Universitat de Barcelona

Papel de las regiones reguladoras de la expresión génica en la adaptación a factores ambientales en procariotas

Leyden Fernandez Vidal

FACULTAD DE BIOLOGÌA

Programa de Doctorado en Biomedicina

Papel de las regiones reguladoras de la expresión génica en la adaptación a factores ambientales en procariotas

Memoria presentada por Leyden Fernandez Vidal para optar al grado de doctora por la universidad de Barcelona

Barcelona, Septiembre de 2014

Esta tesis ha sido realizada bajo la dirección del Dr. David Torrents Arenales y la tutoría del Dr. Josep Lluis Gelpí, en el departamento de Ciencias de la vida en el Centro Nacional de Supercomputación.

Dr. David Torrents Arenales
Director

Dr. Josep Lluis Gelpí
Tutor

Leyden Fernández Vidal

ABREVIATURAS

ADN	ácido desoxiribonucleico
ADNc	ADN codificante
ARN	ácido ribonucleico
ARNm	ARN mensajero
ARNr	ARN ribosomal
ARNt	ARN transferencia
aTBP	del inglés archaeal TATA Binding Protein
BRE	del inglés, transcription factor B Recognition Element
EST	acrónimo del inglés expressed sequence tag
RNAP	ARN polimerasa
TF	acrónimo del inglés Transcription Factor
TFBSs	acrónimo del inglés TF Binding Sites
TSS	del inglés Transcription Start Site
UTR	del inglés UnTranslated Region
PWM	del inglés Position Weight Matrix

Contenido

I. INTRODUCCIÓN

Muchas cuestiones referentes a los mecanismos moleculares que permiten la adaptación de los organismos vivos al ambiente permanecen sin resolver; aun, cuando la descripción de este problema es anterior a la publicación de El Origen de las Especies por Charles Darwin. Es precisamente, "El origen de las especies", la fuente que muchos citan por contener las primeras anotaciones en biología evolutiva; pero su legado también radica en la persistente vigencia de sus observaciones, por ejemplo, aún no están claros todos los fundamentos biológicos detrás de archiconocidas inferencias como la siguiente:

"Individuals less suited to the environment are less likely to survive and less likely to reproduce; individuals more suited to the environment are more likely to survive and more likely to reproduce and leave their inheritable traits to future generations, which produces the process of natural selection"

Charles Darwin , The origin of species. (Darwin, 1909)

A pesar que en múltiples trabajos científicos se describe el fenotipo de la "especie exitosa", (Edwards et al, 1999) (Darwin, 1909) poco se conoce sobre las bases genéticas, moleculares y fisiológicas que infuyen en la adaptación para cada uno de los diferentes organismos. Por consiguiente, aún permanece sin resolverse el misterio que esconde la selección natural como fuerza impulsora de la supervivencia de un grupo dado en detrimento del resto. La buena noticia al respecto es, que en parte, esta cuestión es posible contestarla a través del estudio de las bases genéticas que condicionan la evolución y la adaptación.

1. La era post-genómica: Contribuciones al estudio de las bases evolutivas de la adaptación a través del análisis de genes y su expresión.

Aunque, la búsqueda de información sobre mecanismos específicos de adaptación a través de la dotación génica de entes biológicos ha sido constante; el primer análisis genómico completo no fue posible hasta 1976, con la secuenciación del Bacteriófago MS2 (Fiers et al, 1976). En los años posteriores los métodos de secuenciación de ADN (Maxam and Gilbert, 1977;Sanger et al, 1977) fueron perfeccionándose, en paralelo con sistemas informáticos de identificación de genes cada vez más robustos (Altschul et al, 1997). De la secuencia de genomas virales se pasa a la de bacterias y eucariotas unicelulares (ensayos a gran escala para conocer la dotación génica de *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae* comienzan en la década de los 90) y se continúa con la secuenciación de eucariotas complejos hasta la publicación en el año 2001 del primer borrador del genoma humano (Lander et al, 2001).

Grandes proyectos siguieron a estos análisis como la secuenciación de diferentes líneas celulares tumorales (International Cancer Genome et al, 2010) y se comenzó también la búsqueda de marcadores genéticos asociados a enfermedad (da Silva Xavier et al, 2013;Reddy et al, 2011). Durante este período también se empezó a conocer la secuencia de ADN de comunidades enteras de microorganismos procariotas, sin aplicar ninguna técnica de cultivo previa. Este nuevo método, basado en la secuenciación de ADN de Bacterias, Arqueas y Virus de muestras tomadas directamente de un entorno natural se le denominó metagenómica (Tringe et al, 2005). La metagenómica tiene amplia aplicabilidad en biomedicina y ecología (Arumugam et al, 2011) (Tringe et al, 2005) y se complementa con otras especialidades como la metatranscriptómica y la metaproteómica. Estás técnicas en su conjunto permiten conocer tanto la secuencia de ADN de los microorganismos, como los perfiles de expresión dentro de comunidades microbianas en momentos específicos.

En la actualidad, la subsecuente secuenciación y el análisis de genomas completos ha proporcionado, junto con la identificación de genes específicos una visión general sobre pérdidas y ganancias de funciones que se correlacionan con las características particulares de las especies o sus hábitats. (International Chicken Genome Sequencing, 2004) (Dooley, 2001).

1.1 Principales enfoques para evaluar los mecanismos de adaptación de los procariotas a sus hábitats naturales.

Hasta hace pocos años el estudio de los microorganismos estuvo limitado a cepas que se pueden cultivar en el laboratorio bajo condiciones controladas. Estos trabajos han aportado datos y han permitido avances importantes en el conocimiento de la biología básica de los microorganismos (Uruburu, 2003). Pero, por otro lado, la ausencia de la interacción con el entorno natural donde estos microorganismos coexisten con otras comunidades bacterianas y de arqueas simplifica y desvía la información que estas aproximaciones y estudios de laboratorio pueden obtener sobre como afectan diferentes variables medioambientales a la biología molecular y la bioquímica de los procariotas. Por ejemplo, el conocimiento e incorporación de varios aspectos en el estudio, tales como la disponibilidad de nutrientes, las características físico-químicas del entorno y, sobretodo, la interacción con otras especies dentro de la comunidad, son imprescindibles para responder a un gran número de preguntas relevantes sobre la supervivencia de los microorganismos en diferentes hábitats (Tringe et al, 2005).

El avance tecnológico en el campo de la secuenciación impulsó la disciplina de la metagenómica, herramienta que permite el estudio de comunidades enteras dentro de su propio nicho. Las etapas de un análisis metagenómico típico consisten en: toma y procesamiento de la muestra, secuenciación, ensamblaje, ordenamiento de secuencias de ADN dentro de grupos que pueden representar un genoma individual o genomas de organismos cercanos evolutivamente. Posteriormente se procede a la anotación de funciones y al análisis estadístico (Thomas et al, 2012). En la Figura 1 se representa cada una de estas etapas de un análisis metagenómico típico y a conticuación se detalla en que consiste cada una de ellas.

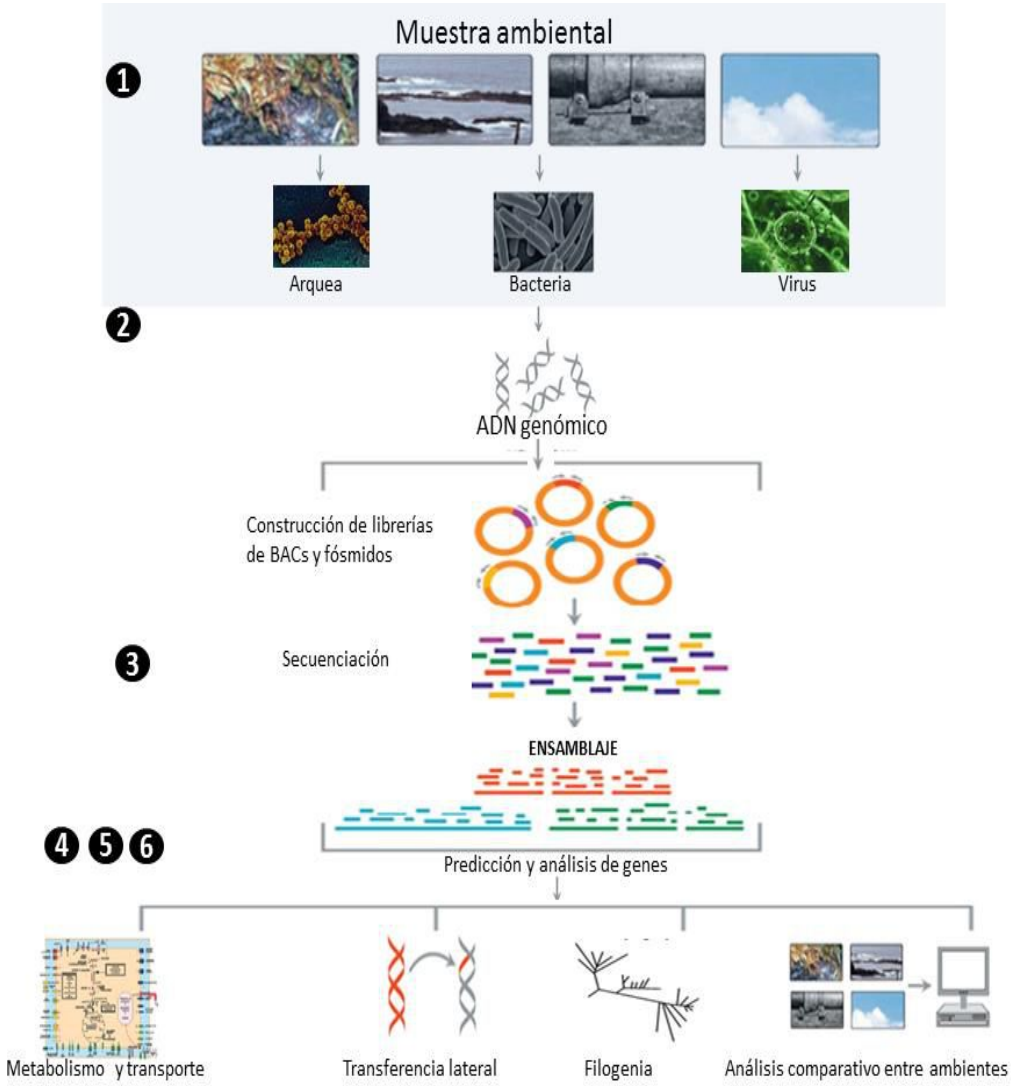


Figura 1. Etapas de un análisis metagenómico. (1) Toma y procesamiento de la muestra, (2) secuenciación, (3) ensamblaje, (4) ordenamiento de secuencias de ADN dentro de grupos que pueden representar un genoma individual o genomas de organismos cercanos evolutivamente (Análisis filogenético y transferencia lateral de genes), (5) anotación de funciones que permiten descifrar, por ejemplo, adaptaciones metabólicas típicas de un ambiente dado (6) análisis estadístico para comparar entre ambientes. Imagen modificada tomada de http://sgugenetics.pbworks.com/.

(1) Toma y procesamiento de muestra: Es la etapa crucial de cualquier proyecto en metagenómica. El ADN extraído debe ser representativo de todas las células presentes en la muestra y además existir una cantidad suficiente de ácidos nucleicos de alta-calidad para la subsecuente producción de la biblioteca genómica. Posteriormente, el procesamiento requiere protocolos específicos para cada tipo de muestra (por ejemplo si es asociado a huesped o de un ambiente externo: marino o terrestre) (Burke et al, 2009;Delmont et al, 2011;Venter et al, 2004).

(2) Secuenciación: Inicialmente los estudios de metagenómica estaban enfocados en la construcción de librerías de BACs y fósmidos, para la posterior búsqueda y selección de clones con interés funcional y evolutivo. Aunque esta aproximación es todavía útil, está siendo reemplazada por la secuenciación indiscriminada de fragmentos de ADN (shotgun), debido a que es una técnica más barata y que incrementa el rendimiento (Hugenholtz and Tyson, 2008). En la actualidad es posible la secuenciación de microorganismos en casi cualquier ambiente mediante las técnicas de nueva generación (NGS, por Next Generation Sequencing) (Thomas et al, 2012).

(3) Ensamblaje: se aplica cuando el estudio está enfocado en la recuperación del genoma de organismos no cultivados hasta el momento y también para la obtención de ADN codificante de larga longitud para una posterior caracterización de la comunidad. El ensamblaje se basa en la unión de pequeños fragmentos para construir otros más grandes llamados contigs. En la actualidad, debido a que la mayoría de programas de ensamblaje han sido diseñados para genomas individuales, su utilidad para mezclas de pan-genomas complejos debe ser evaluada con precaución. Dos estrategias se emplean para muestras metagenómicas: el ensamblaje basado en un genoma de referencia (co-ensamblaje) y el ensamblaje *de novo* (Thomas et al 2012). El co-ensamblaje puede ser realizado con programas tales como, Newbler (Roche), AMOS (http://sourceforge.net/projects/amos/), o MIRA (Chevreux et al, 2004). En general, los programas basados en genomas de

referencia funcionan bien, si los datos metagenómicos contienen secuencias parecidas a genomas conocidos. No obstante, las diferencias entre el genoma de la muestra y el de referencia, tales como: largas inserciones, polimorfismos o deleciones, ocasionan que el ensamblaje se fragmente o que no cubra todas las regiones de los nuevos genomas. A diferencia del co-ensamblaje, el *de novo*, requiere un mayor poder computacional y la mayoría de herramientas actuales como Velvet (Zerbino and Birney, 2008) o SOAP (Li et al, 2008) están basados en el algoritmo de Bruijn, que permite la manipulación de gran cantidad de datos (Miller et al, 2010;Pevzner et al, 2001). El hecho que muchas (si no todas), las comunidades microbianas incluyan variaciones significativas al nivel de cepa y especie, hace difícil el uso de algoritmos de ensamblaje que asuman como referencia genomas secuenciados a través de organismos cultivados en el laboratorio. Asumir que los genomas de organismos aislados de ambientes naturales "se parecen" al de los organismos que crecen bajo condiciones controladas, puede traer consigo la supresión de información de grupos heterogéneos cuando se ensamblan los contigs. El ensamblaje también conduce a otro problema, causado por la presencia de fragmentos de ADN de especies menos abundantes dentro de la comunidad y que tendrán una calidad menor que aquellos fragmentos de ADN de especies muy abundantes (Thomas et al 2012). Esta limitación hace que en metagenómica si la cobertura o profundidad de la secuenciación es baja el ensamblaje de los organismos menos representados dentro de la comunidad sea imposible.

(4) Ordenamiento de los fragmentos de ADN ensamblados (contigs) dentro de grupos de genomas individuales o especies evolutivamente cercanas: Los algoritmos más usados en la actualidad incluyen: a Phylopythia (McHardy et al, 2007), IMG/M (Markowitz et al, 2008;Markowitz et al, 2012), MG-RAST (Glass et al, 2010), MEGAN (Huson et al, 2007), CARMA (Krause et al, 2008), MetaPhyler (Liu et al, 2011), PhymmBL (Brady and Salzberg, 2009) y MetaCluster (Leung et al, 2011). Todas estas herramientas emplean diferentes métodos para agrupar secuencias que van desde mapas autoorganizados hasta agrupaciones jerárquicas.

(5) Anotación: La anotación de metagenomas está basada en clasificar las secuencias usando funciones y taxonomías conocidas. La clasificación se realiza a través de búsquedas de homología empleando bases de datos de secuencias conocidas. Para estas búsquedas el algoritmo más empleado es el BLAST, pero tiene el inconveniente que es computacionalmente costoso. Con el crecimiento de las bases de datos, cada vez es más urgente la necesidad de crear algoritmos más rápidos y algunos programas de búsquedas de similitud de secuencias han sido desarrollados para resolver este problema (Edgar, 2010;Wang et al, 2009;Ye et al, 2011).

(6) Análisis estadístico: Debido al alto costo de la secuenciación, los primeros metagenomas analizados, no fueron replicados o el estudio fue diseñado de manera simple para la exploración de genomas de organismos específicos, como por ejemplo, especies resistentes a las condiciones de cultivos (Tyson et al, 2004). La reducción del costo de la secuenciación aparejado con una mayor apreciación por parte de la comunidad científica de las ventajas de la metagenómica para responder preguntas fundamentales en ecología microbiana, hizo que los diseños experimentales de estudios posteriores se tornaran más complejos. Así, uno de los más recientes objetivos en la metagenómica consiste en unir la información funcional y filogenética con parámetros físicos, químicos y biológicos del entorno (Prosser, 2010). La medición de estos parámetros requiere mucho tiempo y un costo monetario elevado, lo que ocasiona que hasta la fecha, no se registre información suficiente sobre ellos (Markowitz et al, 2012;Tossici-Bolt et al, 2011).

La introducción de la secuenciación de alto rendimiento (en inglés, High Throughput Sequencing), está revolucionando el campo de la ecología microbiana. La gran cantidad de datos y la complicada estadística necesaria para su análisis ha llevado a la generación de nuevos paquetes de programas como Mothur (Schloss et al, 2009) y Qiime (Caporaso et al, 2010). Ambos permiten el uso de datos derivados de la secuenciación de comunidades sin ensamblar previamente. Por ejemplo, Qiime, el cual está disponible en http://qiime.sourceforge.net/ permite el procesamiento de un amplio rango de comunidades microbianas (tanto secuencias de 16S ARNr como “shotgun”). Además incluye la

visualización de los datos analizados mediante histogramas, redes y arboles filogenéticos (Caporaso et al, 2010). Debido a su capacidad de análisis rápida y robusta, Qiime y Mothur son cada día más empleados en ecología microbiana.

En resumen, los resultados de estudios metagenómicos han provisto un mayor conocimiento acerca de la estructura y potencial genómico de comunidades microbianas así como sobre la evolución de los microorganismos y sus peculiaridades ecológicas. En este sentido, métodos basados en el análisis de secuencia y función de fragmentos de ADN metagenómico han resultado en la identificación de nuevos genes y sus productos (Daniel, 2005;Handelsman, 2004). Conocer el genoma da información sobre lo que potencialmente puede expresarse, pero, también es importante conocer las funciones que responden a determinada condición ambiental. Así, el éxito derivado de la secuenciación de muestras extraídas de ambientes naturales ha provocado un cambio de enfoque hacia el estudio de la expresión génica y la identificación de distintos perfiles de transcripción de genes, análisis estos enfocados también en ahondar más en el conocimiento sobre los factores biológicos que afectan el fenotipo de las especies (Khaitovich et al, 2004a;Khaitovich et al, 2004b).

La metatranscriptómica, disciplina basada en el estudio de los niveles de expresión génica, conlleva a un análisis experimental más complejo que la metagenómica. No obstante, para la completa comprensión acerca del modo empleado por comunidades de organismos procariotas para responder a fluctuaciones rápidas de parámetros medio-ambientales; la mejor herramienta consiste en la exploración de los perfiles de ARNm de genes que constituyan marcadores taxonómicos o funcionales. Los estudios de metatranscriptómica han empleado inicialmente el uso de microarreglos o bibliotecas de ADNc (también conocidos como EST, acrónimo del inglés *e*xpressed *s*equence *t*ag) obtenidos a partir de ARNm de la muestra. Algunos factores desde el punto de vista práctico han enlentecido el desarrollo de la metatranscriptómica como por ejemplo, los ARNm de bacterias y arqueas no son poliadenilados. Esto ocasiona que los métodos típicos empleados en Eucariotas para la captura de ARNm no sean aplicables en ecología ambiental microbiana. Por tanto, una de las opciones es remover el ARNr del total de ARN presente en la extracción, lo cual resulta en un bajo rendimiento de genes expresados. Hay que sumar a este problema otro reto relacionado con el tiempo de vida media del ARN, que puede llegar a ser sólo de minutos incluso bajo condiciones óptimas. Para solventar estas limitaciones se han descrito

diferentes metodologías, el resumen de algunas de ellas aplicadas a estudios metatranscriptómicos en muestras oceánicas (Frias-Lopez et al, 2008;Gilbert et al, 2008;Poretsky et al, 2009) se ilustra en la Figura 2.

A nivel conceptual existe otro inconveniente en la aplicación de la metatranscriptómica, ocasionado porque la abundancia de proteínas en la célula no se correlaciona siempre con su actividad. Algunos genes se expresan de manera constitutiva y la actividad enzimática se regula también postranscripcionalmente. Lo que hace necesario la aplicación de otros métodos complementarios para conocer la gama de estrategias de adaptación al ambiente presentes en diferentes microorganismos. Por ejemplo, la metaproteómica es la disciplina que estudia el comportamiento de las proteínas expresadas en una comunidad de microorganismos, ofreciendo una mejor descripción de las funciones metabólicas. Su basamento general consiste en la extracción y separación de proteínas para su posterior detección e identificación a través de técnicas como MS/MS (del inglés Tandem mass spectrometry). Problemas derivados de la aplicación de este método hacen que la metaproteómica sea una estrategia aún más onerosa que la metatranscriptómica (Jagtap et al, 2012;Wilmes and Bond, 2006).

A continuación se detallan los resultados obtenidos al aplicar las nuevas estrategias de metagenómica y metatranscriptómica enfocadas en la dilucidación de los principales mecanismos de adaptación de los procariotas a medios naturales.

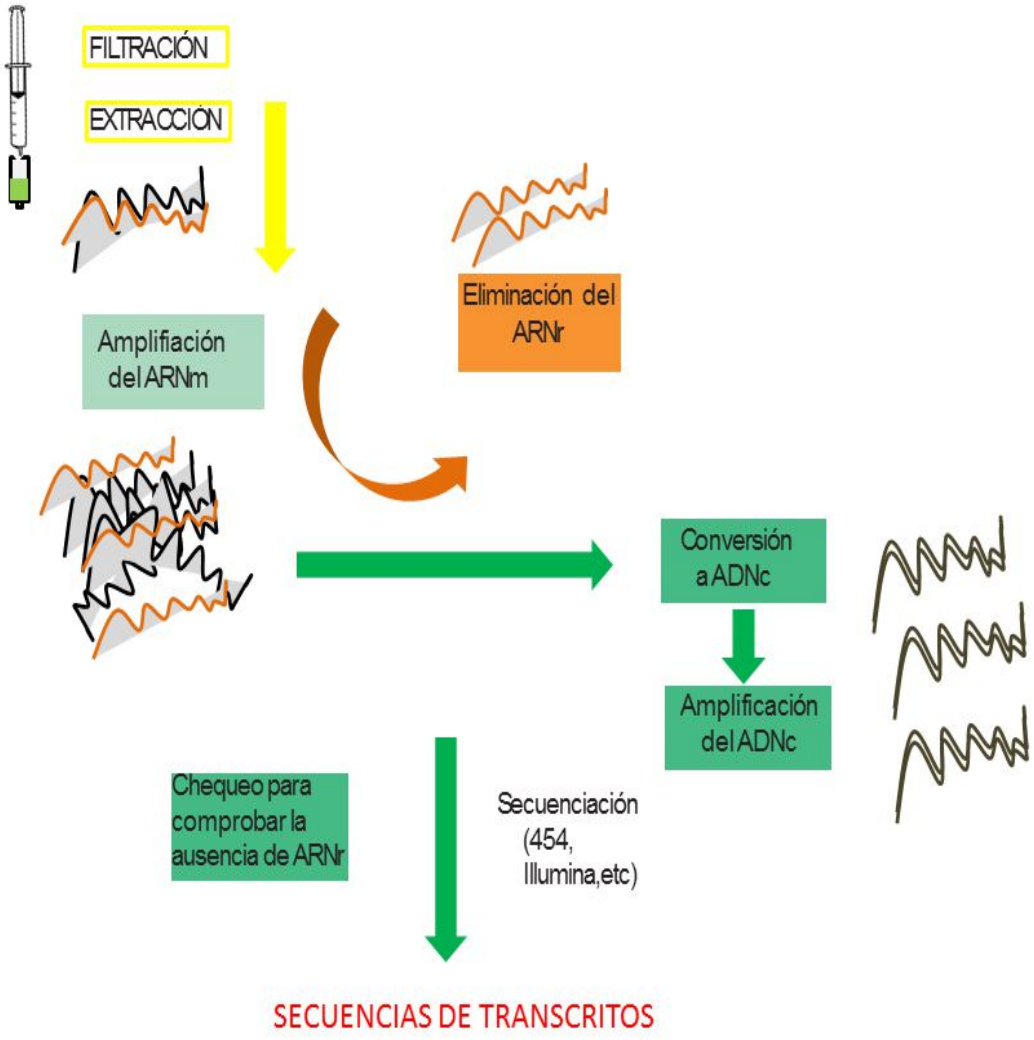


Figura 2. Protocolo para la secuenciación del metatranscriptoma de comunidades microbianas. La recolección y estabilización rápida del ARN reduce su degradación. Así como la rápida eliminación del ARNr incrementa la proporción final de ARNm. Los transcritos pueden ser amplificados antes o después de la conversión a ADNc siendo la pirosecuenciación una técnica efectiva para eliminar desviaciones asociadas a la clonación. Este protocolo resume varios métodos descritos por Frias-Lopez, Gilbert, Poretsky y colaboradores (Frias-Lopez et al, 2008;Gilbert et al, 2008;Poretsky et al, 2009)

1.2 Aportaciones de los estudios de metagenómica y metatranscriptómica al campo de la ecología microbiana.

Diferentes nichos ecológicos han sido caracterizados a través de técnicas basadas en la secuenciación de muestras tomadas directamente del ambiente. En uno de los primeros estudios comparativos en el campo de la metagenómica, se analizaron muestras de suelo y aguas marinas superficiales y profundas. Mediante la secuenciación de microorganismos que cohabitan estos nichos se comprobó la existencia de vías metabólicas diferentes asociadas a la disponibilidad de nutrientes y la presencia de ortólogos[1] exclusivos de determinado ambiente, así como divergencias en el número de operones involucrados en el transporte de iones y componentes inorgánicos dependiendo de la composición química del entorno (Tringe et al, 2005).

Más adelante siguiendo la misma línea comparativa, pero sólo considerando ambientes acuáticos, se observó gran flexibilidad en las vías de conversión de energía y otras relacionadas con el metabolismo de los aminoácidos. Provocado por el alto coste asociado a la obtención de ciertos cofactores, como la cobalamina, que contiene cobalto, un metal escaso en ambientes marinos (Gianoulis et al, 2009).

Los microorganismos que cohabitan los sedimentos de ríos y mares, constituyen otro caso de estudio interesante. Un nicho muy bien caracterizado son los restos de ballena en el fondo oceánico. Varias muestras de este particular tipo de sedimento se han tomado desde diferentes puntos, por ejemplo, en la bahía Santa Susana en California y también cerca de la península Antártica. El análisis de estas comunidades demostró que carecen de bacteriorodopsina, que sí está presente en las aguas superficiales expuestas a la luz solar (Tringe et al, 2005).

---

[1] Las secuencias ortólogas son aquellas que tienen un alto grado de similitud entre sí debido a que se han originado de un ancestro común, pero se encuentran en diferentes especies (Fitch, 1970).

A diferencia del hábitat oceánico, las comunidades de ambientes terrestres, por ejemplo suelos dedicados a la agricultura, presentan gran cantidad de ortólogos relacionados con el metabolismo de los carbohidratos (genes de celobiosa fosforilasa, necesarios para la degradación de material proveniente de plantas sólo se encontraron en suelo dedicado a la agricultura en un estudio comparativo entre ambientes terrestres y acuáticos) (Tringe et al, 2005). Sin embargo, las comunidades microbianas que habitan suelos desérticos tienen pocos genes asociados al catabolismo de compuestos orgánicos derivados de plantas, pero los genes asociados a osmoregulación y dormancia son muy abundantes (Fierer et al, 2012). La metagenómica es también una herramienta útil para conocer como el cambio climático afecta a comunidades biológicas. Un análisis metagenómico de una comunidad microbiana de aguas subterráneas, expuesta de manera prolongada a altas concentraciones de metales pesados y solventes orgánicos, presentó una reducción significativa en el número de especies, diversidad alélica y metabólica. Después de 50 años de exposición a tóxicos las especies sobrevivientes se enriquecieron en genes que confieren resistencia a metales pesados y acetona, siendo la transferencia lateral de genes un factor esencial en la adaptación y supervivencia de la comunidad (Hemme et al, 2010).

Otro ambiente extremo, cuyas poblaciones de microorganismos han sido secuenciadas se encuentra en Richmond Mountain, California. La principal característica de este nicho es la presencia del mineral pirita, así como los bajos valores de pH (Tyson et al, 2004). Las especies más abundantes en este nicho pertenecen a *Ferroplasma* y *Leptospirillum* y su energía metabólica se genera fundamentalmente de la oxidación del hierro. Otra característica común a todos los microorganismos que habitan este nicho es el enriquecimiento en genes asociados con la detoxificación. Por ejemplo, los sistemas de eflujo de protones son probablemente los responsables del mantenimiento del pH intracelular cercano al neutro y de extraer metales del interior celular que puedan ser tóxicos para la misma Los genomas de ambos grupos también contienen gran número de cadenas transportadoras de electrones, pero su secuencia y estructura difieren dependiendo de la especie (Handelsman, 2004).

Unido a la búsqueda de funciones hábitat-específicas, el estudio de los perfiles de expresión génica está ganando protagonismo. En particular, el efecto de la radiación solar sobre comunidades acuáticas, ha permitido conocer como la expresión se modifica en

correspondencia a los cambios en la luz solar recibida (Poretsky et al, 2009). La información provista a través de un análisis de aguas superficiales del océano Pacífico Norte, mostró que algunas vías metabólicas estuvieron muy activas durante la noche y otras durante el día. Se demostró así gran flexibilidad en la regulación génica siguiendo patrones diurnos-nocturnos. En ausencia de luz, funciones relacionadas con el metabolismo de los carbohidratos estuvieron sobre-representadas en los perfiles de expresión. A diferencia del horario diurno, cuando se hizo más abundante la presencia de transcritos relacionados con fotosíntesis y metabolismo energético en general (Poretsky et al, 2009).

Con estos ejemplos se ilustra como el ambiente condiciona la presencia y abundancia de determinados genes, interfiere también en la organización del genoma y modifica la expresión génica. Aunque existen algunos estudios, todavía queda mucho por esclarecer sobre la regulación de la expresión en la adaptabilidad a determinadas propiedades físico-químicas del entorno.

2. Regulación de la expresión génica en procariotas.

Las especies bacterianas y de arqueas reaccionan ante las condiciones ambientales mediante diferentes modos de regulación, por ejemplo, metabólica, traduccional y transcripcional. A través de la regulación transcripcional se modifican sus patrones de expresión génica por la exposición a factores ambientales dinámicos en sus nichos naturales. Las alteraciones controladas de los patrones de expresión garantizan la supervivencia frente a parámetros variables del entorno, como las fluctuaciones en la concentración de nutrientes. Los genes se organizan dentro de redes jerárquicas interconectadas conocidas también como regulones. La expresión de estos regulones está, a su vez, controlada por proteínas reguladoras de la expresión o factores de transcripción (TF, acrónimo del inglés Transcription Factor), que reconocen sitios de unión en el ADN (siglas TFBSs, acrónimo del inglés TF Binding Sites) (van Hijum et al, 2009). La regulación de la expresión génica o regulación de la transcripción es un fenómeno muy complejo que se lleva a cabo en el promotor. La estructura de esta región del ADN se espera que también se modifique por la acción de los parámetros físico-químicos propios de cada ambiente; para permitir así la expresión de determinados genes u operones de manera controlada.

El mecanismo de transcripción permite que una cadena de ADN sea reescrita a ARN (por

ejemplo, ARNm, ARNt, ARNr y pequeños ARNs) y es llevado a cabo por la ARN polimerasa (RNAP). Este proceso consta de 5 fases: (1) pre-iniciación, (2) iniciación, (3) disgregación del promotor (donde se encuentran los sitios de unión de la RNAP y TFs), (4) elongación y (5) terminación (Browning and Busby, 2004;van Hijum et al, 2009). A continuación se detallan las características de las estructuras biológicas que condicionan las bases de la regulación génica, tales como: promotores, factores de transcripción y los mecanismos de regulación en Bacterias y Arqueas.

2.1 Estructura de los promotores de Bacterias.

Algunos genes se transcriben frecuentemente mientras otros rara vez o nunca. La competición entre sus promotores se debe a la baja concentración de la RNAP y factores sigma dentro de la célula. Este hecho ocasiona que las etapas críticas de la regulación transcripcional ocurran principalmente al inicio de la unión de la RNAP al ADN, durante el proceso de isomerización y en los estadios tempranos donde la enzima comienza su desplazamiento sobre la doble cadena de ADN. Los factores sigma también son importantes en la regulación de la expresión por varias razones: (1) su rol en el proceso de reconocimiento de la RNAP y el promotor (2) su efecto sobre el posicionamiento de la RNAP a su secuencia diana y (3) su participación en el proceso de desenrrollamiento del ADN cerca del TSS (del inglés Transcription Start Site) (Wosten, 1998). El genoma de un organismo puede codificar diferentes factores sigmas que unido a factores de transcripción específicos son necesarios para guiar a la RNAP a un grupo diana de genes, de forma precisa.

El promotor bacteriano está formado también por las regiones -10 y -35 (posiciones relativas al TSS en pares de bases) y constan de una secuencia consenso. La fuerza del promotor (o lo que es lo mismo el nivel al que un gen se transcribe) depende de cuan parecida sea la secuencia real a la consenso; unido al efecto que provoque la unión de factores de transcripción adicionales (Kobayashi et al, 1990). Otros sitios ubicados en el promotor con un determinado papel en la regulación son el elemento -10 extendido y el elemento UP (Figura 3A) (Browning and Busby, 2004).

La célula bacteriana tiene mecanismos adicionales de regulación de la expresión génica liderados por los factores de transcripción, Los cuales se unen a secuencias en el ADN que

estos reconocen (TFBSs) Los sitios de unión para un mismo TF difieren en la composición de la secuencia nucleotídica, pero a pesar de este hecho, pueden ser representados por motivos consensos, y su longitud más común está entre 12 y 30 pares de bases. La localización y la composición nucleotídica de estos sitios determinan en gran medida si un TF reprime o activa la expresión de cierto gen. Es muy común en bacterias que las secuencias a las que se une la proteína reguladora sean repeticiones directas o palíndromas para facilitar la unión de los TFs en forma de dímeros (Rodionov, 2007). Debido a que muchos TF en bacterias tienen un dominio hélice-giro-hélice y actúan como homodímeros, luego los TFBSs se estructuran como una diada o motivo con un brazo espaciador constituido por un número dado de nucleótidos sin información relevante (van Helden et al, 2000).

Aunque la existencia de restricciones espaciales para la ubicación de los TFBSs en los promotores es un hecho obvio, no se encuentran estudios experimentales suficientes que las especifiquen. Por ejemplo, algunos de estos estudios acerca de los TFBSs de represores han reportado una mayor densidad de sitios entre las posiciones -60 a +60 relativas al TSS (Figura 3B) (Collado-Vides et al, 1991;Espinosa et al, 2005;Madan Babu and Teichmann, 2003), aunque también se pueden encontrar sitios de unión más allá de la posición -60 (Lanzer and Bujard, 1988). Por otro lado, la mayoría de los activadores Clase 1 se sitúan generalmente entre las posiciones -60 y -95, mientras que los activadores clase 2 se encuentran adyacentes o solapando el sitio -35 (Barnard et al, 2004) (Figura 3 A,C).

En un estudio de promotores de *E. coli* se estimó la longitud relativa de la región 5′ UTR (del inglés UnTranslated Region). La mayoría de estas regiones que se extienden desde el TSS al primer codón ATG varían fundamentalmente entre 20 y 40 nucleótidos, aunque en algunos promotores puede llegar hasta 290 nucleótidos. Las regiones 5′ UTR de larga longitud pueden contener elementos reguladores de ARN conservados o riboswitches (Mendoza-Vargas et al, 2009).

