UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Doctorado en Ciencias Biomédicas
Instituto de Ecología

ANÁLISIS DE PATRONES DE DIVERSIDAD TAXONÓMICA Y FUNCIONAL EN METAGENOMAS

T E S I S

QUE PARA OBTENER EL GRADO ACADÉMICO DE

DOCTOR EN CIENCIAS

P R E S E N T A

GERMAN BONILLA ROSSO

TUTOR PRINCIPAL DE TESIS: DRA. VALERIA SOUZA SALDÍVAR

COMITÉ TUTOR:
DR. DAVID ROMERO CAMARENA
DR. LORENZO SEGOVIA FORCELLA

MÉXICO, D.F. **SEPTIEMBRE, 2012**

Análisis de Patrones de Diversidad Taxonómica y Funcional en Metagenomas

por

Germán Bonilla Rosso

Tutor

Dr. Valeria Souza Saldívar

Comité Tutoral

Dr. David Romero Camarena

Dr. Lorenzo Segovia Fourcella

Artículos publicados en revistas especializadas que se desprenden del trabajo realizado en ésta tesis doctoral

Rusch et al. **2007**. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* 5(3): e77
http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.0050077

Bonilla-Rosso G, Souza V y Eguiarte LE. **2008**. Metagenómica, Genómica y Ecología Molecular: La Nueva Ecología en el Bicentenario de Darwin. *TIP Revista Especializada en Ciencias Químico-Biológicas* 11(1):41-51.
http://www.artemisaenlinea.org.mx/articulo.php?id=2224&arte=a

Bonilla-Rosso G, Eguiarte LE, Romero DR, Travisano M y Souza V. **2012**. Understanding microbial community diversity: use of simulated datasets to evaluate the performance of diversity metrics derived f rom metagenomes. *FEMS Microbiology Ecology* 82(1):37-49.
http://onlinelibrary.wiley.com/doi/10.1111/j.1574-6941.2012.01405.x/full

Bonilla-Rosso G, Peimbert M, Alcaraz LD, Hernandez I, Eguarte LE, Olmedo G y Souza V. **2012**. Comparative metagenomics of two microbial mats at Cuatro Ciénegas Basin II: Community Structure and Composition in Oligotrophic Environments. *Astrobiology* 12(7):659-673.
http://online.liebertpub.com/doi/abs/10.1089/ast.2011.0724

Peimbert M, Alcaraz LD, Bonilla-Rosso G, Olmedo G García-Oliva F, Eguiarte LE y Souza V. **2012**. Comparative metagenomics of two microbial mats at Cuatro Ciénegas Basin I: Ancient Lessons on How to Cope with an Environment under Severe Nutrient Stress.
Astrobiology 12(7):659-673.
http://online.liebertpub.com/doi/abs/10.1089/ast.2011.0694

A mis hermanas Daphne, Nadia y Dalia.

"La ciencia está compuesta de errores que, a su vez, son los pasos hacia la verdad"
--Julio Cortázar

"La fiction a précédé la science. Tout ce que nous rêvons, c'est-à-dire tout ce que nous désirons, est vrai. Le mythe d'Icare a précédé l'aviation. Il n'y a de vrai que le mythe : l'histoire, tentant de le réaliser, le défigure, le rate à moitié; elle est imposture, mystification, quand elle prétend avoir "réussi". Tout ce que nous rêvons est réalisable. La réalité n'a pas à être réalisable : elle n'est que ce qu'elle est. C'est le rêveur, ou le penseur, ou le savant, qui sont les révolutionnaires, ce sont eux qui tentent de changer le monde."
--Eugéne Ionesco

"Things need not have happened to be true. Tales and adventures are the shadow truth that will endure when mere facts are dust and ashes and forgotten"
--Neil Gaiman

[*En los monasterios, seminarios y sinagogas, temen el infierno y añoran el paraíso. Aquellos que estudian los misterios de Dios nunca dejan que dicha semilla sea sembrada en sus almas.*]
--Omar Khayyam

Falsch heiße uns jede Wahrheit, bei der es nicht ein Gelächter gab
--Friederich Nietzche

Agradecimientos

A Frédérique Reverchon y la familia Olarte-Rosso por el apoyo incondicional y constante a lo largo de la última media década, y simplemente por acompañarme, especialmente porque reconozco lo difícil que fue lidiar conmigo durante este doctorado.

A mi familia extendida del laboratorio Esmeralda, Nuria, René, Enrique, Lau, Ana, Rodrigo, Eria, Morena, Andrea, Jaime, Ricardo, Santiago, Julia, Silvia y Christine, por acompañarme en seminarios, extracciones y peceerres vespertinos, muestras en la cineteca, idas al campo, terapias grupales en Cuernavaca, mariscos, pozoles, birrias y chelas asociadas. A Fred, René, Tania, Xitla y Jorge por ayudarme a tener un lugar amable donde regresar en la noche.

A Ana Laura, Edgardo, Fred, Coatlicue, Alejandra, Saúl, Jimena, Luis David y Eduardo que me ayudaron a sobrevivir especialmente al inicio de la tesis. Y al final.

Al Clan Souza-Eguiarte por recibirme como en casa, y también nada más por ser como son.

A Esmeralda y Laura por enseñarme lo poco que sé hacer en el laboratorio, y si digo poco no es porque sus enseñanzas hayan sido escasas sino porque definitivamente soy un mal alumno. A Mark Olson, Mike Travisano y Christine Rooks por la paciencia y perseverancia para enseñarme a ordenar, escribir y revisar manuscritos en inglés.

A mi comité tutoral Lorenzo Segovia y David Romero, por regresar constantemente a la chiva al corral. Especialmente a David Romero, que se convirtió en un ejemplo no sólo académico sino también personal. También a mi comité tutoral honorario Luis Eguiarte, Gaby Olmedo, Felipe García-Oliva y Mike Travisano, que siempre estuvieron dispuestos a discutir y aportar al proyecto, aún bajo el sol de 40°C en el desierto.

A los revisores de éste documento, Valeria Souza, Luis Eguiarte, Pablo Vinuesa, Enrique Merino, Gabriela Olmedo y Xavier Soberón, que a pesar del susto del volúmen de la tesis y mi lenguaje obscuro tuvieron el tiempo y ganas de corregirla.

A todos los 40+ miembros presentes y pasados del Laboratorio de Evolución Molecular y Experimental sede Calcuta, pos su apoyo, tolerancia y retroalimentación, y porque casi siempre hicimos lo posible por llevarnos casi bien. Agradecimientos especiales a Rodrigo y a todos los que apoyaron los muestreos en campo.

A Felipe, Celeste, Yunuén y el resto del Laboratorio de Ecosistemas Terrestres del CIECO. A Gaby, Luis David, Varinia, Panchito y el Laboratorio de Genética Bacteriana, y a Gustavo, Luis Herrera, Luis David, Víctor y los que olvide del CINVESTAV-Irapuato y LANGEBIO. A Janet, Mike, Mya, David y sus equipos de trabajo. A Jim, Zarraz, Everett, Marcia, Peter, Jeremy, Mateo, Amisha y Michelle, y en especial a Jess, Jessie, Jeff y Jeff Dick de ASU. A Doug, Aaron, Shibu, Karla, Jeff y Granger del JCVI.

A Amir Rima y Ana Laura Ramos por su ayuda con las traducciones.

A todos los que se me olvidaron.

A los que lean ésta tesis.

Agradecimientos de Adultos

Principalmente a la Fundación Olarte-Rosso y a la Fundación Eguiarte-Souza por sus apoyos económicos voluntarios extraordinarios y sin afán de lucro. Ésta tesis literalmente no hubiera sido posible sin la beca para realizar estudios de doctorado de CONACyT (196814), y los proyectos "Metagenómica de un tapete microbiano en Cuatrociénegas" (SEP-CONACyT 50507), "Cuatro Ciénegas Coahuila como sistema modelo para determinar el efecto del cambio climático global en los ciclos de C y N" (SEMARNAT 2006-C01-23459), "Stoichiometry of Life, Task 2b: Field Studies – Cuatro Ciénegas" (NASA-NAI / ASU 2009-2011) y "Conocimiento y conservación de la biodiversidad del Sistema Churince, Cuatrociénegas, Coahuila" (Alianza WWF-Fundación Carlos Slim L039). La escritura final de la tesis fue posible gracias a las generosas aportaciones de la Fundación Reverchon y la Fundación Carlos Slim.

A los miembros de la Comisión Nacional de Áreas Naturales Protegidas en Cuatrociénegas, Pronatura Noreste A.C., y las autoridades y habitantes de Cuatrociénegas de Carranza, Coahuila, por permitirnos trabajar en su casa.

Índice

INTRODUCCIÓN

La característica más sorprendente del planeta Tierra es la extraordinaria variedad de seres vivos que la habitan, con una estimación conservadora de 8.7 millones de especies eucariontes (Mora et al. 2011). La consecuencia natural de tener tal diversidad de organismos coexistiendo en tiempo y espacio es el establecimiento de un número exponencial de interacciones entre diferentes organismos, así como entre los organismos y su entorno abiótico. Ésto genera a su vez diversidad no sólo en el número y abundancia de especies, sino también de genes, variantes, funciones, asociaciones, poblaciones, comunidades y ecosistemas, que de manera colectiva se agrupan bajo el término biodiversidad (Naeem et al. 1999). La tarea monumental de la ecología es entender la distribución y abundancia de los organismos así como las interacciones entre ellos y su medio ambiente, por lo que los estudios ecológicos se categorizan en tres niveles de organización: el organísmico que estudia la interacción entre los individuos y su ambiente; el poblacional que estudia las interacciones entre grupos de individuos de la misma especie que coexisten en tiempo y espacio; el de comunidades que estudia ensambles de especies que interactúan en el mismo tiempo y espacio y el ecosistémico que estudia la interacción entre las comunidades y su entorno abiótico (Begon et al. 2006).

