswitch families per genome, although there are organisms in this phylum, such as *Geosporobacter ferrireducens*, which contain 14 out of 36 families of the transcriptional active riboswitches reported in the Rfam database. *Firmicutes* are enriched in c-dl-GMP-II, *glmS*, and AdoCbl riboswitches, which are almost exclusively found in this phylum. The T-box riboswitch appears almost exclusively in this phylum as well, where it represents 16% of the riboswitches. Moreover, in other phyla (e.g., *Actinobacteria*, *Bacteroidetes*, *Euryarchaeota*, *Planctomycetes*, *Spirochaetes*, and *Verrucomicrobia*), T-box riboswitches do not exist. This riboswitch is not commonly found in members of *Proteobacteria*. However, we found that the gene that codes for the 2-isopropylmalate synthase enzyme is regulated by a T-box in 12 out of 73 organisms in the *Deltaproteobacteria* class (see Table S8 in the supplemental material). The T-box riboswitch might have been acquired by horizontal transfer from some *Firmicutes* to the common ancestor of these *Deltaproteobacteria*.

As shown in Fig. 3E on *Planctomycetes* and *Verrucomicrobia*, members of the PVC superphylum have the lowest number of riboswitch families regardless of the number of CDS in their genomes. We only found three families of riboswitches in *Verrucomicrobia* among eight studied organisms that belong to the TPP (1 riboswitch), SAM (1 riboswitch), and SAM I-IV (2 riboswitches) families. From 18 *Planctomycetes* organisms studied here, we found transcriptional riboswitches in 8 of them, which belonged to 5 families of riboswitches, the TTP (3 riboswitches), SAM (1 riboswitch), cobalamin (4 riboswitches), c-di-GMP-I (1 riboswitch), and mini-*ykkC* (1 riboswitch) riboswitches. Interestingly, there are some *Planctomycetes* with up to 2 or 4 riboswitch copies per genome and many others with none. From the 156 *Euryarchaeota* organisms analyzed here, only 9 of them have riboswitches from the FMN and *crcB* families. We did not find any transcriptional acting riboswitches in *Chlamydiae* or *Crenarchaeota*. We were also able to detect riboswitches that were previously reported as predominant in specific phylogenetic groups. For example, the SAM alpha riboswitch was only found in 137 of the 1,271 *Proteobacteria* organisms, with 125 of them in *Alphaproteobacteria*, 7 in *Betaproteobacteria*, 4 in *Gammaproteobacteria*, and 1 in *Deltaproteobacteria*. In addition, we only found transcriptional riboswitches from the SAM IV family in *Actinobacteria* as previously reported (Table S5).

The analysis of the expression platform for all of the identified riboswitches allowed us to distinguish the level at which they exerted their regulation, whether transcriptional or translational (see Materials and Methods for further details). In certain organisms, the lack of transcriptional acting riboswitches seems to be compensated for by the corresponding translational acting versions of riboswitches and vice versa. To confirm this hypothesis, we looked for riboswitches in their expression platforms that could form secondary RNA structures for sequestering the Shine-Dalgarno sequence and thereby inhibit the start of the translation process (see Table S9 in the supplemental material and Materials and Methods for further details). Figure 4 shows the average number of each of the families of riboswitches per organism within each phylum as well as the percentage that could regulate gene expression at the transcriptional versus translational level. It is noteworthy that specific trends in the riboswitch gene regulation levels can be observed depending on the phylum. For example, *Firmicutes* and *Tenericutes* use predominantly transcriptional riboswitches; however, *Actinobacteria* and *Proteobacteria* use translational riboswitches from a wide range of Rfam families. All of the riboswitches in *Crenarchaeota* and most of the riboswitches in *Euryarchaeota*, as representatives of *Archaea*, belong to translational acting riboswitches. Notably, some phyla, e.g., *Planctomycetes*, *Verrucomicrobia*, *Bacteroidetes*, and *Spirochaetes*, possess both transcriptional and translational riboswitches. Moreover, the

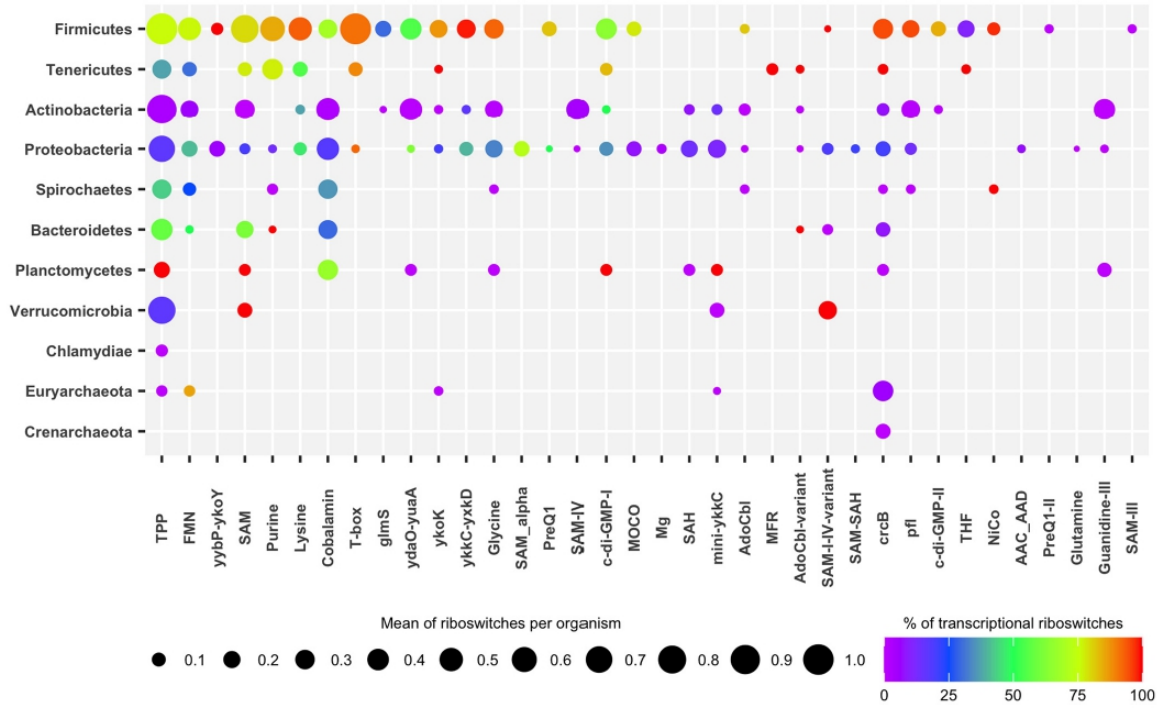

FIG 4 Frequency comparison of transcriptional and translational acting riboswitches. The point size represents the mean of either the transcriptional or translational total riboswitches found per organism for each Rfam family and phylum. The color scale shows the mean percentage of transcriptional riboswitches found for each Rfam family and phylum.