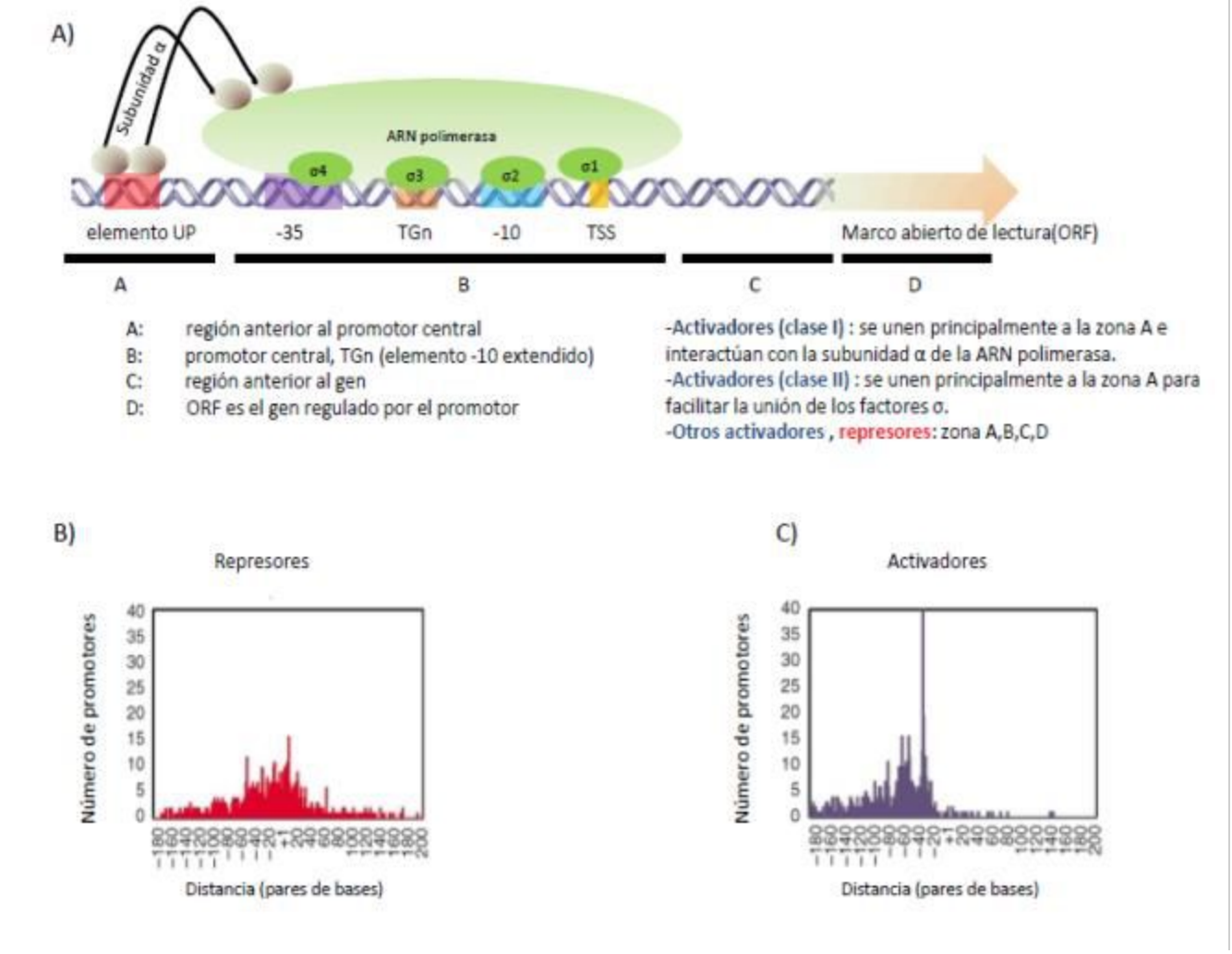


Figura 3. A) Mecanismo molecular y estructuras del promotor, que participan en la modulación de la transcripción en bacterias (van Hijum et al, 2009). B) y C) Distribución de los sitios de unión de factores de transcripción en promotores de *E.Coli* (1102 promotores) en la base de datos RegulonDB (Gama-Castro et al, 2008). Las localizaciones de los sitios son relativas al TSS (+1). En la gráfica B) se representa la localización de los sitios de unión de los represores, mientras que en C) se muestran activadores (Cox et al, 2007;Gama-Castro et al, 2008). Imágenes tomadas de (van Hijum et al, 2009).

2.2 Estructura de los promotores de Arqueas.

Los promotores de arqueas son tripartitos. Contienen una región rica en A+T parecida a los elementos TATA de eucariotas y ubicados relativamente en la misma posición en el promotor. Estos elementos son reconocidos por la aTBP (del inglés archaeal TATA Binding Protein). Su localización es anterior al TSS a una distancia relativa de 25 pares de bases. Este elemento fue llamado anteriormente como caja A (Thomm, 1996), pero después fue renombrado como caja TATA. La segunda partición del promotor de arquea es el elemento BRE (del inglés, transcription factor B Recognition Element). Se localiza inmediatamente adyacente a la caja TATA y es importante para la fortaleza de expresión del promotor y como guía orientativa para el complejo de iniciación de la transcripción (Bell and Jackson, 1998). En la región alrededor del TSS se encuentra el elemento iniciador conocido por las siglas INR y constituye la tercera partición del promotor de arqueas (Figura 4).

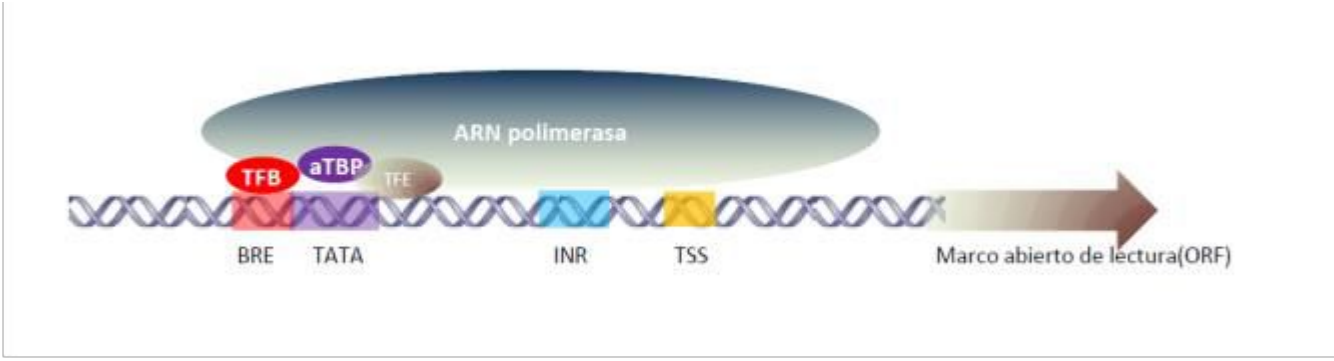


Figura 4. Estructura general del promotor en arqueas y factores de transcripción generales (basales y adicionales. El elemento BRE y la caja TATA se ubican cerca de la posición -35 mientras que el elemento INR se ubica cercano al TSS (Soppa, 1999).

La transcripción en arqueas requiere la formación de un multi-complejo de subunidades de ARN polimerasa y dos factores de transcripción el aTBP y el TFB (siglas provenientes de la denominación en inglés Transcription Factor B). El factor aTBP se une a la caja TATA y su interacción es estabilizada por la unión de la proteína TFB (Rowlands et al, 1994) que también reconoce al elemento BRE.

Además de los factores de transcripción basales mencionados en los párrafos anteriores algunas especies de arqueas necesitan factores adicionales como el TFE (siglas provenientes del inglés Termed Transcription factor E). TFE actúa como estimulador de

algunos promotores a través de facilitar la unión de la proteína TBP a la caja TATA. Otro descubrimiento fascinante en la transcripción de arqueas es su capacidad de codificar múltiples homólogos de la proteína TFB y/o TBP. Esta característica le permite reconocer diferentes secuencias BRE y por tanto discriminar entre varios subtipos de promotores. Aunque el uso selectivo de diferentes isoformas en la maquinaria de transcripción basal permite regular la expresión, es improbable que provea suficiente capacidad reguladora para la adaptación de la células de arquea. De hecho, se ha demostrados que los genomas de las arqueas presentan muchos homólogos de genes que codifican para otros factores de transcripción específicos de genes y no globales o generales como los mencionados anteriormente (Bell and Jackson, 2001).

A pesar que la compleja ARN polimerasa de arqueas se asemeja mucho a la ARN polimerasa II de linajes eucarióticos la mayoría de sus reguladores transcripcionales están más relacionados con los de bacteria (Bell and Jackson, 2001;Geiduschek and Ouhammouch, 2005). Por ejemplo, una de las proteínas reguladoras más ampliamente distribuidas dentro del reino de arqueas pertenece a la familia Lrp/AsnC (Figura 5). Las proteínas de esta familia se caracterizan por poseer un dominio hélice-giro-hélice semejante al de muchos reguladores de bacteria y muy extendido dentro de arqueas. Similar también a los sitios de unión de bacterias son muchos de los sitios de arqueas formando estructuras palíndromas tal como se observa en la Figura 5 (Geiduschek and Ouhammouch, 2005).

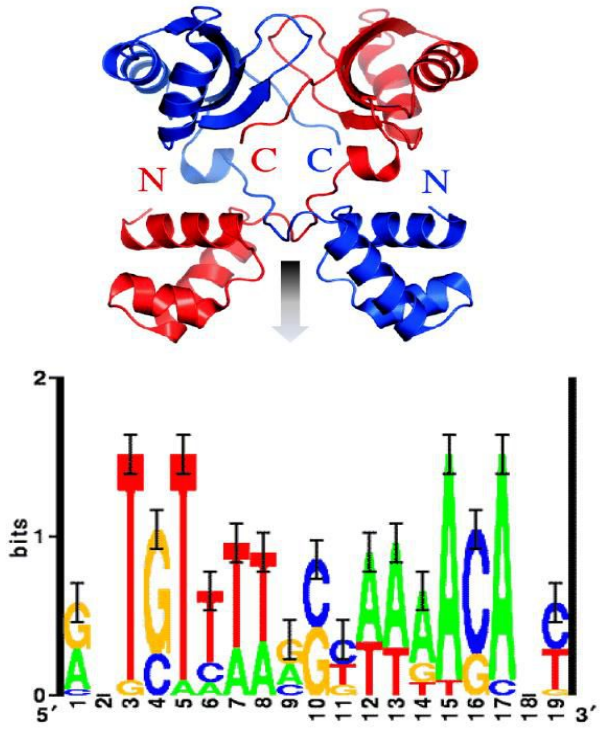


Figura 5. Regulador transcripcional Lpr y sus sitios de unión. (Geiduschek and Ouhammouch, 2005). Los sitios de unión más frecuentes de este regulador siguen una estructura palíndroma de la forma (TGTTTNNNNAAACA). Los sitios para Lrp fueron tomados de http://www.wadsworth.org/. Para más información acerca del significado de la notación de los sitios de unión ver epígrafe 2.4.1.

Algunos de los TFs reprimen la transcripción, otros la activan, pero también pueden tener una acción dual dependiendo de la ubicación del sitio de unión relativa al TSS. A continuación se detallan los mecanismos moleculares responsables de la regulación de la expresión en las células procariotas.

2.3 Mecanismos de regulación de la transcripción en procariotas.

Aunque los mecanismos de activación y represión son similares en procariotas, existen peculiaridades entre bacterias y arqueas.

2.3.1 Bacterias.

Los factores mencionados en la sección 2.1 (como, TFs y factores σ) no son los únicos involucrados en el mecanismo de regulación de la expresión génica en bacterias. El enlace o liberación de represores y activadores se controla en algunas ocasiones por los cofactores. Los cofactores son moléculas que varían en tamaño y en su naturaleza. Por lo que, pequeños iones o nucleótidos hasta azúcares o grupos fosfatos unidos covalentemente a péptidos o proteínas (Aki et al, 1996;Fabret et al, 1999) pueden actuar como cofactores. Aunque muchos activadores ejercen su acción al enlazarse al promotor, otros pueden unirse directamente a la RNAP en el citosol, por ejemplo, el factor de transcripción soxS de *E. coli* (Griffith et al, 2002). A pesar de estos mecanismos alternativos, el papel protagónico en la orquestación de la regulación de la expresión lo tienen los factores de transcripción y su unión al promotor del gen.

Hay descritos cuatro mecanismos principales que permiten a un TF ejercer una acción represiva sobre la transcripción: 1) represión por hibridación estérica, frecuentemente por unión del represor entre o sobre el promotor central; 2) represión mediante el bloqueo de la elongación de la transcripción, cuando el TF se enlaza cerca del inicio de la región codificadora; 3) represión por formación de un lazo en el ADN (en este caso, una interacción entre dos monómeros del mismo TF es posible si los sitios de unión en el ADN están separados por la distancia adecuada) y 4) represión por la modulación de un activador. En este último mecanismo el sitio de unión de un represor parcialmente solapa al de un activador. Por tanto, cuando el represor se enlaza, previene la unión del activador a su respectivo TFBS. Un ejemplo de esta interacción ocurre entre los factores de transcripción CytR y CRP (Browning and Busby, 2004).

Para la activación también se han descrito cuatro mecanismos fundamentales (Barnard et al, 2004;Browning and Busby, 2004). 1) Activación clase I, ocurre cuando los TFs se

enlazan delante del promotor central e interactúan con la subunidad-α de la ARN polimerasa (Figura 3A); 2) la activación clase II en la cual el TF se une al ADN directamente adyacente al promotor central y promueve a su vez el enlace de los factores σ; 3) activación por cambio conformacional del ADN, donde el TF se enlaza al promotor central y lo capacita para la unión de los factores σ, mediante la torsión de la hélice del ADN y 4) activación por modulación de la represión. Un ejemplo de este último modo de activación se ha visto en *B. subtilis*, para el activador ComK, una pequeña proteína que se enlaza cerca del represor Rok y CodY en el promotor de su propio gen comK. Aunque el enlace directo de ComK al ADN no resulta en el desplazamiento físico de los represores, sí que anula su acción represiva, activando de esta manera la expresión del gen (Smits et al, 2007).

2.3.2 Arqueas.

Numerosos represores de arqueas ejercen su acción acorde a las reglas de los reguladores bacterianos: se enlazan a sus sitios en el ADN ocultando el acceso a la caja TATA y al elemento BRE o simplemente bloquean el reclutamiento de la RNAP (Ouhammouch, 2004). Similar ocurre con los activadores, que mimetizan en gran medida los mecanismos usados por los activadores bacterianos.

Un ejemplo concreto de represión *in vitro* es llevado a cabo por una proteína codificada por el gen mdr1 de *Archaeoglobus fulgidus.* Esta proteína (Mdr1) es un homólogo de la familia bacteriana de DtxR (represor dependiente de la concentración de metales). mdr1 es el primer gen que se transcribe de manera co-lineal de un grupo formado por cuatro genes. Los otros tres adyacentes a mdr1 codifican a transportadores ABC de iones metálicos. Así, el represor Mdr1, que depende de los iones (Fe^{2+}, Mn^{2+} o Ni^{2+}), se enlaza a sitios múltiples que solapan el TSS del promotor de su propio gen y evita el reclutamiento de la ARN polimerasa. Al añadir aniones al medio de cultivo, que secuestren los iones metálicos, el represor Mdr1 es liberado con la consecuente restauración de la transcripción de su gen y de los transportadores ABC colindantes (Bell et al, 1999).

El factor LysM de la familia Lrp (Figura 5) ha sido identificado como regulador positivo de la transcripción en *S. solfataricus.* Esta proteína es codificada por el gen lysM y pertenece a un grupo que codifica enzimas para la vía metabólica de biosíntesis de la lisina.

La transcripción del segmento de 4 genes que forman este grupo es inducida en un medio carente de lisina. LysM se enlaza inmediatamente adyacente al elemento BRE y la caja TATA de la unidad transcripcional y el aminoácido lisina disminuye la fuerza de la interacción. Por consiguiente, LysM debe actuar como activador, mientras que la lisina es un efector negativo. Este mecanismo no ha sido visto *in vitro* quizás por la ausencia de algunos elementos importantes en el extracto usado como medio de cultivo (Brinkman et al, 2002). Una razón más del peligro que acarrea los estudios utilizando medios de cultivo donde muchos procesos de regulación transcripcional al adaptarse a condiciones controladas se alejan de lo que ocurre en los hábitats naturales. También se debe tener en cuenta que dependiendo de la especie y el contexto, factores de la familia Lrp pueden actuar como represores (Bell and Jackson, 2000) y este comportamiento está extendido a muchos reguladores de células de bacterias y arqueas.

Otros mecanismos de regulación de la expresión génica en procariotas comprenden por ejemplo, interferencia transcripcional, metilación del ADN y superenrollamiento del cromosoma. También existen formas de regulación post-transcripcional como, la degradación del ARNm, los riboswitches y los pequeños ARN (van Hijum et al, 2009). A pesar de la existencia de estos mecanismos adicionales, la mayor parte de la regulación transcripcional celular ocurre en el estado de iniciación: cuando la ARN polimerasa se enlaza y empieza su recorrido por la doble cadena de ADN, permitiendo así la transcripción de la información. Por tanto, es la interacción de los TFs y sus sitios de unión lo que permite una orquestación adecuada y precisa de la transcripción: guiando a la ARN polimerasa a sus sitios diana y permitiendo o evitando su desplazamiento a través de la doble cadena de ADN. Por esta razón, son numerosos los estudios abocados en conocer cuáles son las secuencias que los TFs reconocen y los genes regulados por estos. Los métodos experimentales y computacionales desarrollados para la identificación de sitios de unión, TFs y regulones se detallan en los párrafos siguientes.

3. Identificación de los sitios de unión de factores de transcripción y sus genes diana en procariotas.

La determinación de los genes diana sobre los que actúan los reguladores transcripcionales es un campo que ha evolucionado en los últimos años de manera rápida. Su avance se ha visto impulsado por metodologías emergentes como los micro-arreglos de ADN (Dufva, 2009), la secuenciación de ARNm (metatranscriptómica) (Wilhelm and Landry, 2009) y los estudios de inmuno-precipitación de la cromatina (ChIP) (Pillai and Chellappan, 2009). Los trabajos más recientes se han enfocado, en la asociación de los TFs con sus genes dianas (que en su conjunto forman el regulón). Unido a los métodos experimentales se han desarrollado algoritmos informáticos, pero todavía su mejora es un reto para los biólogos computacionales.

3.1 Métodos experimentales para la identificación de regulones y TFBSs.

La identificación de los regulones se realiza por ejemplo, mediante la comparación de la cepa salvaje con otra que se le ha inactivado el gen que codifica para determinado TF manteniéndose ambas cepas bajo las mismas condiciones (Zhou and Yang, 2006). Estudios más recientes emplean series temporales de análisis transcriptómico (Zomer et al, 2007) también para la identificación de regulones. Mediante la realización de estos experimentos se pueden detectar grupos de genes u operones que respondan a perturbaciones específicas de factores ambientales, a lo que se le ha llamado como estimulones (Rodionov, 2007). Para identificar los estimulones se compara, la respuesta celular mediante micro-arreglos de ADN, de un control con el grupo sometido a una alteración de algún parámetro físico-químico. Si el ARNm es aislado de una muestra que ha estado influenciada por una condición ambiental particular, se obtiene también información del rol de los TFs bajo estas condiciones específicas (Zhou and Yang, 2006).

Los experimentos conocidos como ChIP permiten la identificación de los sitios de unión del TF. A través de esta técnica el ADN es entrecruzado con un proteína reguladora marcada, después es sonicado para producir pequeños fragmentos y posteriormente se inmunoprecipita con un anticuerpo que reconoce al TF o a su marcaje (Pillai and Chellappan, 2009). Los estudios más recientes utilizan una técnica novel conocida como ChIP-Seq, donde se combinan experimentos ChIP con técnicas de secuenciación de nueva

generación (Johnson et al, 2007). Lo más típico al usar este último análisis es obtener fragmentos de 25 a 50 nucleótidos (conocidos como "tags"). Estos fragmentos se secuencian y las regiones que aparecen con mayor densidad corresponden, probablemente, a sitios de unión (Jothi et al, 2008).

Un grupo de sitios que corresponden al mismo TF pueden ser agrupados en un motivo. Es difícil representar estos motivos debido a que sus secuencias no son exactamente iguales. Aunque existen diferentes maneras de representarlos (Figura 6), escoger una u otra depende del nivel de exactitud, simplicidad o conveniencia a la hora de usar algoritmos computacionales (MacIsaac and Fraenkel, 2006). El modo más simple de representación es a través de secuencias consensos (Figure 6, A), aunque la forma más usada es mediante matrices ponderadas de posición (conocidas también como PWM, del inglés Position Weight Matrix) (Figura, 6, C). Las matrices de posición son la forma más precisa que se emplea en la actualidad, para la identificación de motivos de TFBSs (Tompa et al, 2005). En las PWMs, se asume que el nucleótido que se observa en una posición es independiente del resto de los nucleótidos observados a lo largo de la secuencia. Los motivos se visualizan de manera conveniente a partir de los logotipos de secuencias (Figura 6, D; Figura 5), donde la altura de la letra indica cuan probable es encontrar determinado nucleótido en esa posición (Crooks et al, 2004).

A. Sitios de unión para CodY

ctrA	A	A	T	T	G	T	C	T	G	A	C	A	A	T	T
dppA	A	T	T	T	T	T	C	T	G	A	C	A	A	T	T
.															
.															
.															
Consenso	A	A	T	T	T	T	C	**W**	G	A	A	A	A	T	T

B. Matriz de posición y frecuencia (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	12	13	0	2	3	1	0	8	3	17	12	16	18	7	3
C	0	0	0	2	0	1	18	0	0	0	4	0	0	0	0
G	0	1	0	2	3	0	0	2	15	0	0	0	0	0	0
T	6	4	18	12	12	16	0	8	0	1	2	2	0	11	15

C. Matriz de posición ponderada. (Position Weight Matrix, PWM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0,90	1,01	-2,39	-1,09	-0.71	-1,60	-2,39	0,39	-0,71	1,36	0,90	1,28	1,44	0,23	-0,71
C	-2,39	-2,39	-2,39	-0,53	-2,39	-1,18	2,,24	-2,39	-2,39	-2,39	0,26	-2,39	-2,39	-2,39	-2,39
G	-2,39	-1,18	-2,39	-0,53	-0,08	-2,39	-2,39	-0,53	1,99	-2,39	-2,39	-2,39	-2,39	-2,39	-2,39
T	0,01	-0,41	1,44	0,90	0,90	1,28	-2,39	0,39	-2,39	-1,60	-1,09	-1,09	-2,39	0,70	1,19

D. Logotipo de la secuencia

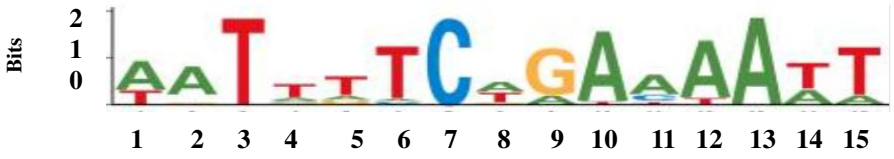


Figura 6. Diferentes representaciones de las secuencias consenso del factor CodY de *L. lactis*. (A) Datos colectados para un grupo de TFBSs para un factor específico (CodY). (B), (C) y (D) se obtienen a partir de los datos representados en (A) (den Hengst et al, 2005).

3.2 Predicción *in silico* de TFBSs.

La predicción *in silico* de sitios de unión que forman un motivo es aún muy complicado. Por ejemplo, la mayoría de algoritmos computacionales se basan en la búsqueda de sitios sobre-representados en promotores de genes que se conoce que están co-regulados. Una vez se tenga la secuencia consenso del motivo; se busca éste por todo el genoma del microorganismo en cuestión, o de otros relacionados, para predecir nuevos posibles TFBSs. Aunque este es el basamento fundamental los algoritmos computacionales se dividen en dos grupos, 1) enumerativo o basado en palabras y 2) probabilísticos (Das and Dai, 2007;GuhaThakurta, 2006;MacIsaac and Fraenkel, 2006).

Los métodos enumerativos catalogan exhaustivamente secuencias de ADN (también llamadas "palabras") y estas se ordenan por la puntuación obtenida de acuerdo al número de veces que aparecen en el grupo de secuencias de referencia (MacIsaac and Fraenkel, 2006). De esta manera las cadenas de letras sobre-representadas emergen como un motivo o motivos de acuerdo a cuan similares sean entre sí. El primer algoritmo basado en este precepto fue desarrollado por van Helden *et al* (van Helden et al, 1998) y más tarde adaptado para el análisis de motivos bipartitos o diadas (van Helden et al, 2000). Estos métodos basados en la búsqueda de diadas palíndromas, debido a la naturaleza dimérica de los TFs, están muy extendidos en la predicción de sitios de unión en procariotas (Bi and Rogan, 2006;Chakravarty et al, 2007;Laing et al, 2008;Li et al, 2002;Li, 2009;Wijaya et al, 2007).

Los métodos probabilísticos, de manera general, empiezan por el desarrollo de modelos (frecuentemente PWMs) a partir de los datos de las secuencias. Después se optimizan estas matrices para encontrar motivos comunes en un grupo múltiple de secuencias de entrada. Dos algoritmos muy conocidos para la optimización son: EM (expectation- maximization) (Cardon and Stormo, 1992) y muestreo de Gibbs (Lawrence et al, 1993). A diferencia de los métodos enumerativos, los probabilísticos, pueden llegar a ser computacionalmente más costosos y no siempre encuentran el mínimo global en su espacio de búsqueda (van Hijum et al, 2009).

Para construir los regulones correctamente, además de la predicción precisa de los TFBSs es necesario conocer también como se organizan los genomas en los procariotas. Toda esta

información es importante para identificar el TF, su promotor diana y a los genes regulados por éste.

4. Organización del genoma bacteriano y de arqueas

Los microorganismos han convertido su característica más obvia, sus pequeños genomas, en un gran avance evolutivo. El poder de adaptación que ellos poseen se ve reflejado en la inmensa variedad de nichos en que habitan, además que conforman la mayoría de la diversidad filogenética encontrada en la Tierra (Bourret et al, 2002). Pero, ¿Qué hace a los procariotas tan diversos y adaptables?

4.1 Ventajas evolutivas de los genomas de organismos procariotas que condicionan su adaptación y plasticidad.

Después de la publicación de miles de genomas bacterianos y de arqueas completos, muy pocos se atreven a subestimar el papel que tiene la genómica en la microbiología molecular contemporánea. La secuenciación del ADN facilita la manipulación génica y promete convertirse en el esquema básico para conocer las características de los microorganismos resistentes a las condiciones de cultivo. Los datos genómicos han permitido conocer el modo peculiar a través del cual ocurre la evolución en los procariotas comparados con los eucariotas. Por ejemplo, en eucariotas las nuevas funciones se adquieren fundamentalmente por duplicación génica. Sus unidades transcripcionales incluyen un solo gen y los procesos celulares están altamente compartimentados en los orgánulos. En los procariotas el repertorio génico se incrementa por transferencia horizontal de genes y no por duplicación. Los cromosomas son relativamente uniformes en términos de densidad de genes y composición de secuencias, los genes se co-transcriben como operones y los procesos celulares están acoplados. Mientras que dos cepas de *E. coli* tienen más genes no relacionados que dos genomas típicos de mamíferos; *E. coli* y *Bacillus subtilis* (que divergieron hace billones de años) tienen genomas más similares que los genomas de las levaduras (que divergieron hace sólo ciento de millones de años). Como consecuencia los genomas de los procariotas en cuanto a la arquitectura de sus cromosomas tienen ambas

características: complejidad y plasticidad. Aunque las bacterias y arqueas sorprenden por su flexibilidad en términos de repertorios de genes; su organización está muy conservada (Rocha, 2008).

4.2 Co-localización génica y los operones.

Desde hace más de medio siglo se conoce que los genes que codifican para enzimas tienden a estar co-localizados en el cromosoma bacteriano. Además que siguen el orden en el que actúan las correspondientes enzimas en la vía metabólica y se transcriben en la misma unidad policistrónica, el operón (Jacob and Monod, 1961). El regulador (o reguladores) del promotor del operón también es frecuente hallarlo en la misma unidad policistrónica (Korbel et al, 2004). La palabra sintenia es la más aceptada para referirse a múltiples regiones génicas donde el ADN tiene secuencia similar y el orden de los genes se conserva entre genomas diferentes. El mapa genético detallado de *E. coli* y *Bacillus subtilis* indica que los genes no tienen necesariamente que ubicarse en la misma posición relativa en todos los genomas de bacteria, pero sí se conoce que ciertos grupos de genes son sinténicos. La sintenia en procariotas ha sido una ventaja para evaluar la evolución genómica y también se aprovecha para la predicción de la función de los genes. Varios son los ejemplos de grupos resistentes a la dispersión, como por ejemplo, el operón de la proteína ribosomal y NADH deshidrogenasa (Bentley and Parkhill, 2004).

4.3 Relación entre el tamaño del genoma y el ambiente.

Los genomas más pequeños en procariotas tienden a pertenecer a organismos, cuyo hábitat está restringido a nichos estables o asociados a un hospedero. Mientras que las bacterias con genomas muy grandes tienden a ocupar nichos complejos como, los suelos (Das et al, 2006;Kaneko et al, 2002).

En el caso de los microorganismos que habitan ambientes complejos, un gran número de genes son adquiridos para poder sobrevivir. Los parásitos obligados o asociados a hospederos no necesitan un gran repertorio de genes, debido a que no están bajo la

influencia de tantas variables ambientales (que incluyen: parámetros físico-químicos, nutricionales o de índole biológico); a diferencia de lo que ocurre con los microorganismos de hábitats expuestos. En los suelos, los tiempos de replicación no parecen ser críticos y el crecimiento lento no es un problema. Un ejemplo es *S. coelicolor*, conocida también como una bacteria "boy-scout" porque parece estar preparada para todo tipo de condiciones ambientales (Hopwood, 2003). Su genoma es rico en genes que permiten la degradación de carbohidratos complejos, encontrados en restos de plantas y hongos. Además gran parte de su proteoma se dedica a la producción de metabolitos secundarios, que funcionan como antibióticos o protegen contra la desecación y las bajas temperaturas. A su genoma también se le suman los genes encargados de la esporulación, cuando las condiciones ambientales son muy drásticas para un crecimiento normal.

La reducción en el tamaño del genoma se puede ver como una adaptación o respuesta a un ambiente más simplificado o más estable. Esta adaptación se debe diferenciar de la pérdida de genes que ocurre en organismos de vida libre cuando el nicho experimenta cambios físico-químicos. Lawrence y Ochman (Lawrence and Ochman, 1998) fueron los primeros en postular que en las bacterias la adquisición de genes está balanceada a largo plazo con la pérdida de los mismos, de tal manera que el tamaño del genoma permanece estable.

El consecuente balance entre pérdida y ganancia de genes, se observa de manera clara, entre genomas de microorganismos que han cambiado recientemente de nicho, tales como, *Salmonella enterica serovar Typhi* (Parkhill et al, 2001a), *Shigella flexneri* (Wei et al, 2003) y *Yersinia pestis* (Parkhill et al, 2001b). Estos tres patógenos humanos contienen un gran número de pseudogenes, que alcanza hasta el 5% de sus capacidades codificantes. Se cree que estos pseudogenes se han quedado en sus genomas como consecuencia de una rápida evolución al cambiar de nicho. Muchos genes que han sido inactivados eran importantes en la adaptación al antiguo nicho pero en el nuevo no se les requiere más (o son desventajosos). Entre los ejemplos específicos se encuentran: el aparato flagelar y varias adhesinas de *Y. pestis* (Parkhill et al, 2001b). Todas estas proteínas son requeridas por el microorganismo cuando actúa como patógeno intestinal, pero no las necesita más al convertirse en un patógeno sistémico.

En resumen, importantes evidencias se han acumulado en los últimos años sobre la adaptación y evolución de los microorganismos. El gran avance en la adquisición de nuevos

conocimientos se debe (en gran parte) al estudio de las regiones codificantes. En la actualidad se ha ido más allá del simple contaje de genes y han surgido técnicas como la metatranscriptómica que permite la detección y el análisis de las proteínas que se expresan en cada momento. A pesar de los avances logrados, existe una estructura genómica con un papel clave en la adaptación y que ha sido muy poco explorada: el promotor.

OBJETIVOS

Esta tesis parte de la hipótesis:

"Así como el ambiente condiciona la ganancia y pérdida de genes y afecta las regiones codificadoras en procariotas, también actuará sobre la estructura del promotor"

Por lo tanto, el objetivo general, derivado de la hipótesis anterior es: **Evaluar la correlación entre el ambiente y la estructura de las regiones reguladoras en procariotas.** Para esta tesis nos planteamos también las siguientes preguntas: ¿Qué bases sustentan el hecho que la arquitectura de un promotor en bacterias o arqueas pueda estar afectada por el ambiente? ¿Qué estructuras en particular se verían afectadas? ¿Cuál sería el mejor modelo o modelos para probar nuestra hipótesis?

Con este trabajo buscamos conocer si la estructura de los promotores en procariotas también responderá a factores ambientales, haciéndose más compleja a medida que las características físico-químicas de los nichos sean más variables. Como medida de complejidad usamos una estimación de los sitios de unión de factores de transcripción por promotor. Para poder conocer que sucede realmente en los nichos naturales donde habitan los microorganismos, usamos como modelos genomas, o sus fragmentos, provenientes de secuenciación metagenómica; aunque también estudiamos diferentes cepas aisladas de *E. coli* y los clados SAR11 y LD12 para evaluar nuestro método. Hasta el momento no se han reportado estudios similares al aquí descrito en organismos procariotas.

Los **objetivos específicos** que seguimos para probar nuestra hipótesis fueron:

- Desarrollo de metodología para la identificación de promotores y su caracterización a nivel de sitios de unión a TF y poder inferir potenciales de regulación.

- Analizar el espacio funcional de comunidades bacterianas en relación a los niveles de complejidad del potencial regulador que presentan.

-Encontrar puntos de relación entre funciones altamente reguladas en comunidades microbianas y variaciones de parámetros ambientales.

-Identificar nuevos sitios de unión a factores de transcripción que pudieran estar implicados en la adaptación al medio.

-Analizar el comportamiento del potencial regulador y la presencia de determinados sitios de unión según la bio-distribucion de nueve cepas de *E. coli.*

-Estudiar como se estructuran los promotores, atendiendo a la dirección de los mismos y también con respecto a la ubicación y densidad de TFBSs, en bacterias con genomas compactos (ej. Clados de SAR11 y LD12).

II. METODOLOGIA

1. Datos.

Los genomas individuales así como las secuencias de metagenomas que hemos analizado se han tomado de diversas bases de datos.

1.1 Metagenomas.

Las secuencias metagenómicas analizadas en este trabajo fueron tomadas de la base de datos Camera (Sun et al, 2011). La secuenciación de los microorganismos de los tres nichos estudiados se realizó mediante el método de Sanger y 454 pirosecuenciación. La longitud de los fragmentos de ADN para estas comunidades varía entre 100-1000 pares de bases, pero más del 90% de los fragmentos de las tres comunidades tienen una longitud mayor de 800 pares de bases (ver Tabla 1).

Tabla 1. Longitud de los fragmentos de ADN secuenciados a partir de comunidades microbianas colectadas de 1) sedimentos marinos formados por restos de ballenas , 2) una granja en Minnesota y 3) una mina acidificada en Richmond Mountain.

Ambiente	Total de fragmentos	Fragmentos > 800nt
Sedimentos marinos	117326	106961
Granja en Minnesota	138347	131103
Mina acidificada	319166	282250

1.1.1 Comunidades microbianas encontradas en sedimentos marinos formados por restos de ballenas.

Las muestras de los sedimentos marinos se tomaron de tres bibliotecas independientes nombradas como:

- CAM_SMPL_WHALEFALLBONE: Estos restos de ballena fueron encontrados en la península antártica y se secuenciaron las comunidades de microrganismos de los huesos de la carcasa. La recolección se realizó a una profundidad de 560 metros, y las coordenadas son (S65.10 W64.47).

- CAM_SMPL_WHALEFALLMAT: También tomados de los huesos de la carcasa pero de una ballena gris hundida en Santa Cruz Basin (Océano Pacífico), coordenadas (N33.30 W119.22) y una profundidad de 1674 metros.

- CAM_SMPL_WHALEFALLRIB: corresponde a la misma carcasa de donde se colectó CAM_SMPL_WHALEFALLMAT, pero esta vez se secuenció la comunidad de microorganismos encontrada en el hueso de la costilla.

Todas estas comunidades se caracterizan por el crecimiento en una fuente rica en lípidos (proporcionada por los restos de ballena).

1.1.2 Comunidades microbianas encontradas en una mina acidificada en Richmond Mountain, California, USA.

Las muestras analizadas corresponden a dos locaciones diferentes dentro de la misma mina y se nombran:

-5-Way (CG): Es un bio-filme de microbios de baja complejidad. Este bio-filme creció a cientos de metros bajo tierra y estaba rodeado de mineral pirita (FeS_2).

- UBA: es una muestra sub-aérea colectada en la base de una pila de sedimentos de mineral pirita (de aproximadamente 2 metros).

Ambas comunidades microbianas crecieron a pH muy bajos y la composición de especies del nicho fue limitada por las condiciones drásticas tanto de acidez como de temperatura y disponibilidad de nutrientes características de este hábitat.

1.1.3 Comunidad microbiana secuenciada a partir de suelo de granja en Waseca, Minnesota, USA.

Esta comunidad se colectó de suelo superficial (0-10 cm) de una granja en Waseca en Septiembre de 2001. El área alrededor había sido usada para la ganadería, incluyendo ovejas, cabras y cerdos. También el área muestreada se encuentra en la vía de drenaje de un búnker de almacenamiento y había sido utilizado para operaciones de ensilado de residuos de maíz dulce y guisante desde 1990-1997. Las muestras se colectaron y se sellaron posteriormente en bolsas de polietileno y se almacenaron a 4ºC para ser procesadas más adelante. Un análisis bioquímico de 20g de este mismo suelo, llevado a cabo en los laboratorios Wallace, reveló que se trataba de suelos arcillosos con muy poco material orgánico y alto niveles de muchos minerales esenciales (Tringe et al, 2005).

1.2 Genomas de E. coli, LD12 y SAR11.

1.2.1 Genomas de *E. coli* secuenciados a partir de cepas colectadas de diferentes sistemas de órganos de mamíferos.

Los datos de los genomas de *E. coli* analizados en este trabajo fueron tomados de la base de datos IMG/JGI (http://img.jgi.doe.gov/). A continuación se muestran los identificadores correspondientes a cada una de las cepas analizadas (según la anotación taxonómica usada por IMG) así como su fuente de aislamiento.

- Intestino humano y animal: 2513237219, 2518645559, 2513237251, 2506520037
- Sistema urinario: 2512047041, 2511231170, 2511231198
- Fluido cerebroespinal: 2511231131

También hemos utilizado como control (“outgroup”, en un análisis de agrupamiento) una cepa aislada por vez primera en 1977, que fue posteriormente modificada para la producción de etanol en 1990. El identificador de esta cepa es: 2513237200.

1.2.2 Genomas de LD12 y *Pelagibacter Ubique* (SAR11).

Los genomas de SAR11 y LD12 fueron tomados también de IMG/JGI (http://img.jgi.doe.gov/) y son los siguientes:

- Pangenoma combinado de nueve especies de LD12 reconstruido a partir de genomas individuales amplificados (Para la secuenciación de estos genomas se usó una técnica, conocida en inglés como Single Cell Genomics y se basa en el aislamiento de una célula y posterior amplificación y secuenciación de su ADN, sin necesidad de usar ninguna técnica de cultivo previa (Kvist et al, 2007). Los identificadores taxonómicos de estos genomas según IMG se detallan en la Tabla 2.