1. Biodiversidad

El término *biodiversidad* es en la actualidad uno de los más utilizados en publicaciones científicas del área de las ciencias biológicas, y su uso ha trascendido la academia para formar parte del vocabulario popular contemporáneo. Dado que su etimología sugiere inequívocamente que el término abarca la totalidad de la variedad de la vida, biodiversidad es uno de los conceptos más difíciles de definir. Una de las definiciones más aceptadas es la utilizada por el Convenio de Diversidad Biológica (United Nations Environmental Programme 1992), que reconoce la diversidad biológica en tres niveles diferentes: dentro de las poblaciones (diversidad genética), entre poblaciones (diversidad de especies) y entre hábitats y comunidades (diversidad ecológica) (Norse & McManus 1980). La cuantificación de la diversidad genética fue ampliamente estudiada en el siglo pasado, y se centra en el análisis de polimorfismos y variantes alélicas en poblaciones; sin embargo, los otros dos niveles han demostrado ser más difíciles de cuantificar. A continuación explicare porqué, empezando por la diversidad taxonómica.

A primera vista el conteo del número de especies en una comunidad podría parecer una buena medida de diversidad de especies, pero al hacerlo de manera implícita asumimos que todas las especies medidas son ecológicamente equivalentes y contribuyen en la misma magnitud al ecosistema, independientemente de su afiliación taxonómica, función ecológica y abundancia en el número de individuos de cada especie (Hawksworth 1995). Una mejor definición de diversidad de especies sigue el concepto de **estructura** de las comunidades y considera simultáneamente la *riqueza específica* (el número de especies) y la *abundancia relativa* de cada una de éstas especies. La abundancia relativa puede ser considerada de dos formas complementarias: en la *dominancia* se pondera la proporción de las especies con mayor número de individuos, mientras que la *equitabilidad* considera qué tan homogéneamente se encuentran distribuidos los individuos en todas las especies (Margalef 1958; Lloyd & Ghelardi 1964; Pielou

1966; Hurlbert 1971). La generación de un gran número de métricas y estadísticos usualmente llamados índices (de diversidad, dominacia, equitatividad, etc.) en la literatura ecológica, que cuantifican y ponderan de manera diferencial la riqueza y dominancia de las especies (Preston 1948; Simpson 1949; MacArthur 1955; MacArthur 1957; Margalef 1958; Rényi 1961; Whittaker 1965; MacArthur 1965; Paine 1966; Pianka 1966; Pielou 1967; McIntosh 1967; Hill 1973) ha conducido a largas discusiones sobre la naturaleza de la biodiversidad y la elección de la métrica más apropiada para medirla, cuando en realidad distintos índices revelan aspectos diferentes sobre la estructura de las comunidades y varios son redundantes o equivalentes (Hurlbert 1971; Tóthmérész 1995; Ricotta 2003; Keylock 2005; Jost 2006). En el Apéndice 1 se presenta un resúmen de las métricas de diversidad utilizadas en éste trabajo, con la manera de calcularlos y una breve descripción de su interpretación.

Sin embargo, una tercera dimensión que se había encontrado ausente de las discusiones sobre diversidad en comunidades naturales es la *distancia filogenética*. Ésta hace notar que además del número de especies y el número de individuos dentro de cada especie, la historia evolutiva de cada especie (y la divergencia entre ellas) afecta también la equivalencia entre ellas. Es decir, es más probable que un grupo de especies estrechamente relacionadas entre sí sean ecológicamente equivalentes que especies que pertenecen a grupos taxonómicos divergentes. De ésta forma, se incorpora no sólo el número y abundancia relativa de las especies en una comunidad, sino también la **composición** taxonómica de las comunidades (Faith 1992; Webb et al. 2002; Cadotte et al. 2010; Kembel et al. 2011). Muchos de los estudios de ecología de comunidades simplemente dividían a la diversidad mediante la delimitación de *taxocenos*: subgrupos de especies filogenéticamente relacionadas entre sí (Chodorowski 1959; Hutchinson 1967; Hurlbert 1971). Éste enfoque permite hacer la comparación de la diversidad de grupos taxonómicos particulares en condiciones diferentes, manteniendo cierta coherencia taxonómica en la comparación. La definición del grupo filogenético es de la mayor importancia, pues si el grupo resulta demasiado inclusivo, las especies no serán ecológicamente equivalentes por sus diferencias biológicas, obstaculizando la interpretación de los parámetros de diversidad a comparar (Hurlbert 1971).

Por otra parte, la partición de las comunidades en subgrupos no sólo responde a un interés teórico, sino también resulta de la solución de un problema de carácter metodológico, pues a pesar de que las definiciones de biodiversidad tratan de ser lo más amplias e incluyentes posibles, en la práctica no es factible contabilizar y categorizar todas las especies de todos los grupos taxonómicos comprendidas en una muestra. Es por eso que la mayor parte de los estudios de ecología de comunidades se centran en taxocenos (Chodorowski 1959; Hutchinson 1967; Hurlbert 1971) o en grupos funcionales particulares (Root 1967; Simberloff & Dayan 1991; Wilson 1999). Un *gremio funcional* es un conjunto de especies que explotan una misma clase de recursos naturales de manera similar (Root, 1967), lo que presenta una manera alternativa de categorizar a los miembros de una comunidad independientemente de su afiliación taxonómica al tiempo que se utilizan explícitamente los caracteres ecológicos funcionales para categorizar y subsecuentemente cuantificar a los miembros de una comunidad.

Una tercera alternativa para el análisis de comunidades naturales incorporando al mismo tiempo su taxonomía y su función ecológica es mediante el estudio de *rasgos funcionales*: propiedades medibles y bien definidas de los organismos que influyen fuertemente en el desempeño del mismo (McGill et al. 2006). Algunos rasgos funcionales se encuentran filogenéticamente conservados, de manera que su análisis incorpora simultáneamente el aspecto ecológico y evolutivo de las especies en las comunidades.

Dadas sus dimensiones espaciales y sus cortos tiempos generacionales respecto a los macroorganismos, las comunidades microbianas presentan una oportunidad sin precedentes para el análisis de diversidad y ecología de comunidades. Lo que es más, a la fecha no se han encontrado comunidades de macroorganismos que estén libres de microorganismos, pues éstos se hallan en todos los ambientes habitables por los macroorganismos e incluso habitan dentro y sobre ellos. En contraste, se han detectado un gran número de comunidades compuestas exclusivamente por arqueas y bacterias, debido a las condiciones ambientales extremas en las que se encuentran.

Asimismo, el análisis de diversidad de especies en comunidades naturales se ve particularmente beneficiado del desarrollo de técnicas para el análisis de comunidades microbianas, pues la manera actual de identificar microorganismos es mediante *marcadores moleculares* que genotipifican al organismo al que pertenecen, de manera que la categorización de los individuos de la comunidad está directamente relacionada con un componente filogenético. La manera de evaluar la presencia de ciertos atributos funcionales es de igual forma mediante marcadores moleculares, permitiendo analizar simultáneamente los componentes taxonómicos y funcionales de una misma muestra. Y dado que una comunidad microbiana compleja puede ocupar una dimensión espacial pequeña, es posible apuntar al análisis del total de las especies contenidas en una muestra, independientemente de su filiación taxonómica o funcional.

2. Metagenómica como herramienta para entender la biodiversidad.

La **metagenómica** es una herramienta derivada de la biología molecular y la bioinformática que permite el análisis de secuencias taxonómico y funcional de comunidades microbianas mediante la secuenciación del complemento genómico total (DNA ambiental) de todos los organismos presentes en una comunidad natural. El procedimiento general comprende la extracción del DNA ambiental de una comunidad tratándola como si fuera un sólo tejido, de manera que contiene el DNA genómico combinado de todos los organismos presentes en la muestra. Éste DNA es fragmentado y subsecuentemente secuenciado, y anotado, revelando el potencial genómico integrado pero fragmentado de toda la comunidad.

En el proyecto metagenómico más ambicioso de la década pasada (Global Ocean Sampling; GOS), se muestrearon 200 lts de aguas marinas superficiales cada 300 millas náuticas a lo largo de un transecto que pretendía circunscribir el planeta (Rusch et al. 2007). Dado que un proyecto metagenómico genera una gran cantidad de secuencias de marcadores filogenéticos, estos datos representan una buena oportunidad para la estimación de la diversidad microbiana total. Sin embargo, dado que las regiones del DNA son secuenciadas aleatoriamente, muchos de los fragmentos resultantes de los marcadores filogenéticos no son comparables entre sí, impidiendo los análisis clásicos de ecología tradicional. Al final

de éste capítulo se encuentra un artículo de revisión que profundiza sobre el alcance y las limitaciones de las herramientas metagenómicas para el análisis ecológico de comunidades.

En el proyecto piloto del GOS, el grupo de Venter analizó datos del mar de los Sargassos (Venter et al. 2004), en el que se señaló el problema adicional que representa el uso del 16SrRNA como marcador filogenético para la estimación de la diversidad, dado que presenta un número de copias en el genoma diferente en cada especie. Debido a este sesgo, se analizó la diversidad utilizando un juego de cuatro marcadores filogenéticos codificantes de proteínas (*recA, HSP70, EF-Tu, EF-G*) y utilizando alineaciones y filogenias de referencia para asignar cada fragmento a un grupo taxonómico (Venter et al. 2004; Wu & Eisen 2008).

La idea central detrás de éstas aproximaciones es la construcción de alineaciones filogenéticamente inclusivas y cuidadosamente curadas que permitan la rápida incorporación de nuevos fragmentos de manera confiable. Una vez alineada, se realiza una reconstrucción filogenética rápida o un algoritmo de posicionamiento filogenético para incrustar el fragmento en el árbol, asignando así la categorización taxonómica correspondiente a la posición filogenética en el árbol. Algunos programas que utilizan ésta aproximación son PhylOTU (Sharpton et al. 2011) para SSU-rRNAs y el paquete *picante* en R (Kembel et al. 2011) para proteínas.