use of particular families as transcriptional or translational attenuators seems to vary depending on the phylum. This is the case for the mini-*ykkC* and TPP riboswitches that act as transcriptional riboswitches in *Planctomycetes* but as translational riboswitches in *Actinobacteria*, *Chlamydiae*, *Euryarchaeota*, and *Verrucomicrobia*. In addition, Fig. 4 shows that in some cases, only a subgroup of the organisms from a given phylum possess one element of the type riboswitch (e.g., T-box and glutamine in *Proteobacteria* and FMN in *Euryarchaeota*). Given their sparse distribution in the phylum, we suggest that these riboswitches might have been acquired by horizontal gene transfer.

Finally, there are two riboswitches that deserve mention, due to the type of molecule they recognize and their wide distribution in certain types of bacteria. The first of these is the riboswitch T-box, which, instead of identifying a small metabolite, recognizes uncharged tRNAs as a signal to sense the intracellular levels of amino acids. According to our results, the T-box riboswitch is the most common in *Firmicutes* and *Tenericutes*, regulating the expression of genes that code for enzymes responsible for charging the amino acids on their corresponding tRNAs and the biosynthesis and transport of amino acids. As an example, we found that the firmicute *Bacillus cereus* has 68 genes in 37 operons that are regulated by the T-box riboswitch. The second family of riboswitches that deserves special mention is the c-di-GMP riboswitches (I and II). As their name indicates, these riboswitches can recognize cyclic di-GMP, which is a secondary messenger used to signal metabolic states in bacteria. In response, many genes involved in cellular processes, such as virulence, motility, or biofilm formation, are regulated by this family of riboswitches.

In summary, we report a tendency for organisms to compensate for the low frequencies of a particular type of regulatory element (i.e., transcription factors) with a high frequency of other types of regulatory element (i.e., sigma factors), providing a

comprehensive description of the most abundant COG, KEGG, and Rfam families of transcriptional regulators present in prokaryotic genomes according to their genome size and phylogenetic origin.

MATERIALS AND METHODS

Selection of representative organisms. Genome sequences were retrieved from the KEGG database. This database contained 4,852 bacterial and 277 archaeal genomes. To avoid redundant species, we selected the organism with the highest number of open reading frames per species for each representative. We also chose the phylogenetic groups with at least nine organisms. In total, we selected 2,518 bacterial and 202 archaeal genomes belonging to 11 phyla.

Quantification of transcription factors and sigma factors. Using COG and KEGG orthology classification in bacteria and archaea, we obtained all gene IDs and their corresponding COG/KO descriptions for the selected genomes. In accordance with their COG/KO classifications, we quantified the genes corresponding to sigma and transcriptional factors.

All data obtained for the TFs and sigma factors are summarized in the tables containing the number of regulators per genome, the number of open reading frames (CDS), phylum, and class for each organism. For better visualization, we expressed the genome size as the number of CDS divided by 100. We then plotted the number of regulators per genome versus the CDS and split them by phylum.

Theil-Sen estimator analysis. The total numbers of sigma factors and transcription factors per organism were compared and split by phylum. The linear regression models per phylum were adjusted using the Theil-Sen estimator, which is available in the *mblm* R package (https://cran.r-project.org/package=mblm). This method counts the median of all possible slopes in the data, resulting in an outlier-resistant model (56, 57).

Riboswitch prediction. We obtained the covariance model for all riboswitches reported in Rfam and concatenated them into a compiled matrix. We used a region that is 400 bp upstream of the coding region of genes within available genomes from bacteria and archaea. We evaluated the putative presence of riboswitches by comparing these regions versus the covariance matrix using the CMsearch program (43), which takes advantage of both sequence and secondary structural conservation in the search for RNA homologs. The parameters used to run CMsearch were as follows: E value $\leq$ 1e$-$3, cpu = 32, and toponly = TRUE (to only search for riboswitches in the top strand). The parameter $-$nohmm, which skips all HMM filter stages, was not used when running the CMsearch program. The results from the CMsearch analysis were filtered, taking into account only those sequences with score values greater than the trusted cutoff value reported for each riboswitch in Rfam.

To identify transcriptional acting riboswitches, we developed a local program written in Perl, which was based on the previously reported computer approach by Merino and Yanofsky in 2005 (17). In brief, using this method, a 50-nt analysis window was considered downstream of each region identified as a riboswitch. In this analytical window, a run of 6 nucleotides in length was searched for those that contained at least 5 T residues. Afterward, using the RNAfold program version 2.4.0 (58), we searched for the most stable RNA secondary structure that could be formed, which was composed of a stem and loop structure and whose Gibbs free energy was less than $-$10 kcal/mol. In cases in which the analyzed region could fold into more than one stem and loop structure, only the one closest to the run of T residues was considered. We allowed for two nucleotides between the base of the secondary structure and the run of T residues.

Unlike bacteria, archaea do not require a stable secondary structure in RNA for the end of transcription, but they are stimulated by the presence of oligo(T) sequences. Therefore, for archaeal genomes, we considered a riboswitch to be acting at the transcription termination level if its sequence is followed by a consecutive run of 6 T residues and its distance to the downstream-regulated gene is greater than 50 nt. These results are summarized in Table S5 in the supplemental material.

To identify translational acting riboswitches, we took riboswitches that do not form a predicted transcriptional attenuator and whose predicted 5′ position is up to 100 nt downstream from the translational start site of their corresponding genes. We then evaluated whether they could form secondary structures that sequester the Shine-Dalgarno sequences with Gibbs free energy ≤ -7 kcal/mol and a 3′ position up to 7 nt downstream from the translational start site. The resulting riboswitches are summarized in Table S9 in the supplemental material.