Tabla 2. Descripción taxonómica de los genomas del clado LD12

Taxón oid	Dominio	Nombre taxonómico
2236347068	Bacteria	alpha proteobacterium SCGC AAA487-M09
2236347069	Bacteria	alpha proteobacterium SCGC AAA028-D10
2236661000	Bacteria	alpha proteobacterium SCGC AAA023-L09
2236661008	Bacteria	alpha proteobacterium SCGC AAA028-C07
2236876027	Bacteria	alpha proteobacterium SCGC AAA024-N17
2236876029	Bacteria	alpha proteobacterium SCGC AAA280-P20
2236876030	Bacteria	alpha proteobacterium SCGC AAA027-J10
2236876031	Bacteria	alpha proteobacterium SCGC AAA027-L15
2236876032	Bacteria	alpha proteobacterium SCGC AAA280-B11

- Pangenoma combinado de tres especies de *Pelagibacter Ubique* (SAR11) (ver Tabla 3).

Tabla 3. Descripción taxonómica de los genomas del clado SAR11.

Taxón oid	Dominio	Nombre taxonómico
637000058	Bacteria	Candidatus Pelagibacter ubique SAR11 HTCC1062
638341056	Bacteria	Candidatus Pelagibacter ubique SAR11 HTCC1002
647533122	Bacteria	Candidatus Pelagibacter sp. HTCC7211

2. Identificación de regiones intergénicas y promotores.

La predicción de regiones intergénicas se hizo siguiendo diferentes protocolos para los metagenomas y los genomas individuales, debido a los diferentes métodos de secuenciación usados. Para hacer más clara la explicación del protocolo de identificación se detallarán a continuación los pasos seguidos en dos secciones, una para metagenomas y otra para genomas individuales.

2.1 Protocolo para la identificación de regiones intergénicas y promotores en metagenomas.

La metodología para la identificación y extracción de promotores a partir de datos metagenómicos fue desarrollada por nuestro grupo. El primer paso del procedimiento se basa en la extracción de fragmentos de ADN anteriores a genes, estos últimos detectados a través de búsquedas de homología usando como entrada las secuencias de ADN, tal como se reportan en la base de datos Camera (Sun et al, 2011)

Para este protocolo seleccionamos varios filtros conservativos para asegurar que el método de selección de promotores sea robusto y confiable.

(1) Filtrar los fragmentos más cortos, menos de 800 pares de bases. Este filtro mantiene hasta el 90% de todos los fragmentos (ver Tabla 1). Este paso se puede modificar si se usan secuencias de Illumina, en cuyo caso habría que añadir una etapa previa de ensamblaje y filtros de calidad, eliminación de quimeras que en este estudio no fue necesario aplicar.

(2) Detección de secuencias que constituyan potenciales genes codificadores de proteínas. Esto se realizó a través de una comparación de todas las secuencias de los metagenomas que pasaron el filtro (1) contra todas las proteínas reportadas para bacterias y arqueas en el NCBI (versión de 2009-02 de la base de datos, NCBI-FTP: http://www.ncbi.nlm.nih.gov/Ftp/). Para la búsqueda de homología usamos BLASTx (e-value=0.001, los otros parámetros por defecto) (Altschul et al, 1990). Posteriormente, seleccionamos aquellas secuencias de ADN que coincidían con una proteína conocida con un 50% de identidad y una longitud de 150 aminoácidos (Figura 1).

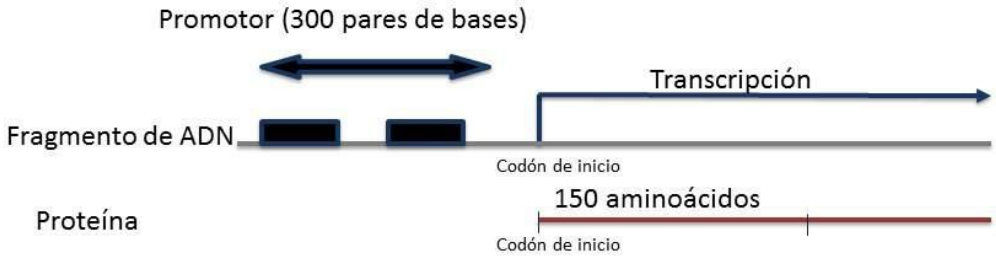


Figura 1. Criterio de selección para los posibles promotores.

(3) Eliminar las secuencias, que aunque coincidían con el criterio (2), la región anterior al fragmento que codifica para la proteína tuviera una longitud inferior a 300 pares de bases (Figura 1). Este filtro enriquece la muestra en secuencias con un posible potencial regulador alto, evitando así la posibilidad de incluir regiones inter-operónicas. Al escoger 300 nucleótidos como longitud de los posibles promotores esperamos no afectar la selección de regiones reguladoras de bacterias o arqueas de genomas grandes (con posibles regiones reguladoras de gran longitud) en detrimento de las especies de pequeños genomas. Por ejemplo, las especies de vida libre con los genomas más pequeños conocidos (pertenecientes a *Pelagibacter*) poseen regiones intergénicas de longitud mayor o igual a 300 nucleótidos (Giovannoni et al, 2005), lo que asegura que al utilizar este filtro ninguna especie de vida libre quede fuera del análisis. En Resultados y Discusión se ampliara mucho más sobre este tema.

(4) Para evitar posibles desviaciones que favorezcan a especies abundantes en cada entorno y para poder realizar análisis cualitativos y cuantitativos de comparación de las comunidades atendiendo al potencial regulador; aplicamos un segundo filtro que elimina las secuencias de promotores repetidas. Los posibles promotores extraídos en el paso (3) se volvieron a comparar entre ellos (mediante BLASTn, parámetros por defecto) y se consideraron duplicados aquellos que tenían un 98% de identidad (sólo de la región de 300 pares de bases correspondiente al promotor, aquí no se

tuvo en cuenta la región codificadora para la búsqueda de homología). De los duplicados encontrados mantuvimos los fragmentos cuyas regiones codificadoras adyacentes eran más informativas (entiéndase por más informativa, la secuencia más larga con mayor posibilidad de ser identificada en la asignación de taxonomía y función).

(5) También eliminamos los fragmentos correspondientes a ADN eucariótico usando el programa MEGAN (Huson et al, 2007). (Una explicación más detallada de la asignación taxonómica se puede consultar en la sección 3.1).

(6) Para descartar la presencia de CRISPRs, ARNt, ARNr, ARN no codificadores de larga longitud y otros tipos de pequeños ARN usamos la base de datos Rfam 11.0. La base de datos Rfam categoriza ARN no codificantes y su secuencia primaria conservada, además de la estructura secundaria a través de alineamientos múltiples de secuencia, anotación de estructura secundaria consenso y modelos de covarianza. Cada familia consiste en un grupo de secuencias de ARN que se supone que comparten un ancestro común. Esta base de datos también provee alineamientos significativos para la familia que ha sido anotada con una secuencia consenso responsable de la estructura secundaria que adopta el ARN. Los modelos de covarianza se construyen para describir la familia y para obtener un alineamiento completo que represente todas las coincidencias del modelo de covarianza y las secuencias en la base de datos subyacente (Burge et al, 2013). Por tanto, usando esta base de datos, todas las secuencias que coincidían con un dominio ARN de bacteria o arquea con un e-value <0.05 de acuerdo al programa cmscan (Versión 1.1rc1, June 2012) fueron eliminadas de nuestro grupo de posibles promotores.

(7) Por último para asegurarnos que el grupo de promotores estaba limpio de secuencias codificadoras, usamos un algoritmo alternativo a las búsquedas por homología. Este algoritmo denominado Prodigal (Hyatt et al, 2012) no usa Hidden Markov Model, ni Interpolated Markov Model, sino una serie de funciones log-likelihood simples. Estas funciones permiten que Prodigal tenga una exactitud aceptable a la hora de identificar los genes. Por ejemplo tiene una efectividad del 96% de localización de los TSS cuando se compara con los datos verificados experimentalmente de

Ecogene. Incluso puede emplearse en la identificación de genes de genomas desconocidos, por lo que es apropiado para metagenomas.

2.2 Protocolo para la identificación de regiones intergénicas y promotores en los genomas de *E. coli* , LD12 y SAR11.

Las secuencias de los genomas individuales se obtuvieron a través de Illumina o 454 por lo que primero se ensamblaron los "contigs" o "scaffolds" fundamentalmente usando el programa Velvet (Zerbino and Birney, 2008) (por ejemplo, este programa fue empleado para el ensamblaje de los genomas de LD12 y algunos de *E.Coli* (más información en la página web de IMG/JGI, http://img.jgi.doe.gov/ de donde descargamos los datos ya ensamblados). Para la identificación de las regiones intergénicas se usó el procedimiento estándar de DOE-JGI (Mavromatis et al, 2009)**.**

El procedimiento desarrollado por DOE-JGI permite la identificación de genes, pequeños ARN, ARN no codificantes, elementos CRISPR o repetitivos y también las regiones intergénicas en los genomas pre-ensamblados. Los métodos se basaron de manera general en una combinación de modelos Hidden Markov y de similitud de secuencia. Los programas específicos usados en cada paso se listan a continuación:

Paso #1: Identificación de ARN no codificante (ARNt, ARNr y otros ARN no codificantes):

- ARNt: para su predicción se usó el programa tRNAScan-SE (Lowe and Eddy, 1997).
- ARN ribosomal (5S, 16S, 23S): para su predicción se usó el programa RNAmmer (Lagesen et al, 2007).

Paso #2: Identificación de elementos CRISPR.

- Los programas usados fueron: CRT (Bland et al, 2007) y PILERCR (Edgar, 2007).

Paso #3: Identificación de genes que codifican para proteínas.

- Se llevó a cabo usando GeneMark (v.2.6r) (Besemer et al, 2001) o Metagene (v. Aug08) (Noguchi et al, 2006). Ambos son programas de predicción de genes *ab initio.*

Las secuencias identificadas en cada paso (correspondientes a ARN. CRISPR, genes que codifican proteínas y las regiones intergénicas fueron descargadas directamente del sitio de IMG-ER (Markowitz et al, 2008)).

Las posibles regiones reguladoras o promotores se extrajeron del grupo de regiones intergénicas identificadas en IMG/JGI. Con este objetivo desarrollamos un programa escrito en Perl para la extracción de regiones intergénicas con las características que se describen en la Figura 2, que al estar bordeadas por una o dos ORF clasifican como posibles promotores. En el primer ejemplo de la Figura 2 la región intergénica corresponde con el promotor que regula la expresión del gen ubicado a su derecha, mientras que el promotor B (segundo caso) regula la expresión del gen de la izquierda. El tercer caso corresponde a una región intergénica que posiblemente incluya dos promotores; en el caso de regiones intergénicas de este tipo y cuya longitud sobrepasase los 300 nucleótidos, se escogieron dos posibles regiones reguladoras. El número máximo de nucleótidos que consideramos por promotor fue 300 nucleótidos.

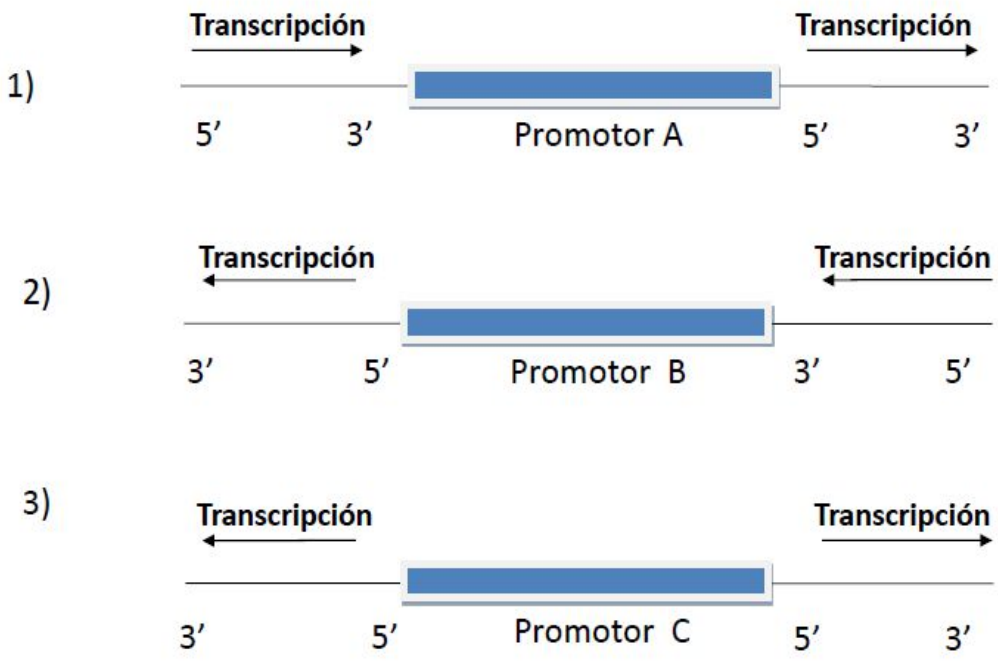


Figura 2. Posibles regiones reguladoras (color azul).

3. Metodologías para la asignación de taxonomía y función.

Una vez fue identificado el promotor, el segundo paso consistió en asignar una función a la región adyacente (o gen), regulado por ese promotor en cuestión. La asignación de función se realizó a través de metodologías diferentes para los genomas individuales y metagenomas. En el caso de los metagenomas también fue necesaria la asignación taxonómica.

3.1 Asignación funcional y taxonómica en metagenomas.

La asignación funcional y taxonómica de las regiones codificantes en los metagenomas se hizo mediante el programa MEGAN V4.70.4 (Huson et al, 2011). Para preparar los ficheros de entrada al programa primero comparamos las secuencias identificadas previamente (con

el procedimiento descrito en la sección 2.1) contra la base de datos NCBI-NR (ftp://ftp.ncbi.nlm.nih.gov/blast/db/). Para realizar esta comparación usamos BLASTx (evalue <0.001). La salida del programa BLASTx la importamos a MEGAN. El programa automáticamente calcula la clasificación taxonómica y funcional mediante el uso de SEED (http://www.theseed.org/wiki/Main_Page) (Figura 3).

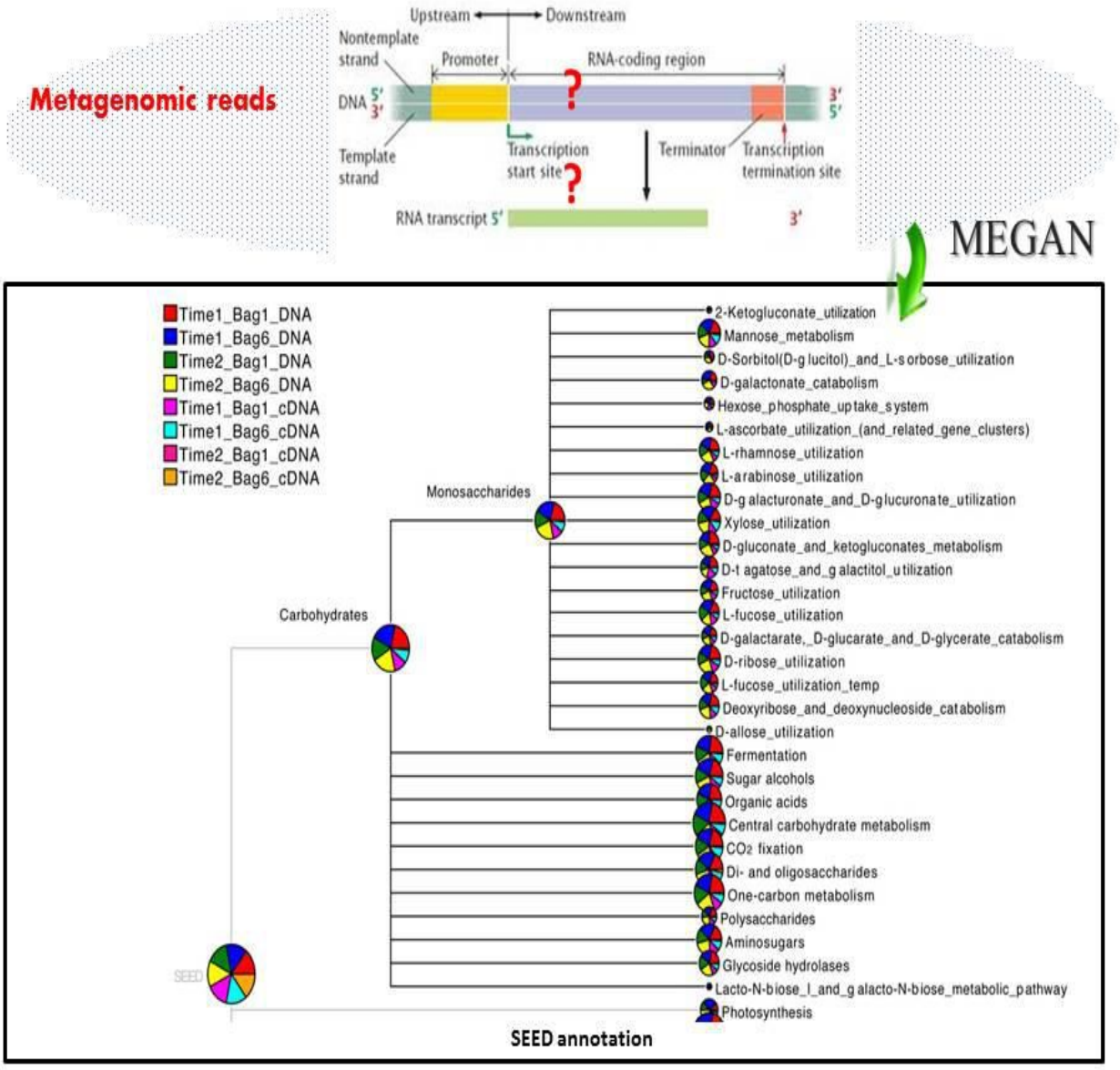


Figura 3. Salida del programa MEGAN para la asignación funcional. Imagen modificada de http://ab.inf.uni-tuebingen.de/software/megan5/.

Para la asignación de taxonomía, primero cargamos en el programa MEGAN la información al respecto completa de NCBI (descargado en Julio de 2012). Durante el análisis el programa colecta todas las coincidencias encontradas en la búsqueda hecha previamente mediante BLASTx (usando el corte para el bit-score que establece MEGAN por defecto para fragmentos largos). A través del algoritmo simple LCA cada secuencia se ubica en su posible taxón. Luego, aplicamos un tercer parámetro denominado "min-support threshold" que permite reasignar los taxones a un parental de la siguiente manera: si el número de fragmentos de ADN asignados a un taxón es más bajo que el umbral ("min-support threshold") entonces todos los fragmentos de ADN asignados a ese taxón se reasignan a un parental. De esta manera los nodos con insuficiente soporte no aparecen en la salida.

El primer paso del análisis funcional fue cargar en el programa los datos de la clasificación SEED, en el año 2012 cuando realizamos el análisis había aproximadamente 1.3 millones de entradas. Para cada fragmento de ADN usamos un corte del BLASTx bit-score igual a 35 bits. Mediante esta clasificación funcional a cada fragmento se le asigna un rol en el subsistema SEED. El subsistema SEED (Figura 3) es una estructura en forma de árbol donde las funciones van de parental a hijos y por consiguiente un mismo fragmento de ADN puede aparecer en varios nodos de un mismo subsistema, hasta la clasificación más interna que es el nombre del gen. Un mismo gen puede tener diferentes roles parentales por lo que puede ser asignado también a un subsistema diferente (Huson et al, 2011).

3.2 Asignación de función en los genomas de *E. coli*, LD12 y SAR11.

Para los genomas individuales usamos la anotación taxonómica de IMG/JGI, (http://img.jgi.doe.gov/), obtenida a través de diferentes bases de datos como: COGs y KEGG. El procedimiento seguido por DOE-JGI (Mavromatis et al, 2009) para obtener estas anotaciones fue el siguiente.

- Búsqueda en la base de datos COGs: Las secuencias de genes que codifican proteínas identificadas en los genomas individuales (acorde al método descrito en la sección 2.2) se compararon con matrices de posición específica de secuencia (siglas

en inglés, PSSMs de Position Specific Sequencing Matrix). Las PSSMs que contenían la información de las COG se obtuvieron de la base de datos CDD (Marchler-Bauer et al, 2007) y se usó para su selección el programa RPS-BLAST con un e-value de corte igual a 0.02.

- Búsqueda en la base de datos KEGG: Para esta búsqueda se usó BLASTp con un e-value de corte igual a 1e-5. Se asigna un rango de ortólogos de 5 o más alto, usando las opción del BLAST (-F 'm S') y una longitud de alineamiento entre la secuencia de entrada y las de la base de datos mayor que el 70%.

Las anotaciones obtenidas al aplicar esta metodología pueden descargarse directamente de IMG-ER.

4 Predicción de sitios de unión de factores de transcripción.

Una vez extraídos los promotores y asignada una función a la región codificante regulada por éstos, el siguiente paso fue la predicción de los sitios de unión de factores de transcripción (TFBSs). Para la predicción de los TFBSs seguimos un método probabilístico basado en la búsqueda de los sitios a través de PSSM de organismos conocidos, y cuando no fue posible usar este método empleamos un algoritmo para la identificación de sitios de unión *de novo*. Una explicación de ambas metodologías y cuando fue adecuada su aplicación se detalla a continuación.

4.1 Estimación de la distribución de sitios de unión a través del uso de matrices de posición específica de secuencia.

El uso de un método probabilístico basado en el escáner de matrices de posición fue aplicado a los genomas de *E. coli* (sección 1.2.1). El uso de este método en metagenomas no fue posible por la poca información que existe en la literatura sobre los sitios de unión de las especies presentes en los metagenomas que fueron analizados en este trabajo.

Las matrices de posición para *E. coli* fueron descargadas del sitio RegulonDB (http://regulondb.ccg.unam.mx). Como algoritmo de búsqueda de sitios de unión en nuestro grupo de promotores usamos MATSCAN (Blanco et al, 2007). Al aplicar este programa se obtiene como resultado la probabilidad de ocurrencia de un sitio de unión dado en una posición específica en el promotor. La posición va variando en +1 nucleótido, comenzando desde el inicio del promotor hasta el último nucleótido de su secuencia. En este estudio se escanearon las secuencias completas sin aplicar ningún umbral de corte y después se calculó el promedio de ocurrencia de cada sitio de unión por promotor como se ilustra en la Ecuación 1, donde: *TFBSscore* es la probabilidad de ocurrencia de un sitio de unión dado en cada posición específica, *n* es la longitud del promotor (en pares de bases) y *x* el número de promotores.

$$Ocurrency(1..x) = \frac{\sum_{i=0}^{n} TFBSscore}{n} \qquad \textbf{Ecuación 1}$$

Usando como entrada las matrices de ocurrencias de sitios de unión por promotor (Figura 4) para cada genoma de *E Coli* estudiado, se procedió a realizar un análisis jerárquico de agrupamiento. Para este análisis se utilizó la función de R hclust (Willett, 1987) y el método “ward”.

	TF1	TF2	TF3…….	TFn
Promotor 1	***TFBSscore(i=1)***	***TFBSscore(i=2)***	***TFBSscore(i=3)…***	***TFBSscore(i=n)***
Promotor 2	***TFBSscore(i=1)***	***TFBSscore(i=2)***	***TFBSscore(i=3)…***	***TFBSscore(i=n)***
Promotor 3	***TFBSscore(i=1)***	***TFBSscore(i=2)***	***TFBSscore(i=3)…***	***TFBSscore(i=n)***
Promotor 4	***TFBSscore(i=1)***	***TFBSscore(i=2)***	***TFBSscore(i=3)…***	***TFBSscore(i=n)***
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Promotor x	***TFBSscore(i=1)***	***TFBSscore(i=2)***	***TFBSscore(i=3)…***	***TFBSscore(i=n)***

Figura 4. Matriz de TFBSscore por promotor. Para cada genoma de E.Coli analizado se obtuvo una matriz con estas características.

4.2 Estimación *de novo* de sitios de unión de factores de transcripción por promotor.

El método *de novo* aplicado en este estudio fue desarrollado para la búsqueda de sitios de unión en procariotas y validado por primera vez en *E. coli* (Li et al, 2002) y años después se usó en la predicción de operones en *Streptomyces Coelicolor* (Laing et al, 2008). Este método se basa en buscar sitios palíndromos o diadas sobrerrepresentados en promotores. Este tipo de estructura es típica de los sitios de unión en procariotas donde usualmente los factores de transcripción tienen dos regiones de enlace a ADN, tanto por la dimerización del TF o por la presencia de dos dominios de unión a ADN en la misma proteína en el caso de los factores sigma (Laing et al, 2008). Otras características predominantes en los sitios de unión en procariotas, que hemos tenido en cuenta para aplicar este método, es la longitud de los mismos. Estos sitios son de longitud variable alrededor de 30 pares de bases. Sin embargo, frecuentemente la secuencia que da la señal de unión al TF está formada por dos subregiones conservadas cada una de ellas de 6 pares de bases como longitud máxima. Es, precisamente, en esta región de 6 pares de bases donde ocurre el contacto con el TF. En numerosos estudios previos (Iqbal et al, 2012;Laing et al, 2008;Li et

al, 2002;Li, 2009) se ha usado esta característica de los TFBSs de bacteria o arqueas para la búsqueda de patrones de la forma W_1NxW_2 (llamados dímeros), donde W_1 y W_2 son oligonucleótidos cortos separados por x bases arbitrarias (ver Figura 5). En particular, en este trabajo, hemos adaptado el algoritmo a la búsqueda de estructuras de este tipo que estén sobrerrepresentadas en los metagenomas. Donde W_1 y W_2 tienen una longitud de 3-5 nucleótidos y W_2 es la repetición complementaria-inversa o palíndroma de W_1 y Nx representa el número de nucleótidos entre W_1 y W_2, para este estudio consideramos una longitud de 0-30 pares de bases para Nx.

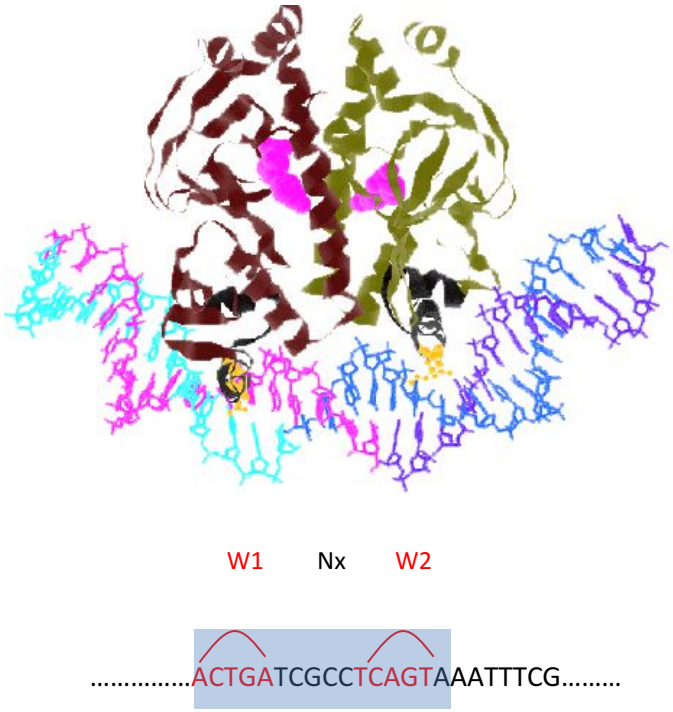


Figura 5. Estructura predominante en los sitios de unión de procariotas.

El primer paso del algoritmo consistió en identificar todos los dímeros, que cumplían con las características mencionadas anteriormente, y estaban presentes en nuestro grupo de promotores. Luego contamos las veces que aparecía cada dímero (o motivo) hallado. Para conocer si el número de veces era significativa aplicamos una estadística Poisson. Para asignar un valor probabilístico a cada uno de estos motivos aplicamos la Ecuación 2, donde

se calcula la probabilidad de observar *n(D)* copias del dímero D por azar, mediante el agrupamiento de todos los promotores analizados en cada ambiente y el cálculo de la frecuencia de aparición de cada dímero en nuestro grupo de datos como sigue:

$$y(D) = Leff(D)\frac{n(W1)}{Leff(W1)}\frac{n(W2)}{Leff(W2)} \qquad \text{Ecuación 2}$$

donde, $n(W_1)$ y $n(W_2)$ representan el número total de ocurrencias de W_1 y W_2 en todo el grupo de datos (en los tres ambientes analizados en este trabajo). Además, Leff(D) = $\sum$r(L(r) − L(D) + 1), representa el número de posiciones independientes donde un motivo D de longitud L(D) puede ser encontrado. La sumatoria se realiza considerando todas las ocurrencias que aparecen en los 9575 promotores analizados que corresponden a los tres ambientes estudiados. Cada uno con una longitud de 300 pares de bases que representa a L(r). Por último, se le asigna un p-valor a cada uno de estos motivos, asumiendo que siguen una distribución de Poisson, este p-valor se calcula:

$$P = \sum_{n \geq n(D)} \frac{y^{n}(D)}{n!} e^{-y(D)} \qquad \text{Ecuación 3}$$

Según la ecuación 3 un dímero es considerado como significativo si P < $1/N_{motif}$, donde N_{motif} representa el número total de dímeros encontrados (Laing et al, 2008;Li et al, 2002). La implementación de este algoritmo fue hecha en nuestro grupo, y los scripts fueron escritos en Perl.

Una vez identificamos los posibles sitios de unión, contamos cuantas veces aparecían los TFBS que encontramos significativos en los promotores, a lo que llamamos potencial regulador. Luego usamos este valor como estimador de la complejidad de la regulación génica en los metagenomas. Además, mediante la asignación de función a la región regulada por el promotor pudimos conocer cuántos reguladores se requieren para la modificación de los patrones de expresión por gen.

5. Caracterización de los sitios *de novo* encontrados en los metagenomas.

Los sitios identificados a través del algoritmo descrito en la sección 4.2 fueron comparados con los TFBSs de la bases de datos RegPrecise (Novichkov et al, 2010). En esta base de datos se almacenan un gran número de sitios de unión de diferentes grupos taxonómicos de bacteria y arqueas. La inclusión de varios linajes, así como, la calidad de los sitios que se reportan; hacen el uso de RegPrecise adecuado para comprobar nuestras predicciones *de novo* y para caracterizar los TFBSs según el posible regulador al que podrían pertenecer.

Por tanto, esta comparación se basó en la búsqueda de las estructuras diméricas (encontradas sobrerrepresentadas de manera significativa, sección 4.2) en sitios de unión identificados experimentalmente o predichos por otros métodos computacionales que se encuentran en la base de datos RegPrecise. Además a través de esta comparación encontramos a que posibles factores de transcripción pertenecían los sitios identificados por el método *de novo*. Luego, hicimos un análisis de la densidad relativa de TFBSs por factor de transcripción por ambiente.

La densidad relativa D(x) de sitios de un factor de transcripción en particular, se calculó a través de la ecuación 4.

$$D(x) = \frac{\sum_{i=0}^{N} TFBS}{N * Tbp}$$ Ecuación 4

Donde, *TFBS* es igual al número sitios que tiene un factor de transcripción en un promotor, *N* representa el número de promotores encontrados en el metagenoma *X* y *Tbp* es igual al número de pares de bases por promotor (en este caso 300). La comparación entre los tres ambientes de la densidad relativa de sitios para determinados TF, se realizó a través de la prueba de Kruskal Wallis utilizando el programa R.

6. Validación y determinación del poder computacional del método para la predicción *de novo* de sitios de unión en promotores extraídos de secuencias de metagenomas.

El método escogido para la predicción de sitios de unión *de novo* descrito en la sección 4.2 ha sido aplicado y validado con anterioridad en genomas de un solo organismo usando como referencia datos de TFBSs identificados experimentalmente (Laing et al, 2008;Li et al, 2002); aunque en este trabajo es la primera vez que se extiende su uso a comunidades de microorganismos. Por esta razón fue necesario validarlo de forma exhaustiva para conocer su confiabilidad en la identificación de nuevos sitios con poca o ninguna información previa. La validación de nuestras predicciones se llevó a cabo a través de dos estrategias: (1) comparación de los resultados obtenidos a partir del método *de novo* y al aplicar una metodología independiente para la predicción de TFBS usando el genoma de *E. coli* como referencia; y (2) mediante una prueba de permutación usando los datos de promotores extraídos como se explica en la sección 2.1.

Para la primera estrategia, primero descargamos las regiones intergénicas de *Escherichia coli K12-W3110* de la base de datos IMG (https://img.jgi.doe.gov). Luego se extrajo los promotores según el método descrito en la sección 2.1. Sobre el grupo de promotores extraídos se realizó la identificación de los sitios de unión a través del programa MotifClick (Zhang et al, 2011) usando una longitud de motivo (k) igual a 14 nucleótidos.

MotifClick es un algoritmo polinomial basado en grafos, que permite la identificación de sitios de unión de factores de transcripción. Para encontrar estos sitios con una longitud *k,* a través de este algoritmo se construye un grafo de 2(*k*-1)-mers (ver Figura 6). Con los datos de la secuencia de entrada como los vértices, se conectan estos mediante una arista si el número máximo de coincidencias de los alineamientos locales (sin espacios en blancos) entre los dos 2(k-1)-mers, supera un valor de corte dado. Por último el motivo es identificado cuando un grupo similar de k-mers emerge por tener un número máximo de clics asociados a la unión de sus vértices.

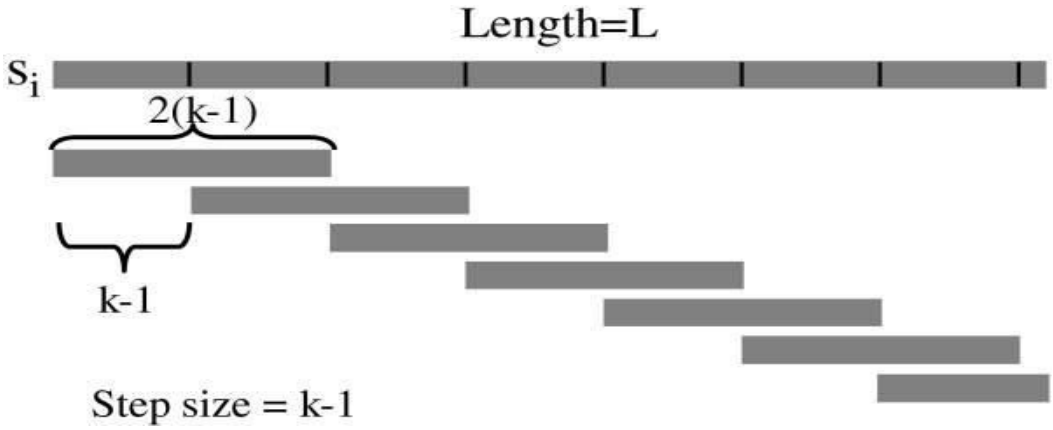


Figura 6. Selección de los kmers para la posterior construcción de los vértices en el algoritmo MotifClick. Figura tomada del artículo (Zhang et al, 2011).

Además de la identificación de sitios de unión mediante MotifClick usamos la metodología descrita en la sección 4.2 sobre el mismo grupo de promotores. Luego, hicimos un análisis de correlación (programa R, función para correlación de Pearson) entre el número de sitios por promotor predichos por el método descrito en la sección 2.2 y MotifClick.

Usando los promotores identificados en el genoma *Escherichia Coli K12-W3110*, también comprobamos el tiempo de ejecución del algoritmo descrito en la sección 2.1 y MotifClick en un ordenador de escritorio con cuatro procesadores Intel Core y 16 GB RAM.

La segunda validación del método consistió en evaluar el nivel de predicción falsa debido al azar, a través de la búsqueda de TFBS sobre ADN sin información biológica. Para ello, primero se desordenó la secuencia de los promotores identificadas en los tres metagenomas analizados en este trabajo (mina acidificada, sedimentos marinos y suelo de granja). La desorganización de las secuencias se realizó mediante ventanas de veinte nucleótidos con solapamiento de cinco y luego se varió el orden de los nucleótidos contenidos en los fragmentos de secuencias de forma aleatoria, manteniendo así, la información local.

Cada promotor se desordenó 1000 veces y el algoritmo de identificación *de novo* de sitios de unión (sección 4.2) se ejecutó también este mismo número de veces sobre los tres metagenomas con sus secuencias de promotores desordenadas.

7. Descripción de los análisis estadísticos para el estudio del potencial regulador en diferentes comunidades de microorganismos.

Después de la asignación de funciones (cuando fue posible) a los genes que flanqueaban los promotores (seleccionados según el método de la sección 2.1); procedimos a la comparación de la distribución de ortólogos según su potencial regulador (número de sitios de unión por promotor). Este estudio permitió conocer que funciones o familias génicas tenían un mayor potencial regulador. La clasificación usada para los ortólogos fue acorde a la base de datos SEED y el método esta descrito en la sección 3.1. Para este análisis estadístico primero dividimos las distribuciones de sitios de unión (de los metagenomas descritos en sección 1.1) en tres cuantiles: (1) bajo potencial regulador (1-8 sitios de unión por promotor), (2) potencial regulador medio (8-13 sitios de unión por promotor) y (3) potencial regulador alto (más de 13 sitios de unión por promotor). Para cada cuantil se determinó cuantas funciones eran compartidas entre los tres ambientes estudiados, entre dos y cuantas eran especificas o estaban presentes en sólo uno (lo cual nos da una primera idea general sobre la diversidad y especificidad de funciones entre ambientes según diferentes niveles de potencial regulador). Con el objetivo de conocer la relación entre el potencial regulador y las funciones específicas o compartidas usamos una prueba de Pearson $\chi 2$ para conteos, usando la función chisq.test incluida en el paquete "stats" de R (http://www.r-project.org/). Además, para el análisis de correspondencia usamos la biblioteca "ca" también del programa R.