Otras aproximaciones para la clasificación taxonómica, menos precisas pero mucho más veloces son la exportación directa de la taxonomía de los mejores hits de BLAST por debajo de un nivel de corte determinado, como el utilizado por MGRAST (Meyer et al. 2008) y MEGAN (Huson et al. 2007), que aplica un algoritmo del último ancestro común a una lista de los mejores hits hallados con BLAST. Las limitaciones de éste tipo de aproximaciones radican en que la precisión en la asignación taxonómica depende de la calidad y resolución de las alineaciones de referencia y por lo tanto el algoritmo será más eficiente con grupos bien muestreados y con abundantes representantes en las bases de datos, mientras que será poco confiable con grupos poco muestreados.

En cuanto a la clasificación funcional de muestras metagenómicas, existen dos aproximaciones prevalentes. La primera consiste en buscar cada fragmento metagenómico contra bases de datos no redundantes de secuencias con anotaciones funcionales confiables y extrapolar la función del mejor hit de BLAST por debajo de un nivel de corte determinado. Idealmente la base de datos de referencia contiene secuencias completas de genes y proteínas como GenBank (Benson et al. 2005) o SwissProt (Magrane et al. 2011). Sin embargo, la mayor parte de las veces los fragmentos metagenómicos son demasiado cortos como para ser asignados de manera confiable a un gen completo, o los fragmentos son demasiado divergentes de las secuencias de referencia y contienen tan sólo dominios o motivos reconocibles. En consecuencia, los fragmentos sólo pueden ser clasificados en funciones metabólicas generales y por lo tanto resulta más eficaz utilizar bases de datos de motivos o dominios como Pfam (Finn et al. 2008) o InterPro (Hunter et al. 2012), o bases de datos de familias de proteínas y grupos de ortólogos que comparten funciones generales bien caracterizadas como Pfam (Finn et al. 2008) o la base de datos de COGs (Tatusov et al. 2003).

La segunda aproximación consiste en incorporar la función de cada fragmento a un mapa metabólico como el KEGG, (Kanehisa et al. 2008) o de subsistemas funcionales como el SEED (Overbeek et al. 2005) o un mapa de interacciones (Jensen et al. 2009) que permita integrar los resultados para una mejor interpretación de la información contenida en el metagenoma. El principal problema de esta segunda aproximación es que la precisión en la asignación de función al fragmento metagenómico depende enormemente de nuestro conocimiento de las funciones, o de otra manera de la presencia de un buen número de representantes de la familia funcional a la que pertenece y sus familias hermanas en las bases de datos de referencia, de manera que sea posible discernir adecuadamente entre las funciones de proteínas estrechamente emparentadas. Más aún, existen proteínas con funciones diferentes que difieren en tan sólo un aminoácido como las proteorhodopsinas (Man et al. 2003).

Dado que nuestra ignorancia respecto al complemento funcional de las comunidades bacterianas es enorme, una alternativa para el análisis de diversidad funcional es la que se basa simplemente en construir **conglomerados** *de novo* de secuencias similares a partir de los mismos juegos de datos metagenómicos y las bases de datos de referencia. La ventaja radica en que también se construirán conglomerados de secuencias que están presentes en los metagenomas, pero no tienen representantes en las bases de datos, y que la similitud de secuencia y distancia filogenética está implícita en el método de construcción y los valores de corte utilizados. Los conglomerados han sido delimitados con valores de corte de similitud entre secuencias, como está implementado en el programa CD-HIT (Li 2009), o con algoritmos de teoría de gráficas como el utilizado para la base de datos de proteínas del GOS (Yooseph et al. 2007). Ésta aproximación permite obtener con mayor precisión una mejor representación de la diversidad de secuencias funcionales presentes en una muestra metagenómica, pero resulta en un gran número de conglomerados con función desconocida.

3. Planteamiento del problema

La metagenómica aplicada a la ecología microbiana representa al mismo tiempo una oportunidad sin precedentes para ampliar el conocimiento de las comunidades microbianas naturales y probar teorías ecológicas, y un reto teórico y metodológico para el análisis correcto de la información contenida en los juegos de datos metagenómicos. Por estas razones, en éste trabajo se plantea explorar a profundidad la diversidad taxonómica y funcional en muestras metagenómicas. En breve, la pregunta general alrededor de la cual se construyó ésta tesis es la siguiente:

¿Cuáles son las técnicas de análisis de diversidad que podemos aplicar a juegos de datos metagenómicos de comunidades microbianas naturales complejas para revelar patrones en la diversidad taxonómica y funcional y las posibles relaciones entre ellas?

Particularmente, se plantearon los siguientes objetivos:

1. Determinar la aplicabilidad de las **medidas de diversidad taxonómica** clásicas de la ecología de comunidades de macroorganismos a muestras con datos metagenómicos, para eventualmente seleccionar las métricas más adecuadas y con mayor precisión.

2. Caracterizar la **diversidad taxonómica** de dos tapetes microbianos del valle de Cuatrociéngas e identificar posibles patrones de diversidad compartidos por comunidades de la misma naturaleza.

3. Caracterizar la **diversidad funcional** de dos tapetes microbianos del valle de Cuatrociéngas e identificar posibles patrones de diversidad compartidos por comunidades de la misma naturaleza.

4. Explorar la **relación** entre la diversidad taxonómica y la diversidad funcional en las comunidades microbianas del valle de Cuatrociénegas.

4. Sitio de Estudio.

La exploración de patrones de diversidad taxonómica y funcional se realizó en metagenomas de tapetes microbianos y estromatolitos de cuerpos acuáticos del valle de Cuatrociénegas. El Valle de Cuatrociénegas de Carranza es un humedal del desierto Chihuahuense dentro del estado de Coahuila, México [26°59′10″N 102°3′59″W] que se extiende 40 km E-O y 30 km N-S (Minckley & Cole 1968). El valle se encuentra a una altitud de 740 msnm y presenta un clima árido con menos de 200 mm de precipitación anual y una evaporación potencial de 2000mm, clima en parte debido a que se encuentra rodeado por cordilleras de más de 3000 msnm, correspondientes a las sierras Madre Oriental y Occidental (INE-SEMARNAP, 2000).

El valle tiene más de 300 pozas de diferente tamaño, profundidad y permanencia, que surgen a partir de manantiales de cuerpos de agua subterráneos que fluyen a lo largo del valle y que parecen estar comunicados con cuerpos de agua al otro lado de las cordilleras (Johannesson et al. 2004; Wolaver et al. 2008). Aunque hay una gran variabilidad en la composición química particular de las pozas, todas son aguas duras con sales de carbonatos de calcio y sulfato de magnesio en altas concentraciones debido a la naturaleza evaporítica de las pozas, y presentan también cloruros de sodio y potasio pero a menor concentración. El pH oscila entre neutro (7.2) cerca de los manantiales, a alcalino (~8) en las lagunas de desecación (Calegari 1997).

Una de las características más sorprendentes de las pozas de Cuatrociénegas es su baja concentración nitrógeno (N) y fósforo (P), que al ser nutrimentos esenciales para toda la vida posicionan a estos ambientes acuáticos como oligotróficos (Elser et al. 2005; Breitbart et al. 2009). A pesar de ser pobres en nutrientes, los ecosistemas acuáticos de Cuatrociénegas albergan una enorme diversidad microbiana (Souza et al. 2006; Escalante et al. 2008). Aparentemente, las bajas concentraciones de fósforo excluyen la presencia de macroorganismos eucariontes en las comunidades, de manera que la productividad primaria parece depender exclusivamente de microorganismos fotosintéticos (Souza et al. 2006). Asimismo, la exclusión de macroorganismos perturbadores permiten el establecimiento y desarrollo de comunidades microbianas complejas como tapetes microbianos y estromatolitos (Minckley & Cole 1968; Cohen 1989).

Los **tapetes microbianos** son estructuras laminares organosedimentarias que se desarrollan en la interfase agua-sedimento y están compuestos por una comunidad de organismos que interactúan estrechamente entre ellos, conformando un sistema complejo autosustentado. El estrecho tejido de bandas multicolor representa también una separación y organización funcional a lo largo de un gradiente fisicoquímico abrupto de pH, oxígeno disuelto y concentraciones de nitratos y sulfatos (Jorgensen et al. 1986; van Gemerden 1993). Las combinaciones de las concentraciones variables en puntos particulares genera una enorme diversidad de micronichos a lo largo del gradiente en una dimensión espacial reducidad (2-50 mm), permitiendo el establecimiento de estrechas e intrincadas redes de interacciones metabólicas entre una gran diversidad de gremios funcionales (Paerl & Yannarell 2010).

A la fecha, el único trabajo publicado sobre metagenomas de tapetes microbianos es el correspondiente a los tapetes halófilos de las lagunas hipersalinas de la salinera Exportadora de Sal S.A. en Guerrero Negro, Baja California, México (Kunin et al. 2008). Dichos tapetes contienen una sorprendente diversidad tanto funcional (Kunin et al., 2008) como de especies (Ley et al. 2006). Sin embargo, el ecosistema en el que se desarrollan es radicalmente diferente a los tapetes de Cuatrociénegas, porque es de pozas de desecación de aguas marinas costeras ricas en cloruro de sodio. Ésta diferencia permite identificar organismos centrales al desarrollo de tapetes microbianos independientemente del medio ambiente en el que se desarrollen.

Otras estructuras organosedimentarias similares a los tapetes microbianos son los estromatolitos. Los estromatolitos aparentemente son formados cuando las condiciones fisicoquímicas permiten la precipitación de carbonato de calcio en la matriz circundante a las comunidades bacterianas que

conforman el tapete (Breitbart et al. 2009), generando estructuras sólidas mineralizadas que pueden perdurar como fósiles por millones de años (Allwood et al. 2006).

Durante el desarrollo de éste trabajo de tesis, se publicó también el metagenoma de dos estromatolitos del valle de Cuatrociénegas, que revelaron dos comunidades muy diferentes entre sí, una dominada por Cyanobacteria y la otra dominada por Alpha y Gammaproteobacteria y Planctomycetes (Breitbart et al. 2009), pero ambas con una gran parte de su metabolismo destinado a la producción de expolisacáridos que componen la matriz extracelular en la que se encuentran embebidos. La existencia de éstos dos metagenomas me permitió también explorar las similitudes entre comunidades microbianas organosedimentarias de la misma región y comparar sus patrones de diversidad.