Enrichment analysis. For each phylum, we compared the frequency of each COG, KO, or Rfam family within each phylum versus the rest of the phyla. We used these frequencies to perform a Fisher test with a standard Bonferroni correction. We then used the logarithm ($\log_{10}$) of the enrichment value (odds ratio) for clustering and plotted the enrichment of each family into a heatmap by using the default parameters in the "Heatmap" function from the R package ComplexHeatmap version 2.4.3, which uses the Euclidean method for hierarchical clustering of rows and columns in the heatmaps (59).

Software used. The pipelines generated for counting transcriptional regulators from the COG/KEGG/Rfam database were written in Perl version 5.18.2. The forward analysis, plots, and the development of the *erba* package were performed in R (https://cran.r-project.org) version 4.0. This package integrates functions from previously developed packages dplyr version 1.0.2 (https://dplyr.tidyverse.org) and ggplot2 version 3.3.2 (https://ggplot2.tidyverse.org).

Data access. All functions developed for the analysis and plotting, which are contained within the R package *erba*, are available at https://github.com/josschavezf/erba. The pipelines used to create the figures and supplemental figures are located at the GitHub repository https://github.com/josschavezf/Chavez_et_al_2020.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, XLSX file, 0.03 MB.
SUPPLEMENTAL FILE 2, XLSX file, 0.1 MB.
SUPPLEMENTAL FILE 3, XLSX file, 0.1 MB.
SUPPLEMENTAL FILE 4, XLSX file, 3 MB.
SUPPLEMENTAL FILE 5, XLSX file, 0.7 MB.
SUPPLEMENTAL FILE 6, XLSX file, 1 MB.
SUPPLEMENTAL FILE 7, XLSX file, 0.01 MB.
SUPPLEMENTAL FILE 8, XLSX file, 0.01 MB.
SUPPLEMENTAL FILE 9, XLSX file, 0.7 MB.
SUPPLEMENTAL FILE 10, PDF file, 1.8 MB.

ACKNOWLEDGMENTS

We sincerely thank Ricardo Ciria for computer support and Shirley Ainsworth for bibliographical assistance.

We acknowledge the Programa de Maestría y Doctorado en Ciencias Bioquímicas at the Instituto de Biotecnología-UNAM and the Consejo Nacional de Ciencia y Tecnología (CONACyT) for the Ph.D. scholarship, number 565669, awarded to J.C.

We declare no conflicts of interest.

REFERENCES

1. Browning DF, Busby SJW. 2004. The regulation of bacterial transcription initiation. Nat Rev Microbiol 2:57–65. https://doi.org/10.1038/nrmicro787.
2. Werner F. 2013. Molecular mechanisms of transcription elongation in archaea. Chem Rev 113:8331–8349. https://doi.org/10.1021/cr4002325.
3. Kavita K, de Mets F, Gottesman S. 2018. New aspects of RNA-based regulation by Hfq and its partner sRNAs. Curr Opin Microbiol 42:53–61. https://doi.org/10.1016/j.mib.2017.10.014.
4. Van Assche E, Van Puyvelde S, Vanderleyden J, Steenackers HP. 2015. RNA-binding proteins involved in post-transcriptional regulation in bacteria. Front Microbiol 6:141. https://doi.org/10.3389/fmicb.2015.00141.
5. Cain JA, Solis N, Cordwell SJ. 2014. Beyond gene expression: the impact of protein post-translational modifications in bacteria. J Proteomics 97:265–286. https://doi.org/10.1016/j.jprot.2013.08.012.
6. Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637. https://doi.org/10.1126/science.278.5338.631.
7. Galperin MY, Makarova KS, Wolf YI, Koonin EV. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res 43:D261–D269. https://doi.org/10.1093/nar/gku1223.
8. Makarova KS, Wolf YI, Koonin EV. 2015. Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. Life (Basel) 5:818–840. https://doi.org/10.3390/life5010818.
9. 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.
10. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. 2019. New approach for understanding genome variations in KEGG. Nucleic Acids Res 47:D590–D595. https://doi.org/10.1093/nar/gky962.
11. Kanehisa M. 2019. Toward understanding the origin and evolution of cellular organisms. Protein Sci 28:1947–1951. https://doi.org/10.1002/pro.3715.
12. Paget MS. 2015. Bacterial sigma factors and anti-sigma factors: structure, function and distribution. Biomolecules 5:1245–1265. https://doi.org/10.3390/biom5031245.
13. Weirauch MT, Hughes TR. 2011. A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. Subcell Biochem 52:25–73. https://doi.org/10.1007/978-90-481-9069-0_3.
14. Grohmann D, Werner F. 2011. Recent advances in the understanding of archaeal transcription. Curr Opin Microbiol 14:328–334. https://doi.org/10.1016/j.mib.2011.04.012.
15. Henkin TM, Yanofsky C. 2002. Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. Bioessays 24:700–707. https://doi.org/10.1002/bies.10125.
16. Merino E, Yanofsky C. 2002. Regulation by termination-antitermination: a genomic approach, p 323–336. *In* Sonenshein A, Losick R, Hoch J (ed), Bacillus subtilis and its closest relatives. American Society for Microbiology, Washington, DC.
17. Merino E, Yanofsky C. 2005. Transcription attenuation: a highly conserved regulatory strategy used by bacteria. Trends Genet 21:260–264. https://doi.org/10.1016/j.tig.2005.03.002.
18. Naville M, Gautheret D. 2010. Transcription attenuation in bacteria: theme and variations. Brief Funct Genomics 9:178–189. https://doi.org/10.1093/bfgp/elq008.
19. French SL, Santangelo TJ, Beyer AL, Reeve JN. 2007. Transcription and translation are coupled in Archaea. Mol Biol Evol 24:893–895. https://doi.org/10.1093/molbev/msm007.
20. Oxender DL, Zurawski G, Yanofsky C. 1979. Attenuation in the Escherichia coli tryptophan operon: role of RNA secondary structure involving the tryptophan codon region. Proc Natl Acad Sci U S A 76:5524–5528. https://doi.org/10.1073/pnas.76.11.5524.
21. Turner RJ, Lu Y, Switzer RL. 1994. Regulation of the Bacillus subtilis pyrimidine biosynthetic (pyr) gene cluster by an autogenous transcriptional attenuation mechanism. J Bacteriol 176:3708–3722. https://doi.org/10.1128/jb.176.12.3708-3722.1994.
22. Amster-Choder O, Wright A. 1993. Transcriptional regulation of the *bgl* operon of *Escherichia coli* involves phosphotransferase system-mediated phosphorylation of a transcriptional antiterminator. J Cell Biochem 51:83–90. https://doi.org/10.1002/jcb.240510115.
23. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. 2003. Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch? Nucleic Acids Res 31:6748–6757. https://doi.org/10.1093/nar/gkg900.
24. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. 2002. Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. J Biol Chem 277:48949–48959. https://doi.org/10.1074/jbc.M208965200.
25. Winkler W, Nahvi A, Breaker RR. 2002. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature 419:952–956. https://doi.org/10.1038/nature01145.
26. Winkler WC, Cohen-Chalamish S, Breaker RR. 2002. An mRNA structure that controls gene expression by binding FMN. Proc Natl Acad Sci U S A 99:15908–15913. https://doi.org/10.1073/pnas.212628899.
27. Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, Breaker RR. 2002. Genetic control by a metabolite binding mRNA. Chem Biol 9:1043–1049. https://doi.org/10.1016/s1074-5521(02)00224-7.