Otro estudio estadístico permitió tener una visión más específica de la distribución del potencial regulador de las familias génicas en un ambiente determinado. Este nuevo análisis consistió en ordenar los promotores según el número de sitios de unión predichos (acorde al método explicado en la sección 2.1). Luego, separamos los promotores en grupos considerando su potencial regulador: llamamos "top 1%", "top 5%", "top 10%", "top 20%" (y así sucesivamente hasta llegar al 40%) a los grupos formados por los promotores con más cantidad de sitios de unión predichos que correspondía al 1%, 5%, 10% , 20%, 30%, 40% del total de los datos, respectivamente (esto se hizo para cada uno de los tres ambientes). Para este análisis usamos las clasificaciones intermedias de MEGAN acorde a la base de datos SEED. Los gráficos para la visualización de estos datos ("heat maps") se realizaron mediante el paquete "ggplot2" de R. Para comprobar la presencia de funciones

significativas dentro cada uno de los grupos “top” por ambiente, aplicamos una prueba de Fisher para conteos.

Para conocer que funciones marcan la diferencia del potencial regulador entre ambientes, retuvimos las funciones que aparecían sobrerrepresentadas en los grupos “top” con alta significancia estadística (p-valor <<0.5 al aplicar la prueba de Fisher a cada entorno en particular) y que además tenían un número aceptable de ortólogos en los otros dos ambientes que permitiera el uso de la prueba de Fisher de una manera confiable.

8. Construcción de las redes de sitios de unión de factores de transcripción para cada metagenoma.

También construimos redes de TFBSs para conocer que sitios de unión eran más abundantes en determinado grupo funcional con respecto al resto. Para la construcción de las redes usamos la clasificación funcional de MEGAN y la significancia estadística de los sitios de unión dentro de un grupo con respecto al total se estimó a través de una prueba de Fisher. Las redes se construyeron usando como nodos los promotores que correspondían a una determinada función y los vértices representan el número de sitios de unión compartidos entre promotores de un mismo subsistema de interés según la clasificación SEED. El dibujo y visualización de las redes se realizó con el programa Cytoscape (Saito et al, 2012).

III. RESULTADOS
(Primera Parte)

Las nuevas técnicas de secuenciación han traído consigo el desarrollo de disciplinas como la metagenómica (Tringe et al, 2005) y la metatranscriptómica (Frias-Lopez et al, 2008). Ambos estudios han ampliado la visión sobre la interacción de los microorganismos con el medio ambiente. La metagenómica ha permitido conocer la dotación génica de especies de bacterias y arqueas resistentes al cultivo, mientras que con la metatranscriptómica se han estudiado los patrones de expresión de comunidades enteras en respuesta a los parámetros físico-químicos del nicho. En este capítulo describimos una aproximación para el estudio de las regiones reguladoras o promotores (reguloma) de organismos procariotas secuenciados directamente del ambiente. Esto, como alternativa y a la vez complementación de los meta-análisis de comunidades microbianas, permite explorar las características propias de las cominidades que juegan un papel en la adaptación.

A continuación describimos los resultados en relación al papel que juegan las regiones reguladoras en la supervivencia frente a determinados entornos. Para ello, nos hemos basado en la asunción de que los promotores con más sitios de unión a factores de transcripción son aquellos que requieren un nivel mayor de regulación, como se ha demostrado en diversos estudios (Farre et al, 2007). Por ejemplo, los genes que responden a situaciones de estrés en levaduras necesitan la vinculación de más reguladores para adaptar sus patrones de expresión a cambios drásticos del ambiente (Lin et al, 2010).

Siguiendo este razonamiento, nos hemos basado en tres muestras obtenidas a partir de tres nichos ecológicos diferentes (Sun et al, 2011):

-El suelo de una granja ganadera en Waseca, Minessota, USA, (WFS).
-Sedimentos marinos formados por restos de ballena en descomposición colectados en la bahía de Santa Cruz en el océano Pacífico y cerca de la península Antártica, (WhF).
- Una mina acidificada en Richmond Mountain, California, (AM).

Luego, la segunda etapa consistió en la detección de sitios de unión de factores de transcripción como estimador del potencial regulador. Para la búsqueda de posibles TFBSs hemos adaptado, para ser usado en secuencias metagenómicas, un algoritmo para la identificación *de novo*. A continuación se detallan los resultados obtenidos en cada una de las etapas que conformaron este estudio, ilustradas en la Figura 1.

1. Promotores identificados a partir de datos metagenómicos.

La identificación y definición de regiones reguladoras en metagenomas, se llevó a cabo sobre la base de la proximidad y orientación respecto a regiones codificantes, como se describe en sección 2.1 de la Metodología y también se resume en la Figura 1. La fracción de fragmentos de ADN que resultaron tener una posible región reguladora fue similar en los tres ambientes (aproximadamente 4%, 2% y 3% para el suelo de granja, los sedimentos marinos y la mina acidificada, respectivamente.

Estas secuencias pasaron otro filtro, para eliminar la contaminación de genes y promotores de eucariotas que se encontró al asignar taxonomía mediante MEGAN. La contaminación fue fundamentalmente en el suelo de granja, por genes pertenecientes a plantas.

Una vez tuvimos las regiones intergénicas libres de secuencias pertenecientes a organismos eucariotas, a través del programa Prodigal (Hyatt et al, 2012), evaluamos de nuevo la presencia de regiones codificantes entre nuestro grupo de posibles promotores. Luego, además de eliminar las secuencias de regiones intergénicas donde se predijeron regiones codificantes (no detectadas con la búsqueda previa de homología), también eliminamos otros promotores con estructuras como: CRISPR, ARNnc y pequeños ARN. La presencia de CRISPR resulto ser particularmente abundante en la muestra de la mina acidificada en comparación con los otros dos nichos estudiados.

Los resultados finales después de aplicar los filtros anteriores y reasignar taxonomía se observan en la Figura 2, 3 y 4 para muestras tomadas de suelo de granja, sedimentos marinos y la mina acidificada, respectivamente. Este estudio permitió asegurar la presencia de las mismas especies detectadas por otros métodos en nuestro grupo de promotores seleccionados. De esta forma, comprobamos que nuestra metodología no provocaba desviaciones hacia un taxón determinado.

Al realizar la asignación de función a la región codificante adyacente al grupo de promotores identificados; el total con función conocida por nicho fue: 1646, 1514 y 4646 para WFS, WhF y AM, respectivamente (Figura 1).

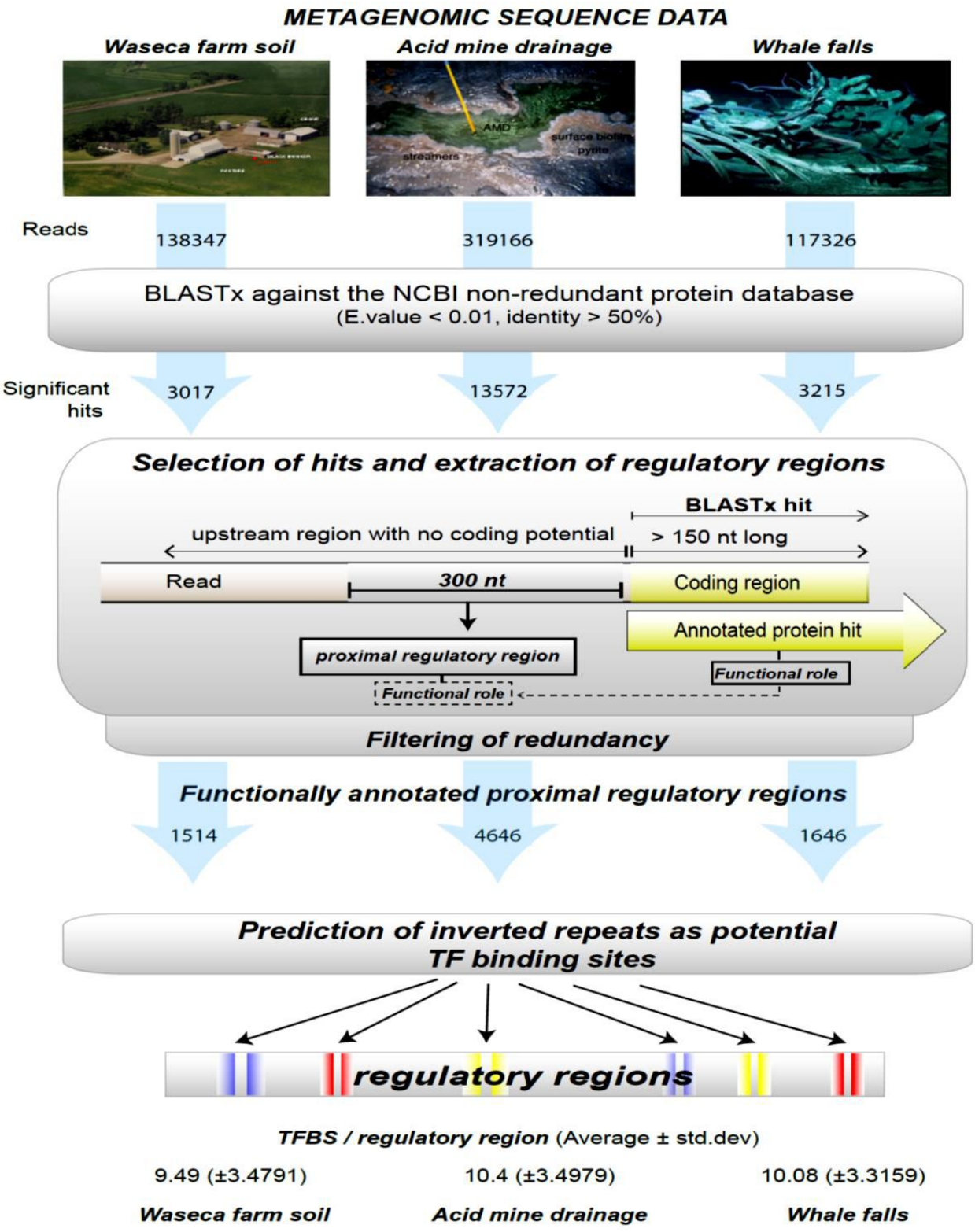


Figura 1. Metodología general seguida para el estudio de la regulación génica en comunidades microbianas.

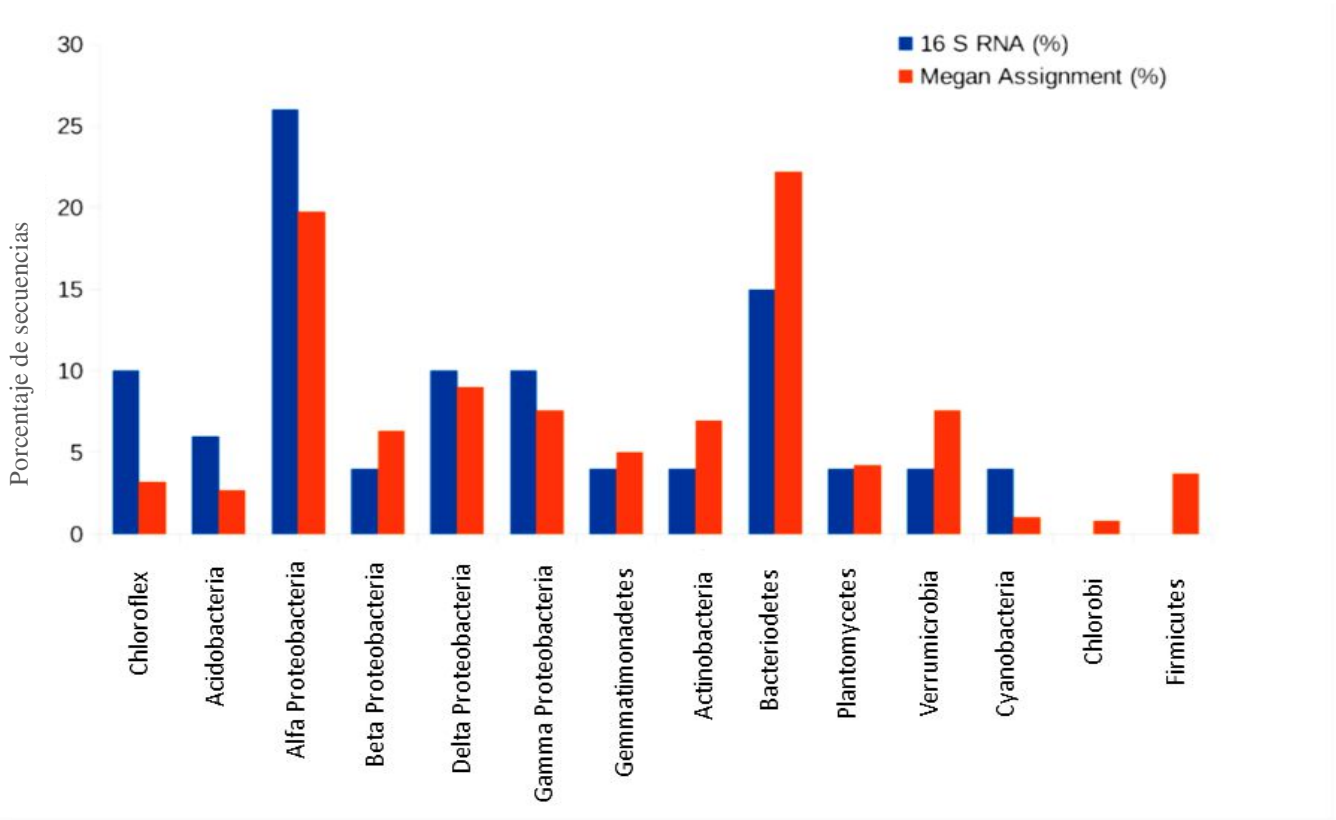


Figura 2. Asignación de taxonomía a promotores de la muestra tomada de suelo de granja comparado con un estudio previo de 16S ARNr. En color rojo se representa el porcentaje asignado a cada taxón con respecto al total en nuestro grupo de promotores seleccionados en el nicho analizado. Por otro lado, en color azul se representa, la asignación de taxonomía (porcentaje con respecto al total) del metagenoma completo sin ningún paso de selección y tomamos como referencia un estudio previo de 16S ARNr (Tringe et al, 2005). La comparación para esta muestra se hizo contra la biblioteca genómica.

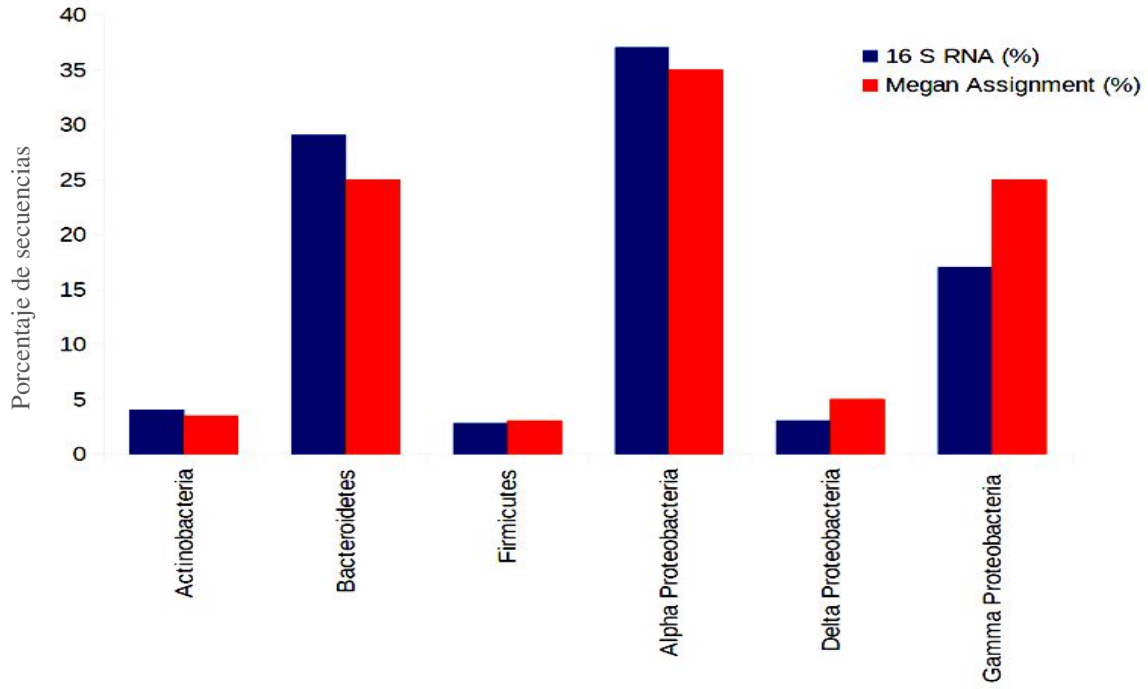


Figura 3. Asignación de taxonomía a promotores de la muestra tomada de restos de ballena comparado con un estudio previo de 16S ARNr. En color rojo se representa el porcentaje asignado a cada taxón con respecto al total en nuestro grupo de promotores seleccionados en el nicho analizado. Por otro lado, en color azul se representa, la asignación de taxonomía (porcentaje con respecto al total) del metagenoma completo sin ningún paso de selección previo y tomamos como referencia un estudio previo de 16S ARNr (Tringe et al, 2005). La comparación para esta muestra se hizo contra la biblioteca de clones de PCR (detectaba más grupos).

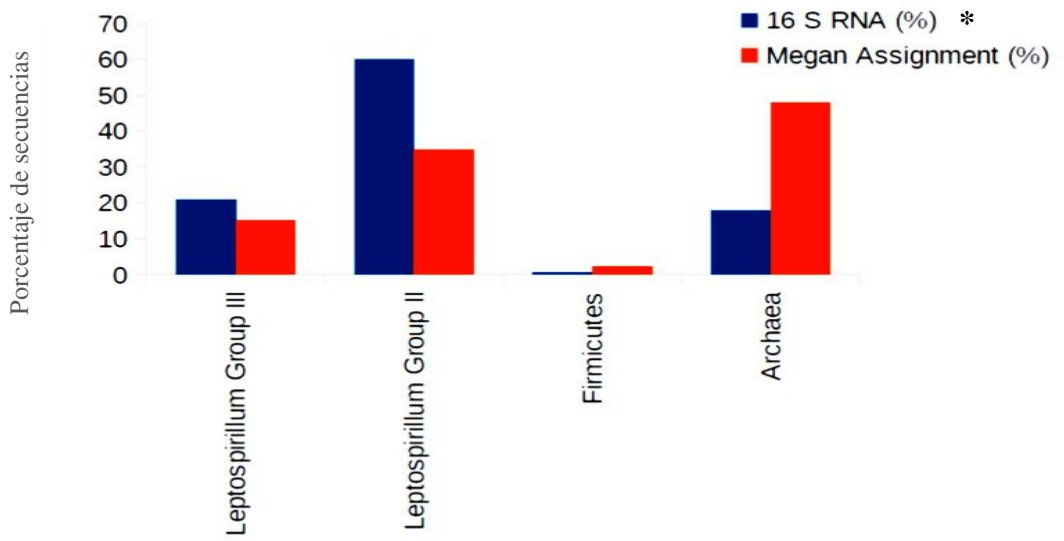


Figura 4. Asignación de taxonomía a promotores de la muestra tomada de una mina acidificada comparado con un estudio previo de 16S ARNr (* combinado con metagenómica). En color rojo se representa el porcentaje asignado a cada taxón con respecto al total en nuestro grupo de promotores seleccionados en el nicho analizado. Por otro lado, en color azul se representa, la asignación de taxonomía (porcentaje con respecto al total) del metagenoma completo sin ningún paso de selección. Los datos de la mina acidificada se tomaron de dos publicaciones diferentes una para la muestra "UBA" (Lo et al, 2007) y otra para la "5WAY" (Tyson et al, 2004) (detalles acerca de estas muestras en Metodología, Sección 1.1). La estimación de taxonomía para "UBA" se hizo a través de metagenómica mientras que para "5WAY" se utilizó 16S ARNr; en el grafico se observa el estimado para las dos muestras combinadas.

2. Predicción del potencial regulador en tres nichos diferentes.

La selección de promotores, permitió su posterior caracterización a nivel de capacidad reguladora o potencial regulador, el cual se infiere y define en este estudio como la densidad y variabilidad de TFBSs en estas regiones. Por ese motivo, se exploraron varias alternativas metodológicas disponibles. Se descartaron todas las estrategias de búsqueda basadas en homología de TFBSs conocidos, ya que favorecería a las especies mejor estudiadas; para finalmente escoger un método de predicción *de novo* que, teóricamente cubriría a todas las especies con el mismo poder predictivo. Teniendo como base los criterios anteriores, calculamos el número de sitios de unión de factores de transcripción por promotor a partir de secuencias metagenómicas (sin ninguna información previa acerca de posibles reguladores o sus sitios). La mejor opción para este estudio fue el uso de un método *de novo*, los detalles del protocolo usado para la predicción de TFBSs están descritos en la Sección 4.2 de Metodología.

Las distribuciones generales de densidad de sitios de unión por promotor para cada nicho se observan en la Figura 5, donde incluimos todos los promotores con y sin función conocida. Estas distribuciones generales no manifiestan cambios drásticos entre nichos. La media fue ~10 sitios de unión por promotor para los tres ambientes (ver Figura 1 para valores exactos de media y desviación estándar por nicho).

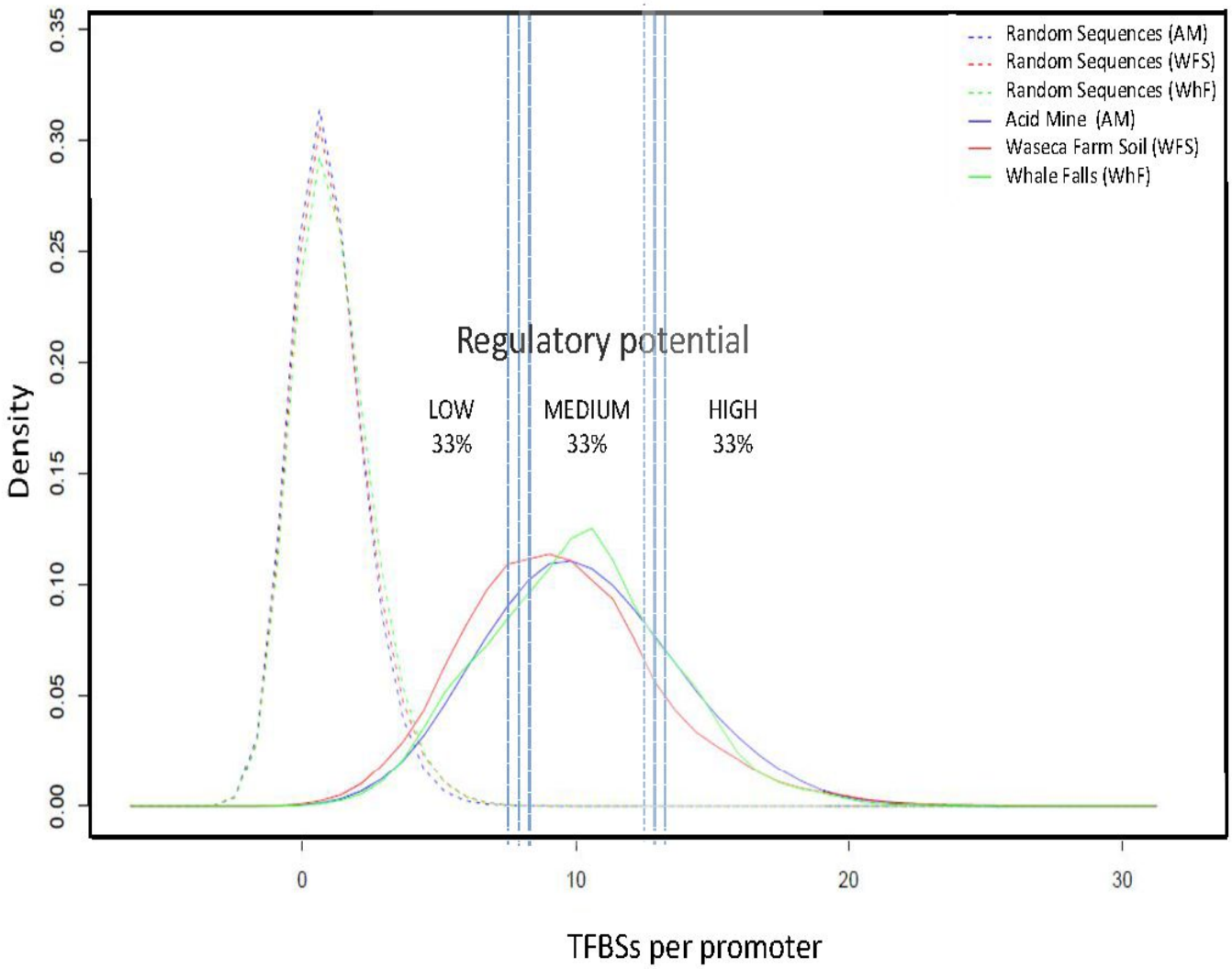


Figura 5. Perfiles de densidad de sitios de unión por promotor en tres nichos diferentes. En líneas discontinuas se muestran las estimaciones de TFBSs por promotor después de un análisis de permutación. Por otro lado, en líneas continuas se muestran las estimaciones de TFBSs en los datos reales de promotores identificados en cada uno de los tres nichos. Las líneas verticales muestras los cuantiles de división de las estimaciones, el primer tercil se considera como bajo potencial regulador al tener estos promotores pocos sitios de unión para TFs, el segundo tercil se considera que tiene un potencial de regulación moderado, mientras el tercer tercil se considera como alto.

3. Evaluación de los métodos para la identificación de promotores y sitios de unión de factores de transcripción en metagenomas.

Considerando que el método propuesto para la identificación de TFBSs no se basó en ninguna información previa de datos experimentales fue necesario investigar de manera cuidadosa la presencia de falsos positivos dentro de los sitios predichos. Varios análisis, tanto cuantitativos como cualitativos, se hicieron para evaluar nuestro método y además reducir la presencia de falsos positivos.

Desde un punto de vista cuantitativo aplicamos el mismo método de extracción de promotores y búsqueda de sitios de unión de factores de transcripción al genoma de *E. coli K12* y lo comparamos con otro método de identificación de TFBSs (MotifClick). En Metodología, Sección 6, se explica la diferencia entre los diferentes algoritmos usados y su aplicación para la predicción de sitios en *E. coli K12*. Escogimos este genoma en particular porque ha sido ampliamente estudiado además pudimos comparar diferentes métodos de extracción de promotores.

En la Figura 6 se observa la correlación entre las predicciones de sitios de unión por promotor a través de MotifClick y el método que implementamos para su aplicación en metagenomas. A diferencia del estudio que se realizó sobre otras cepas de *E. coli* y que se mostrará en la parte II de Resultados, aquí aplicamos el protocolo completo de la misma manera que fue usado para los metagenomas (Metodología, Sección 2.1) mientras que para las otras cepas de *E. coli* caracterizadas usamos el método de extracción de regiones intergénicas de DOE-JGI (Metodología, Sección 2.2 y Resultados parte II).

Desde el punto de vista cualitativo, primero aseguramos la significancia biológica de los sitios de unión encontrados a través de un análisis de permutación. Después de desorganizar las secuencias de los promotores, tal como se explica en la Sección 6 de Metodología, fue aplicado sobre las nuevas secuencias generadas el mismo método de predicción de TFBSs desarrollado para su uso en metagenomas. En la Figura 5 en líneas discontinuas se observa el resultado obtenido sobre las secuencias desorganizadas.

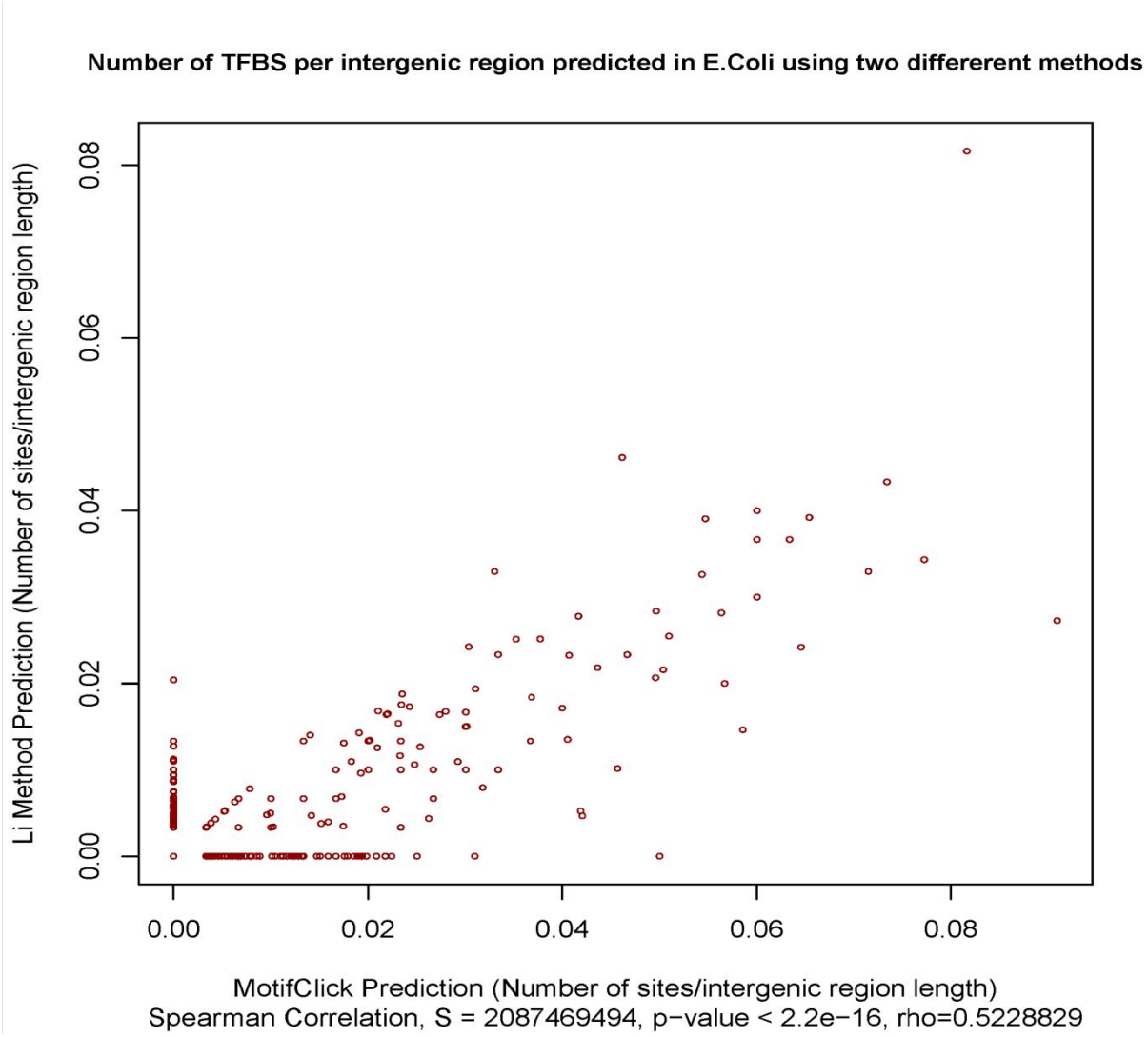


Figura 6. Correlación entre la capacidad predictiva del método implementado para la predicción de sitios de unión en metagenomas (Método de Li) y otra metodología independiente (MotifClick). El análisis se realizó usando como modelo el genoma de *E.Coli K12*.

Otro análisis de validación alternativo consistió en evaluar las coincidencias entre los patrones de sitios de unión predichos en las tres muestras de metagenomas analizadas y todos aquellos TFBSs conocidos para bacterias y arqueas. Para este estudio utilizamos como referencia la base de datos RegPrecise (Novichkov et al, 2010). La relación entre los sitios predichos en los metagenomas para los 38 factores de transcripción analizados se muestran en la Tabla 1 de Anexos. En total el 28% de los sitios detectados en los datos analizados en este trabajo coinciden con la secuencia de aquellos que habían sido reportados previamente. Como se observa en la Tabla 1 de Anexos, el método implementado para los metagenomas fue capaz de detectar TFBSs para todos los factores de transcripción reportados en RegPrecise.

La presencia de CRISPR fue de nuevo analizada, por la interferencia que podrían causar estas estructuras con el método de predicción de TFBSs. Esta vez empleamos un método diferente al explicado en la sección anterior, mediante el uso de la herramienta CRISPRFinder (Grissa et al, 2007). A pesar de los filtros aplicados en etapas previas, 1% de los promotores que habían sido seleccionados dieron positivo a la presencia de CRISPRs. Como consecuencia, estos fueron removidos, así como los TFBSs que contenían el mismo patrón detectado como posible CRISPR.

En resumen, la validación del método permitió comprobar cuan confiable era el grupo de sitios de unión predichos para poder ser usados como estimadores del potencial regulador.

4. Organizacional funcional del potencial regulador dentro del nicho.

De acuerdo a la Figura 5 los tres nichos estudiados tienen distribuciones globales semejantes, a pesar de la diferencia entre los parámetros físicos químicos en cada entorno y la diversidad en la composición de especies entre los nichos.

Por esta razón exploramos los datos más allá de simples conteos de sitios de unión por promotor, y estudiamos el comportamiento de grupos de funciones de acuerdo a su potencial regulador. En este análisis, primero se identificaron aquellos genes que están controlados por el grupo de regiones reguladoras que seleccionamos previamente. La asignación de las funciones de las regiones codificadoras se realizó a través de la anotación SEED mediante el programa MEGAN (Huson et al, 2011). Posteriormente fueron

organizados todos los promotores para cada nicho de acuerdo a su número de sitios de unión y se agruparon dentro de tres categorías: bajo, medio o alto atendiendo a los terciles en que se divide la distribución de sitios de unión por promotor (Figura 5). Para cada uno de estos grupos se calculó la proporción de funciones que se comparten entre los tres o entre dos nichos, así como las que son únicas o específicas de un entorno en particular. Es interesante, como el grupo de funciones, cuyos promotores clasifican con alto potencial regulador mostraron significativamente menos coocurrencias entre nichos que aquellas controladas por promotores con potencial regulador medio o bajo (Figura 7). Este resultado provee información sobre el papel de los promotores con redes complejas de reguladores en la adaptación a condiciones ambientales específicas del nicho.

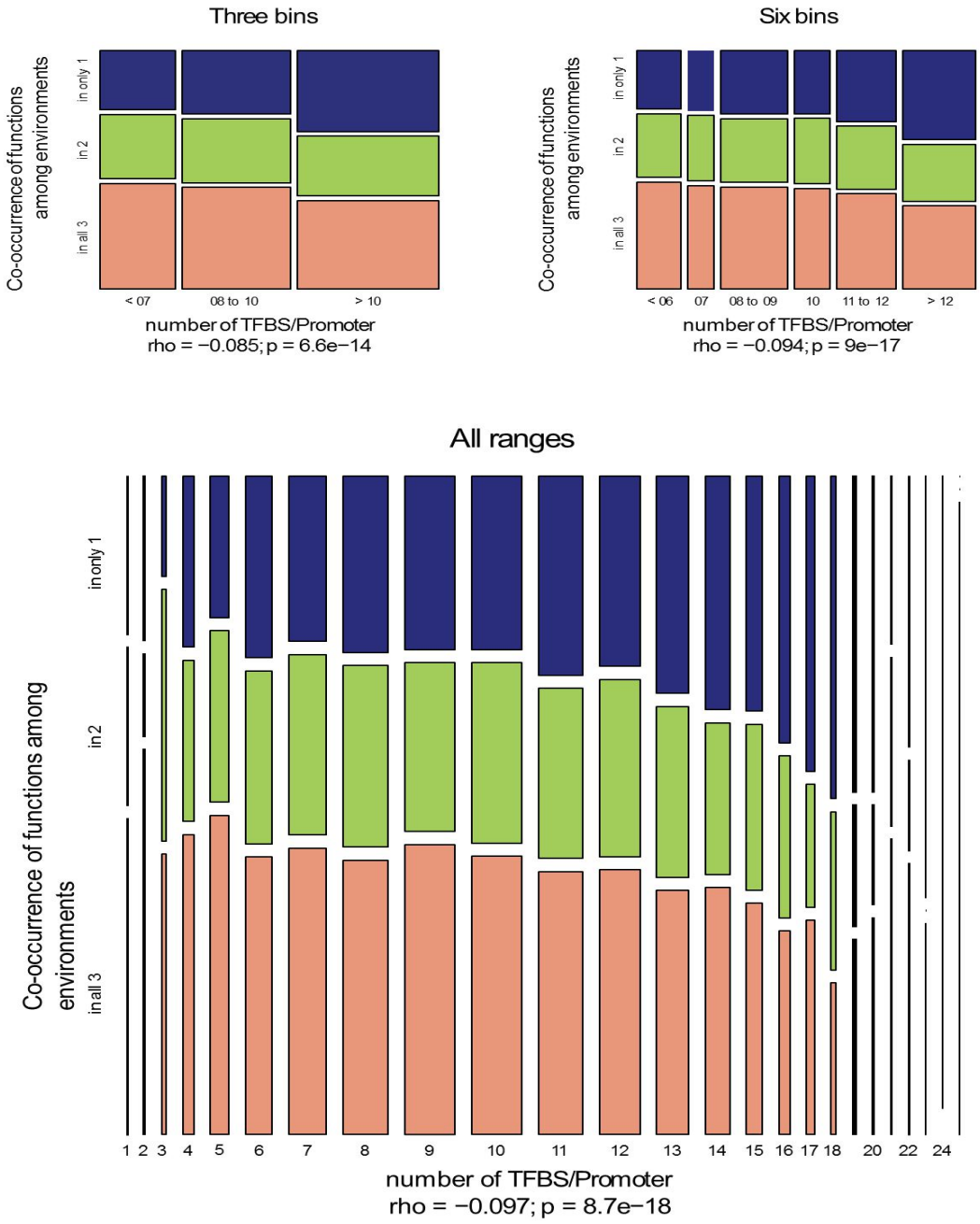


Figura 7. Análisis del potencial regulador de acuerdo a como se comparten las funciones entre nichos a diferentes niveles resolutivos: three bins (usando los terciles de la figura 5), six bins and all ranges. A) El área de cada rectángulo representa el número de promotores que se incluyen dentro de las categorías siguientes: *In only 1*, solo se encontró la función regulada por el promotor en un entorno; *In 2*, se encontró en dos entornos; *In all 3*, se encontró presente en los tres entornos.