5. Organización de la Tesis.

Ésta tesis se encuentra organizada en cuatro capítulos que corresponden a los cuatro objetivos planteados anteriormente. Las secciones que ya han sido publicadas incluyen la versión final del artículo publicado. Incluyo también la referencia a un artículo de colaboración publicado al inicio del doctorado, que me ayudó a delimitar las preguntas de trabajo y me permitió identificar los problemas metodológicos encontrados durante la estimación de diversidad a partir de datos metagenómicos (Rusch et al., 2007).

De esta manera, al final de ésta Introduccción, que es el Primer Capítulo se encuentra un artículo de revisión que profundiza respecto al desarrollo de la metagenómica y presenta nuestras perspectivas sobre la aplicación de ésta herramienta a la ecología molecular (Bonilla-Rosso et al., 2008).

El Segundo Capítulo de la tesis esta conformado por un artículo que explora teóricamente las limitaciones metodológicas para estimar la diversidad, contrastando diferentes métricas de diversidad aplicadas a la estimación de diversidad a partir de un juego de datos metagenómicos simulados (Bonilla-Rosso et al. 2012a).

El Tercer Capítulo corresponde al artículo en donde se aplican los resultados del capítulo anterior a la estimación de diversidad y análisis de la composición taxonómica de los dos metagenomas de tapetes microbianos analizados (Bonilla-Rosso et al. 2012b).

En el Cuarto Capítulo incluyo un artículo que contiene parte de la caracterización funcional de los dos metagenomas, complementado con el análisis de los patrones de diversidad bajo un contexto biológico que no está incluido en el manuscrito publicado (Peimbert et al. 2012).

En el Quinto Capítulo se integran los análisis de diversidad taxonómica y funcional y se discute de manera global los resultados de los capítulos anteriores. El último Capítulo representa las conclusiones globales de esta tesis.

Bonilla-Rosso, Germán;Souza, Valeria;Eguiartea, Luis E.

METAGENÓMICA, GENÓMICA Y ECOLOGÍA MOLECULAR: LA NUEVA ECOLOGÍA EN EL BICENTENARIO DE DARWIN

Tip Revista Especializada en Ciencias Químico-Biológicas, Vol. 11, Núm. 1, junio, 2008, pp. 41-51
Universidad Nacional Autónoma de México
México

Tip Revista Especializada en Ciencias Químico-Biológicas
ISSN (Versión impresa): 1405-888X
revistatip@yahoo.com
Universidad Nacional Autónoma de México
México

D.R. © TIP Revista Especializada en Ciencias Químico-Biológicas, 11(1):41-51, 2008

Metagenómica, Genómica y Ecología Molecular: La Nueva Ecología en el bicentenario de Darwin

Germán Bonilla-Rosso, Valeria Souza y Luis E. Eguiarte[a]

Depto. de Ecología Evolutiva, Instituto de Ecología, UNAM. Ciudad Universitaria, Apdo. Postal 70-275, México, D.F., 04510, México. [a]Autor de correspondencia. E-mail: [a]fruns@servidor.unam.mx

Resumen

En años recientes hemos presenciado un avance enorme en el desarrollo de tecnologías moleculares, en particular en las relacionadas con la secuenciación del ADN. El impacto de la genómica y metagenómica es de tal magnitud que revolucionarán las ciencias ecológicas, dando origen, junto con la Ecología Molecular, a una Nueva Ecología, con preguntas y metodologías diferentes a las abordadas por la ecología clásica. Estas nuevas metodologías abren la puerta a la exploración de nuevas aproximaciones al estudio de las comunidades microbianas, pues permiten el análisis simultáneo de la caracterización taxonómica de las especies contenidas en la comunidad y las funciones que pueden desempeñar. Aquí revisamos algunos de los avances de esta Nueva Ecología, describiendo el potencial que puede alcanzar si se usa junto con otras disciplinas. En general, nos permite analizar comunidades microbianas naturales, incluyendo organismos que escapan a nuestras posibilidades de cultivo en el laboratorio. Asimismo, brinda información sobre la diversidad y estructura trófica de las comunidades al tiempo que permite ligarlas con las funciones ecológicas que llevan a cabo en el ciclo de nutrientes del ecosistema que las contiene. Éste representa un salto enorme para resolver cómo es que cambian los organismos a lo largo del tiempo en respuesta a su entorno ecológico, mientras que permite analizar a una escala más fina cómo es que el entorno ecológico afecta los patrones evolutivos dentro de los linajes filogenéticos contenidos en las comunidades. Finalmente, esbozamos nuestra perspectiva sobre el futuro de la Nueva Ecología.

***Palabras Clave**: ADN, Bacterias, Bioinformática, Comunidades Microbianas, Cuatrociénegas, Diversidad, Ecología Bacteriana, Ecología de Ecosistemas, Filogenia, Microbiología Ambiental.*

Abstract

In recent years, we have witnessed an incredible advance in the development of molecular techniques, in particular new DNA sequencing technologies. Genomics and Metagenomics along with Molecular Ecology will revolutionize ecological sciences, giving birth to a New Ecology, with new questions and approaches to those studied by classical ecology. These new methodologies open the possibilities of new approaches to the study of microbial communities, allowing the simultaneous analysis of taxonomical and functional characterization of the community. In this review we discuss some of the advances of this New Ecology, describing its explanatory potential when used along with other disciplines. In general, it allows us to analyze natural microbial communities, including unculturable organisms. In addition, it produces information on the diversity, trophic structure of the community and functioning and nutrient cycling in the ecosystem. This represents a huge leap towards the understanding of organism evolution in its environment, and how this environment shapes evolutive patterns within phylogenetic lineages. Finally, we present our perspectives on the future development of this nascent discipline.

***Key Words**: DNA, Bacteria, Bioinformatics, Microbial Communities, Cuatrocienegas, Diversity, Bacterial Ecology, Ecosystem Ecology, Phylogeny, Environmental Microbiology.*

Nota: Artículo recibido el 20 de mayo de 2008 y aceptado el 18 de junio de 2008.

Introducción

Hace 10 años, en el primer número de esta revista describimos lo que llamamos el "sueño Darwiniano"[1], esto es, cómo con la ayuda de los marcadores moleculares podemos describir las relaciones genealógicas (filogenéticas) entre todos los seres vivos, o sea, reconstruir el gran árbol de la vida que tanto interesaba a Charles Darwin. En los últimos 10 años estas ideas han avanzado notablemente, pero aún nuestras predicciones más optimistas estaban lejos de concebir los espectaculares avances recientes en genómica[2] . Además de poder obtener con relativa facilidad los genomas completos de todo tipo de organismos, desde virus y bacterias (ver Alcaraz *et al*.[3] para el estudio más reciente publicado por mexicanos) hasta árboles y ornitorrincos[4,5], se ha comenzado a analizar el genoma de TODOS los organismos de una comunidad dada. Estos estudios se conocen como de metagenómica, y representan un avance cuántico en nuestro entendimiento de cómo funciona la naturaleza. Estos datos nos permiten no sólo conocer a todos los organismos de las comunidades, especialmente los microscópicos, sino también todas sus funciones. Esta información promete abrir la "caja negra" que representan estos compartimientos en los ecosistemas y entender, por fin, la verdadera ecología de nuestro planeta. Estas herramientas, junto con la Ecología Molecular clásica que describimos en detalle en nuestro libro publicado recientemente[6] constituyen lo que llamamos la Nueva Ecología, que promete ser una de las ramas más importantes de la Ciencia en este siglo. En este artículo queremos revisar algunos de los avances más interesantes en esta Nueva Ecología, en particular los relacionados a metagenómica ecológica, para que sirvan como una de las primeras guías a esta rama del conocimiento en español.

La ecología es la ciencia que estudia los factores que determinan la distribución y abundancia de los seres vivos. Tradicionalmente su estudio se divide en cuatro niveles de complejidad: La Autoecología o Ecofisiología, que estudia cómo responden los individuos de una especie frente a cambios en su ambiente físico y cómo lo modifican, es decir, estudian la fisiología de los organismos en su entorno abiótico, la Ecología de Poblaciones estudia cómo se comportan los grupos de organismos de una misma especie en un mismo tiempo y espacio, la Ecología de Comunidades que trata de entender cómo se estructuran e interaccionan las poblaciones de diferentes especies dentro de una comunidad y la Ecología de Ecosistemas, que estudia los flujos de materia y energía entre una comunidad y su entorno abiótico.

El análisis ecológico se ve frecuentemente entorpecido por la complejidad intrínseca de los problemas que estudia. En particular, los tamaños muestrales necesarios para realizar comparaciones estadísticas entre poblaciones y comunidades pueden representar un reto formidable, como lo sabe cualquiera que haya ayudado en un estudio de este tipo y haya tenido que contar y medir miles de muestras. Así podemos mencionar la complejidad en número y abundancia de especies que conforman una comunidad (por ejemplo en una selva tropical) y el problema que puede representar el sólo hecho de identificar muchas especies muy parecidas coexistiendo, la ausencia de bordes naturales que delimiten las comunidades y los ecosistemas, la extensa escala temporal a la que suceden cambios mesurables en poblaciones, comunidades y ecosistemas, y la magnitud espacial de los fenómenos ecológicos en los ecosistemas (por ejemplo, los ciclos biogeoquímicos y transferencia de energía entre niveles tróficos que a veces implican toda o amplias extensiones de la biosfera).

A pesar de que la importancia de los microorganismos procariontes (arqueas y bacterias) fue reconocida desde mediados del siglo pasado cuando Kluyver y van Niel[7] estimaron que 95% del CO_2 disponible en la atmósfera era producido por la mineralización bacteriana, la aproximación ecológica a los microorganismos se ha mantenido tradicionalmente enfocada a la fisiología bacteriana con la microbiología clásica. Actualmente se reconoce que el dominio con mayor biomasa (aunque mucha gente insiste en que son los insectos, o los hongos, según con lo que trabajen) y mayor diversidad taxonómica y funcional es del dominio Bacteria.