28. Mandal M, Boese B, Barrick JE, Winkler WC, Breaker RR. 2003. Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. Cell 113:577–586. https://doi.org/10.1016/s0092-8674(03)00391-x.
29. Mandal M, Breaker RR. 2004. Adenine riboswitches and gene activation by disruption of a transcription terminator. Nat Struct Mol Biol 11:29–35. https://doi.org/10.1038/nsmb710.
30. Grundy FJ, Lehman SC, Henkin TM. 2003. The L box regulon: lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. Proc Natl Acad Sci U S A 100:12057–12062. https://doi.org/10.1073/pnas.2133705100.
31. Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, Kenneth Wickiser J, Breaker RR. 2004. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proc Natl Acad Sci U S A 101:6421–6426. https://doi.org/10.1073/pnas.0308014101.
32. Grundy FJ, Henkin TM. 1993. tRNA as a positive regulator of transcription antitermination in B. subtilis. Cell 74:475–482. https://doi.org/10.1016/0092-8674(93)80049-K.
33. Vitreschak AG, Mironov AA, Lyubetsky VA, Gelfand MS. 2008. Comparative genomic analysis of T-box regulatory systems in bacteria. RNA 14:717–735. https://doi.org/10.1261/rna.819308.
34. Gutiérrez-Preciado A, Henkin TM, Grundy FJ, Yanofsky C, Merino E. 2009. Biochemical features and functional implications of the RNA-based T-box regulatory mechanism. Microbiol Mol Biol Rev 73:36–61. https://doi.org/10.1128/MMBR.00026-08.
35. Winkler WC, Nahvi A, Roth A, Collins JA, Breaker RR. 2004. Control of gene expression by a natural metabolite-responsive ribozyme. Nature 428:281–286. https://doi.org/10.1038/nature02362.
36. Cheah MT, Wachter A, Sudarsan N, Breaker RR. 2007. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. Nature 447:497–500. https://doi.org/10.1038/nature05769.
37. Cases I, de Lorenzo V, Ouzounis CA. 2003. Transcription regulation and environmental adaptation in bacteria. Trends Microbiol 11:248–253. https://doi.org/10.1016/S0966-842X(03)00103-3.
38. Pérez-Rueda E, Collado-Vides J, Segovia L. 2004. Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. Comput Biol Chem 28:341–350. https://doi.org/10.1016/j.compbiolchem.2004.09.004.
39. Pérez-Rueda E, Janga SC, Martínez-Antonio A. 2009. Scaling relationship in the gene content of transcriptional machinery in bacteria. Mol Biosyst 5:1494–1501. https://doi.org/10.1039/b907384a.
40. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, Bateman A, Finn RD, Petrov AI. 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res 46:D335–D342. https://doi.org/10.1093/nar/gkx1038.
41. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, Petrov AI. 2018. Non-coding RNA analysis using the Rfam database. Curr Protoc Bioinformatics 62:e51. https://doi.org/10.1002/cpbi.51.
42. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. 2003. Rfam: an RNA family database. Nucleic Acids Res 31:439–441. https://doi.org/10.1093/nar/gkg006.
43. Cui X, Lu Z, Wang S, Jing-Yan Wang J, Gao X. 2016. CMsearch: simultaneous exploration of protein sequence space and structure space improves not only protein homology detection but also protein structure prediction. Bioinformatics 32:i332–i340. https://doi.org/10.1093/bioinformatics/btw271.
44. Kersters K, De Vos P, Gillis M, Swings J, Vandamme P, Stackebrandt E. 2006. Introduction to the proteobacteria, p 3–37. *In* The prokaryotes. Springer, New York.
45. Wagner M, Horn M. 2006. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr Opin Biotechnol 17:241–249. https://doi.org/10.1016/j.copbio.2006.05.005.
46. Pinto D, Liu Q, Mascher T. 2019. ECF σ factors with regulatory extensions: the one-component systems of the σ universe. Mol Microbiol 112:399–409. https://doi.org/10.1111/mmi.14323.
47. Alba BM, Leeds JA, Onufryk C, Lu CZ, Gross CA. 2002. DegS and YaeL participate sequentially in the cleavage of RseA to activate the sigma(E)-dependent extracytoplasmic stress response. Genes Dev 16:2156–2168. https://doi.org/10.1101/gad.1008902.
48. Rowley G, Spector M, Kormanec J, Roberts M. 2006. Pushing the envelope: extracytoplasmic stress responses in bacterial pathogens. Nat Rev Microbiol 4:383–394. https://doi.org/10.1038/nrmicro1394.
49. Wiegand S, Jogler M, Boedeker C, Pinto D, Vollmers J, Rivas-Marín E, Kohn T, Peeters SH, Heuer A, Rast P, Oberbeckmann S, Bunk B, Jeske O, Meyerdierks A, Storesund JE, Kallscheuer N, Lücker S, Lage OM, Pohl T, Merkel BJ, Hornburger P, Müller RW, Brümmer F, Labrenz M, Spormann AM, Op den Camp HJM, Overmann J, Amann R, Jetten MSM, Mascher T, Medema MH, Devos DP, Kaster AK, Øvreås L, Rohde M, Galperin MY, Jogler C. 2020. Cultivation and functional characterization of 79 planctomycetes uncovers their unique biology. Nat Microbiol 5:126–140. https://doi.org/10.1038/s41564-019-0588-1.
50. Lage OM, Van Niftrik L, Jogler C, Devos DP. 2019. Planctomycetes, p 614–626. *In* Encyclopedia of microbiology. Elsevier, Philadelphia, PA.
51. Bentley SD, Chater KF, Cerdeño-Tárraga A-M, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang C-H, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream M-A, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA. 2002. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature 417:141–147. https://doi.org/10.1038/417141a.
52. Dalebroux ZD, Swanson MS. 2012. ppGpp: magic beyond RNA polymerase. Nat Rev Microbiol 10:203–212. https://doi.org/10.1038/nrmicro2720.
53. Hauryliuk V, Atkinson GC, Murakami KS, Tenson T, Gerdes K. 2015. Recent functional insights into the role of (p)ppGpp in bacterial physiology. Nat Rev Microbiol 13:298–309. https://doi.org/10.1038/nrmicro3448.
54. Kazmierczak MJ, Wiedmann M, Boor KJ. 2005. Alternative sigma factors and their roles in bacterial virulence. Microbiol Mol Biol Rev 69:527–543. https://doi.org/10.1128/MMBR.69.4.527-543.2005.
55. Feklístov A, Sharon BD, Darst SA, Gross CA. 2014. Bacterial sigma factors: a historical, structural, and genomic perspective. Annu Rev Microbiol 68:357–376. https://doi.org/10.1146/annurev-micro-092412-155737.
56. Theil H. 1992. A rank-invariant method of linear and polynomial regression analysis, p 345–381. *In* Henri Theil's contributions to economics and econometrics. advanced studies in theoretical and applied econometrics, vol 23. Springer, Dordrecht, Netherlands.
57. Sen PK. 1968. Estimates of the regression coefficient based on Kendall's tau. J Am Stat Assoc 63:1379–1389. https://doi.org/10.1080/01621459.1968.10480934.
58. Lorenz R, Luntzer D, Hofacker IL, Stadler PF, Wolfinger MT. 2016. SHAPE directed RNA folding. Bioinformatics 32:145–147. https://doi.org/10.1093/bioinformatics/btv523.
59. Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32:2847–2849. https://doi.org/10.1093/bioinformatics/btw313.