Para estudiar con más detalle las funciones que involucran promotores con redes complejas de reguladores, se dividieron los datos en subgrupos más reducidos. Los subgrupos cubrieron las categorías de potencial regulador medio y alto: divididos de manera tal que involucraron las siguientes gamas: 1%, 5%, 10%, 20%, 30%, 40% (que representan el porcentaje de los promotores con mayor potencial regulador con respecto al total). Luego, se evaluó el enriquecimiento de funciones dentro de cada gama, mediante una prueba de Fisher. En las Figuras 8, 9 ,10 se ilustran, de acuerdo a la intensidad del color, las proporciones de funciones que pertenecen a cada gama (eje x) con respecto al total de genes que presentan la misma clasificación funcional SEED (eje y). Este análisis exploratorio permitió la identificación de funciones enriquecidas en las gamas evaluadas para cada nicho estudiado (p-valor < 0.5). Más detalles sobre las funciones enriquecidas en cada nicho se pueden consultar en la Figura 8, 9 y 10 para WFS, WhF y AM, respectivamente; donde, las funciones con enriquecimiento significativo se muestran en color rojo.

De manera general las funciones encontradas enriquecidas cubrieron diferentes procesos. Muchos se relacionan con la detección y control de factores externos, tales como: receptores y transportadores, por ejemplo, del tipo Ton and Tol en AM (Figura 10). Para el caso del entorno marino (Figura 9), la característica más distintiva es la presencia de abundantes sitios de unión en promotores que regulan genes que responden a situaciones de estrés oxidativo y nitrosativo.

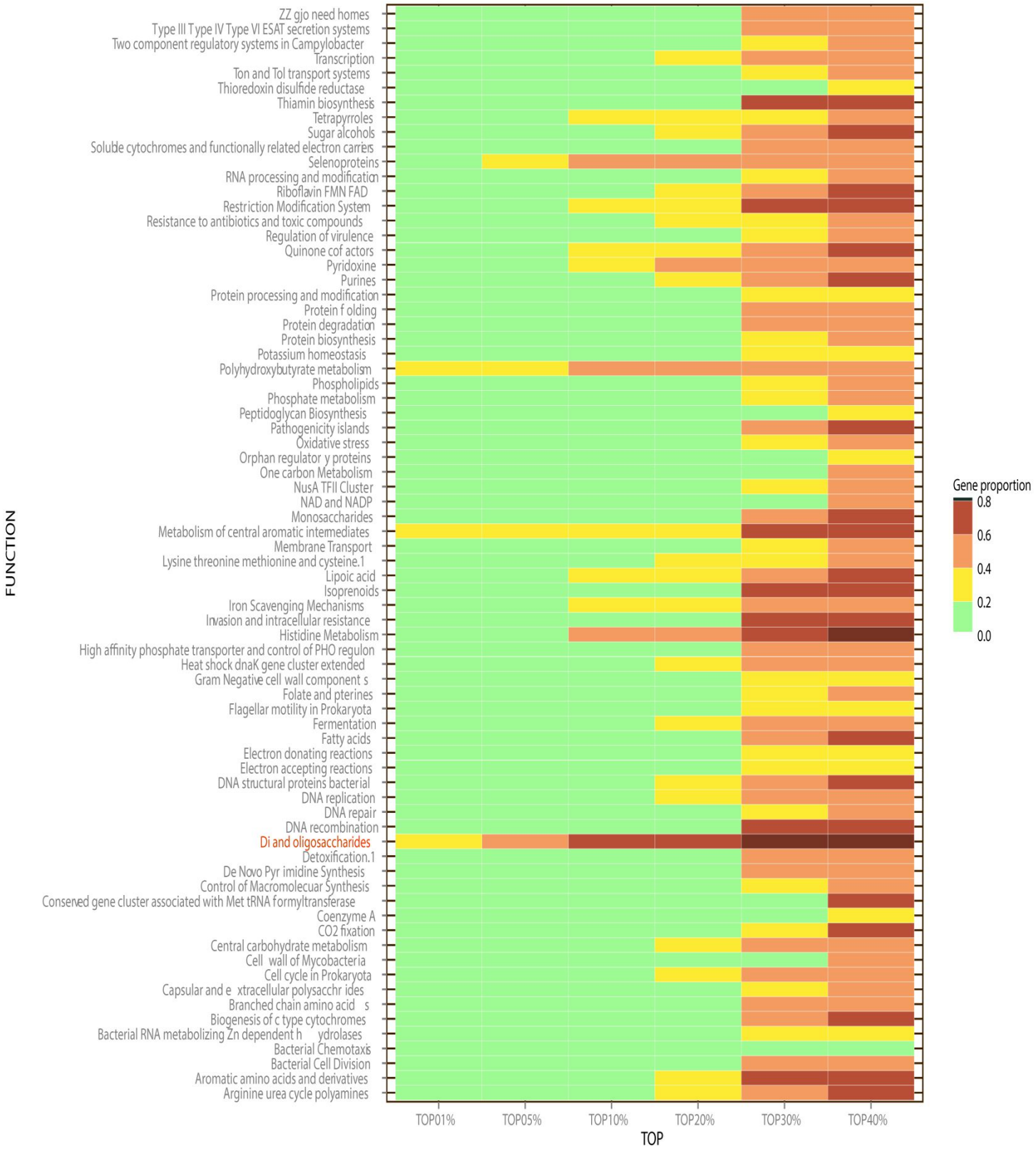


Figura 8. Organizacion funcional de los promotores con potencial regulador medio y alto en la muestra de suelo de granja

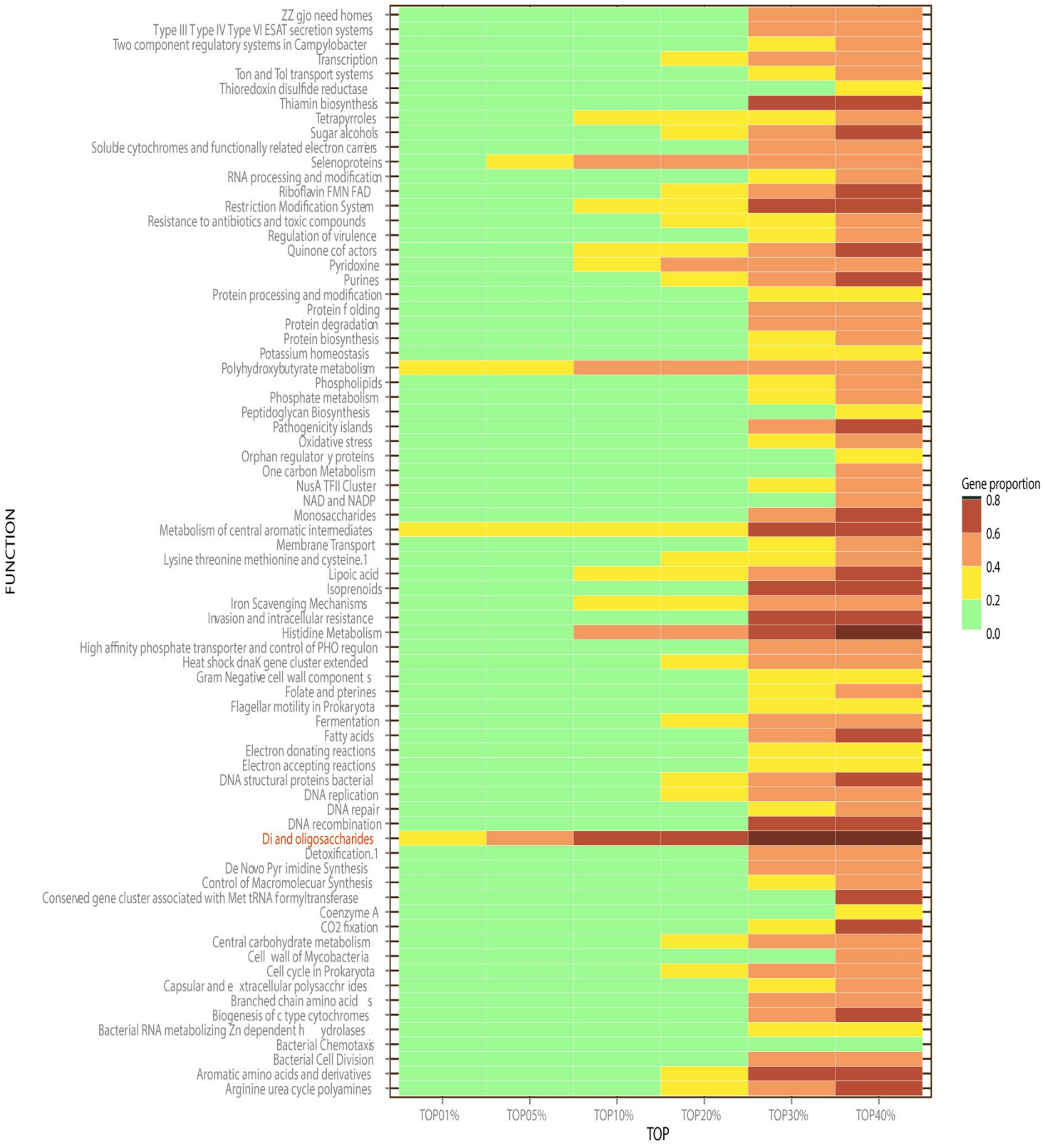


Figura 9. Organizacion funcional de los promotores con potencial regulador medio y alto en las muestras de restos de ballena.

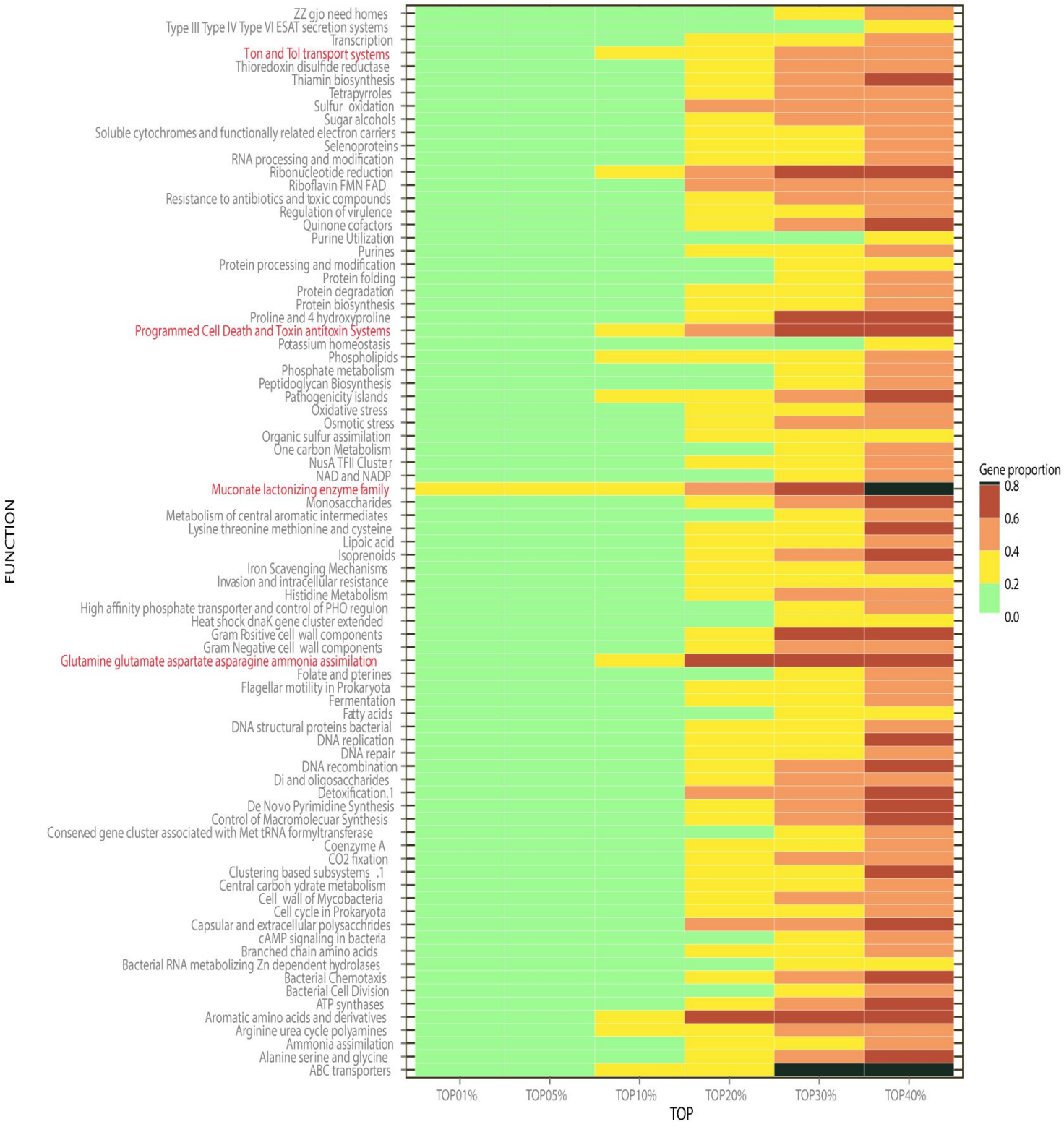


Figura 10. Organizacion funcional de los promotores con potencial regulador medio y alto en la muestra de una mina acidificada.

Con los análisis explicados en esta sección, se observa el comportamiento del potencial regulador dentro de cada ambiente; pero aún es necesario evaluar en los otros entornos, los mismos procesos biológicos encontrados con potencial regulador medio y alto dentro de un nicho en particular.

5. Comparación entre nichos según el comportamiento del potencial regulador.

Enfocados en entender el papel del ambiente en la regulación génica, comparamos las funciones más reguladas entre los tres nichos analizados. Como los entornos de los que se tomaron las muestras tienen parámetros físico-químicos muy diferentes entre sí; constituyen modelos excelentes para conocer el papel de las condiciones ambientales en la modulación de la expresión génica a través de la estructura primaria de los promotores.

El siguiente estudio se basó en la selección de aquellos genes y sus promotores adyacentes, cuyas funciones mostraron un enriquecimiento considerable (p-valor <0.05) en las categorías de alto potencial regulador (gamas correspondiente al 1, 5, 10 y 20 %). Además estas funciones para ser adecuadas para el estudio debían tener suficientes ortólogos en los otros ambientes para dar poder estadístico a la comparación. En total se encontraron cinco funciones que cumplían con los requisitos expuestos que incluyen los procesos relacionados con: virulencia, ciclo celular, metabolismo de los carbohidratos, respuesta a estrés y metabolismo de cofactores. Todos estos genes y promotores fueron comparados contra sus ortólogos en los otros nichos. Luego analizamos la posible asociación de las regiones reguladoras con los factores ambientales; mediante la evaluación de los mecanismos bioquímicos de adaptación donde participan los genes analizados y los posibles factores ambientales que los modulan o desencadenan. A pesar de la información limitada que se tiene acerca de los parámetros físicos y químicos del entorno en estudios de metagenómica (incluidos los nichos estudiados aquí), aun se pudo encontrar posibles escenarios donde la adaptación juega un papel en la modulación de la expresión génica a través de la estructura de los promotores. A continuación se explican los resultados obtenidos en este análisis y en la sección de discusión se detalla más como el ambiente podría influir en los patrones observados para cada nicho.

Por ejemplo, en la Figura 11 se observa como en la muestra tomada de suelo de granja los genes relacionados con metabolismo de los carbohidratos (fundamentalmente di – y oligosacáridos) muestran un alto potencial regulador; incluso si se compara con los otros nichos analizados (p-valor = $1x10^{-16}$ para la prueba de Kruskal Wallis entre ambientes mientras que el p-valor usando la corrección de Bonferroni = 9×10^{-13}).

Un comportamiento diferente se observa en las muestras de sedimentos marinos, donde los genes regulados por promotores con abundantes sitios de unión se le asignaron funciones biológicas relacionadas al ciclo celular y al crecimiento, por ejemplo, genes pertenecientes a operones que controlan la síntesis macromolecular básica (Figura 11). Esta diferencia es más marcada entre las muestras de sedimentos marinos y suelo. En el caso de los mecanismos de respuesta a estrés, las comunidades que habitan los sedimentos conformados por restos de ballena presentan un mayor potencial regulador comparado con los otros dos nichos. Además, los genes ortólogos que codifican proteínas que protegen contra estrés oxidativo como, Dps (proteína de unión a ADN no-específica) y NnrS (proteína que responde a las concentraciones externas de NO) mostraron, entre los tres nichos estudiados, una variación abrupta del potencial regulador de sus correspondientes promotores.

En cuanto al entorno más extremo analizado, la mina acidificada, no mostró de manera general una familia génica o grupo funcional con considerablemente alto potencial regulador (con una sola excepción, al comparar con suelo en funciones relacionadas con virulencia, p-valor < 0.05, Figura 11). Por otra parte, sí se observó variación en ortólogos, por ejemplo, en el receptor Ton B, cuyo potencial regulador fue variable entre los tres nichos.

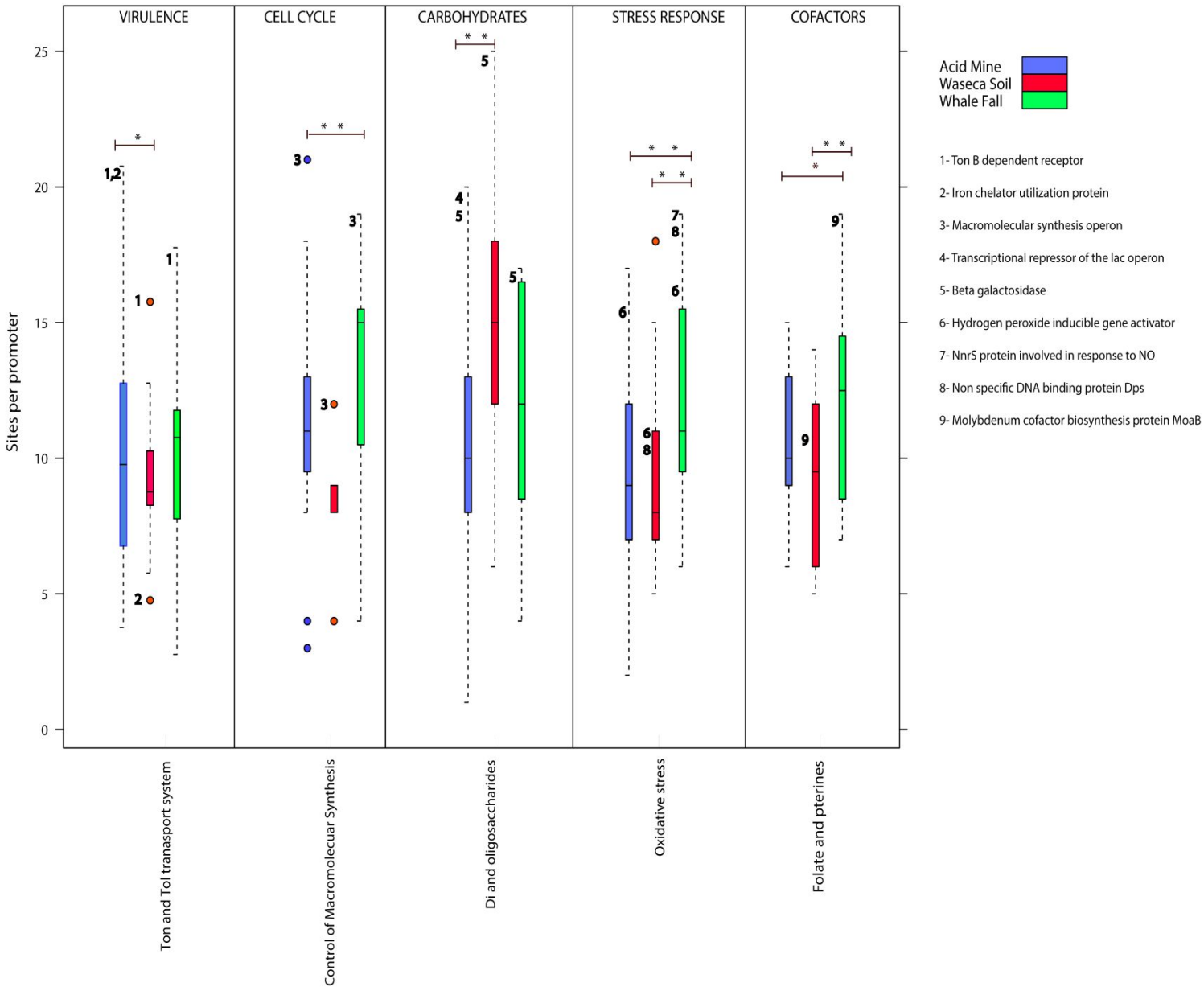


Figura 11. Diagramas de cajas que muestran la densidad de sitios de unión por promotor de procesos biológicos relacionados con funciones que mostraron un alto potencial regulador en alguno de los nichos analizados (AM se representa en color azul, WS en color rojo y WhS en color verde). El símbolo * indica un p-valor ≤ 0.05, mientras que ** indican un p-valor ≤ 0.01 (Prueba de Fisher). Los grupos que mostraron un p-valor significativo se encuentran bajo la línea paralela dibujada sobre las cajas. Los números y su posición representan el nivel de regulación observado para el promotor de determinado gen, que a su vez es nombrado en la lista ubicada a la derecha de la figura.

6. Distribución de sitios de unión conocidos en las muestras de: la mina acidificada, los restos de ballena y el suelo de granja.

Otro análisis alternativo fue llevado a cabo para conocer cuantos TFBSs, de los predichos para los tres nichos, presentan una estructura que coincide con la de sitios descritos previamente, reportados en RegPrecise (Novichkov et al, 2010). El procedimiento usado para determinar la abundancia relativa de sitios de unión conocidos por metagenoma se describe en Metodología, Sección 5. En la Figura 12 se observa aquellos factores de transcripción que, 1) sus TFBSs coincidían con la estructura de los predichos en este estudio y 2) sus sitios de unión por metagenoma fueron significativamente diferentes entre los tres nichos analizados.

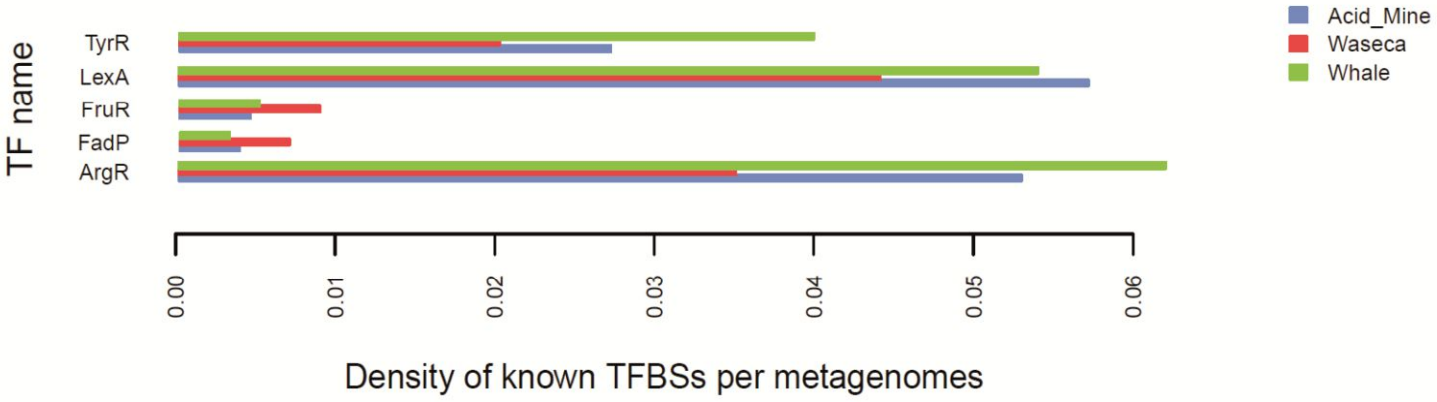


Figura 12. Abundancia relativa de TFBSs reportados previamente en la base de datos RegPrecise. En azul, rojo y verde se representan la mina acidificada, el suelo de granja y los sedimentos marinos, respectivamente.

En la muestra de suelo de granja se encontró abundantes sitios para FruR y FadP. El factor de trascripción FruR, es conocido por modular la expresión de genes involucrados en el metabolismo de los carbohidratos (Ramseier et al, 1995), mientras que FadP es posible que participe en la regulación de la expresión de proteínas que participan en la degradación de ácidos grasos (Cao et al, 2013).

Para las muestras de microorganismos secuenciadas a partir de sedimentos marinos los reguladores: TyrR y ArgR, implicados en el metabolismo de aminoácidos, tienen una abundancia relativa mayor de sitios comparado con los otros dos entornos. TyrR es un regulador transcripcional dual del regulón TyrR. Este regulón involucra genes que son esenciales para la biosíntesis y el transporte de aminoácidos aromáticos (Pittard and Davidson, 1991). El factor ArgR reprime la transcripción de varios genes involucrados en la biosíntesis y el transporte de la arginina, transporte de histidina y su propia síntesis, además activa genes del catabolismo de la arginina (Kiupakis and Reitzer, 2002).

En el caso del regulador LexA se observan abundantes sitios potenciales tanto en las muestras de sedimentos marinos, como en las de la mina acidificada, en comparación con suelo. LexA tiene múltiples funciones que van desde la activación de la respuesta a estrés SOS en bacteria hasta tener un papel en la replicación y reparación del ADN (Janion, 2008;Pruteanu and Baker, 2009).

7. Recopilación de datos sobre parámetros físico-químicos.

Para conocer como el ambiente podría influir en la estructura de los promotores intentamos recopilar parámetros físico-químicos del entorno. Lamentablemente, para las comunidades de suelo de granja y los restos de ballena no existen suficientes datos en la literatura sobre las condiciones ambientales. Debido a esto, sólo pudimos hacer un análisis más exhaustivo en la mina acidificada. Los datos recopilados se muestran en la Figura 13 A estos incluyen, conductividad, concentración de Fe^{2+} / Fe^{3+}, volumen de lluvia caída y pH para las dos muestras de la mina ácidificada analizadas en este trabajo.

8. Redes de regulación de la expresión génica en metagenomas.

A través de la clasificación funcional SEED y del procedimiento descrito en la Sección 8 de Metodología fueron construidas las redes de regulación de promotores para cada función estudiada (listadas en la Figura 8, 9, 10). Para cada red se evaluó mediante la prueba de Fisher que sitios de unión se encontraron sobrerrepresentados para una función determinada. Por ejemplo, como ilustra la Figura 13 B para la mina acidificada en los

promotores relacionados con transportadores Ton y Tol, varios sitios de unión conocidos fueron encontrados como, LexA, ModR, ModE. El estudio sobre las funciones de la muestra de suelo permitió también identificar sitios de unión sobrerrepresentados que coincidían en su estructura con otros reportados en RegPrecise. Por ejemplo, para el factor NikR y PurR se encontraron sitios abundantes con alta significancia estadística en promotores relacionados con el metabolismo del potasio y la histidina, respectivamente. PurR es un factor de transcripción que en *B. subtilis* y *E. coli* actúa como regulador en el metabolismo de las purinas (He and Zalkin, 1994;Sinha et al, 2003). El otro factor, NikR participa en la regulación de numerosos genes para mantener la homeostasia celular (Phillips et al, 2010). Las redes de regulación de la expresión génica para las funciones donde NikR y PurR están sobrerrepresentados se encuentran en Anexos, Figura 1 y Figura 2. Para las muestras de sedimentos marinos no se encontró ningún regulador conocido que fuera abundante de manera significativa en ninguna red de regulación.

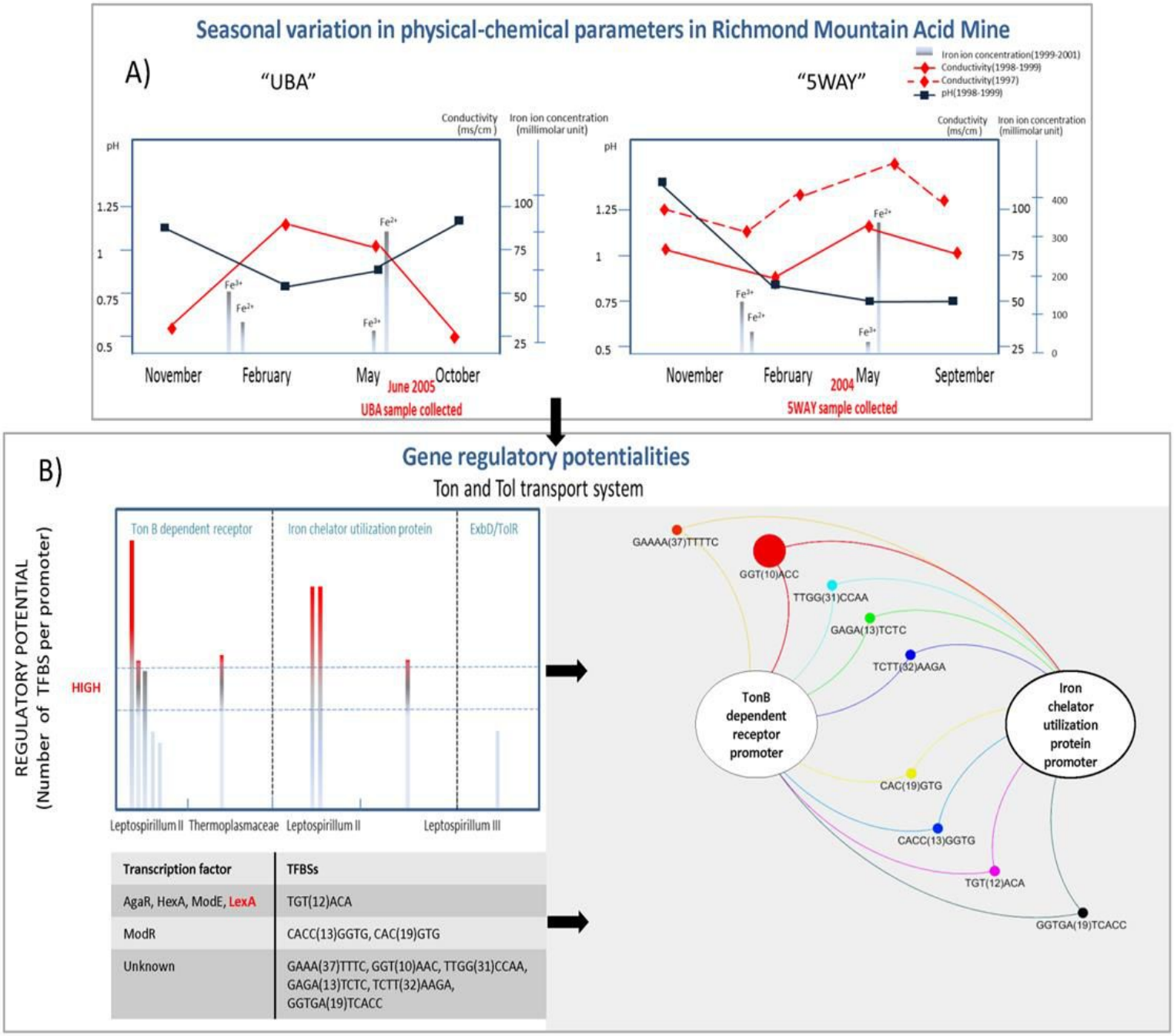


Figura 13 A) Variación estacional del pH, conductividad y concentración de iones de hierro en las muestras "UBA" y "5WAY" en la mina acidificada (Edwards et al, 1999). B) El panel izquierdo ilustra el potencial regulador relacionado con los genes del sistema de transporte Ton y Tol en la muestra "UBA". Los promotores se agrupan por función y especies (eje de las X). Cada barra representa un promotor, el color rojo en las barras sugiere un alto potencial regulador (mayor o igual a 15 sitios de unión por promotor). La tabla representa los sitios de unión que coinciden con factores de transcripción conocidos. El panel derecho representa una red de sitios de unión compartidos entre los promotores de dos genes pertenecientes al grupo de transportadores Ton y Tol. Los genes se representan con círculos blancos. Mientras que los TFBSs se representan por círculos coloreados (el tamaño es proporcional al número de TFBSs encontrados con la estructura ilustrada en la figura).

IV. RESULTADOS
(Segunda Parte)

Una vez habiendo aportado evidencias a la hipótesis sobre la influencia del ambiente sobre las regiones reguladoras y como éstas se modifican en las comunidades microbianas; continuamos la misma línea investigativa, pero centrados en una sola especie. Estudiamos, así, diferentes cepas de *E. coli* localizadas en diversos nichos para analizar los cambios, con más detalle, que se producen a nivel funcional. Conocer más sobre la regulación génica y las funciones que son sensibles a los parámetros ambientales es beneficioso para el futuro diseño de colonias bacterianas con fines biomédicos.

En este capítulo también se describe los resultados obtenidos en colaboraciones externas realizadas con la Universidad de Uppsala y nuestra participación en el proyecto Metahit.

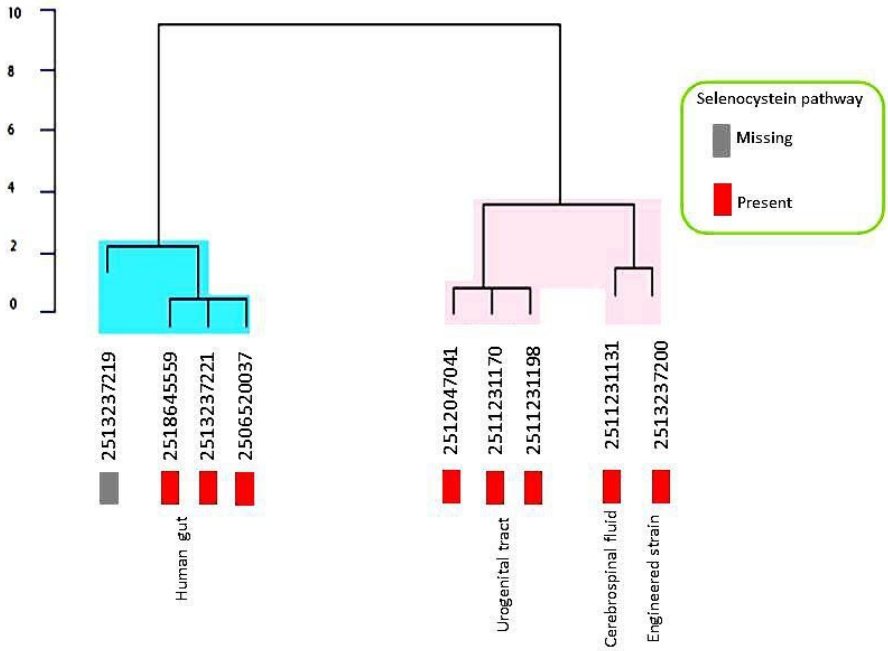
9. Distribución bio-espacial de sitios de unión de factores de transcripción en diferentes cepas de *E. coli*.

Mucho se ha estudiado acerca de la diferencia en la composición génica entre las cepas de *E. coli* (Iguchi et al, 2009), pero poco se conoce de como varían las regiones reguladoras entre las mismas según su habitat. En este trabajo realizamos varios estudios acerca del efecto del nicho sobre las características de los promotores. Primero, caracterizamos las regiones reguladoras según la presencia/ausencia de sitios de unión. Luego, calculamos la abundancia relativa de TFBSs en cada genoma para 86 factores de transcripción (Gama-Castro et al, 2008) (ver Metodología, Sección 1.2 para conocer las cepas escogidas y Sección 4.1 para una información más detallada del protocolo).

En la Figura 14 A se observa el resultado de un análisis de agrupamiento jerárquico, donde se usó como entrada las matrices de TFBSs "scores" (calculados usando el programa MATSCAN) (Blanco et al, 2007). El análisis de agrupamiento muestra una separación, según la localización de la cepa en el cuerpo humano. Por ejemplo, todas las cepas aisladas de intestino se encuentran en la misma rama, separadas de las aisladas del tracto urogenital y fluido cerebroespinal. Otra observación interesante es la separación de las cepas aisladas de intestino, según la capacidad de sintetizar seleno-cisteína (considerado como el aminoácido número 21) (Johansson et al, 2005)

En cuanto a las distribuciones de TFBSs "scores" por promotor, no hay muchas diferencias entre cepas. El caso más relevante es el del factor de transcripción FhlA y el comportamiento observado parece estar marcado por la capacidad o no de sintetizar seleno-cisteína (Figura 14B). Sorprendentemente, la cepa incapaz de sintetizar seleno-cisteína muestra una abundancia mayor de sitios de unión para este factor de transcripción (p-valor corregido por Bonferroni, Kruskall Wallis test = 7.39×10^{-9}). FhlA es un activador transcripcional requerido para la inducción de la expresión de la enzima formato deshidrogenasa H (FDH-H) (Schlensog et al, 1994). FDH-H, a su vez, contiene selenio que se incorpora como seleno-cisteína, en el mismo proceso de transcripción (Sawers, 1994;Zinoni et al, 1986).