Asimismo, han sido las bacterias las causantes del cambio atmosférico más importante (hasta ahora) en el planeta Tierra: la producción de oxígeno diatómico y con ello la conversión de una atmósfera reductora a una oxidante. Hoy en día, sólo los organismos procariontes son capaces de realizar la fijación de nitrógeno atmosférico, mineralizar nitrógeno hasta nitrato y respirar sulfatos o hierro, por mencionar algunos de los múltiples pasos de los ciclos biogeoquímicos globales.

Sin embargo, el estudio de los microorganismos, dado su pequeño tamaño y sus tasas generacionales muy variables (pero que pueden ser muy rápidas), implica una problemática natural. Actualmente, en general se acepta que se conoce tan sólo el 1% de la diversidad total de bacterias en el planeta[8], y que en un gramo de suelo existen 10 mil millones de células bacterianas que pertenecen, al menos, a miles (según algunos autores millones) de especies[9]. Dado que las escalas espaciales y temporales de los organismos procariontes presentan una oportunidad única para el estudio de fenómenos ecológicos, el presente artículo pretende revisar los estudios recientes en ecología microbiana con un enfoque molecular alcanzado gracias al desarrollo de nuevas tecnologías.

La revolución genómica actual (entendida como el desarrollo de tecnologías de secuenciación de genomas completos) le ha dado un nuevo sentido al análisis de la ecología bacteriana, ya que al conocer el total de los genes contenidos dentro del genoma de una especie es posible resolver incógnitas fisiológicas y

ecológicas de ciertos organismos e inferir las necesidades de crecimiento de otros nuevos.

Por ejemplo, *Geobacter sulfurreducens* es una delta proteobacteria que se había identificado por la amplificación de su 16SrARN en suelos contaminados con material radiactivo y que podía cultivarse con dificultad en medios específicos. Tras la secuenciación de su genoma completo, resultó posible el diseño de medios de crecimiento específicos al descubrir su aerobiosis facultativa y su posibilidad de reducir sulfato como fuente de energía, así como su capacidad para producir electricidad[10,11]. Asimismo, el genoma del planctomicete *Kuenenia stturtgartensis* fue obtenido a partir de la secuenciación de ADN ambiental de lodos activados de una planta de tratamiento antes de poder cultivarlo, y su genoma reveló adaptaciones metabólicas relacionadas con la oxidación anaeróbica del amonio, proceso descrito bioquímicamente pero del cual se desconocía hasta entonces qué proteínas podrían realizarlo[12].

En la Nueva Ecología se utilizan una serie de métodos moleculares y estadísticos para mejorar nuestro entendimiento en los cuatro niveles clásicos de la Ecología (Tabla I). La Nueva Ecología nos ayuda por ejemplo a describir la diversidad total, con énfasis en los organismos microscópicos (arqueas, bacterias, "protistas" y hongos microscópicos), junto con métodos genéticos y geonómico/proteómicos para describir la funciones y las interacciones, en una perspectiva claramente filogenética y evolutiva.

Comunidades bacterianas, genética y metagenómica

Se puede sugerir que la ecología de comunidades microbianas moderna surgió cuando Pace *et al.*[13,14] amplificaron y secuenciaron el gen de ARN ribosomal 16S (utilizado como marcador taxonómico universal en arqueas y bacterias) del ADN extraído directamente de comunidades microbianas naturales (ADN ambiental: se toma una muestra de suelo o agua y se extrae TODO el ADN de la muestra, y se amplifica con métodos estándar de PCR[15]), encontrando una gran diversidad de organismos pertenecientes a grupos taxonómicos no descritos anteriormente a nivel de *phylum*. Antes de este estudio, la ecología microbiana se basaba en describir funciones (a veces enriqueciendo el ambiente) utilizando un enfoque de "caja negra"(o sea, se sabe lo que entra y lo que sale, pero no se sabe ni cómo se realiza el proceso ni quién lo lleva a cabo), o de tratar de cultivar las bacterias, un proceso muy ineficiente, ya que sólo una proporción muy baja de ellas son cultivables.

Desde entonces se ha descrito la composición taxonómica microbiana de una gran cantidad de ambientes distintos con una diversidad impresionante de grupos taxonómicos no descritos, en el extremo de los cuales se encuentra la definición del Reino Korarchaeota dentro del dominio Archaea, que no posee ningún representante cultivable[16]. Aunque las primeras filogenias universales de bacterias son sumamente recientes y reconocían alrededor de una decena de phyla[17], actualmente se reconocen entre 40 y 82 phyla (dependiendo del autor) sin representantes cultivables y, por lo tanto, sin conocimiento sobre su biología y ecología[18,19].

La metagenómica representa una aproximación totalmente nueva al estudio de las comunidades microbianas, definida como el análisis funcional y de secuencias de los genomas microbianos colectivos contenidos en una muestra ambiental, basándose ya sea en expresión o secuenciación[20]. El primer trabajo con datos tipo metagenoma, aunque parciales, fue el de Rondon *et al.*[21], donde se describe una librería de clonas en BACs (Bacterial Artificial Chromosomes) que contenía fragmentos de ADN relativamente grandes obtenidos directamente de muestras de suelos agrícolas de la Estación de Investigación Agrícola de

Nivel	Métodos tradicionales	Nueva Ecología
Ecofisiología/ Autoecología	Curvas de tolerancia Medidas en campo y laboratorio	Genómica: Genes-procesos. Expresión Genes: ARN Proteómica y microchips.
Poblaciones	Estimación tamaño de población: Captura/recaptura, crecimiento. Estructura/demografía. Mecanismos de regulación.	Marcadores moleculares para análisis de paternidad. Estimaciones de migración. Estimación de parámetros históricos con coalescencia y filogeografía.
Comundiades	Listados de especies. Número de individuos por especie.	ADN ambiental. Librerías de clonas de16 y 18S.
Ecosistemas	Análisis de nutrientes Análisis de tasas y de entradas de energía, recursos. Ciclos biogeoquímicos	Metagenomas Metaproteomas Genomas de especies clave

Tabla I. Los niveles clásicos de estudio y complejidad de la Ecología, los métodos como se han enfrentado tradicionalmente y las propuestas de la Nueva Ecología.

West Madison, Wisconsin. Cada una de las clonas fue sometida a diversos análisis diferenciales para la identificación de nuevas proteínas involucradas en actividades hemolíticas, lipolíticas y amilolíticas, demostrando así su potencial en el descubrimiento de nuevos genes.

Recientemente, el desarrollo y abaratamiento de tecnologías de secuenciación como el shotgun o la pirosecuenciación (Cuadro 1) ha brindado la posibilidad de análisis metagenómicos amplios. Las principales ventajas de estos métodos sobre las técnicas tradicionales de amplificación genética son: 1) un menor sesgo taxonómico en la secuenciación, y 2) la capacidad de analizar simultáneamente el complemento taxonómico y funcional de una misma comunidad. Lo anterior permite que el desarrollo tecnológico sea recibido por el marco teórico robusto de la ecología de comunidades tradicional, por ejemplo, permitiendo implementar los análisis tradicionales utilizados por ecólogos vegetales y animales para evaluar la riqueza de especies, la estructura de comunidades y las diversidades alfa, beta y gamma[15,22]; a la par que es posible analizar el reconocimiento de gremios funcionales, inclusive con una mayor precisión y resolución que la permitida por los estudios macroecológicos tradicionales, pues la presencia de un gen permite medir el potencial taxonómico de la comunidad incluso en condiciones donde dicho gen no se encuentre en uso.

Las tecnologías actuales se enfocan en la secuenciación no de genes u operones individuales, sino de grandes fragmentos de ADN como plásmidos o genomas completos. Todas involucran un paso previo en el cual el ADN total es fragmentado en pedazos más cortos y manejables, y por lo tanto todas involucran un proceso computacional posterior a la secuenciación que intenta reensamblar estos fragmentos en secuencias lineares individuales.

Shotgun: Rompe el ADN total en fragmentos y los separa por tamaños (3kb, 8kb y 40kb). Los fragmentos pequeños son insertados en cromosomas bacterianos artificiales (BAC) y los grandes en fósmidos, para luego ser clonados en *Escherichia coli* para amplificarlos. Después los fragmentos son extraídos y purificados para ser secuenciados. Si las secuencias de los BACs no pueden ser ensambladas, se buscan en los fósmidos para secuenciación completa. Este método es caro y lleva mucho tiempo construir y purificar las librerías, y es posible que regiones enteras no puedan ser clonables por contener genes letales. Sin embargo, produce fragmentos de 600-800 bp de longitud con una precisión de 99.97%. La cantidad de producto depende del número de clonas que se elija secuenciar. La secuenciación de un genoma menor a 5Mb, desde su preparación hasta la recuperación de datos, puede tomar un mes. La cobertura total es aproximadamente un 30x.

Pirosecuenciación: Nebuliza el ADN en fragmentos de 900bp. Toma una sola hebra de ADN de cada fragmento y añade "adaptadores" a cada extremo de la secuencia (un primer de 44 bp). Estos adaptadores fijan la secuencia a una superficie sólida donde duplica cada fragmento a partir del primer adaptador. Un lector óptico lee los fotones producidos por la hidrólisis de cada uno de los nucleótidos añadidos, generando la secuencia de ADN. El método prescinde de la construcción de librería de clonas, reduciendo drásticamente el precio y el tiempo de secuenciación. Un genoma menor a 5Mb, desde su preparación hasta la recuperación de datos, toma alrededor de 3 días con una cobertura aproximada de 5-8x (cada corrida de 4hrs produce 5 millones de bases). Sin embargo, la longitud promedio de las secuencias producidas es de 100-200 bp.

Cuadro 1. Metodologías de Secuenciación a Gran Escala.