Anexo 2: Artículo publicado como colaboración externa

Como parte de una colaboración con otros grupos de trabajo, durante el periodo del doctorado se publicó el siguiente artículo titulado "*Programmatic access to bacterial regulatory networks with regutools"* publicado en la revista Bioinformatics en Agosto de 2020.

Bioinformatics, 36(16), 2020, 4532–4534
doi: 10.1093/bioinformatics/btaa575
Advance Access Publication Date: 23 June 2020
Applications Note

OXFORD

Databases and ontologies

Programmatic access to bacterial regulatory networks with *regutools*

Joselyn Chávez[1,†], Carmina Barberena-Jonas[2,3,4,†], Jesus E Sotelo-Fonseca[2,3,4,†], José Alquicira-Hernández[2,5,6], Heladia Salgado [2], Leonardo Collado-Torres [7,*] and Alejandro Reyes[8,9,*]

[1]Departamento de Microbiología Molecular, Instituto de Biotecnología, [2]Programa de Genómica Computacional, Centro de Ciencias Genómicas and [3]Licenciatura de Ciencias Genómicas, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico 62210, [4]National Laboratory of Genomics for Biodiversity, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Irapuato, Guanajuato 36821, Mexico, [5]Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia, [6]Garvan Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Darlinghurst, Sydney, NSW 2010, Australia, [7]Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD 21205, USA, [8]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA and [9]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: Peter Robinson

Received and revised on April 29, 2020; editorial decision on June 9, 2020; accepted on June 15, 2020

Abstract

Summary: *RegulonDB* has collected, harmonized and centralized data from hundreds of experiments for nearly two decades and is considered a point of reference for transcriptional regulation in *Escherichia coli* K12. Here, we present the *regutools R* package to facilitate programmatic access to *RegulonDB* data in computational biology. *regutools* gives researchers the possibility of writing reproducible workflows with automated queries to *RegulonDB*. The *regutools* package serves as a bridge between *RegulonDB* data and the *Bioconductor* ecosystem by reusing the data structures and statistical methods powered by other *Bioconductor* packages. We demonstrate the integration of *regutools* with *Bioconductor* by analyzing transcription factor DNA binding sites and transcriptional regulatory networks from *RegulonDB*. We anticipate that *regutools* will serve as a useful building block in our progress to further our understanding of gene regulatory networks.

Availability and implementation: *regutools* is an *R* package available through *Bioconductor* at bioconductor.org/packages/regutools.

Contact: lcolladotor@gmail.com or alejandro.reyes.ds@gmail.com

Referencias

Alba, B. M., Leeds, J. A., Onufryk, C., Lu, C. Z., & Gross, C. A. (2002). DegS and YaeL participate sequentially in the cleavage of RseA to activate the sigma(E)-dependent extracytoplasmic stress response. *Genes & Development*, *16*(16), 2156–2168. https://doi.org/10.1101/gad.1008902

Balleza, E., López-Bojorquez, L. N., Martínez-Antonio, A., Resendis-Antonio, O., Lozada-Chávez, I., Balderas-Martínez, Y. I., Encarnación, S., & Collado-Vides, J. (2009). Regulation by transcription factors in bacteria: Beyond description. *FEMS Microbiology Reviews*, *33*(1), 133–151. https://doi.org/10.1111/j.1574-6976.2008.00145.x

Barrick, J. E., Corbino, K. A., Winkler, W. C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I., Kenneth Wickiser, J., & Breaker, R. R. (2004). New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. In *PNAS* (Vol. 101). www.pnas.orgcgidoi10.1073pnas.0308014101