A)



B)

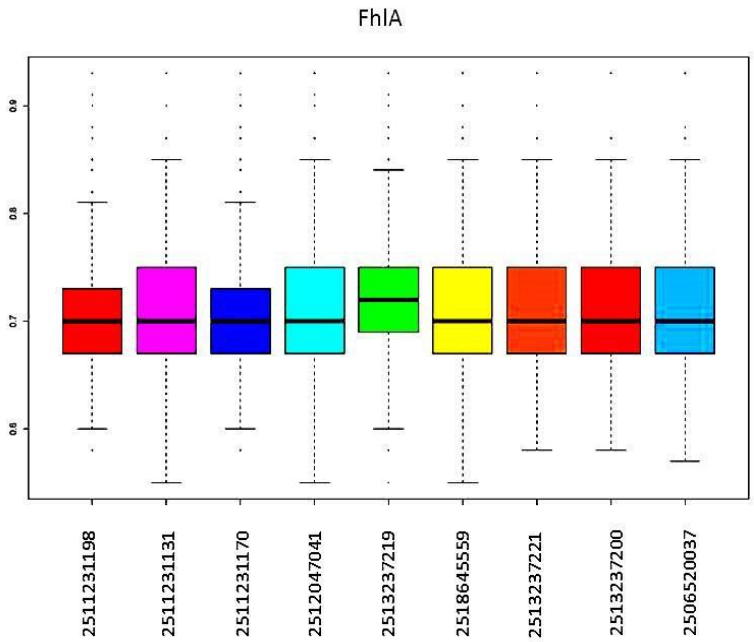


Figura 14. Características de los sitios de unión en nueve cepas de E. coli. A) Análisis de agrupamiento según los TFBSs "scores". B) Distribuciones de los sitios de unión del factor FhlA.

10. Estimación del potencial regulador en *E. coli*.

El potencial regulador como se explica en Metodología Sección 4.2, se estimó a partir de la predicción del número de sitios de unión por promotor. Para este estudio utilizamos las mismas nueve cepas de *E. coli* del análisis explicado en la sección anterior. Los promotores fueron extraídos según la metodología descrita en la Sección 2.2.

Para la predicción de sitios de unión escogimos dos métodos *de novo,* para así evitar desviaciones hacia cepas más cercanas en cuanto a similitud de secuencia a *E. coli K12 (*a partir de la cual se predijeron los sitios para los 86 factores de transcripción usados previamente). Los métodos empleados fueron: 1) MotifClick (Zhang et al, 2011) y 2) el algoritmo descrito en Sección 4.2 que llamaremos de forma simplificada método de Li, por ser éste el nombre del autor de la primera publicación sobre la aplicación de este protocolo para la identificación de sitios de unión en *E.Coli* (Li et al, 2002).

Al aplicar ambos métodos, los sitios predichos que contienen exactamente la misma secuencia y posición en el promotor, para la mayoría de las cepas, constituyeron menos del 10% de la suma total de sitios predichos por MotifClick y el algoritmo de Li en conjunto. Por ejemplo, para la cepa 2511231170 la coincidencia de sitios alcanzó un valor equivalente al 5.9% del total, mientras que para la cepa 2512047041, fue aproximadamente del 8.5% (ver Figura 15A para las cepas 2511231170 y 2512047041). Atendiendo al número total de sitios, MotifClick predice mayor cantidad comparado con el método de Li.

En cuanto a la correlación entre ambos métodos según el número de sitios de unión predichos por promotor se observa la misma tendencia (Figura 15B, resultados de la predicción para las cepas 2511231170 y 2512047041). Los coeficientes de correlación (método de Pearson) para 8 de las 9 cepas estudiadas fueron mayores de 0.5, excepto para 2513237200; cuyos promotores en la predicción por el método de Li muestra una anómala gran cantidad de sitios palíndromos sobrerrepresentados.

Los tiempos de ejecución el algoritmo que implementamos usando el método de Li tardó 30 minutos como promedio, mientras que al usar MotifClick el tiempo de ejecución fue de aproximadamente 4-5 horas por genoma, excepto para la cepa 2513237200, que incrementó el uso de memoria hasta los 25 GB, y fue necesario un mayor poder computacional y un tiempo de ejecución de más de 24 horas.

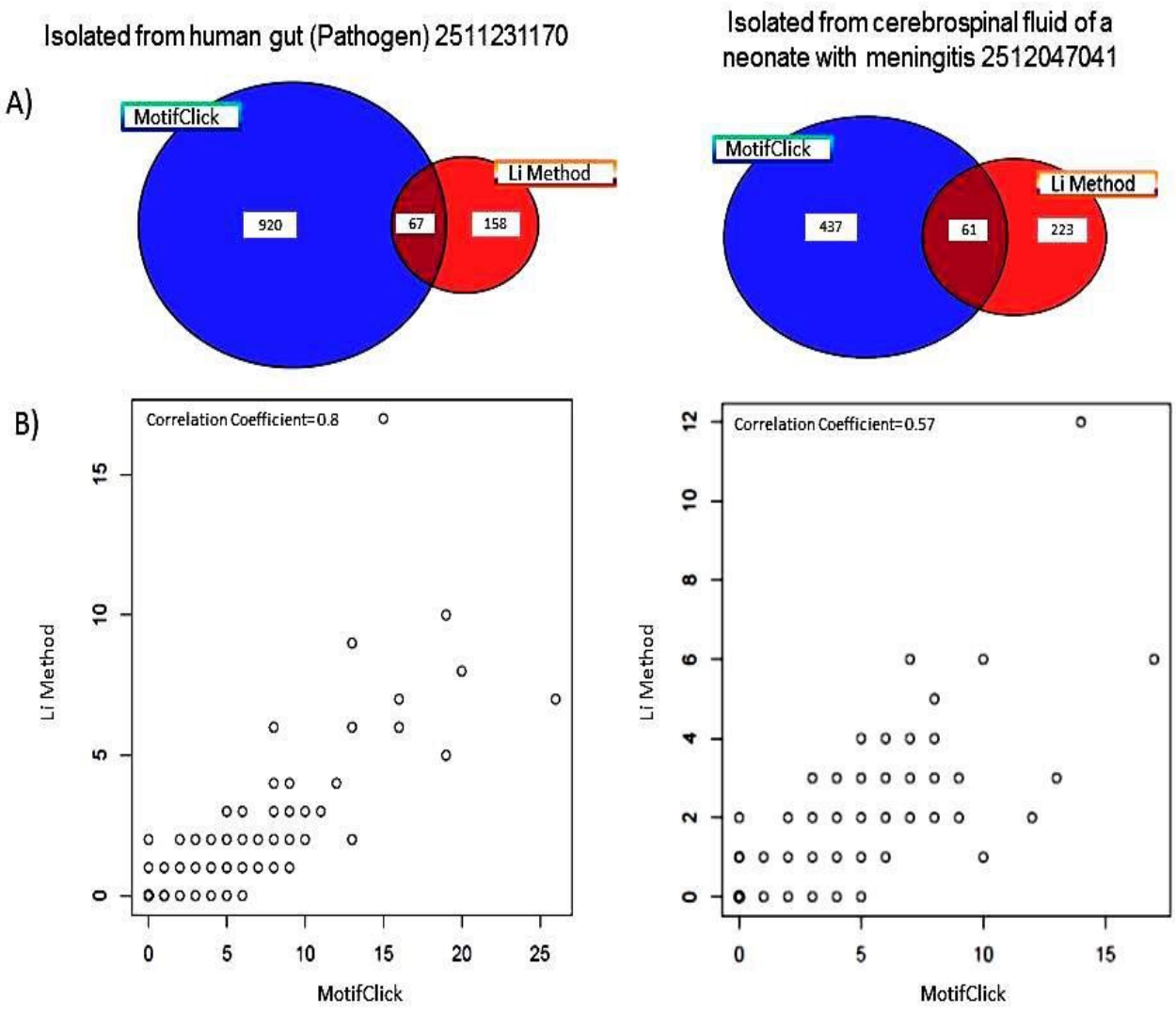


Figura 15. Predicción de sitios de unión por promotor a través de dos métodos *de novo* (MotifClick y el método de Li). A) Diagrama de Venn que ilustra los sitios que coinciden en cuanto a secuencia y ubicación en el promotor en ambas predicciones. B) Análisis de correlación de los sitios de unión predichos por promotor al usar estos dos algoritmos.

Una comparación de las funciones con más sitios de unión por promotor para cada cepa contra funciones pertenecientes a la misma vía metabólica en los otros genomas; reveló que el potencial regulador no seguía los mismos patrones. Para este análisis usamos las estimaciones por ambos métodos (MotifClick y el método de Li) teniendo solo en cuenta aquellas en que las predicciones eran coherentes y mantenían el mismo comportamiento. Por ejemplo, en la Figura 16 se observa las estimaciones del potencial regulador en tres

cepas, donde funciones involucradas en una misma vía metabólica (metabolismo del ácido colánico y metabolismo de las pirimidinas) mostraron una variación drástica de su potencial regulador.

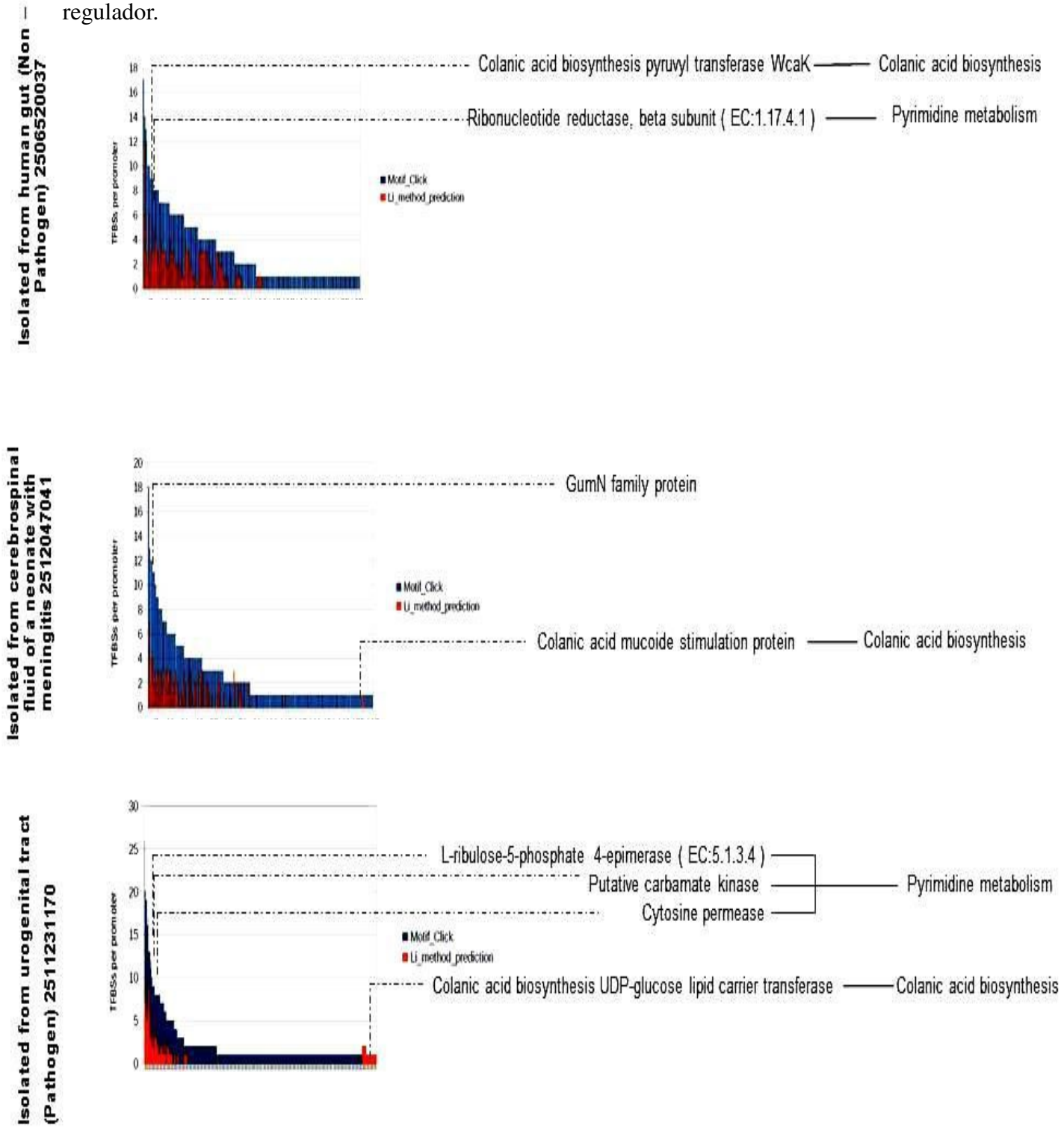


Figura 16. Comportamiento del potencial regulador en tres cepas de E. Coli.

11. Colaboraciones externas como parte del estudio sobre la regulación génica en Procariotas.

Al margen del eje central de esta tesis, también tuvimos la oportunidad de estudiar otros aspectos de la regulación bacteriana en especies de hábitats no asociados a hospederos, en el marco de una colaboración con el Dr. Stefan Bertilsson de la Universidad de Uppsala, Suecia. Este grupo tenía interés en poder identificar diferencias a nivel genómico y funcional entre bacterias de ambientes acuíferos marinos o de agua dulce. Con el objetivo de aplicar nuestro conocimiento y protocolos, y como contribución a este estudio, nos dispusimos a buscar posibles diferencias a nivel de regulación entre estos hábitats. Además del objetivo mencionado, también contribuimos con el análisis de los genes en estas bacterias.

Otra colaboración realizada durante esta tesis, en conjunto con el Dr. Peer Bork, EMBL, Alemania; consistió en la asignación de funciones a los genes secuenciados de metagenomas de tracto intestinal (Arumugam et al, 2011). Los datos obtenidos en este estudio también los analizaremos posteriormente para conocer como varía el potencial regulador en función de los diferentes enterotipos hallados por Arumugam y colaboradores.

11.1 Características de las regiones intergénicas y de los promotores en los clados de SAR11 Y LD12 comparados con *E. coli*.

La colaboración con el Dr. Stefan Bertillson comenzó con el análisis de las regiones intergénicas y promotores de los clados SAR11 y LD12 que poseen los genomas más pequeños conocidos hasta la fecha para bacterias de vida libre (Giovannoni et al, 2005).

Nuestro interés por el estudio de las regiones intergénicas de SAR11 y LD12 es debido a la compactación genómica necesaria para poder almacenar la información usando menos pares de bases si se comparan con otras Proteobacterias (Giovannoni et al, 2005). Por lo que constituyen un interesante punto de partida para el estudio de la regulación génica dentro del reino bacterias, particularmente en las de vida libre que no utilizan la maquinaria biológica de los hospederos para llevar a cabo la transcripción. En la Figura 17 se observa como varía la mediana de longitud de regiones intergénicas entre *E. coli* y varias bacterias pertenecientes a los clados de SAR11 y LD12. La mediana de *E. coli* de longitud de

regiones intergénicas sobrepasa de manera significativa (p-valor << 0.5, Kruskall Wallis test) la de LD12 y SAR11. Tres estrategias fundamentales podrían ayudar a que se alcance mayor compactación en un genoma bacteriano, 1) mediante la inclusión de más genes dentro de la unidad operónica, 2) sobrelapamiento de las regiones codificantes y 3) a través del uso de promotores bidireccionales (Adachi and Lieber, 2002;Korbel et al, 2004).

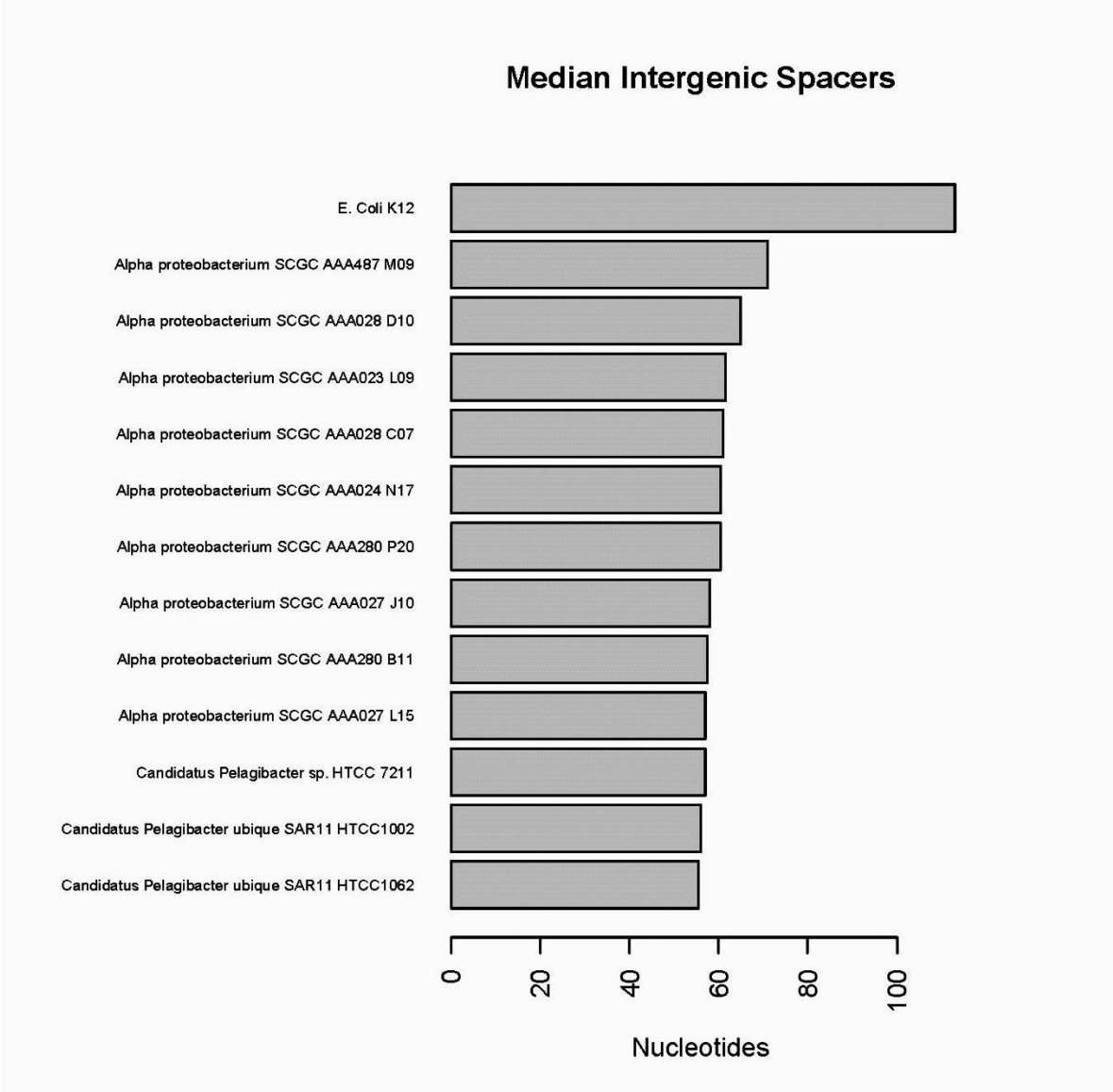


Figura 17. Mediana de la longitud de las regiones intergénicas en LD12 y SAR11 en comparación con *E. coli*. Para *E. coli* se usó el genoma de *Escherichia Coli K12-W3110.*

Las dos primeras estrategias han sido ampliamente estudiadas en SAR11 y LD12 (Carini et al, 2013;Giovannoni et al, 2005;Smith et al, 2010;Smith et al, 2013) pero sobre la presencia o abundancia de promotores bidireccionales en estos genomas tan compactos no hay ninguna información hasta el momento.

Los promotores bidireccionales o divergentes tienen el potencial para regular los dos genes que los flanquean debido a la presencia de dos extremos 5' en las proximidades y se ha visto alta conservación en las funciones reguladas por estos entre diferentes especies de bacterias y arqueas. Además las funciones adyacentes a promotores divergentes están relacionadas en muchas ocasiones con la misma vía metabólica (Korbel et al, 2004;Yang et al, 2013). Pero, indudablemente, este tipo de estructuras permiten también la regulación mediante el uso de menos factores de transcripción y su abundancia dentro del reino Bacteria ayuda a la compactación genómica en general.

En la Figura 18 se observa la estimación del número de promotores bidireccionales o divergentes (dos extremos 5' adyacentes al promotor), promotores con dos extremos 3' adyacentes y promotores con extremos 5' y 3' adyacentes. Para el caso de SAR11 (tres genomas, especies marinas) y LD12 (nueve genomas) se observa que el número de promotores divergentes se incrementa en estas especies con respecto a *E. coli*. Por el contrario, el número de regiones flanqueadas por dos extremos 3' se reduce el doble en el caso de los genomas más compactos de SAR11 y LD12.

En cuanto a los reguladores transcripcionales también se hizo una estimación del número de sus sitios de unión por promotor para SAR11 (tres genomas, especies marinas), LD12 (nueve genomas) y se comparó con la predicción de sitios de unión en *E. coli K12*. El perfil de densidad de sitios de unión por promotor en cada uno de los clados estudiados se observa en la Figura 19. Si usamos el programa MotifClick o el método descrito en la Sección 4.2 (Metodología), se observa la misma tendencia (*E.Coli*>>LD12>SAR11) tanto para el número máximo como para el promedio de sitios de unión por promotor predichos.

Debido a la poca información sobre los TFBSs de especies cercanas o de los propios clados de SAR11 y LD12; no se pudo hacer un mapaje a través de matrices de posición. Por esta razón, en la caracterización de los promotores, sólo se usaron los métodos *de novo* mencionados en el párrafo anterior.

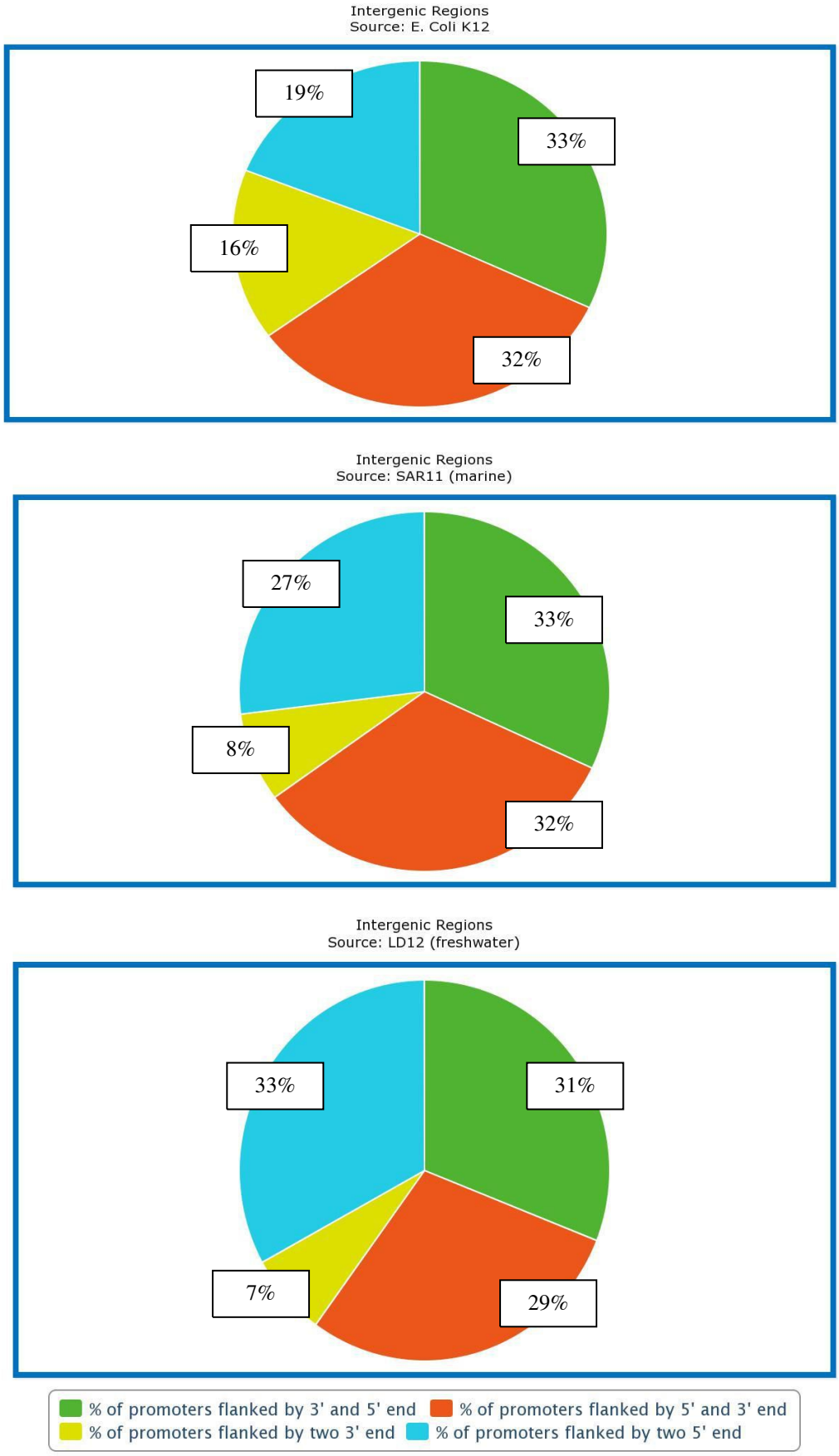


Figura 18. Dirección de los promotores.

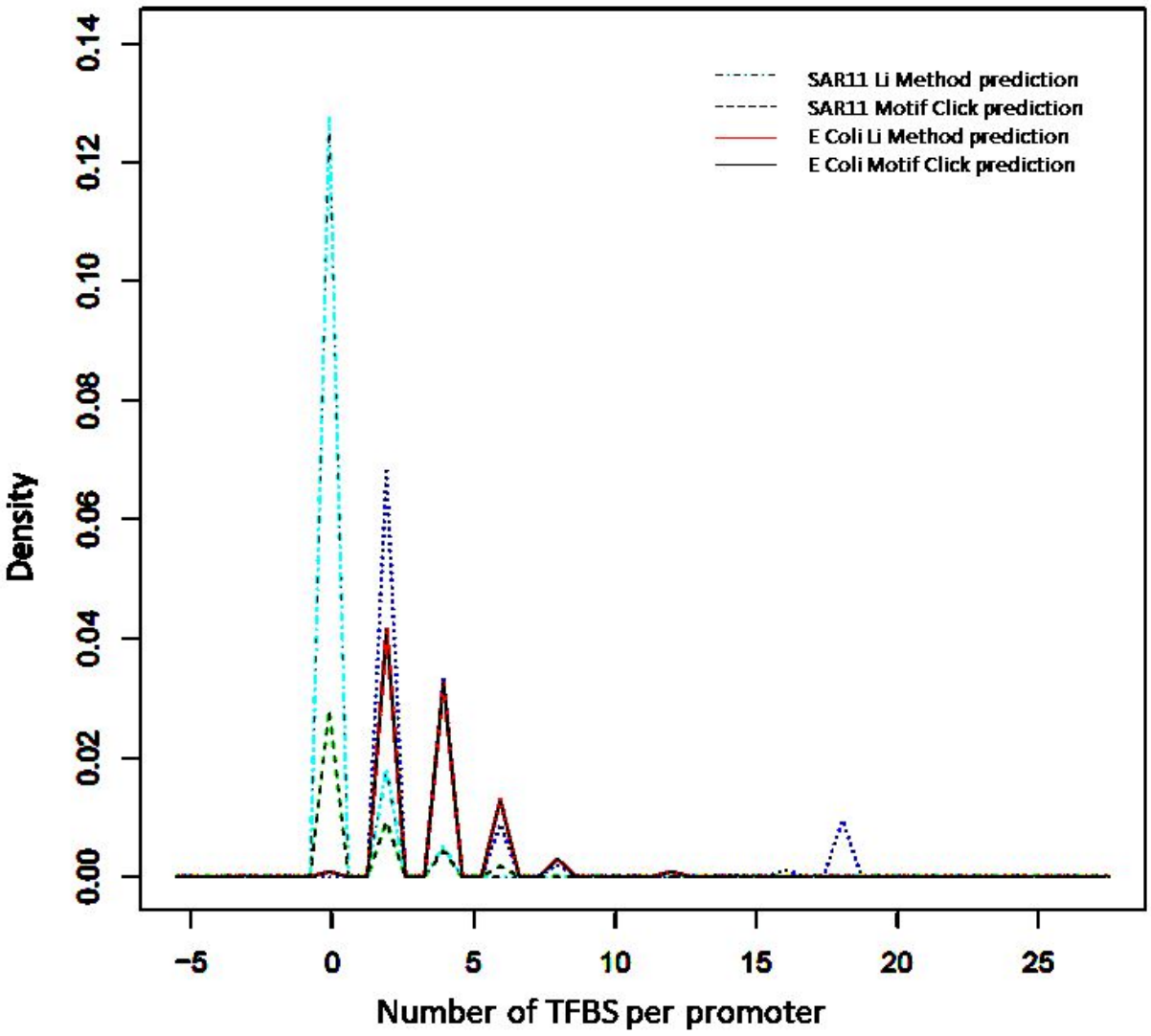


Figura 19. Predicción de sitios de unión en SAR11, LD12 y *E.Coli* según dos métodos *de novo:* 1) un algoritmo adaptado a procariotas usado por primera vez por Li y colaboradores y 2) el programa MotifClick descritos en la sección 4.2 y 6, respectivamente (capítulo de Metodología).

A pesar de la cercanía en términos taxonómicos de los clados LD12 y SAR11 (especies marinas) (Heinrich et al, 2013), la variación en los parámetros físico-químicos de los entornos que habitan no era tan drástica como para poder hacer un estudio de adaptación a condiciones ambientales fluctuantes. Por tanto, la mejor candidata para ese análisis fue, *E. coli;* debido a la amplia distribución espacial de sus cepas y su capacidad de adaptación a condiciones ambientales diversas. Además, por ser una de las especies de bacteria más estudiada desde el punto de vista de la regulación génica y que cuenta con información confiable (Collado-Vides et al, 1991;Espinosa et al, 2005;Gama-Castro et al, 2008;Mendoza-Vargas et al, 2009;van Helden et al, 2000).

V. DISCUSIÓN

1. Análisis comparativo de los resultados obtenidos en la identificación de promotores y TFBSs en metagenomas.

Para estudiar y caracterizar como los procariotas de vida libre modifican las regiones reguladoras de la expresión génica, para así adaptarse al medioambiente, diseñamos metodologías que permitieron tanto la identificación de promotores como de TFBSs. Dada la novedad que tienen las secuencias de comunidades de bacterias y arqueas obtenidas a través de estudios metagenómicos, los procedimientos usados dependieron lo menos posible de los datos sobre microrganismos conocidos; para así evitar desviar la información hacia sus genomas o similares.

El método propuesto en este trabajo para la identificación de promotores, aunque primero se basa en una búsqueda de homología; después emplea un segundo algoritmo (Prodigal) para eliminar aquellas regiones codificadoras que se puedan enmascarar como posibles promotores, debido al bajo porciento de identidad (ID%) con las proteínas reportadas en la base de datos del NCBI. A diferencia de otros métodos que se basan en la identificación de los sitios -10 y -35 o sitios de unión de factores sigma 70 (Dekhtyar et al, 2008), nuestro método no se afecta por el contenido de A+T. También permite la identificación de promotores de arqueas, que se conoce que no tienen la misma estructura de los sitios -10 y -35 de bacterias. Con la asignación taxonómica (Figura 2, 3 y 4 de Resultados) se comprobó

que de manera efectiva nuestro método es capaz de cubrir toda la diversidad de microorganismos encontrada mediante otros estudios previos (Lo et al, 2007;Tringe et al, 2005;Tyson et al, 2004). Además en los nichos analizados, se mantuvo estable la abundancia relativa, con sólo una excepción para el caso de la mina acidificada (Figura 4); donde la abundancia relativa de arqueas fue mayor que la encontrada por otros estudios (Lo et al, 2007;Tyson et al, 2004). Este hecho pudiera deberse a la escaza información, para el diseño de los iniciadores (primer en inglés) de arqueas, que se tenía hace diez años cuando fueron realizados los estudios de 16S ARNr en la mina acidificada (Gantner et al, 2011).

La predicción de sitios de unión por promotor en metagenomas se realizó a través de una metodología que no incluía algoritmos probabilísticos basados en matrices de posición específicas. Este hecho permitió que pudiésemos abordar la totalidad de la diversidad de las muestras estudiadas, sin desviar la predicción hacia las especies de las que se tienen más datos hasta el momento. Al utilizar 300 nucleótidos como longitud estándar de los promotores para realizar la búsqueda de TFBSs dentro de ese rango, se pudo explorar también los posibles sitios de unión de todos los microorganismos de vida libre presentes en la muestra. Por ejemplo, el organismo procariota de vida libre con el genoma más compacto conocido hasta la fecha, que pertenece a Pelagibacter Ubique (clado SAR 11), posee regiones intergénicas que alcanzan o sobrepasan los 300 nucleótidos. Una estimación de la longitud de las regiones intergénicas en otros microorganismos dentro de este clado se muestra en la segunda parte de Resultados, Figura 17. Estudios realizados en *E. coli* han demostrado que la mayoría de los reguladores reconocen sitios a una distancia entre 0 y 300 nucleótidos del TSS (Collado-Vides et al, 2009), además se ha usado esta misma longitud en otros estudios de predicción de TFBSs en bacterias (Li et al, 2002).

En cuanto al número de sitios de unión identificados por promotor (Figura 5), se observa que la media global de 10 sitios (con 0 como mínimo y 25 como máximo número de TFBSs encontrados), concuerda con otras estimaciones en procariotas, obtenidas mediante diferentes metodologías. Por ejemplo, usando un análisis genómico comparativo, un promedio de 11 a 13 motivos relacionados con TFBS fueron predichos en promotores de *Shewanella* (Liu et al, 2008). Otro estudio que corrobora los resultados obtenidos en este trabajo fue realizado en *E. coli*, donde hallaron un máximo de 16 sitios de unión posibles

por promotor (Sun et al, 2007) y hasta 20 mediante la identificación de motivos parciales (Leuze et al, 2012).

La comparación de los sitios de unión predichos en los tres nichos estudiados contra todos los sitios de unión conocidos para bacteria y arqueas, reportados en RegPrecise también tuvo resultados positivos. Aún más, las redes de factores de transcripción que participan en la regulación de funciones, tales como, metabolismo de la histidina y del potasio permitieron la identificación de TFs que juegan un posible papel en la regulación de estos procesos (las redes para promotores de suelo de granja se pueden ver en Anexos, Figura 1 y Figura 2).

Para el metabolismo de la histidina identificamos como sobrerrepresentado para esta función al dímero, CCG(10)CGG, sitio reconocido por PurR. Los promotores que contienen este dímero regulan los genes que codifican las siguientes enzimas: imidazol-glicerol fosfato deshidratasa, imidazol-glicerol fosfato sintasa e histidina-amonio liasa. Aunque PurR ha sido descrito como regulador del metabolismo de las purinas, pudiera estar también relacionado con estas funciones donde encontramos sus sitios de unión; por ejemplo, la enzima imidazol-glicerol fosfato sintasa, participa también de manera indirecta en la biosíntesis de las purinas a través de la generación del compuesto amino-imidazol-ribonucleótido-carboxamida. De manera similar la enzima imidazol-glicerol fosfato deshidratasa (EC 4.2.1.19), utiliza como sustrato un derivado de un intermediario común (5-[(5-fosfo-1-deoxi-β-D-ribulosa-1-ylamino)metilideneamino]-1-(5-fosfo-β-D-ribosil)imidazol-4-carboxamida) que participa en la biosíntesis de la histidina y en las etapas tardías de la biosíntesis de las purinas (Figura 1, Discusión).

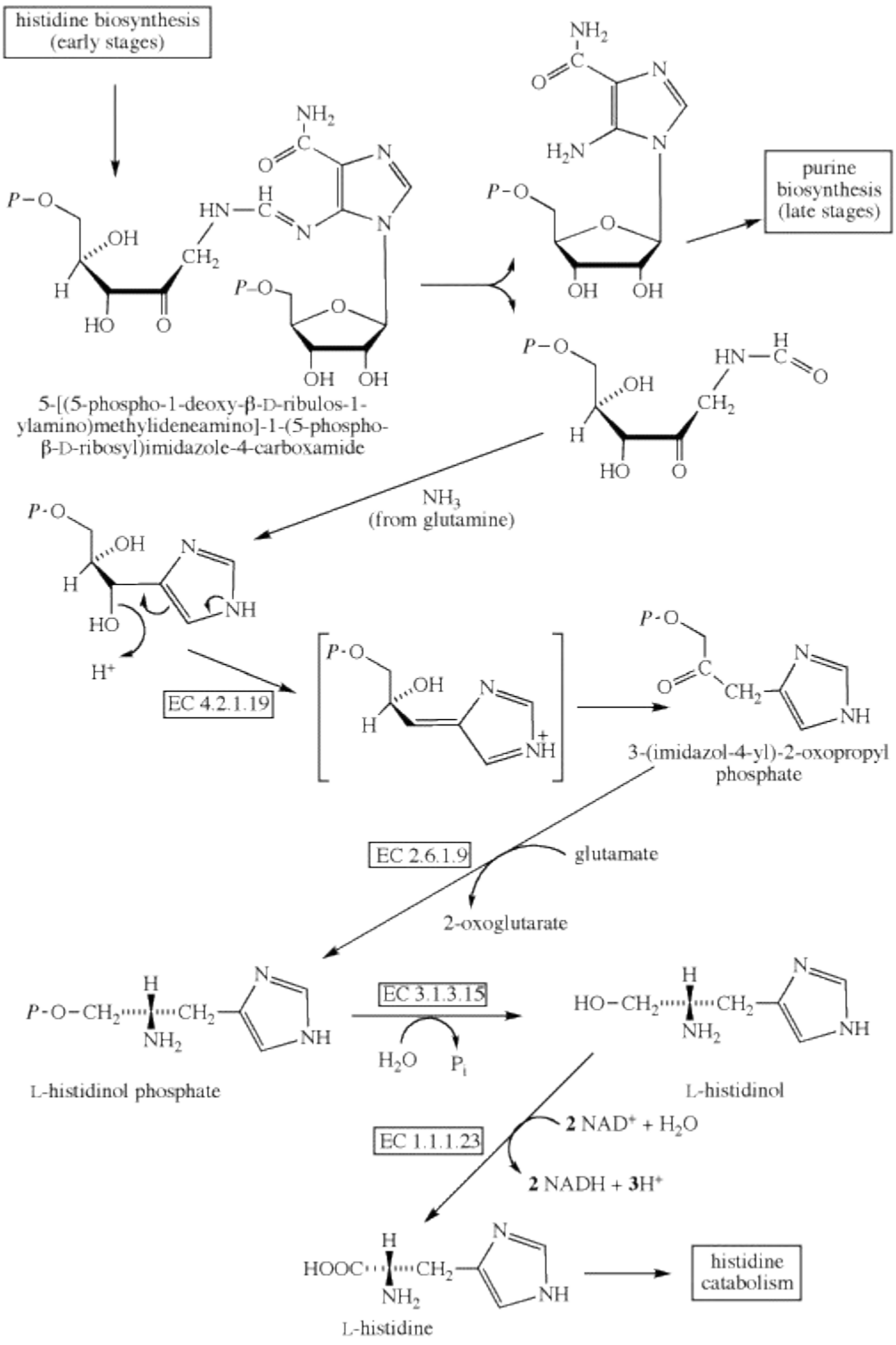


Figura 1. Etapas finales de la vía metabólica de biosíntesis de la histidina.