Un ejemplo: el estudio del Mar de los Sargazos

Uno de los primeros trabajos publicados de metagenómica resulta de la secuenciación de la comunidad microbiana de aguas superficiales oceánicas del Mar de los Sargazos (Bermudas) realizado por Venter *et al.*[23] utilizando el ADN de 100 litros de agua en cada una de las 10 estaciones de muestreo, estas muestras se separan con tres filtros de diferentes tamaños (0.1-0.8 μm; 0.8-3 μm y 3-20 μm). Se analizaron la fracciones entre 0.8 y 3 micras de tamaño, que no incluyen ni las cosas relativamente grandes (como eucariontes planctónicos y cianobacteiras filamentosos), ni las muy chicas (bacterias muy chicas a veces llamadas picoplancton y virus). Dado que el Mar de los Sargazos es un océano oligotrófico (pobre en nutrientes) se esperaba que la comunidad bacteriana que mantiene sería relativamente "simple", pero el metagenoma resultó en una cantidad aproximada de 1800 genomas diferentes y más de 1.2 millones de genes codificantes. La reconstrucción de genomas completos fue totalmente infructuosa por múltiples causas: i) La increíble diversidad genómica encontrada reduce la posibilidad de encontrar fragmentos correspondientes a la misma especie (o población u organismo), impidiendo ensamblar fragmentos razonablemente grandes de genoma bajo una cobertura estadísticamente confiable; ii) La abundancia natural relativa y diferencial de cada especie en la muestra la sesga hacia una mayor cobertura de las pocas especies más abundantes; iii) La variabilidad intrapoblacional (es decir, de los miembros de una misma especie) complica el ensamble, incrementando la probabilidad de ensamblar fragmentos que no se encuentran juntos en la naturaleza.

Los segmentos ensamblados con mayor cobertura (hasta 14x, lo que significa que el segmento se encuentra confirmado por el ensamble de al menos 14 fragmentos, ver Cuadro 2) pudieron ser exitosamente alineados con segmentos del genoma secuenciado de la cianobacteria unicelular *Prochlorococcus marinus* MED4, aislado originalmente de esta región oceánica. Sin embargo, los segmentos no pudieron ser ensamblados entre sí, y ninguno presenta una identidad del 100% con la cepa secuenciada, sugiriendo la presencia de una población altamente diversa y abundante de *P. marinus* con diversas variantes genómicas.

Cualquiera que sea la metodología de secuenciación, el proceso siguiente es el ensamble de las secuencias obtenidas. El ensamble se basa en el sobrelapamiento de diferentes fragmentos de ADN que permitirán alinear los extremos similares y así reconstruir secuencias lineales completas conocidas como contigs y, con un poco de suerte, la secuencia lineal completa del genoma. El problema radica en que muchas veces los fragmentos secuenciados no cubren la totalidad del genoma (existen contigs que no pueden ensamblarse con ningún otro fragmento porque las regiones que los unirían no fueron secuenciadas) o existen regiones muy similares repetidas varias veces a lo largo del genoma, de manera que no es posible saber cuál contig se ensambla con cualquier otro. Obviamente, si los fragmentos producidos por la secuenciación son más grandes es más fácil ensamblar que si los fragmentos son más pequeños. Sin embargo, para saber con precisión si una secuencia es en realidad la secuencia del genoma y no un artefacto de la secuenciación, es necesario que cada base en cada sitio a lo largo de la secuencia lineal se encuentre repetida varias veces, es decir, que cada base en cada fragmento haya sido secuenciado varias veces y al alinear todos los diferentes fragmentos secuenciados de una misma región sean coeherentes entre sí. A esto se le conoce como profundidad de la cobertura de secuenciación, y un genoma completo se considera confiable si tiene más de 6x. Ésto quiere decir que la totalidad de bases secuenciadas es 6 veces el tamaño del genoma inferido, y por lo tanto se esperaría que cada base del genoma completo estuviera representado al menos en seis fragmentos independientes. Cabe señalar que el esfuerzo computacional que requiere este análisis es enorme y, aunque hemos observado enormes avances, el ensamblaje automático dista mucho de ser perfecto y frecuentemente requiere la asistencia humana.

Cuadro 2. Ensamble y Cobertura.

Uno de los descubrimientos más importantes producido por el análisis del metagenoma del Mar de los Sargazos es la revisión de la fotoautotrofía (la capacidad de sintetizar nutrimentos a partir de componentes inorgánicos utilizando la luz solar como fuente de energía) en comunidades marinas bacterianas, pues la enzima principal involucrada en el ciclo de Calvin (RubisCO) se encuentra en una muy baja abundancia y diversidad en las muestras, lo que podía explicarse por la sobredominancia del fotoautótrofo *Prochlorococcus* en la muestra. Sin embargo, una gran diversidad de proteorhodopsinas fue encontrada, con al menos 782 homólogos. Las proteorhodopsinas son una familia de rhodopsinas que son capaces de bombear protones en contra de un gradiente electroquímico transmembranal, inicialmente descritas en una alfaproteobacteria del clado SAR86[24,25], cuya variante en tan sólo una substitución nucleotídica que se traduce en la optimización en la absorbancia espectral de cada variante a diferente profundidad. La presencia de tal número de homólogos sugiere un gremio diverso y abundante de fotoautótrofos diferencialmente adaptado a nichos marinos para optimizar la absorción luminosa.

Ecología de Ecosistemas, la mina de hierro (AMD) y los biorreactores

El análisis funcional de comunidades permite realizar extrapolaciones respecto a la interacción entre el componente biótico y abiótico de un ecosistema. Por ejemplo, la presencia de una gran diversidad de proteorhodopsinas en el metagenoma del Mar de los Sargasos sugiere revisar las estimaciones de productividad primaria en los océanos y con ello los cálculos correspondientes al ciclo global de carbono. Sin embargo, dada la complejidad y versatilidad funcional de las comunidades marinas, las interpretaciones funcionales parecen tener más sentido cuando se analizan sistemas más simples en conjunto con estudios geoquímicos y bioquímicos.

El primer metagenoma "real" secuenciado fue publicado un poco antes que el estudio del Mar de los Sargasos por Tyson *et al.* también en el 2004. Los trabajos anteriores corresponden a metagenomas "dirigidos" o parciales como el que mencionamos del suelo de Wisconsin[21]. Tyson *et al.* analizaron una biopelícula simple que crece sobre el agua acidificada por escurrimiento dentro de una mina de hierro abandonada (AMD: Acid Mine Drainage). La simplicidad de la comunidad debido a la dominancia de dos organismos permitió la reconstrucción con una cobertura 10x de los genomas de una bacteria del grupo II del género *Leptospirillum* (Nitrospiraceae) y de una arquea menos abundante del género *Ferroplasma* tipo II (Euryarchaeota), así como el ensamble parcial de fragmentos del genoma de otras especie de *Leptospirillum* (grupo III,) y *Ferroplasma* (tipo III), mucho menos abundantes pero con alta transferencia horizontal, así como la presencia de otras tres especies procariontes con densidades tan bajas que no fue posible ensamblar sus fragmentos.

La cobertura de las dos especies dominantes (aproximadamente 10x cada una) permitió elucidar las rutas metabólicas involucradas en la oxidación de la pirita como fuente única de energía que permite la manutención de la comunidad. La continuidad de este trabajo mediante una aproximación proteómica[26] mediante espectrometría de masas por shotgun permitió analizar una fracción de las proteínas parcialmente degradadas en la mina, resultando en la identificación de la principal enzima oxidante de hierro en la comunidad: una secuencia nueva ácido-tolerante del *Leptospirillum* grupo II denominada Cyt579, demostrando la capacidad del enfoque combinado metagenómico-proteómico para develar nuevos procesos ecológicos.

Otra aplicación radicalmente diferente de las herramientas metagenómicas reside en el análisis de los biorreactores, como los usados en la eliminación de fósforo de aguas contaminadas (BPR), en los cuales se utilizan lodos con organismos que precipitan fosfatos en un tanque y después se les permite asentarse para su eliminación, dejando el cuerpo de agua

virtualmente libre de fosfatos. Sin embargo, la ignorancia respecto a la funcionalidad biológica de la comunidad BPR ocasiona reducción espontánea en la capacidad de eliminación del biorreactor. La secuenciación del metagenoma del lodo enriquecido[27] permitió la reconstrucción parcial del genoma de la especie dominante en el BPR: *Accumulibacter phosphatis*, un miembro del grupo Rhodocyclales (Betaproteobacteria) que en condiciones anaeróbicas acumula acetato y propionato, utilizando el ATP de sus reservas de polifosfato y glicógeno. Al cambiar a condiciones aeróbicas, el oxígeno disponible le permite respirar las reservas de acetato y propionato y reabastecer sus reservas de polifosfato tomando fósforo inorgánico (Pi) del medio (a través de su membrana plasmática por medio de un juego de transportadores de baja afinidad), eliminando el fósforo del agua. También cuenta con un juego de transportadores de alta afinidad que se espera se expresen sólo hacia el final de la fase aeróbica. De esta manera, fue posible hipotetizar que las condiciones de anaerobiosis o de microanaerobiosis ocasionarán una disminución en la precipitación de fosfatos por las comunidades BPR.

Biogeografía bacteriana y funcionalidad: el Global Ocean Sampling GOS

El proyecto biológico más ambicioso en lo que va del siglo XXI es sin lugar a dudas el Global Ocean Sampling (GOS), llevado a cabo por el J.C. Venter Institute[28]. En éste, se planteó muestrear 200 litros de agua cada 300 millas náuticas alrededor del mundo y secuenciar el metagenoma de la fracción de la comunidad cuyo tamaño estuviese comprendido entre los 0.8 y 1 μm. Este trabajo es comparable con los viajes de colecta de los naturalistas del siglo XIX, como el mítico viaje de Charles Darwin en el Beagle, o las exploraciones botánicas de Alexander Humboldt y Aimé Bonpland en el Continente Americano en cuanto a la cantidad de nueva información obtenida sobre nuevas especies.