Barrios, H., Valderrama, B., & Morett, E. (1999). Compilation and analysis of σ54-dependent promoter sequences. *Nucleic Acids Research*, *27*(22), 4305–4313. https://doi.org/10.1093/nar/27.22.4305

Becker, N. A., Peters, J. P., Lionberger, T. A., & Maher, L. J. (2013). Mechanism of promoter repression by Lac repressor-DNA loops. *Nucleic Acids Research*, *41*(1), 156–166. https://doi.org/10.1093/nar/gks1011

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., & Philips, R. (2005). Transcriptional regulation by the numbers: applications. *Curr Opin Genet Dev*, *15*(2), 125–135. https://doi.org/10.1016/j.gde.2005.02.006

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., & Phillips, R. (2005). Transcriptional regulation by the numbers: Applications. In *Current Opinion in Genetics and Development* (Vol. 15, Issue 2, pp. 125–135). Elsevier Ltd. https://doi.org/10.1016/j.gde.2005.02.006

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., & Phillips, R. (2005). Transcriptional regulation by the numbers: Models. In *Current Opinion in*

Genetics and Development (Vol. 15, Issue 2, pp. 116–124). Elsevier Ltd. https://doi.org/10.1016/j.gde.2005.02.007

Brödel, A. K., Rodrigues, R., Jaramillo, A., Jaramillo, A., Jaramillo, A., & Isalan, M. (2020). Accelerated evolution of a minimal 63-amino acid dual transcription factor. *Science Advances*, *6*(24), 1–10. https://doi.org/10.1126/sciadv.aba2728

Brown, N. L., Stoyanov, J. V., Kidd, S. P., & Hobman, J. L. (2003). The MerR family of transcriptional regulators. In *FEMS Microbiology Reviews* (Vol. 27, Issues 2–3, pp. 145–163). Elsevier. https://doi.org/10.1016/S0168-6445(03)00051-2

Browning, D. F., & Busby, S. J. W. (2004). The regulation of bacterial transcription initiation. *Nature Reviews Microbiology*, *2*, 1–9. https://doi.org/10.1038/nrmicro787

Browning, D. F., Butala, M., & Busby, S. J. W. (2019). Bacterial Transcription Factors: Regulation by Pick "N" Mix. *Journal of Molecular Biology*, *431*(20), 4067–4077. https://doi.org/10.1016/j.jmb.2019.04.011

Busby, S., & Ebright, R. H. (1994). Promoter structure, promoter recognition, and transcription activation in prokaryotes. In *Cell* (Vol. 79, Issue 5, pp. 743–746). Cell Press. https://doi.org/10.1016/0092-8674(94)90063-9

Bush, M. J. (2018). The actinobacterial WhiB-like (Wbl) family of transcription factors. *Molecular Microbiology*, *110*(5), 663–676. https://doi.org/10.1111/mmi.14117

Bush, M. J., Chandra, G., Bibb, M. J., Findlay, K. C., & Buttner, M. J. (2016). Genome-Wide Chromatin Immunoprecipitation Sequencing Analysis Shows that WhiB Is a Transcription Factor That Cocontrols Its Regulon with WhiA To Initiate Developmental Cell Division in Streptomyces. *MBio*, *7*(2), e00523-16. https://doi.org/10.1128/mBio.00523-16

Campagne, S., Allain, F. H. T., & Vorholt, J. A. (2015). Extra Cytoplasmic Function sigma factors, recent structural insights into promoter recognition and regulation. *Current Opinion in Structural Biology*, *30*, 71–78. https://doi.org/10.1016/j.sbi.2015.01.006

Caron, M.-P., Bastet, L., Lussier, A., Simoneau-Roy, M., Massé, E., & Lafontaine, D. A. (2012). Dual-acting riboswitch control of translation initiation and mRNA decay.

Proceedings of the National Academy of Sciences of the United States of America, *109*(50), E3444-53. https://doi.org/10.1073/pnas.1214024109

Cases, I., Lorenzo, V. de, & Ouzounis, C. A. (2003). Transcription regulation and environmental adaptation in bacteria. *Trends in Microbiology*, *11*(6), 248–253. https://doi.org/10.1016/S0966-842X(03)00103-3

Douglas, A. L., & Hatch, T. P. (2000). Expression of the transcripts of the sigma factors and putative sigma factor regulators of Chlamydia trachomatis L2. *Gene*, *247*, 209–214.

Edwards, A. L., & Batey, R. T. (2010). Riboswitches: A Common RNA Regulatory Element. *Nature Education*, *3*(9). https://www.nature.com/scitable/topicpage/riboswitches-a-common-rna-regulatory-element-14262702/

Eichelberg, K., & Galán, J. E. (2000). The flagellar sigma factor FliA (sigma(28)) regulates the expression of Salmonella genes associated with the centisome 63 type III secretion system. *Infection and Immunity*, *68*(5), 2735–2743. http://www.ncbi.nlm.nih.gov/pubmed/10768967

Feklístov, A., Sharon, B. D., Darst, S. A., & Gross, C. A. (2014). Bacterial Sigma Factors: A Historical, Structural, and Genomic Perspective. *Annual Review of Microbiology*, *68*(1), 357–376. https://doi.org/10.1146/annurev-micro-092412-155737

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology*, *19*(2), 99–113. https://doi.org/10.2307/2412448

French, S. L., Santangelo, T. J., Beyer, A. L., & Reeve, J. N. (2007). Transcription and translation are coupled in Archaea. *Molecular Biology and Evolution*, *24*(4), 893–895. https://doi.org/10.1093/molbev/msm007

Galán-Vásquez, E., Sánchez-Osorio, I., & Martínez-Antonio, A. (2016). Transcription factors exhibit differential conservation in bacteria with reduced genomes. *PLoS ONE*, *11*(1). https://doi.org/10.1371/journal.pone.0146901

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2015). Expanded Microbial

genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, *43*(D1), D261–D269. https://doi.org/10.1093/nar/gku1223

Gérard, H. C., Freise, J., Wang, Z., Roberts, G., Rudy, D., Krauß-Opatz, B., Köhler, L., Zeidler, H., Ralph Schumacher, H., Whittum-Hudson, J. A., & Hudson, A. P. (2002). Chlamydia trachomatis genes whose products are related to energy metabolism are expressed differentially in active vs. persistent infection. *Microbes and Infection*, *4*(1), 13–22. https://doi.org/10.1016/S1286-4579(01)01504-0