Cuando analizamos el metabolismo del potasio, no encontramos posibles evidencias sobre la acción directa de NikR (sitio de unión, TCA(21)TGA, Figura 2, Anexos) sobre el promotor de algún gen involucrado en este proceso; aunque sí existen estudios sobre el papel del ion K^+, en la estabilización de la unión de NikR al ADN (Phillips et al, 2010).

En la red de regulación de promotores para la mina acidificada, es más difícil encontrar una asociación entre el papel de los sistemas Ton y Tol y los reguladores encontrados debido a la función general que tienen estos como transportadores de varios compuestos (Tang et al, 2012).

En resumen, tanto la comparación con otras metodologías reportadas en la literatura, como los resultados obtenidos de la evaluación cuantitativa y cualitativa de los métodos usados en este estudio; sugieren alto grado de confiabilidad en su aplicación. Aunque una desventaja del método es que sólo cubre aquellos sitios con una estructura palíndroma, si se tiene en cuenta que nuestro principal objetivo no es predecir TFBSs, sino usarlos como estimadores del potencial regulador, esto no supone un gran problema. Además, se conoce que las estructuras palíndromas son características preponderantes de los TFBSs bacterianos y de arqueas (Gelfand et al, 2000;Huffman and Brennan, 2002;Laing et al, 2008;Li et al, 2002;McCue et al, 2001;Rodionov, 2007), por lo que probablemente, cubrimos una gran parte de los mismos. La presencia de falsos positivos en nuestras predicciones, como pueden ser ARN largos en la predicción de promotores o elementos CRISPRs en la predicción de TFBSs, fue cuidadosamente analizada por varios métodos para así garantizar su remoción de los resultados finales.

2. Posibles puntos de interacción entre el potencial regulador y el ambiente.

A partir del análisis de la Figura 7 se puede apreciar como las funciones con mayor potencial regulador tienden también a ser específicas. Por tanto, se explicará a continuación cuales son los procesos que para cada nicho presentaron un alto potencial regulador y que posibles parámetros físico-químicos del entorno pudieran estar influenciando este comportamiento.

2.1 Muestra de suelo de granja.

El hecho que en este entorno funciones relacionadas con el metabolismo de los di- y oligosacáridos aparezcan entre las más reguladas, pudiera estar en concordancia con las fluctuaciones en la concentración de materia orgánica en suelo, fundamentalmente restos de plantas. El factor anterior ha sido mencionado como el responsable de la variedad de ortólogos relacionados con la degradación de carbohidratos que se puede encontrar en esta muestra comparada con los otros nichos analizados (Tringe et al, 2005). Los resultados obtenidos aquí concuerdan además con otro estudio realizado en organismos eucariotas que habitan suelo, más específicamente en levaduras; donde se encontró alta complejidad en la regulación de los promotores de genes vinculados al metabolismo de los carbohidratos (Chin et al, 2005). En la Figura 12, se observa como un conocido regulador (FruR), vinculado a genes que intervienen en la síntesis y degradación de polisacáridos, mostró significativamente más sitios en los promotores de suelo de granja; lo que constituye otra prueba de la importancia que tiene este proceso biológico para los microrganismos que forman la comunidad.

2.2 Muestras de sedimentos marinos formados por restos de ballenas.

Un escenario distinto se observa en las muestras de restos de ballena, donde a pesar que fueron colectadas en un momento y sitio diferente: los nichos comparten características fisicoquímicas comunes, tales como, los drásticos cambios en la disponibilidad de nutrientes (Tringe et al, 2005). Como se esperaría para los microorganismos que crecen bajo las condiciones encontradas en estos nichos, muchas de las funciones identificadas con alto potencial regulador facilitan la adaptación a periodos de escasez de nutrientes (por ejemplo, en la Figura 11 (Resultados) un operón relacionado con la síntesis macromolecular que participa en el ciclo celular y crecimiento presenta alto potencial regulador en las muestras de restos de ballena en comparación con los otros dos nichos analizados).

Las comunidades de bacterias que habitan aguas templadas y frías están expuestas a concentraciones más altas de agentes oxidantes (Abele and Puntarulo, 2004), que causan un incremento asociado a la activación de las defensas contra la oxidación. En las muestras de

restos de ballenas que analizamos, las funciones relacionadas con el estrés oxidativo aparecieron con promotores específicamente enriquecidas en TFBSs. Por ejemplo, entre las proteínas que codifican los genes identificados con alto potencial regulador se encuentran: 1) NnrS: que responde a las concentraciones externas de NO y 2) DpS involucrada en la maquinaria que protege el ADN genómico durante las fases en que no hay crecimiento (Storz and Imlay, 1999).

La adquisición de cofactores y aminoácidos además de los genes involucrados en su metabolismo han sido señalados como particularmente variables en ambientes marinos; por lo que la adaptación a las fluctuaciones en la concentración de estos nutrientes es esencial para sobrevivir a las condiciones oligotróficas típicas de los entornos acuáticos (Gianoulis et al, 2009). De acuerdo con lo anterior, las funciones relacionadas con el metabolismo de los cofactores mostraron promotores enriquecidos en sitios de unión para reguladores, en particular, enzimas relacionadas con el metabolismo del molibdeno, las pterinas y el folato (Figura 11, Resultados). También se halló en este nicho abundantes sitios de unión para factores de transcripción conocidos por su papel en el control de la síntesis y transporte de aminoácidos (ArgR y TyrR , Figura 12, Resultados).

2.3 Muestras tomadas de una mina acidificada.

El siguiente entorno se caracteriza por condiciones extremas para la vida, como pH variable dentro de un rango muy ácido, además de valores de temperatura, conductividad y acumulados de lluvia fluctuantes (Figura 13 A, Resultados). Entre las funciones con un alto potencial regulador, que es posible que ayuden a los microrganismos a hacerle frente a esos factores ambientales, se encuentran aquellas relacionadas con brindar adaptación a cambios en la osmolaridad externa. Por ejemplo, los sistemas de transportadores Ton y Tol (Figura 13B), que están involucrados en evitar la toxicidad celular mediante el mantenimiento de la homeostasia frente a metales, en particular, hierro (Osorio et al, 2008). También se encontró para este entorno una sobrerrepresentación de sitios de unión para el factor de trascripción LexA (identificado además en los promotores del sistema de Ton y Tol, Figura 13B). LexA es un regulador involucrado en la protección del ADN cuando el pH es muy ácido y variable (Guazzaroni et al, 2013). Es interesante además la presencia de otros reguladores

en los promotores de los sistemas de Ton y Tol como, ModR y ModE, relacionados con el metabolismo de metales.

Los hechos explicados, ilustran como las funciones con alto potencial regulador son diferentes entre los nichos e indica que la organización de la regulación génica está influenciada por las condiciones medioambientales. Estos hallazgos reflejan la presencia de posibles puntos de interacción, donde la regulación génica es capaz de acondicionar la plasticidad de las redes funcionales para la adaptación a parámetros externos variables.

3. Comportamiento de la regulación génica entre diferentes cepas *E.Coli*.

Para hacer un análisis robusto del comportamiento de la regulación génica entre diferentes cepas de *E. coli*, escogimos aquellas que tuvieran un ensamblaje completo y buena calidad según IMG (http://img.jgi.doe.gov/). Además se buscó diferencias en la bio-distribución a través del uso de diferentes fuentes de aislamiento de la cepa. Otra característica que se tuvo en cuenta fue la diferencia en los grados de patogenicidad. Sin embargo, no hubo ninguna relación entre la patogenicidad y las características de los promotores analizadas, tales como, potencial regulador y distribuciones de sitios de unión de factores de transcripción por genomas.

Donde sí se encontró relación fue entre las características de los promotores (usando como descriptores los TFBSs "scores") y la fuente de aislamiento; probablemente por el efecto del entorno en la regulación génica. Las condiciones físico-químicas del ambiente es posible que afecten como los sitios de unión se distribuyen en los promotores y de esta manera garantizar los patrones de expresión adecuados. Es importante resaltar que es la primera vez que un estudio de este tipo es llevado a cabo en *E.Coli.*

En cuanto al análisis de las distribuciones de 86 factores de transcripción por cepa no vimos ningún patrón que apuntara hacia un papel del entorno en la abundancia relativa de sitios de unión. Uno de los hechos más remarcables en este sentido fue la presencia de un mayor número de sitos para el regulador FhlA en la cepa 2513237219, la cual es incapaz de sintetizar seleno-cisteína. Como la enzima FDH-H necesita de la seleno-cisteína para ser funcional, además que su expresión es activada por el factor FhlA (Schlensog et al, 1994); creemos que la sobrerrepresentación de sitios constituye una respuesta para mantener una

concentración siempre constante de FDH-H en la célula. Mantener un suplemento de FDH-H en las cepas analizadas es crucial debido a su crecimiento anaeróbico (Zinoni et al, 1986) por el tipo de entorno del que fueron aisladas; aún más en el caso de 2513237219 que depende del suministro externo de seleno-cisteína.

El estudio del potencial regulador se dificultó en las nueve cepas analizadas por las grandes diferencias existentes entre sus pan-genomas. En este sentido, encontrar ortólogos que permitieran comparar el potencial regulador no fue posible en muchos casos. Por tanto, a pesar que observamos diferencias entre cepas en cuanto a potencial regulador (en algunas funciones la diferencia es drástica, como se observa en la Figura 16); no podemos atribuir de manera segura si estas se deben al entorno o a la variabilidad entre familias génicas. Un hecho interesante que ilustra la Figura 16 es la presencia de la enzima ácido colánico piruvil transferasa (WcaK) entre las funciones más reguladas. Se cree que esta proteína pertenece a un operón involucrado en la síntesis de ácido colánico. Este último operón se ha reportado en estudios previos (Gottesman and Stout, 1991) que necesita de una compleja red de reguladores para modular su expresión, hecho que concuerda con nuestro resultado para la cepa 2506520037.

Un factor importante a tener en cuenta es el medio de cultivo. Las condiciones controladas y el cambio en las condiciones físico-químicas del aislamiento y el cultivo; quizás hayan influido en los patrones de regulación génica observados para estas cepas. Por lo que recomendamos el uso de la metagenómica para poder tener una visión más amplia del papel del ambiente en la estructura de los promotores.

Los análisis del potencial regulador y las distribuciones de sitios de unión discutidos aquí, podrían extenderse a más genomas; para así aumentar el conocimiento de la variedad intra-especifica en la regulación génica y la influencia del ambiente.

4. Relación entre las características de los promotores y el hábitat en Procariotas.

Según el análisis realizado en este trabajo, los clados de SAR11 y LD12, además de poseer los genomas más pequeños para bacterias de vida libre, poseen regiones intergénicas cortas con pocos sitios de unión para factores de transcripción si lo comparamos con *E. coli K12*. Dos posibles explicaciones tienen estos hechos, primero la mayor abundancia de sitios en *E. coli K12* pudiera estar asociada con el carácter cosmopolita de esta especie que se puede encontrar en disímiles hábitats: acuáticos, terrestres o asociados a huésped. Por otro lado, las especies de SAR11 y LD12, habitan sólo en ambientes marinos y de agua dulce, respectivamente, bajo condiciones más controladas. Por tanto, la adaptación a estos entornos acuáticos se espera que no requiera una complejidad de regulación como la que necesitaría *E. coli* para sobrevivir a disímiles nichos con características físico-químicas diferentes.

Otro hecho interesante es la presencia de abundantes promotores bidireccionales en los clados de SAR11 y LD12, que aunque constituye un avance para lograr una mayor compactación genómica; es más bien una desventaja en la respuesta adaptativa a través de la regulación génica. Los promotores bidireccionales regulan más funciones, pero con la limitación espacial que ocasiona la inclusión de menos sitios. Este hecho a su vez, pudiera afectar la capacidad adaptativa porque restringe la versatilidad en los patrones de expresión génica, debido al uso de menos factores de transcripción. A su vez el uso de menos reguladores hace que se restrinja el número de factores ambientales, así como, las fluctuaciones de los mismos, a los que la célula es capaz de responder de manera eficiente.

En resumen, una respuesta eficiente y versátil se lograría a través de redes metabólicas y de TF más complejas que conllevaría a promotores más enriquecidos en sitios de unión que respondan a la acción de más reguladores.

VI. CONCLUSIONES

Con la finalización de este trabajo hemos propuestos nuevos métodos para estudiar las regiones reguladoras en organismos procariotas (mediante la identificación de promotores y la predicción de sitios de unión de factores de transcripción). Además hemos ayudado a responder o a ampliar el conocimiento que se tenía sobre las siguientes cuestiones:

1) Perfiles de distribución de sitios de unión por promotor de acuerdo al hábitat en metagenomas.

Los perfiles de distribución de sitios de unión por hábitat resultaron muy semejantes para los tres nichos analizados: muestras de sedimentos marinos formados por restos de ballena, suelo de granja y una mina acidificada.

2) Comportamiento del potencial regulador (número de sitios de unión por promotor).

Al ir más allá de conteos simples observamos que muchos genes regulados por promotores con alto potencial regulador son específicos del ambiente.

3) Elucidación de las funciones y familias génicas que requieren una regulación más compleja y en qué nicho ecológico. ¿Qué papel juegan los factores físico-químicos del ambiente en la estructura del promotor?

En esta dirección se han encontrado funciones que sugieren puntos de interacción entre la regulación génica y los parámetros dinámicos o fluctuantes del ambiente, como: la disponibilidad de co-factores y aminoácidos en las muestras de restos de ballena, oligosacáridos en suelo y variaciones de pH en la mina acidificada

4) ¿Qué regulones son característicos de los nichos estudiados? ¿Cuáles son los sitios de unión más abundantes por nicho?

A través del método de identificación de TFBS, combinado con la creación de redes de regulación de promotores, es posible identificar nuevos regulones. Por ejemplo, aquí hallamos sitios de unión para LexA, ModE y ModR en funciones relacionadas a los sistemas de transportadores Ton y Tol.

5) Comportamiento de la distribución espacial de sitios de unión de acuerdo al hábitat donde fue aislada la cepa de *E. coli*.

Se comprobó la existencia de variaciones intra-específicas en la regulación génica a través del estudio de los TFBSs de diferentes cepas de *E.Coli* aisladas de disimiles fuentes.

6) Características de los promotores en los clados SAR11 y LD12.

A través de este estudio se comprobó la longitud de los promotores y sus características, en las bacterias de vida libre con los genomas más compactos conocidos hasta la fecha. Al comparar con otras bacterias los clados SAR11 y LD12 presentan mayor número de promotores divergentes (con la capacidad de regular dos genes al mismo tiempo). Además, los resultados obtenidos fueron de gran utilidad para el diseño posterior de una estrategia para la identificación de promotores en metagenomas.

En resumen, los resultados obtenidos en este trabajo destacan el impacto de la regulación génica en la adaptación de los microorganismos al hábitat. Más allá de la contribución al estudio general de como los procariotas interactúan con el ambiente; la metodología expuesta aquí puede ser usada para identificar los factores externos a los que las diferentes poblaciones son sensibles. Con este conocimiento sería posible diseñar comunidades *in silico* de manera eficiente para su uso en bio-remediación o como estrategia terapéutica.

Referencias

Abele D, Puntarulo S (2004). Formation of reactive species and induction of antioxidant defence systems in polar and temperate marine invertebrates and fish. *Comp Biochem Physiol A Mol Integr Physiol* **138:** 405-415.

Adachi N, Lieber MR (2002). Bidirectional gene organization: A common architectural feature of the human genome. *Cell* **109:** 807-809.

Aki T, Choy HE, Adhya S (1996). Histone-like protein HU as a specific transcriptional regulator: co-factor role in repression of gal transcription by GAL repressor. *Genes Cells* **1:** 179-188.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR *et al* (2011). Enterotypes of the human gut microbiome. *Nature* **473:** 174-180.

Barnard A, Wolfe A, Busby S (2004). Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. *Current Opinion in Microbiology* **7:** 102-108.

Bell SD, Jackson SP (1998). Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. *Trends Microbiol* **6:** 222-228.

Bell SD, Cairns SS, Robson RL, Jackson SP (1999). Transcriptional regulation of an archaeal operon in vivo and in vitro. *Mol Cell* **4:** 971-982.

Bell SD, Jackson SP (2000). Mechanism of autoregulation by an archaeal transcriptional repressor. *J Biol Chem* **275:** 31624-31629.

Bell SD, Jackson SP (2001). Mechanism and regulation of transcription in archaea. *Current Opinion in Microbiology* **4:** 208-213.

Bentley SD, Parkhill J (2004). Comparative genomic structure of prokaryotes. *Annual Review of Genetics* **38:** 771-792.

Besemer J, Lomsadze A, Borodovsky M (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29:** 2607-2618.

Bi C, Rogan PK (2006). BIPAD: a web server for modeling bipartite sequence elements. *BMC Bioinformatics* **7:** 76.

Blanco E, Guigo R, Messeguer X (2007). Multiple non-collinear TF-map alignments of promoter regions. *BMC Bioinformatics* **8:** 138.

Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC *et al* (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8:** 209.

Bourret RB, Charon NW, Stock AM, West AH (2002). Bright lights, abundant operons - Fluorescence and genomic technologies advance studies of bacterial locomotion and signal transduction: Review of the BLAST Meeting, Cuernavaca, Mexico, 14 to 19 January 2001. *J Bacteriol* **184:** 1-17.

Brady A, Salzberg SL (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* **6:** 673-676.

Brinkman AB, Bell SD, Lebbink RJ, de Vos WM, van der Oost J (2002). The Sulfolobus solfataricus Lrp-like protein LysM regulates lysine biosynthesis in response to lysine availability. *J Biol Chem* **277:** 29537-29549.

Browning DF, Busby SJ (2004). The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2:** 57-65.

Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP *et al* (2013). Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* **41:** D226-232.

Burke C, Kjelleberg S, Thomas T (2009). Selective extraction of bacterial DNA from the surfaces of macroalgae. *Appl Environ Microbiol* **75:** 252-256.

Cao Y, Tian B, Liu Y, Cai L, Wang H, Lu N *et al* (2013). Genome Sequencing of Ralstonia solanacearum FQY_4, Isolated from a Bacterial Wilt Nursery Used for Breeding Crop Resistance. *Genome Announc* **1**.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7:** 335-336.

Cardon LR, Stormo GD (1992). Expectation Maximization Algorithm for Identifying Protein-Binding Sites with Variable Lengths from Unaligned DNA Fragments. *J Mol Biol* **223:** 159-170.

Carini P, Steindler L, Beszteri S, Giovannoni SJ (2013). Nutrient requirements for growth of the extreme oligotroph 'Candidatus Pelagibacter ubique' HTCC1062 on a defined medium. *Isme Journal* **7:** 592-602.

Chakravarty A, Carlson JM, Khetani RS, DeZiel CE, Gross RH (2007). SPACER: identification of cis-regulatory elements with non-contiguous critical residues. *Bioinformatics* **23:** 1029-1031.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T *et al* (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* **14:** 1147-1159.

Chin CS, Chuang JH, Li H (2005). Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. *Genome Res* **15:** 205-213.

Collado-Vides J, Magasanik B, Gralla JD (1991). Control site location and transcriptional regulation in Escherichia coli. *Microbiol Rev* **55:** 371-394.

Collado-Vides J, Salgado H, Morett E, Gama-Castro S, Jimenez-Jacinto V, Martinez-Flores I *et al* (2009). Bioinformatics resources for the study of gene regulation in bacteria. *J Bacteriol* **191:** 23-31.

Cox RS, 3rd, Surette MG, Elowitz MB (2007). Programming gene expression with combinatorial promoters. *Mol Syst Biol* **3:** 145.

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004). WebLogo: a sequence logo generator. *Genome Res* **14:** 1188-1190.

da Silva Xavier G, Bellomo EA, McGinty JA, French PM, Rutter GA (2013). Animal models of GWAS-identified type 2 diabetes genes. *J Diabetes Res* **2013:** 906590.

Daniel R (2005). The metagenomics of soil. *Nat Rev Microbiol* **3:** 470-478.

Darwin CR (1909). The Origin of species. P.F. Collier & Son
New York.

Das MK, Dai HK (2007). A survey of DNA motif finding algorithms. *BMC Bioinformatics* **8 Suppl 7:** S21.

Das S, Paul S, Bag SK, Dutta C (2006). Analysis of Nanoarchaeum equitans genome and proteome composition: indications for hyperthermophilic and parasitic adaptation. *BMC Genomics* **7:** 186.

Dekhtyar M, Morin A, Sakanyan V (2008). Triad pattern algorithm for predicting strong promoter candidates in bacterial genomes. *BMC Bioinformatics* **9:** 233.

Delmont TO, Robe P, Clark I, Simonet P, Vogel TM (2011). Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J Microbiol Methods* **86:** 397-400.

den Hengst CD, van Hijum SA, Geurts JM, Nauta A, Kok J, Kuipers OP (2005). The Lactococcus lactis CodY regulon: identification of a conserved cis-regulatory element. *J Biol Chem* **280:** 34332-34342.

Dooley EE (2001). New mouse genomics consortium. *Environ Health Perspect* **109:** A421.

Dufva M (2009). Introduction to microarray technology. *Methods Mol Biol* **529:** 1-22.

Edgar RC (2007). PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8:** 18.

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460-2461.

Edwards KJ, Gihring TM, Banfield JF (1999). Seasonal variations in microbial populations and environmental conditions in an extreme acid mine drainage environment. *Appl Environ Microbiol* **65:** 3627-3632.

Espinosa V, Gonzalez AD, Vasconcelos AT, Huerta AM, Collado-Vides J (2005). Comparative studies of transcriptional regulation mechanisms in a group of eight gamma-proteobacterial genomes. *J Mol Biol* **354:** 184-199.

Fabret C, Feher VA, Hoch JA (1999). Two-component signal transduction in Bacillus subtilis: how one organism sees its world. *J Bacteriol* **181:** 1975-1983.

Farre D, Bellora N, Mularoni L, Messeguer X, Alba MM (2007). Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol* **8**.

Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL *et al* (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci U S A* **109:** 21390-21395.

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J *et al* (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260:** 500-507.

Fitch WM (1970). Distinguishing homologous from analogous proteins. *Syst Zool* **19:** 99-113.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW *et al* (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A* **105:** 3805-3810.

Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B *et al* (2008). RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* **36:** D120-124.

Gantner S, Andersson AF, Alonso-Saez L, Bertilsson S (2011). Novel primers for 16S rRNA-based archaeal community analyses in environmental samples. *J Microbiol Methods* **84:** 12-18.

Geiduschek EP, Ouhammouch M (2005). Archaeal transcription and its regulators. *Mol Microbiol* **56:** 1397-1407.

Gelfand MS, Koonin EV, Mironov AA (2000). Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res* **28:** 695-705.

Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I *et al* (2009). Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A* **106:** 1374-1379.

Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P *et al* (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One* **3:** e3042.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D *et al* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309:** 1242-1245.

Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc* **2010:** pdb prot5368.

Gottesman S, Stout V (1991). Regulation of capsular polysaccharide synthesis in Escherichia coli K12. *Mol Microbiol* **5:** 1599-1606.

Griffith KL, Shah IM, Myers TE, O'Neill MC, Wolf RE, Jr. (2002). Evidence for "pre-recruitment" as a new mechanism of transcription activation in Escherichia coli: the large excess of SoxS binding sites per cell relative to the number of SoxS molecules per cell. *Biochem Biophys Res Commun* **291:** 979-986.

Grissa I, Vergnaud G, Pourcel C (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35:** W52-57.

Guazzaroni ME, Morgante V, Mirete S, Gonzalez-Pastor JE (2013). Novel acid resistance genes from the metagenome of the Tinto River, an extremely acidic environment. *Environ Microbiol* **15:** 1088-1102.

GuhaThakurta D (2006). Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res* **34:** 3585-3598.

Handelsman J (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68:** 669-685.

He B, Zalkin H (1994). Regulation of Escherichia-Coli Pura by Purine Repressor, One-Component of a Dual Control Mechanism. *Journal of Bacteriology* **176:** 1009-1013.

Heinrich F, Eiler A, Bertilsson S (2013). Seasonality and environmental control of freshwater SAR11 (LD12) in a temperate lake (Lake Erken, Sweden). *Aquatic Microbial Ecology* **70:** 33-44.

Hemme CL, Deng Y, Gentry TJ, Fields MW, Wu L, Barua S *et al* (2010). Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *ISME J* **4:** 660-672.

Hopwood DA (2003). The Streptomyces genome--be prepared! *Nat Biotechnol* **21:** 505-506.

Huffman JL, Brennan RG (2002). Prokaryotic transcription regulators: more than just the helix-turn-helix motif. *Curr Opin Struct Biol* **12:** 98-106.

Hugenholtz P, Tyson GW (2008). Microbiology: metagenomics. *Nature* **455:** 481-483.

Huson DH, Auch AF, Qi J, Schuster SC (2007). MEGAN analysis of metagenomic data. *Genome Res* **17:** 377-386.

Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Research* **21:** 1552-1560.

Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28:** 2223-2230.

Iguchi A, Thomson NR, Ogura Y, Saunders D, Ooka T, Henderson IR *et al* (2009). Complete Genome Sequence and Comparative Genome Analysis of Enteropathogenic Escherichia coli O127:H6 Strain E2348/69. *Journal of Bacteriology* **191:** 347-354.

International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C *et al* (2010). International network of cancer genome projects. *Nature* **464:** 993-998.

International Chicken Genome Sequencing C (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695-716.

Iqbal M, Mast Y, Amin R, Hodgson DA, Consortium S, Wohlleben W *et al* (2012). Extracting regulator activity profiles by integration of de novo motifs and expression data: characterizing key regulators of nutrient depletion responses in Streptomyces coelicolor. *Nucleic Acids Res* **40:** 5227-5239.

Jacob F, Monod J (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3:** 318-356.

Jagtap P, McGowan T, Bandhakavi S, Tu ZJ, Seymour S, Griffin TJ *et al* (2012). Deep metaproteomic analysis of human salivary supernatant. *Proteomics* **12:** 992-1001.

Janion C (2008). Inducible SOS Response System of DNA Repair and Mutagenesis in Escherichia coli. *International Journal of Biological Sciences* **4:** 338-344.

Johansson L, Gafvelin G, Arner ES (2005). Selenocysteine in proteins-properties and biotechnological use. *Biochim Biophys Acta* **1726:** 1-13.

Johnson DS, Mortazavi A, Myers RM, Wold B (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497-1502.

Jothi R, Cuddapah S, Barski A, Cui K, Zhao K (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* **36:** 5221-5231.

Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, Sasamoto S *et al* (2002). Complete genomic sequence of nitrogen-fixing symbiotic bacterium Bradyrhizobium japonicum USDA110 (supplement). *DNA Res* **9:** 225-256.

Khaitovich P, Muetzel B, She X, Lachmann M, Hellmann I, Dietzsch J *et al* (2004a). Regional patterns of gene expression in human and chimpanzee brains. *Genome Res* **14:** 1462-1473.

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B *et al* (2004b). A neutral model of transcriptome evolution. *PLoS Biol* **2:** E132.

Kiupakis AK, Reitzer L (2002). ArgR-independent induction and ArgR-dependent superinduction of the astCADBE operon in Escherichia coli. *Journal of Bacteriology* **184:** 2940-2950.

Kobayashi M, Nagata K, Ishihama A (1990). Promoter selectivity of Escherichia coli RNA polymerase: effect of base substitutions in the promoter -35 region on promoter strength. *Nucleic Acids Res* **18:** 7367-7372.

Korbel JO, Jensen LJ, von Mering C, Bork P (2004). Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotechnol* **22:** 911-917.

Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F *et al* (2008). Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res* **36:** 2230-2239.

Kvist T, Ahring BK, Lasken RS, Westermann P (2007). Specific single-cell isolation and genomic amplification of uncultured microorganisms. *Applied Microbiology and Biotechnology* **74:** 926-935.

Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35:** 3100-3108.

Laing E, Sidhu K, Hubbard SJ (2008). Predicted transcription factor binding sites as predictors of operons in Escherichia coli and Streptomyces coelicolor. *BMC Genomics* **9**.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J *et al* (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lanzer M, Bujard H (1988). Promoters largely determine the efficiency of repressor action. *Proc Natl Acad Sci U S A* **85:** 8973-8977.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993). Detecting Subtle Sequence Signals - a Gibbs Sampling Strategy for Multiple Alignment. *Science* **262:** 208-214.

Lawrence JG, Ochman H (1998). Molecular archaeology of the Escherichia coli genome. *Proc Natl Acad Sci U S A* **95:** 9413-9417.

Leung HC, Yiu SM, Yang B, Peng Y, Wang Y, Liu Z *et al* (2011). A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics* **27:** 1489-1495.

Leuze MR, Karpinets TV, Syed MH, Beliaev AS, Uberbacher EC (2012). Binding Motifs in Bacterial Gene Promoters Modulate Transcriptional Effects of Global Regulators CRP and ArcA. *Gene Regul Syst Bio* **6:** 93-107.

Li H, Rhodius V, Gross C, Siggia ED (2002). Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc Natl Acad Sci U S A* **99:** 11772-11777.

Li LP (2009). GADEM: A Genetic Algorithm Guided Formation of Spaced Dyads Coupled with an EM Algorithm for Motif Discovery. *Journal of Computational Biology* **16:** 317-329.

Li R, Li Y, Kristiansen K, Wang J (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics* **24:** 713-714.

Lin Z, Wu WS, Liang H, Woo Y, Li WH (2010). The spatial distribution of cis regulatory elements in yeast promoters and its implications for transcriptional regulation. *BMC Genomics* **11:** 581.

Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* **12 Suppl 2:** S4.

Liu J, Xu X, Stormo GD (2008). The cis-regulatory map of Shewanella genomes. *Nucleic Acids Res* **36:** 5376-5390.

Lo I, Denef VJ, Verberkmoes NC, Shah MB, Goltsman D, DiBartolo G *et al* (2007). Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446:** 537-541.

Lowe TM, Eddy SR (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25:** 955-964.

MacIsaac KD, Fraenkel E (2006). Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput Biol* **2:** e36.

Madan Babu M, Teichmann SA (2003). Functional determinants of transcription factors in Escherichia coli: protein families and binding sites. *Trends Genet* **19:** 75-79.

Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M *et al* (2007). CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* **35:** D237-240.

Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D *et al* (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* **36:** D534-538.

Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Grechkin Y *et al* (2012). IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* **40:** D123-129.

Mavromatis K, Ivanova NN, Chen IM, Szeto E, Markowitz VM, Kyrpides NC (2009). The DOE-JGI Standard Operating Procedure for the Annotations of Microbial Genomes. *Stand Genomic Sci* **1:** 63-67.

Maxam AM, Gilbert W (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74:** 560-564.

McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V *et al* (2001). Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res* **29:** 774-782.

McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4:** 63-72.

Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B *et al* (2009). Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. *PLoS One* **4:** e7526.

Miller JR, Koren S, Sutton G (2010). Assembly algorithms for next-generation sequencing data. *Genomics* **95:** 315-327.

Noguchi H, Park J, Takagi T (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* **34:** 5623-5630.

Novichkov PS, Laikova ON, Novichkova ES, Gelfand MS, Arkin AP, Dubchak I *et al* (2010). RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Res* **38:** D111-118.

Osorio H, Martinez V, Nieto PA, Holmes DS, Quatrini R (2008). Microbial iron management mechanisms in extremely acidic environments: comparative genomics evidence for diversity and versatility. *BMC Microbiol* **8:** 203.

Ouhammouch M (2004). Transcriptional regulation in Archaea. *Curr Opin Genet Dev* **14:** 133-138.

Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J *et al* (2001a). Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. *Nature* **413:** 848-852.

Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB *et al* (2001b). Genome sequence of Yersinia pestis, the causative agent of plague. *Nature* **413:** 523-527.

Pevzner PA, Tang H, Waterman MS (2001). An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* **98:** 9748-9753.

Phillips CM, Stultz CM, Drennan CL (2010). Searching for the nik Operon: How a Ligand-Responsive Transcription Factor Hunts for Its DNA Binding Site. *Biochemistry* **49:** 7757-7763.

Pillai S, Chellappan SP (2009). ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications. *Methods Mol Biol* **523:** 341-366.

Pittard AJ, Davidson BE (1991). Tyrr Protein of Escherichia-Coli and Its Role as Repressor and Activator. *Molecular Microbiology* **5:** 1585-1592.

Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, Moran MA (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol* **11:** 1358-1375.

Prosser JI (2010). Replicate or lie. *Environ Microbiol* **12:** 1806-1810.

Pruteanu M, Baker TA (2009). Proteolysis in the SOS response and metal homeostasis in Escherichia coli. *Research in Microbiology* **160:** 677-683.

Ramseier TM, Bledig S, Michotey V, Feghali R, Saier MH, Jr. (1995). The global regulatory protein FruR modulates the direction of carbon flow in Escherichia coli. *Mol Microbiol* **16:** 1157-1169.

Reddy MV, Wang H, Liu S, Bode B, Reed JC, Steed RD *et al* (2011). Association between type 1 diabetes and GWAS SNPs in the southeast US Caucasian population. *Genes Immun* **12:** 208-212.

Rocha EPC (2008). The Organization of the Bacterial Genome. *Annual Review of Genetics* **42:** 211-233.

Rodionov DA (2007). Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem Rev* **107:** 3467-3497.

Rowlands T, Baumann P, Jackson SP (1994). The TATA-binding protein: a general transcription factor in eukaryotes and archaebacteria. *Science* **264:** 1326-1329.

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S *et al* (2012). A travel guide to Cytoscape plugins. *Nat Methods* **9:** 1069-1076.

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74:** 5463-5467.

Sawers G (1994). The hydrogenases and formate dehydrogenases of Escherichia coli. *Antonie Van Leeuwenhoek* **66:** 57-88.

Schlensog V, Lutz S, Bock A (1994). Purification and DNA-binding properties of FHLA, the transcriptional activator of the formate hydrogenlyase system from Escherichia coli. *J Biol Chem* **269:** 19590-19596.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75:** 7537-7541.

Sinha SC, Krahn J, Shin BS, Tomchick DR, Zalkin H, Smith JL (2003). The purine repressor of Bacillus subtilis: a novel combination of domains adapted for transcription regulation. *Journal of Bacteriology* **185:** 4087-4098.

Smith DP, Kitner JB, Norbeck AD, Clauss TR, Lipton MS, Schwalbach MS *et al* (2010). Transcriptional and Translational Regulatory Responses to Iron Limitation in the Globally Distributed Marine Bacterium Candidatus Pelagibacter Ubique. *Plos One* **5**.

Smith DP, Thrash JC, Nicora CD, Lipton MS, Burnum-Johnson KE, Carini P *et al* (2013). Proteomic and Transcriptomic Analyses of "Candidatus Pelagibacter ubique" Describe the First P-II-Independent Response to Nitrogen Limitation in a Free-Living Alphaproteobacterium. *Mbio* **4**.

Smits WK, Hoa TT, Hamoen LW, Kuipers OP, Dubnau D (2007). Antirepression as a second mechanism of transcriptional activation by a minor groove binding protein. *Mol Microbiol* **64:** 368-381.

Soppa J (1999). Transcription initiation in Archaea: facts, factors and future aspects. *Mol Microbiol* **31:** 1295-1305.

Storz G, Imlay JA (1999). Oxidative stress. *Curr Opin Microbiol* **2:** 188-194.

Sun J, Tuncay K, Haidar AA, Ensman L, Stanley F, Trelinski M *et al* (2007). Transcriptional regulatory network discovery via multiple method integration: application to E. coli K12. *Algorithms Mol Biol* **2:** 2.

Sun SL, Chen J, Li WZ, Altintas I, Lin A, Peltier S *et al* (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* **39:** D546-D551.

Tang K, Jiao N, Liu K, Zhang Y, Li S (2012). Distribution and functions of TonB-dependent transporters in marine bacteria and environments: implications for dissolved organic matter utilization. *PLoS One* **7:** e41204.

Thomas T, Gilbert J, Meyer F (2012). Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp* **2:** 3.