La primera parte del proyecto publicada, que comprende 8,000 km entre el Atlántico Norte y el Pacífico Tropical del Este (desde Nova Scotia, Canadá, a través del Golfo de México, cruzando por el Canal de Panamá y hasta la Polinesia Francesa pasando por las Galápagos), representa el mayor aporte de secuencias nuevas y sin caracterizar a la base de datos del GenBank con 7.7 millones de fragmentos de ADN (reads) que corresponden a 6.3 mil millones de pares de bases (bp), 57% del cual representa secuencias únicas. En ellas se detectaron en total 811 ribotipos distintos entre 4,125 fragmentos diferentes del gen 16SrRNA con una similitud de 97%. De éstos, 52% representan ribotipos de los que no se tienen representantes actualmente en las bases de datos, es decir, especies nuevas, y alrededor del 16% son organismos nuevos a nivel de familia por lo menos. Funcionalmente, la predicción de proteínas de la base de datos aportó 6.12 millones de secuencias de proteínas, duplicando la cantidad de éstas conocidas en las bases de datos públicas, de las cuales un 23.4% representan familias de proteínas no descritas previamente y que se encuentran exclusivamente en la base de datos del GOS (es decir, existen otras familias sin caracterizar pero de las que se conocen homólogos en otras bases de datos[29]).

Pero la información aportada por el GOS no se reduce a una cantidad indescriptible de especies y proteínas nuevas, sino que aporta información fundamental sobre la estructura y composición de especies microbianas en las comunidades marinas de aguas oceánicas superficiales:

i) El 73% de los reads de 16S hallados corresponden a 60 de los 811 ribotipos, es decir, que las comunidades marinas se encuentran dominadas por 60 especies diferentes de las cuales sólo una no se había descrito anteriormente, pero la mayoría carece de miembros cultivables.

ii) Ninguno de los organismos dominantes pertenece al dominio Archaea, aunque previamente se había sugerido que eran dominantes en los océanos.

iii) De los 60 organismos dominantes, tan sólo cinco son ubicuos (es decir que están presentes en todas las muestras): dos son representantes del "grupo SAR11" (así llamado originalmente ya que no se habían descrito estas especies, actualmente se sabe que incluye a *Pelagibacter ubique*), otros dos pertenecen al "grupo SAR86" y el otro es cercano a SAR1.

iv) Las muestras tropicales se encuentran dominadas por varios grupos del clado SAR86, un grupo de Rhodospirillaceae y las cianobacterias unicelulares *Synechococcus* y *Prochlorococcus*.

v) Las muestras templadas se encuentran dominadas por grupos relacionados a Roseobacter RCA, SAR11 y otras gamaproteobacterias que se encuentran ausentes en las muestras tropicales.

Lo anterior no sólo describe la estructura de las comunidades marinas microbianas, sino que también sugiere la existencia de regiones biogeográficas bacterianas, fenómeno cuya existencia ha generado polémica desde su postulación[30]. Otro dato que lo fundamenta es el análisis de asociación por *clusters* de todo el contenido génico de la muestra, es decir, la comparación de todos los fragmentos de ADN y no sólo el 16SrARN de un sitio contra otro sitio. A pesar de que el GOS incluye muestras a ambos lados del continente americano, las muestras no se agrupan por distancia geográfica, sino por temperatura, es decir, las muestras de los mares templados del Atlántico se agrupan en un mismo lado, mientras que las muestras del Mar de los Sargasos y del Golfo de México y el Caribe se agrupan junto con las muestras tropicales del Pacífico, sugiriendo la existencia de Regiones Oceánicas Biogeográficas delimitadas.

Asimismo, el estudio presenta la oportunidad única de analizar

simultáneamente la funcionalidad y su distribución espacial, delineando la existencia de límites funcionales biogeográficos: de las 2,674 proteorhodopsinas detectadas en las muestras, se conocen tres variantes fisiológicas correspondientes a la substitución de un solo residuo de aminoácidos: las variantes con Metionina (M) o Leucina (L) presentan absorbancia máxima hacia el espectro verde, mientras que la variante con Glutamina (Q) presenta una absorbancia máxima hacia el espectro azul. En las muestras del GOS, se encontró que mientras la variante M se encuentra dispersa a lo largo de todas las muestras, la variante L domina las aguas templadas del Atlántico cercanas a las costas de US y Canadá, mientras que la variante Q es más abundante en aguas cálidas alejadas de las costas.

El Metagenoma de los simbiontes de un gusano: Interacciones y la nueva definición de organismo (*Olavius algarvensis*)

Una aplicación totalmente diferente del análisis metagenómico es demostrada por el análisis de la comunidad subcuticular asociada al gusano *O. algarvensis*[31], un anélido del grupo de los oligoquetos (el mismo que las lombrices de tierra), que viven en los sedimentos costeros marinos del Mediterráneo, y que tiene la particularidad de carecer de boca, ano y tracto digestivo. Lo anterior sugiere la existencia de una relación simbiótica obligada con una comunidad de microorganismos que habitan por debajo de la cutícula del gusano. Tradicionalmente, el análisis de interacciones simbióticas se limitaba a estudiar la simbiosis a nivel fisiológico de una pareja de organismos (*Rhizobium* en las leguminosas, las hormigas en las acacias, los peces mágidos con las anémonas), o de un par de poblaciones respecto a sus tasas poblacionales en presencia y ausencia del otro simbionte. El estudio metagenómico de la comunidad subcuticular permite analizar el potencial genómico de sus bacterias simbiontes extracelulares simultáneamente. El trabajo reveló la presencia de cuatro especies bacterianas simbióticas diferentes: dos delta y dos gama proteobacterias, todas con capacidad de fijación de carbono que provee al hospedero múltiples fuentes de nutrientes.

Los simbiontes gama-1 y gama-2 poseen ambos los genes necesarios para un estilo de vida quimioautotrófico mediante la oxidación del azufre, lo que representa el primer reporte de una asociación simbiótica con dos organismos metabólicamente equivalentes. La diferencia entre los dos radica en que gama-1 es capaz de almacenar azufre en glóbulos, mientras que gama-2 posee dos genes más para la oxidación del azufre, posiblemente brindándole una mayor versatilidad ecológica, y también posee genes para acoplar la oxidación de azufre a la reducción disimilatoria de nitrato en condiciones de oxígeno limitado.

De esta manera, se sabe que también los dos grupos de simbiontes se encuentran en simbiosis entre sí, generando un sistema cerrado donde los simbiontes gama realizan sulfo-oxidación y fijan carbono vía el ciclo de Calvin, y en oposición los simbiontes delta realizan sulfo-reducción acoplada a la oxidación del hidrógeno y fijan carbono vía la ruta del-acetil coA o el TCA inverso.

Respecto al gusano hospedero, las cuatro especies de bacterias simbiontes pueden aportar compuestos orgánicos sintetizados mediante la fijación de CO_2 y la incorporación de carbono orgánico disuelto del medio. También poseen la capacidad de sintetizar todos los aminoácidos y algunas vitaminas. La toma de los nutrimentos por parte del hospedero seguramente ocurre mediante la lisis y la digestión intracelular de las bacterias, pues no se encontraron exportadores de nutrientes en los genomas bacterianos y se ha corroborado la lisis de los simbiontes en la región epidérmica basal del hospedero. Asimismo, los productos de excreción del gusano son incorporados a las células bacterianas mediante importadores de urea y amonia, no sólo eliminando los compuestos tóxicos para el eucarionte sino reciclando compuestos ricos en nitrógeno. Por último, los genomas codifican una gran variedad de importadores de compuestos osmoreguladores orgánicos (glicinbetaína, taurina, TMAO) que se sabe son sintetizados por organismos osmoconformadores (es decir que cambia las concentraciones de sales dentro de su cuerpo para igualarlas al del medio) como *O. algarvenis*, y que son utilizados por los simbiontes como fuente de carbono y nitrógeno. Finalmente, el trabajo permite construir un modelo de comportamiento ecológico conjunto del sistema simbiótico, en el cual el gusano puede desplazarse de las regiones superficiales (ricas en oxígeno y sulfatos pero carente de ácido sulfídrico) en donde puede realizar sulfooxidación de las reservas de azufre del simbionte gama-1 utilizando oxígeno a la par que realiza respiración aeróbica; hacia la región media (rica en nitrato y sulfato y carente de oxígeno), en donde el simbiote gama-3 puede utilizar nitrato para oxidar compuestos reducidos de azufre que proveen los simbiontes delta mediante la sulforreducción, que en las dos regiones superficiales se comportan como heterótrofos. En las regiones más profundas (anóxicas carentes de nitrato pero ricas en compuestos reducidos de azufre), los simbiontes delta se comportan como autótrofos fijando CO_2 a partir de la oxidación del hidrógeno y los simbiontes gama realizan respiración anaeróbica de fumarato acoplada a la recarga de las reservas de azufre mediante la oxidación parcial de compuestos reducidos de azufre del medio.

Limitaciones y promesas de la metagenómica

Como toda subdisciplina naciente, la metagenómica presenta nuevos retos y problemáticas a resolver dentro del análisis ecológico de comunidades procariontes. El más obvio se refiere a la ausencia de estandarización en las metodologías de obtención de datos, causando una enorme heterogeneidad en la información contenida en los análisis metagenómicos (pero ver Raes *et al.*[32]). Se usan diferentes protocolos de extracción de ADN que varían en sesgos y eficiencia, se usan diferentes volúmenes de muestra, y varía mucho el número de células por muestra y cantidad de ADN total utilizada en la secuenciación, por lo que la comparación confiable entre metagenomas actualmente es muy difícil. Este

problema se amplifica por falta de costumbre para emplear metodologías estadísticas para "normalizar" las bases de datos durante las comparaciones (para evitar problemas relacionados con la cobertura y redundancia de secuencias). Todo lo anterior conduce a la subestimación de la similitud entre muestras, composición de la comunidad y potencial funcional.

Nosotros consideramos que la mayoría de las inconsistencias ecológicas presentadas por la metagenómica se deben a la naturaleza de la abundancia relativa de las especies contenidas, dado que el mismo método de secuenciación por *shotgun* es sensible a diferencias en las abundancias relativas, y muchas veces la especie ecológicamente más importante no es la numéricamente más abundante (dominante). Así, no es imposible que secciones importantes de la comunidad sean pasadas por alto.