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., & Eddy, S. R. (2003). Rfam: an RNA family database. *Nucleic Acids Research*, *31*(1), 439–441. https://doi.org/10.1093/nar/gkg006

Grohmann, D., & Werner, F. (2011). Recent advances in the understanding of archaeal transcription. In *Current Opinion in Microbiology* (Vol. 14, Issue 3, pp. 328–334). Elsevier Current Trends. https://doi.org/10.1016/j.mib.2011.04.012

Grundy, F. J., Lehman, S. C., & Henkin, T. M. (2003). The L box regulon: lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(21), 12057–12062. https://doi.org/10.1073/pnas.2133705100

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Ishimoto, K. S., & Lory, S. (1989). Formation of pilin in Pseudomonas aeruginosa requires the alternative σ factor (RpoN) of RNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(6), 1954–1957. https://doi.org/10.1073/pnas.86.6.1954

Jo, Y.-L., Nara, K. ., & Luscombe, N. M. (1986). Purification and characterization of the OmpR protein, a positive regulator involved in osmoregulatory expression of the ompF and ompC genes in Escherichia coli. *Journal of Biological Chemistry*, *261*(32), 15252–15256. https://doi.org/10.1016/s0021-9258(18)66860-7

Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms.

Protein Science, *28*(11), 1947–1951. https://doi.org/10.1002/pro.3715

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. In *Nucleic Acids Research* (Vol. 28, Issue 1, pp. 27–30). https://doi.org/10.1093/nar/28.1.27

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., & Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, *47*(D1), D590–D595. https://doi.org/10.1093/nar/gky962

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457-62. https://doi.org/10.1093/nar/gkv1070

Köhler, T., Harayama, S., Ramos, J. L., & Timmis, K. N. (1989). Involvement of Pseudomonas putida RpoN sigma factor in regulation of various metabolic functions. *Journal of Bacteriology*, *171*(8), 4326–4333. https://doi.org/10.1128/jb.171.8.4326-4333.1989

Koonin, E. V., Makarova, K. S., & Elkins, J. G. (2007). Orthologs of the small RPB8 subunit of the eukaryotic RNA polymerases are conserved in hyperthermophilic Crenarchaeota and "Korarchaeota." *Biology Direct*, *2*(1), 1–5. https://doi.org/10.1186/1745-6150-2-38

Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics 1. *The Annual Review of Genetics*, *39*, 309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725

Kustu, S., Santero, E., Keener, J., Popham, D., & Weiss, D. (1989). Expression of sigma 54 (ntrA)-dependent genes is probably united by a common mechanism. *Microbiology and Molecular Biology Reviews*, *53*(3).

Lee, D. J., Minchin, S. D., & Busby, S. J. W. (2012). Activating transcription in bacteria. *Annual Review of Microbiology*, *66*(June), 125–152. https://doi.org/10.1146/annurev-micro-092611-150012

Leonhartsberger, S., Huber, A., Lottspeich, F., & Böck, A. (2001). The hydH/G genes from

Escherichia coli code for a zinc and lead responsive two-component regulatory system. *Journal of Molecular Biology*, *307*(1), 93–105. https://doi.org/10.1006/jmbi.2000.4451

Lonetto, M. A., Brown, K. L., Rudd, K. E., & Buttner, M. J. (1994). Analysis of the Streptomyces coelicolor sigE gene reveals the existence of a subfamily of eubacterial RNA polymerase σ factors involved in the regulation of extracytoplasmic functions. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(16), 7573–7577. https://doi.org/10.1073/pnas.91.16.7573

Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2015). Archaeal clusters of orthologous genes (arCOGs): An update and application for analysis of shared features between thermococcales, methanococcales, and methanobacteriales. *Life*, *5*(1), 818–840. https://doi.org/10.3390/life5010818

Mandal, M., Boese, B., Barrick, J. E., Winkler, W. C., & Breaker, R. R. (2003). Riboswitches Control Fundamental Biochemical Pathways in Bacillus subtilis and Other Bacteria. *Cell*, *113*(5), 577–586. https://doi.org/10.1016/S0092-8674(03)00391-X

Mandal, M., & Breaker, R. R. (2004). Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nature Structural & Molecular Biology*, *11*(1), 29–35. https://doi.org/10.1038/nsmb710

Martin, R. G., Gillette, W. K., Martin, N. I., & Rosner, J. L. (2002). Complex formation between activator and RNA polymerase as the basis for transcriptional activation by MarA and SoxS in Escherichia coli. *Molecular Microbiology*, *43*(2), 355–370. http://www.ncbi.nlm.nih.gov/pubmed/11985714

Martinez-Guerrero, C. E., Ciria, R., Abreu-Goodger, C., Moreno-Hagelsieb, G., & Merino, E. (2008). GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways. *Nucleic Acids Research*, *36*(Web Server issue), W176. https://doi.org/10.1093/nar/gkn330

Mascher, T. (2013). Signaling diversity and evolution of extracytoplasmic function (ECF) σ factors. *Current Opinion in Microbiology*, *16*(2), 148–155. https://doi.org/10.1016/j.mib.2013.02.001

Merino, E., & Yanofsky, C. (2002). *Regulation by Termination-Antitermination: a Genomic Approach*.