Thomm M (1996). Archaeal transcription factors and their role in transcription initiation. *FEMS Microbiol Rev* **18:** 159-171.

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E *et al* (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23:** 137-144.

Tossici-Bolt L, Dickson JC, Sera T, de Nijs R, Bagnara MC, Jonsson C *et al* (2011). Calibration of gamma camera systems for a multicentre European (1)(2)(3)I-FP-CIT SPECT normal database. *Eur J Nucl Med Mol Imaging* **38:** 1529-1540.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW *et al* (2005). Comparative metagenomics of microbial communities. *Science* **308:** 554-557.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37-43.

Uruburu F (2003). History and services of culture collections. *Int Microbiol* **6:** 101-103.

van Helden J, Andre B, Collado-Vides J (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* **281:** 827-842.

van Helden J, Rios AF, Collado-Vides J (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* **28:** 1808-1818.

van Hijum SA, Medema MH, Kuipers OP (2009). Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiol Mol Biol Rev* **73:** 481-509, Table of Contents.

Wang W, Zhang P, Liu X (2009). Short read DNA fragment anchoring algorithm. *BMC Bioinformatics* **10 Suppl 1:** S17.

Wei J, Goldberg MB, Burland V, Venkatesan MM, Deng W, Fournier G *et al* (2003). Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T. *Infect Immun* **71:** 2775-2786.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66-74.

Wijaya E, Rajaraman K, Yiu SM, Sung WK (2007). Detection of generic spaced motifs using submotif pattern mining. *Bioinformatics* **23:** 1476-1485.

Wilhelm BT, Landry JR (2009). RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48:** 249-257.

Willett P (1987). Multidimensional Clustering Algorithms - Murtagh,F. *Journal of Classification* **4:** 253-255.

Wilmes P, Bond PL (2006). Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol* **14:** 92-97.

Wosten MM (1998). Eubacterial sigma-factors. *FEMS Microbiol Rev* **22:** 127-150.

Yang S, Sleight SC, Sauro HM (2013). Rationally designed bidirectional promoter improves the evolutionary stability of synthetic genetic circuits. *Nucleic Acids Research* **41**.

Ye Y, Choi JH, Tang H (2011). RAPSearch: a fast protein similarity search tool for short reads. *BMC Bioinformatics* **12:** 159.

Zerbino DR, Birney E (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18:** 821-829.

Zhang S, Li S, Niu M, Pham PT, Su Z (2011). MotifClick: prediction of cis-regulatory binding sites via merging cliques. *BMC Bioinformatics* **12:** 238.

Zhou D, Yang R (2006). Global analysis of gene transcription regulation in prokaryotes. *Cell Mol Life Sci* **63:** 2260-2290.

Zinoni F, Birkmann A, Stadtman TC, Bock A (1986). Nucleotide sequence and expression of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from Escherichia coli. *Proc Natl Acad Sci U S A* **83:** 4650-4654.

Zomer AL, Buist G, Larsen R, Kok J, Kuipers OP (2007). Time-resolved determination of the CcpA regulon of Lactococcus lactis subsp. cremoris MG1363. *J Bacteriol* **189:** 1366-1381.

ANEXOS

Additional Material

Table 1. Number of transcription factor binding sites predicted per phylum in Acid Mine, Waseca Farm Soil and Whale Fall using the manually curated reconstructions of transcriptional regulons in RegPrecise Database.

AgaR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria	AgaR - Caulobacterales	4	1	0
Proteobacteria	AgaR - Pasteurellales	9	2	1
Proteobacteria	AgaR - Psychromonadaceae/Aeromonadales	6	11	0
Proteobacteria	AgaR - Xanthomonadales	4	2	0
Proteobacteria	AgaR2 - Enterobacteriales / Vibrionales	5	15	1
Proteobacteria	AgaR3 - Enterobacteriales / Vibrionales	6	7	0
Proteobacteria/gamma	AgaR - Enterobacteriales	12	15	3
Proteobacteria/gamma	AgaR - Shewanellaceae	16	10	0

AraR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Firmicutes	AraR - Bacillales	11	54	14
Firmicutes	AraR - Lactobacillales	16	19	5
Firmicutes	AraR1 - Clostridiaceae	20	12	5
Firmicutes	AraR2 - Clostridiaceae	20	11	4
Thermotogae	AraR - Thermotogales	11	35	8

ArgR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Firmicutes	ArgR - Bacillales	11	83	30
Firmicutes	ArgR - Staphylococcus	7	65	31
Proteobacteria/gamma	ArgR - Alteromonadales	9	61	34
Proteobacteria/gamma	ArgR - Enterobacteriales	12	259	113
Proteobacteria/gamma	ArgR - Pasteurellales	9	69	35
Proteobacteria/gamma	ArgR - Psychromonadaceae/Aeromonadales	6	70	34
Proteobacteria/gamma	ArgR - Shewanellaceae	16	502	177
Proteobacteria/gamma	ArgR - Vibrionales	10	107	43

BirA

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Firmicutes	BirA - Bacillales	11	36	1
Firmicutes	BirA - Staphylococcus	7	21	16
Proteobacteria	BirA - Pseudomonadaceae	8	15	2
Proteobacteria/Gamma	BirA - Alteromonadales	9	17	4
Proteobacteria/Gamma	BirA - Oceanospirillales/Alteromonadales	12	26	2
Proteobacteria/Gamma	BirA - Psychromonadaceae/Aeromonadales	6	12	0
Proteobacteria/Gamma	BirA - Various betaproteobacteria	12	13	0
Proteobacteria/Gamma	BirA - Vibrionales	10	20	10
Proteobacteria/Gamma	BirA - Xanthomonadales	4	5	0
Proteobacteria/delta	BirA - Desulfovibrionales	13	8	6
Proteobacteria/delta	BirA - Desulfuromonadales	9	15	1
Proteobacteria/gamma	BirA - Enterobacteriales	12	24	2
Proteobacteria/gamma	BirA - Shewanellaceae	16	32	22

ExuR/UxuR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria	UxuR - Oceanospirillales/Alteromonadales	12	4	0
Proteobacteria	UxuR - Pasteurellales	9	19	5
Proteobacteria/alpha	UxuR - Rhodobacterales	15	5	0
Proteobacteria/gamma	ExuR - Enterobacteriales	12	47	8
Proteobacteria/gamma	UxuR - Enterobacteriales	12	42	4
Proteobacteria/gamma	UxuR - Psychromonadaceae/Aeromonadales	6	17	3
Proteobacteria/gamma	UxuR - Vibrionales	10	28	10

FabR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/Gamma	FabR - Alteromonadales	9	45	0
Proteobacteria/Gamma	FabR - Oceanospirillales/Alteromonadales	12	8	0
Proteobacteria/Gamma	FabR - Psychromonadaceae/Aeromonadales	6	27	2
Proteobacteria/Gamma	FabR - Vibrionales	10	67	4
Proteobacteria/Gamma	FabR/DesT - Pseudomonadaceae	8	13	0

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/Gamma	FabR2 - Moraxellaceae	4	6	2
Proteobacteria/Gamma	FabR2 - Xanthomonadales	4	6	6
Proteobacteria/gamma	FabR - Enterobacteriales	12	40	1
Proteobacteria/gamma	FabR - Pasteurellales	9	19	1
Proteobacteria/gamma	FabR - Shewanellaceae	16	118	19

FadP

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/beta	FadP - Burkholderia	8	63	21
Proteobacteria/beta	FadP - Comamonadaceae	11	46	16
Proteobacteria/beta	FadP - Ralstonia	6	85	18

FadR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria	FadR - Pasteurellales	9	17	1
Proteobacteria	FadR - Vibrionales	10	59	33
Proteobacteria/gamma	FadR - Alteromonadales	9	34	24
Proteobacteria/gamma	FadR - Enterobacteriales	12	112	32
Proteobacteria/gamma	FadR - Psychromonadaceae/Aeromonadales	6	10	6
Proteobacteria/gamma	FadR - Shewanellaceae	16	78	49

FruR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/gamma	FruR - Enterobacteriales	12	213	40
Proteobacteria/gamma	FruR - Oceanospirillales/Alteromonadales	12	4	1
Proteobacteria/gamma	FruR - Pasteurellales	9	2	0
Proteobacteria/gamma	FruR - Pseudomonadaceae	8	16	4
Proteobacteria/gamma	FruR - Psychromonadaceae/Aeromonadales	6	22	2
Proteobacteria/gamma	FruR - Vibrionales	10	60	16

HexR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria	HexR - Oceanospirillales/Alteromonadales	12	40	11
Proteobacteria	HexR1 - Pseudomonadaceae	8	30	0
Proteobacteria/beta	HexR - Burkholderia	8	38	0
Proteobacteria/beta	HexR - Comamonadaceae	11	28	6
Proteobacteria/beta	HexR - Ralstonia	6	24	0
Proteobacteria/beta	HexR - Various betaproteobacteria	12	12	0
Proteobacteria/gamma	HexR - Alteromonadales	9	38	20
Proteobacteria/gamma	HexR - Enterobacteriales	12	45	14
Proteobacteria/gamma	HexR - Pseudomonadaceae	8	44	1
Proteobacteria/gamma	HexR - Psychromonadaceae/Aeromonadales	6	79	22
Proteobacteria/gamma	HexR - Shewanellaceae	16	316	151
Proteobacteria/gamma	HexR - Vibrionales	10	190	64
Proteobacteria/gamma	HexR2 - Oceanospirillales/Alteromonadales	12	12	6

Irr

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/alpha	Irr - Rhizobiales	15	122	59
Proteobacteria/alpha	Irr - Rhodobacterales	15	38	13
Proteobacteria/alpha	Irr - Rhodospirillales	9	13	7

IscR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria	IscR - Betaproteobacteria	15	29	11
Proteobacteria/alpha	IscR - Rhodobacterales	10	91	29
Proteobacteria/gamma	IscR - Alteromonadales	18	86	33
Proteobacteria/gamma	IscR - Enterobacteriales	5	22	13
Proteobacteria/gamma	IscR - Pasteurellales	6	24	7
Proteobacteria/gamma	IscR - Pseudomonadales	4	6	1
Proteobacteria/gamma	IscR - Vibrionales	5	24	13

KdgR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/gamma	KdgR - Enterobacteriales	12	155	36
Proteobacteria/gamma	KdgR - Pasteurellales	9	17	9
Proteobacteria/gamma	KdgR - Vibrionales	10	40	4
Thermotogae	KdgR - Thermotogales	11	26	8

LexA

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Actinobacteria	LexA - Corynebacteriaceae	8	100	4
Chloroflexi	LexA - Chloroflexi	5	52	5
Cyanobacteria	LexA - Cyanobacteria	14	25	2
Firmicutes	LexA - Bacillales	11	202	14
Firmicutes	LexA - Staphylococcus	7	85	11
Proteobacteria	LexA - Desulfuromonadales	9	11	2
Proteobacteria	LexA - Rhodobacterales	15	128	2
Proteobacteria/Alphaproteobacteria	LexA - Caulobacterales	4	47	0
Proteobacteria/Alphaproteobacteria	LexA - Rhizobiales	15	212	15
Proteobacteria/Alphaproteobacteria	LexA - Rhodospirillales	9	52	2
Proteobacteria/Alphaproteobacteria	LexA - Sphingomonadales	7	61	1
Proteobacteria/Beta	LexA - Burkholderia	8	75	50
Proteobacteria/Beta	LexA - Comamonadaceae	11	71	49
Proteobacteria/Beta	LexA - Various betaproteobacteria	12	12	6
Proteobacteria/Gamma	LexA - Alteromonadales	9	111	64
Proteobacteria/Gamma	LexA - Oceanospirillales/Alteromonadales	12	91	64
Proteobacteria/Gamma	LexA - Pasteurellales	9	96	54
Proteobacteria/Gamma	LexA - Psychromonadaceae/Aeromonadales	6	77	39
Proteobacteria/Gamma	LexA - Xanthomonadales	4	6	0
Proteobacteria/Gamma	LexA2 - Pseudomonadaceae	9	5	0
Proteobacteria/Gamma	LexA2 - Xanthomonadales	4	2	0
Proteobacteria/beta	LexA - Ralstonia	6	62	34
Proteobacteria/delta	LexA - Desulfovibrionales	10	4	2
Proteobacteria/gamma	LexA - Enterobacteriales	12	253	159
Proteobacteria/gamma	LexA - Pseudomonadaceae	8	92	49
Proteobacteria/gamma	LexA - Shewanellaceae	16	192	103
Proteobacteria/gamma	LexA - Vibrionales	10	184	92

LiuQ

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/beta	LiuQ - Burkholderia	8	32	0
Proteobacteria/beta	LiuQ - Comamonadaceae	11	8	0
oteobacteria/beta	LiuQ - Ralstonia	6	14	6

LiuR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria	LiuR - Sphingomonadales	7	65	4
Proteobacteria/alpha	LiuR - Caulobacterales	4	14	2
Proteobacteria/alpha	LiuR - Rhizobiales	15	84	7
Proteobacteria/alpha	LiuR - Rhodobacterales	15	19	3
Proteobacteria/alpha	LiuR - Rhodospirillales	9	17	1
Proteobacteria/beta	LiuR - Burkholderia	8	6	1
Proteobacteria/beta	LiuR - Comamonadaceae	11	59	0
Proteobacteria/beta	LiuR - Ralstonia	6	57	3
Proteobacteria/beta	LiuR - Various betaproteobacteria	12	31	8
Proteobacteria/gamma	LiuR - Alteromonadales	9	68	15
Proteobacteria/gamma	LiuR - Oceanospirillales/Alteromonadales	12	28	2
Proteobacteria/gamma	LiuR - Pseudomonadaceae	8	32	9
Proteobacteria/gamma	LiuR - Psychromonadaceae/Aeromonadales	6	20	4
Proteobacteria/gamma	LiuR - Shewanellaceae	16	134	11
Proteobacteria/gamma	LiuR - Vibrionales	10	38	12

ModE

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/Alpha	ModE - Caulobacterales	4	5	1
Proteobacteria/Alpha	ModE - Rhizobiales	15	11	6
Proteobacteria/Beta	ModE - Burkholderia	8	8	8
Proteobacteria/Beta	ModE - Comamonadaceae	11	10	3
Proteobacteria/Delta	ModE - Desulfuromonadales	9	10	9
Proteobacteria/Gamma	ModE - Moraxellaceae	4	1	0
Proteobacteria/Gamma	ModE - Oceanospirillales/Alteromonadales	12	2	2
Proteobacteria/alpha	ModE - Rhodobacterales	15	5	2
Proteobacteria/alpha	ModE - Rhodospirillales	9	2	2
Proteobacteria/alpha	ModE - Sphingomonadales	7	2	2
Proteobacteria/beta	ModE - Ralstonia	6	8	5
Proteobacteria/beta	ModE - Various betaproteobacteria	12	5	3
Proteobacteria/delta	ModR - Desulfovibrionales	13	25	15
Proteobacteria/gamma	ModE - Enterobacteriales	12	55	47
Proteobacteria/gamma	ModE - Pasteurellales	9	21	20
Proteobacteria/gamma	ModE - Pseudomonadaceae	8	9	9
Proteobacteria/gamma	ModE - Psychromonadaceae/Aeromonadales	6	3	3
Proteobacteria/gamma	ModE - Shewanellaceae	16	10	10
Proteobacteria/gamma	ModE - Vibrionales	10	2	2

NadQ

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/Alpha	NadQ - Caulobacterales	4	16	9
Proteobacteria/Alpha	NadQ - Rhizobiales	15	22	7
Proteobacteria/Alpha	NadQ - Rhodobacterales	15	5	2
Proteobacteria/Alpha	NadQ - Rhodospirillales	9	10	4
Proteobacteria/Beta	NadQ - Comamonadaceae	11	6	4
Proteobacteria/Beta	NadQ - Various betaproteobacteria	12	2	2
Proteobacteria/Gamma	NadQ - Moraxellaceae	4	6	5

NagC

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/gamma	NagC - Enterobacteriales	12	68	42
Proteobacteria/gamma	NagC - Oceanospirillales/Alteromonadales	12	5	4
Proteobacteria/gamma	NagC - Pasteurellales	9	16	7
Proteobacteria/gamma	NagC - Psychromonadaceae/Aeromonadales	6	16	9
Proteobacteria/gamma	NagC - Vibrionales	10	199	111

NagQ

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/alpha	NagQ - Caulobacterales	4	13	0
Proteobacteria/alpha	NagQ - Rhizobiales	15	16	0
Proteobacteria/alpha	NagQ - Rhodobacterales	15	8	0
Proteobacteria/alpha	NagQ - Rhodospirillales	9	6	0
Proteobacteria/beta	NagQ - Burkholderia	8	7	0
Proteobacteria/beta	NagQ - Ralstonia	6	2	0
Proteobacteria/beta	NagQ - Various betaproteobacteria	12	12	4
Proteobacteria/gamma	NagQ - Oceanospirillales/Alteromonadales	12	11	2
Proteobacteria/gamma	NagQ - Pseudomonadaceae	8	4	0
Proteobacteria/gamma	NagQ - Xanthomonadales	4	2	0

NagR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/gamma	NagR - Alteromonadales	9	27	4
Proteobacteria/gamma	NagR - Oceanospirillales/Alteromonadales	12	19	1
Proteobacteria/gamma	NagR - Shewanellaceae	16	105	0
Proteobacteria/gamma	NagR - Xanthomonadales	4	17	0

NiaR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Firmicutes	NiaR - Bacillales	11	26	15
Firmicutes	NiaR - Clostridiaceae	20	27	9
Firmicutes	NiaR - Lactobacillaceae	15	7	0
Firmicutes	NiaR - Streptococcaceae	15	23	8
Thermotogae	NiaR - Thermotogales	11	13	0

NikR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/alpha	NikR - Rhizobiales	15	11	5
Proteobacteria/alpha	NikR - Rhodobacterales	15	2	1
Proteobacteria/alpha	NikR - Rhodospirillales	9	2	0
Proteobacteria/beta	NikR - Burkholderia	8	2	0
Proteobacteria/beta	NikR - Comamonadaceae	11	3	1
Proteobacteria/beta	NikR - Various betaproteobacteria	12	2	0
Proteobacteria/delta	NikR - Desulfovibrionales	13	19	7
Proteobacteria/delta	NikR - Desulfuromonadales	9	7	0
Proteobacteria/gamma	NikR - Enterobacteriales	12	9	6
Proteobacteria/gamma	NikR - Pseudomonadaceae	8	2	0
Proteobacteria/gamma	NikR - Shewanellaceae	16	8	8

NorR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/beta	NorR - Burkholderia	8	5	3
Proteobacteria/beta	NorR - Comamonadaceae	11	25	6
Proteobacteria/beta	NorR - Ralstonia	6	35	8
Proteobacteria/gamma	NorR - Alteromonadales	9	18	4
Proteobacteria/gamma	NorR - Enterobacteriales	12	33	28
Proteobacteria/gamma	NorR - Oceanospirillales/Alteromonadales	12	13	5
Proteobacteria/gamma	NorR - Pseudomonadaceae	8	40	6
Proteobacteria/gamma	NorR - Psychromonadaceae/Aeromonadales	6	21	4
Proteobacteria/gamma	NorR - Shewanellaceae	16	91	21
Proteobacteria/gamma	NorR - Vibrionales	10	40	15
Proteobacteria/gamma	NorR2 - Vibrionales	10	19	3

NrdR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Actinobacteria	NrdR - Actinobacteria	15	57	10
Chlamydiae	NrdR - Chlamydia	5	10	2
Cyanobacteria	NrdR - Cyanobacteria	5	9	1
Deinococcus-Thermus	NrdR - Deinococcus-Thermus	2	5	3
Firmicutes	NrdR - Firmicutes	33	167	56
Proteobacteria	NrdR - Alphaproteobacteria	8	27	5
Proteobacteria	NrdR - Betaproteobacteria	9	20	5
Proteobacteria	NrdR - Deltaproteobacteria	6	18	6
Proteobacteria	NrdR - Gammaproteobacteria	35	151	28
Thermotogae	NrdR - Thermotogales	11	60	29
unclassified	NrdR - Mixture	5	11	3

NrtR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Actinobacteria	NrtR - Actinobacteria-1	6	9	5
Actinobacteria	NrtR - Actinobacteria-2	5	5	1
Actinobacteria	NrtR - Actinobacteria-3	3	5	1
Actinobacteria	NrtR - Actinobacteria-4	5	11	4
Bacteroidetes	NrtR - Bacteroidetes-1	2	2	0
Bacteroidetes	NrtR - Bacteroidetes-2	4	15	5
Bacteroidetes	NrtR - Bacteroidetes-3	3	9	3
Bacteroidetes	NrtR - Cytophaga	2	4	4
Chloroflexi	NrtR - Chloroflexi	1	4	1
Cyanobacteria	NrtR - Cyanobacteria	7	14	4
Firmicutes	NrtR - Firmicutes-1	4	6	1
Firmicutes	NrtR - Firmicutes-2	3	3	0
Firmicutes	NrtR - Firmicutes-3	1	3	2
Firmicutes	NrtR - Firmicutes-4	1	2	2
Planctomycetes	NrtR - Pirellula	1	2	0
Proteobacteria	NrtR - Gammaproteobacteria-1	6	14	4
Proteobacteria	NrtR - Gammaproteobacteria-2	3	5	1
Proteobacteria	NrtR - Gammaproteobacteria-3	3	4	3

NsrR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Firmicutes	NsrR - Bacillales	11	30	17
Proteobacteria/alpha	NsrR - Caulobacterales	4	3	2
Proteobacteria/alpha	NsrR - Rhizobiales	15	2	0
Proteobacteria/alpha	NsrR - Rhodobacterales	15	17	4
Proteobacteria/alpha	NsrR - Rhodospirillales	9	13	5
Proteobacteria/alpha	NsrR - Sphingomonadales	7	5	4
Proteobacteria/beta	NsrR - Burkholderia	8	23	12
Proteobacteria/beta	NsrR - Comamonadaceae	11	15	10
Proteobacteria/beta	NsrR - Neisseriales	12	5	4
Proteobacteria/beta	NsrR - Ralstonia	6	23	18
Proteobacteria/beta	NsrR - Various betaproteobacteria	12	16	9
Proteobacteria/gamma	NsrR - Alteromonadales	9	2	1
Proteobacteria/gamma	NsrR - Enterobacteriales	12	73	39
Proteobacteria/gamma	NsrR - Moraxellaceae	4	7	0
Proteobacteria/gamma	NsrR - Oceanospirillales/Alteromonadales	12	8	6
Proteobacteria/gamma	NsrR - Psychromonadaceae/Aeromonadales	6	5	2
Proteobacteria/gamma	NsrR - Shewanellaceae	16	51	22
Proteobacteria/gamma	NsrR - Vibrionales	10	40	14

PdhR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/gamma	PdhR - Alteromonadales	9	15	6
Proteobacteria/gamma	PdhR - Enterobacteriales	12	36	25
Proteobacteria/gamma	PdhR - Oceanospirillales/Alteromonadales	12	22	12
Proteobacteria/gamma	PdhR - Psychromonadaceae/Aeromonadales	6	8	6
Proteobacteria/gamma	PdhR - Shewanellaceae	16	109	37
Proteobacteria/gamma	PdhR - Vibrionales	10	16	12

PsrA

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria	PsrA - Xanthomonadales	4	7	5
Proteobacteria/alpha	PsrA - Caulobacterales	4	70	28
Proteobacteria/alpha	PsrA - Rhizobiales	15	12	6
Proteobacteria/beta	PsrA - Burkholderia	8	16	9
Proteobacteria/beta	PsrA - Ralstonia	6	9	4
Proteobacteria/beta	PsrA - Various betaproteobacteria	12	34	8
Proteobacteria/gamma	PsrA - Alteromonadales	9	26	4
Proteobacteria/gamma	PsrA - Oceanospirillales/Alteromonadales	12	67	21
Proteobacteria/gamma	PsrA - Pseudomonadaceae	8	69	18
Proteobacteria/gamma	PsrA - Psychromonadaceae/Aeromonadales	6	19	4
Proteobacteria/gamma	PsrA - Shewanellaceae	16	295	105
Proteobacteria/gamma	PsrA - Vibrionales	10	49	15

PurR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/gamma	PurR - Enterobacteriales	12	298	71
Proteobacteria/gamma	PurR - Pasteurellales	9	77	14
Proteobacteria/gamma	PurR - Vibrionales	10	149	29

RbsR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Firmicutes	RbsR - Bacillales	11	12	11
Firmicutes	RbsR - Staphylococcus	7	2	2
Firmicutes	RbsR - Streptococcus	8	4	1
Firmicutes	RbsR2 - Staphylococcus	7	4	4
Proteobacteria/beta	RbsR - Burkholderia	8	7	0
Proteobacteria/beta	RbsR - Comamonadaceae	11	2	0
Proteobacteria/beta	RbsR - Desulfovibrionales	10	1	0
Proteobacteria/beta	RbsR - Ralstonia	6	2	0
Proteobacteria/gamma	RbsR - Enterobacteriales	12	17	1
Proteobacteria/gamma	RbsR - Pasteurellales	9	6	0
Proteobacteria/gamma	RbsR - Pseudomonadaceae	8	5	0
Proteobacteria/gamma	RbsR - Psychromonadaceae/Aeromonadales	6	5	0
Proteobacteria/gamma	RbsR - Shewanellaceae	16	4	0
Proteobacteria/gamma	RbsR - Vibrionales	10	15	0
Thermotogae	RbsR - Thermotogales	11	15	3

Rex

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Actinobacteria	Rex - Actinomycetales	18	88	4
Chloroflexi	Rex - Chloroflexi	5	39	5
Deinococcus-Thermus	Rex - Deinococcus-Thermus	5	10	0
Firmicutes	Rex - Bacillales	11	62	26
Firmicutes	Rex - Clostridiaceae	20	191	122
Firmicutes	Rex - Lactobacillaceae	15	70	24
Firmicutes	Rex - Staphylococcus	7	74	18
Firmicutes	Rex - Streptococcaceae	15	160	52
Firmicutes	Rex - Thermoanaerobacterales	3	42	29
Proteobacteria/delta	Rex - Desulfovibrionales	10	100	37
Thermotogae	Rex - Thermotogales	11	114	41
Thermotogae	Rex1 - Thermotogales	11	10	2

RutR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/Alpha	RutR - Caulobacterales	4	4	0
Proteobacteria/Alpha	RutR - Rhizobiales	15	44	12
Proteobacteria/Alpha	RutR - Rhodobacterales	15	62	19
Proteobacteria/Alpha	RutR - Rhodospirillales	9	2	0
Proteobacteria/Beta	RutR - Burkholderia	8	12	2
Proteobacteria/Beta	RutR - Ralstonia	6	4	0
Proteobacteria/Gamma	RutR - Alteromonadales	9	19	5
Proteobacteria/Gamma	RutR - Moraxellaceae	4	8	4
Proteobacteria/Gamma	RutR - Oceanospirillales/Alteromonadales	12	32	6
Proteobacteria/Gamma	RutR - Pseudomonadales	8	29	13
Proteobacteria/Gamma	RutR2 - Pseudomonadales	8	7	3
Proteobacteria/Gamma	RutR3 - Pseudomonadales	8	23	13
Proteobacteria/Gamma	RutR3 - Psychromonadaceae/Aeromonadales	6	4	0
Proteobacteria/Gamma	RutR3 - Vibrionales	10	4	2
Proteobacteria/Gamma	RutR4 - Alteromonadales	9	2	2
Proteobacteria/Gamma	RutR4 - Pseudomonadales	9	4	0
Proteobacteria/gamma	RutR - Enterobacteriales	12	13	0

TrpR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/gamma	TrpR - Enterobacteriales	12	44	4
Proteobacteria/gamma	TrpR - Moraxellaceae	4	3	3
Proteobacteria/gamma	TrpR - Oceanospirillales/Alteromonadales	12	3	1
Proteobacteria/gamma	TrpR - Pasteurellales	9	24	8
Proteobacteria/gamma	TrpR - Psychromonadaceae/Aeromonadales	6	2	0
Proteobacteria/gamma	TrpR - Shewanellaceae	16	30	8
Proteobacteria/gamma	TrpR - Vibrionales	10	25	5
Proteobacteria/gamma	TrpR - Xanthomonadales	4	3	2

TyrR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Proteobacteria/gamma	PhhR - Pseudomonadales	8	32	19
Proteobacteria/gamma	TyrR - Alteromonadales	9	42	20
Proteobacteria/gamma	TyrR - Enterobacteriales	12	107	57
Proteobacteria/gamma	TyrR - Pasteurellales	9	29	13
Proteobacteria/gamma	TyrR - Psychromonadaceae/Aeromonadales	6	17	11
Proteobacteria/gamma	TyrR - Shewanellaceae	16	309	135
Proteobacteria/gamma	TyrR - Vibrionales	12	67	43

XylR

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Firmicutes	XylR - Bacillales	11	41	32
Firmicutes	XylR - Enterococcaceae	2	4	2
Firmicutes	XylR - Lactobacillaceae	15	15	9
Firmicutes	XylR1 - Clostridiaceae	20	22	16
Firmicutes	XylR2 - Clostridiaceae	20	9	5
Proteobacteria/alpha	XylR - Rhizobiales	15	19	18

Zur

Phylum	Regulog	Genomes	Sites RegPrecise	Sites Metagenomic TFBS Searching
Actinobacteria	zur - Actinobacteria	9	23	22
Cyanobacteria	zur - Cyanobacteria	9	15	13
Firmicutes	zur - Bacilli	14	69	43
Firmicutes	zur - Clostridia	6	12	7
Proteobacteria	zur - Alphaproteobacteria-1	8	19	12
Proteobacteria	zur - Alphaproteobacteria-2	4	4	2
Proteobacteria	zur - Alphaproteobacteria-3	12	19	17
Proteobacteria	zur - Alphaproteobacteria-4	8	17	9
Proteobacteria	zur - Betaproteobacteria-1	6	18	9
Proteobacteria	zur - Gammaproteobacteria	34	118	69
Proteobacteria	zur - Proteobacteria	11	21	12
Thermotogae	zur - Thermotoga	5	6	4

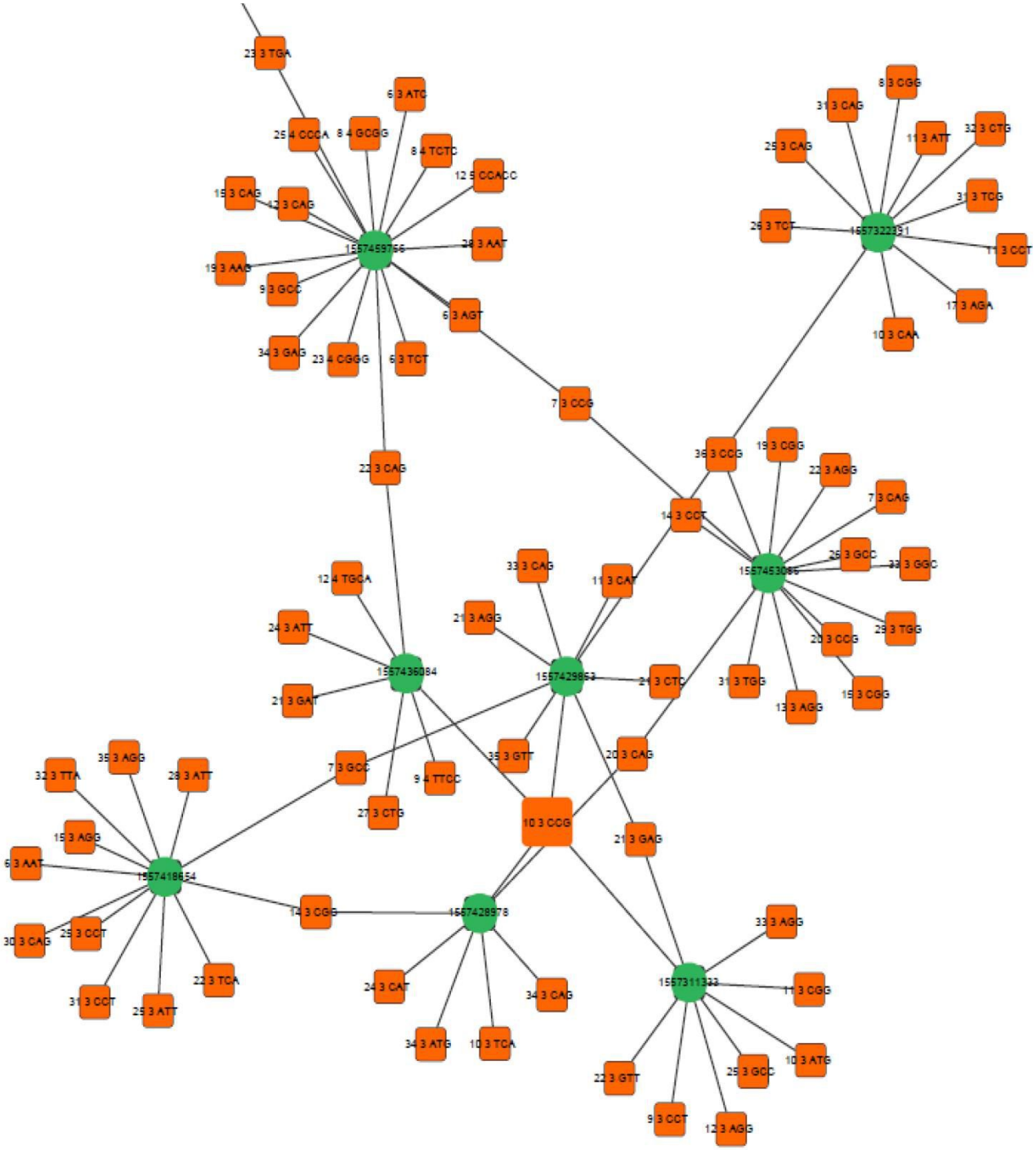


Figura 1. Red de regulación de la expresión génica en el suelo de granja, esta red ilustra los promotores cuyos genes se relacionan con metabolismo de la histidina. En color verde se representan los promotores y en naranja los factores de transcripción El sitio de unión marcado como 10 3 CCG (o CCG(10)CGG usando la nomenclatura explicada en Metodología) coincide en estructura con sitios reportados previamente para PurR.

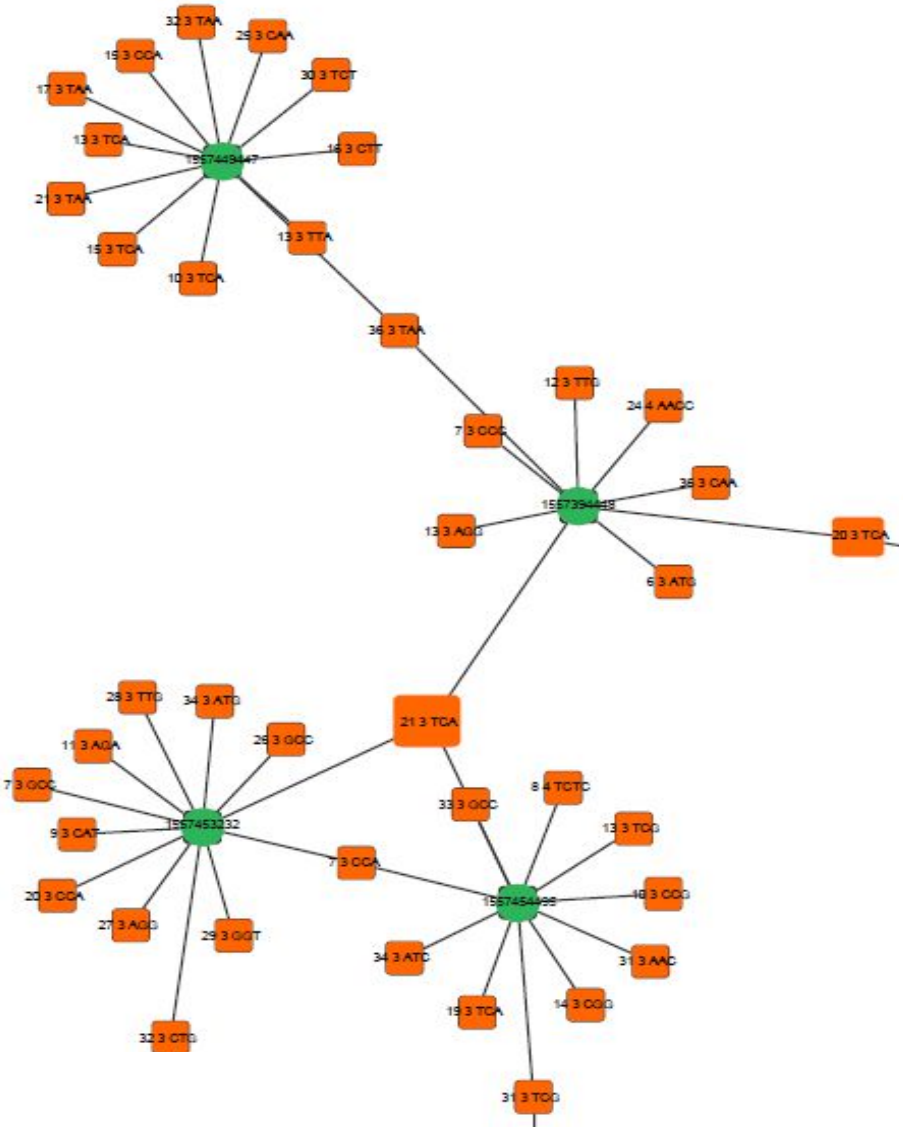


Figura 2. Red de regulación de la expresión génica en el suelo de granja, esta red ilustra los promotores cuyos genes se relacionan con metabolismo del potasio. En color verde se representan los promotores y en naranja los factores de transcripción El sitio de unión marcado como 21 3 TCA (o TCA(21)TGA usando la nomenclatura explicada en Metodología) coincide en estructura con sitios reportados previamente para NikR.