Merino, E., & Yanofsky, C. (2005). Transcription attenuation: A highly conserved regulatory strategy used by bacteria. In *Trends in Genetics* (Vol. 21, Issue 5, pp. 260–264). Elsevier Ltd. https://doi.org/10.1016/j.tig.2005.03.002

Nahvi, A., Sudarsan, N., Ebert, M. S., Zou, X., Brown, K. L., & Breaker, R. R. (2002). Genetic Control by a Metabolite Binding mRNA. *Chemistry & Biology*, *9*(9), 1043–1049. https://doi.org/10.1016/S1074-5521(02)00224-7

Naville, M., & Gautheret, D. (2009). Transcription attenuation in bacteria: Theme and variations. *Briefings in Functional Genomics and Proteomics*, *8*(6), 482–492. https://doi.org/10.1093/bfgp/elp025

Ohnishi, K., Kutsukake, K., Suzuki, H., & Iino, T. (1990). Gene fliA encodes an alternative sigma factor specific for flagellar operons in Salmonella typhimurium. *Molecular & General Genetics : MGG*, *221*(2), 139–147. http://www.ncbi.nlm.nih.gov/pubmed/2196428

Paget, M. S. (2015). Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution. *Biomolecules*, *5*(3), 1245–1265. https://doi.org/10.3390/biom5031245

Pérez-Rueda, E., Collado-Vides, J., & Segovia, L. (2004). Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. *Computational Biology and Chemistry*, *28*(5–6), 341–350. https://doi.org/10.1016/j.compbiolchem.2004.09.004

Perez-Rueda, E., Hernandez-Guerrero, R., Martinez-Nuñez, M. A., Armenta-Medina, D., Sanchez, I., & Ibarra, J. A. (2018). Abundance, diversity and domain architecture variability in prokaryotic DNA-binding transcription factors. *PLoS ONE*, *13*(4), e0195332. https://doi.org/10.1371/journal.pone.0195332

Riordan, J. T., & Mitra, A. (2017). Regulation of Escherichia coli Pathogenesis by Alternative Sigma Factor N. *EcoSal Plus*. https://doi.org/10.1128/ecosalplus.ESP-0016-2016

Rodionov, D. A., Vitreschak, A. G., Mironov, A. A., & Gelfand, M. S. (2002). Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. *The Journal of Biological Chemistry*, *277*(50), 48949–48959. https://doi.org/10.1074/jbc.M208965200

Rodionov, D. A., Vitreschak, A. G., Mironov, A. A., & Gelfand, M. S. (2003). Regulation of lysine biosynthesis and transport genes in bacteria: Yet another RNA riboswitch? *Nucleic Acids Research*, *31*(23), 6748–6757. https://doi.org/10.1093/nar/gkg900

Rowley, G., Spector, M., Kormanec, J., & Roberts, M. (2006). Pushing the envelope: extracytoplasmic stress responses in bacterial pathogens. *Nature Reviews Microbiology*, *4*(5), 383–394. https://doi.org/10.1038/nrmicro1394

Seshasayee, A. S. N., Sivaraman, K., & Luscombe, N. M. (2011). An Overview of Prokaryotic Transcription Factors. In *Sub-Cellular Biochemistry* (Vol. 52). Springer, Dordrecht. https://doi.org/10.1007/978-90-481-9069-0_2

Shen, L., Li, M., & Zhang, Y.-X. (2004). Chlamydia trachomatis s 28 recognizes the fliC promoter of Escherichia coli and responds to heat shock in chlamydiae. *Microbiology*, *150*, 205–215. https://doi.org/10.1099/mic.0.26734-0

Singh, A., Crossman, D. K., Mai, D., Guidry, L., Voskuil, M. I., Renfrow, M. B., & Steyn, A. J. C. (2009). Mycobacterium tuberculosis WhiB3 Maintains redox homeostasis by regulating virulence lipid anabolism to modulate macrophage response. *PLoS Pathogens*, *5*(8). https://doi.org/10.1371/journal.ppat.1000545

Smith, A. M., Fuchs, R. T., Grundy, F. J., & Henkin, T. (2010). Riboswitch RNAs: Regulation of gene expression by direct monitoring of a physiological signal. *RNA Biology*, *7*(1), 104–110. https://doi.org/10.4161/rna.7.1.10757

Smith, L. J., Stapleton, M. R., Fullstone, G. J. M., Crack, J. C., Thomson, A. J., Le Brun, N. E., Hunt, D. M., Harvey, E., Adinolfi, S., Buxton, R. S., & Green, J. (2010). Mycobacterium tuberculosis WhiB1 is an essential DNA-binding protein with a nitric oxide-sensitive iron-sulfur cluster. *The Biochemical Journal*, *432*(3), 417–427. https://doi.org/10.1042/BJ20101440

Soules, K. R., Labrie, S. D., May, B. H., & Hefty, P. S. (2020). Sigma 54-regulated

transcription is associated with membrane reorganization and type iii secretion effectors during conversion to infectious forms of chlamydia trachomatis. *MBio*, *11*(5), 1–19. https://doi.org/10.1128/mBio.01725-20

Taboada, B., Estrada, K., Ciria, R., & Merino, E. (2018). Operon-mapper: A web server for precise operon identification in bacterial and archaeal genomes. *Bioinformatics*, *34*(23), 4118–4120. https://doi.org/10.1093/bioinformatics/bty496

Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. In *Nucleic Acids Research* (Vol. 28, Issue 1). https://doi.org/10592175

Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, *278*(5338), 631–637. https://doi.org/10.1126/science.278.5338.631

Wagner, M., & Horn, M. (2006). The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Current Opinion in Biotechnology*, *17*(3), 241–249. https://doi.org/10.1016/J.COPBIO.2006.05.005

Weinstein-Fischer, D., & Altuvia, S. (2007). Differential regulation of *Escherichia coli* topoisomerase I by Fis. *Molecular Microbiology*, *63*(4), 1131–1144. https://doi.org/10.1111/j.1365-2958.2006.05569.x

Werner, F. (2013). Molecular Mechanisms of Transcription Elongation in Archaea. *Chemical Reviews*, *113*, 8331–8349. https://doi.org/10.1021/cr4002325

Winkler, W. C., Cohen-Chalamish, S., & Breaker, R. R. (2002). An mRNA structure that controls gene expression by binding FMN. *Proceedings of the National Academy of Sciences*, *99*(25), 15908–15913. https://doi.org/10.1073/pnas.212628899

Winkler, W., Nahvi, A., & Breaker, R. R. (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, *419*(6910), 952–956. https://doi.org/10.1038/nature01145

Yang, Y., Darbari, V. C., Zhang, N., Lu, D., Glyde, R., Wang, Y. P., Winkelman, J. T., Gourse, R. L., Murakami, K. S., Buck, M., & Zhang, X. (2015). Structures of the RNA

polymerase-σ54 reveal new and conserved regulatory strategies. *Science*, *349*(6250), 882–885. https://doi.org/10.1126/science.aab1478

Yanofsky, C. (1981). Attenuation in the control of expression of bacterial operons. *Nature*, *289*(5800), 751–758. https://doi.org/10.1038/289751a0