Un problema similar ocurre con el análisis de variabilidad intraespecífica, que comprende la presencia de polimorfismos y elementos móviles, complicando el ensamble confiable de segmentos; a pesar de que la variabilidad es la esencia del estudio biológico y los biólogos y ecólogos microbianos debemos desarrollar nuevas aproximaciones para el manejo y valoración de la variabilidad dentro de una comunidad.

La secuencia del metagenoma de una comunidad representa el potencial funcional a disposición de la comunidad y no la verdadera funcionalidad de la misma, dado que la presencia de genes en un genoma no implica su expresión ni su traducción. Para una comprensión total de la funcionalidad bioquímica y biogeoquímica de una comunidad, es necesario realizar análisis de expresión de ARN, medir las tasas de traducción con variaciones temporales y finalmente la cinética enzimática de las proteínas más abundantes en condiciones naturales para poder realizar conclusiones respecto a la funcionalidad real de la comunidad. Sin embargo, el metagenoma es el único análisis de todos los mencionados que permite intuir cómo responderá la comunidad en condiciones medioambientales diferentes a aquéllas en las que se encuentra.

Otra dificultad técnica para el desarrollo de la metagenómica es que la secuenciación *en masse* dista mucho de ser barata. Todo proyecto metagenómico debe contemplar la secuenciación de grandes cantidades de ADN porque representan varias especies diferentes. Afortunadamente, los costos de secuenciación se abaratan a una velocidad impresionante gracias al desarrollo de nuevas tecnologías de secuenciación masiva como la pirosecuenciación (454 Life Sciences) o la secuenciación reversible basada en términos (Solexa/Illumina) (ver Cuadro 1). Sin embargo, la longitud de las secuencias que estas tecnologías producen es muy pequeña en comparación con los métodos de secuenciación tradicionales (Sanger). Esto representa un enorme problema cuando se intenta ensamblar las secuencias, pues si son cortas (de menos de 100 pares de bases) las regiones cortas no poseen suficiente información para asignarlas a una secuencia lineal única confiable.

El proceso de ensamblado de un genoma o peor, de un metagenoma, es análogo a reconstruir un rompecabezas: considerando que todas las secuencias de la comunidad son un rompecabezas completo, la extracción y secuenciación de ADN produce una imagen fragmentada y necesariamente incompleta de la imagen total. Las nuevas tecnologías de secuenciación permiten obtener más piezas individuales del rompecabezas, pero estas piezas son más pequeñas. Si tenemos una comunidad muy compleja, es probable que no podamos reconstruir ni siquiera algunas regiones del rompecabezas y nos quedemos solamente con piezas individuales que no son muy informativas por sí solas. Por ejemplo, la secuenciación de 71 millones de pares de bases del metagenoma de la comunidad simbiótica del tracto digestivo de las termitas (*Nasutitermes*)[33] se intuía sencillo, sin embargo resultó en una comunidad compleja con alrededor de 216 filotipos (grupos filogenéticos) dentro de 12 *phyla* diferentes. Esto ocasiona que los fragmentos individuales no puedan ser ensamblados pues de los 71Mbp, las secuencias lineales más largas tenían tan sólo 15 kbp (un genoma de una bacteria individual es de alrededor de 4 Mb). Por lo tanto, los esfuerzos tecnológicos deben dirigirse no sólo al mejoramiento de las tecnologías de secuenciación para obtener fragmentos más largos y de mejor calidad, sino también al desarrollo de algoritmos de cómputo alternativos para el ensamble de los fragmentos obtenidos, pues actualmente no tenemos la capacidad ni de cómputo ni de recursos humanos para realizar esta tarea. Esto nos lleva a mencionar otro posible problema: la capacidad de secuenciación aumenta mucho más rápido que la capacidad de analizar los datos producidos, de manera que hoy en día tenemos suficientes datos como para emplear los próximos diez años en su análisis.

La Metagenómica y la Nueva Ecología en Cuatrociénegas

Cuatrociénegas es un oasis en el Desierto Chihuahuense que presenta una serie de cuerpos de agua conocidos localmente como pozas, más de 300 cristalinas, muchas termales a 30-40°C, algunas con ricas en sales, en un valle que alberga una alta diversidad de todo tipo de organismos[34-38], sobresaliendo las comunidades formadoras de tapetes microbianos y estromatolitos, que en este ecosistema en particular son la base de la cadena trófica como lo eran hace 500 millones de años[37,39]. Esto probablemente se debe al aislamiento del valle y a que la pobreza extrema de fósforo en el ecosistema lo mantuvo pobre en algas y otros organismos cosmopolitas modernos. Los estromatolitos fósiles son la evidencia más antigua que tenemos de vida sobre nuestro planeta y que en este tipo de comunidad microbiana primigenia se dieron las condiciones metabólicas para la cohesión de los grandes ciclos biogeoquímicos (S, N, C, O) que hicieron posible la evolución de los eucariontes. Un hecho aún mas notable es que las redes tróficas (constituidas por

varios niveles que van desde los caracoles herbívoros hasta los peces carnívoros) se mantienen en este sitio al "filo de la navaja" nutricional debido a estas condiciones limitantes[39]. En nuestra búsqueda de la biodiversidad a nivel de comunidades y poblaciones encontramos que la microbiota constituye un ecosistema similar a ambientes del pasado de la tierra[36]. Sus organismos endémicos aparentemente han sobrevivido historias de cambios climáticos, manteniendo a las comunidades de estromatolitos como base de la cadena alimenticia. La mayor parte de los microorganismos secuenciados no sólo tienen una filiación marina, sino que son muy diversos y divergentes a todo lo conocido[36,38]. El análisis de sus comunidades acuáticas nos indica que no hay migración significativa entre las pozas, aun si éstas son similares químicamente y están a pocos kilómetros de distancia y hay un viento constante en el valle que podrían generar la migración. Los organismos de estas pozas tienen un alto endemismo y diferenciación ecológica (diversidad beta), es decir, cada poza tiene diversidad a nivel de comunidad diferente a la que sigue y la diversidad total debe de ser enorme. El que existan tantas especies en un sitio muy pobre en nutrientes libres y con muchas sales probablemente se debe a que no sólo el ambiente es fluctuante, sino a que los virus están eliminando predominantemente a los linajes más abundantes[40] explicando la distribución equitativa de las comunidades y el gran recambio observado entre comunidades, las cuales estarían sujetas a una selección dependiente de la frecuencia donde los linajes raros son exitosos, mientras no se vuelvan abundantes.

Consideramos que la obtención de datos metagenómicos en Cuatrociénegas es una excelente apuesta para poder comprender no sólo la diversidad, sino cómo la diversidad se liga a funciones ecosistémicas básicas como ciclaje de nutrientes (C, N, S, P, H y O) y la estructuración de la cadena alimenticia. La posibilidad de que el tapete microbiano de Cuatrociénegas sea un "ecosistema completo" (o sea que contiene muchos de los procesos o ciclos biogeoquímicos importantes), lo hace un sitio privilegiado para la metagenómica, ya que los metagenomas anteriores, incluido el GOS, nos han revelado la complejidad de la mayor parte de los ecosistemas, donde los ciclos biogeoquímicos requieren de grandes áreas y diversos compartimentos (por ejemplo: zonas con oxígeno y anóxicas). Por otra parte, creemos que en un sitio aislado y relativamente simple (dada su extrema oligotrofía) es más fácil entender la compleja red de la vida. Si a esto agregamos que en el valle de Cuatrociénegas es especialmente interesante en términos de diversidad y por sus características ambientales únicas que nos hablan de la tierra primitiva, un metagenoma, junto con genomas de sus especies dominantes/clave, nos puede dar elementos clave no sólo sobre su ecología, sino también sobre su evolución. En colaboración con el Laboratorio Nacional para la Genómica y Diversidad (LANGEBIO), el CINVESTAV-Irapuato, el Instituto de Biotecnología de la UNAM y la UAM-Cuajimalpa, estamos usando nuestra experiencia en el GOS[28,30], en el genoma de *B. coahuilensis*[3] y en las comunidades bacterianas de la región[36-40] para realizar un detallado estudio metagenómico de Cuatrociénegas. El proyecto ha resultado complicado por los problemas que representa extraer ADN ambiental de buena calidad en esta comunidad, pero creemos que va a constituir un estudio muy redondo dentro de la metagenómica y la Nueva Ecología, que esperamos poder revisitar en unos años en esta revista.

El futuro de la Nueva Ecología

A pesar de sus costos y problemas técnicos, la Nueva Ecología, y en particular la Metagenómica ecológica nos promete por fin lograr entender cómo funciona realmente el planeta y su biósfera: cuáles son los ciclos y procesos dominantes, los indispensables para la vida y los globales o cuáles son sus procesos más locales, qué organismos son los realmente importantes, especies clave para que operen los ciclos biogeoquímicos y para poder mantener la vida en la tierra y cómo funciona. Éste es un sueño que hasta ahora sólo lo imaginábamos, el abrir y entender qué hay dentro de la "caja negra" y qué representaban los microbios en los ecosistemas. En el 2009 vamos a celebrar el bicentenario del nacimiento de Charles Darwin, junto con otras celebraciones, como los 200 años de la publicación de la "Filosofía zoológica" de Lamarck, los 150 años de la publicación del Origen de las Especies y los 150 años de la muerte de Alexander Humboldt. La aplicación de las herramientas genómicas y metagenómicas nos permiten actualmente abordar el estudio de la biosfera desde el enfoque que propuso Darwin: el entendimiento de la naturaleza a partir de la visualización de la historia evolutiva de los organismos como producto de su funcionamiento con su entorno natural tanto biótico como abiótico. Es decir, nos permite comprender simultáneamente que la diversidad biológica actual es producto de las interacciones funcionales de los organismos con su entorno (el teatro ecológico) que se mantiene y que cambia a lo largo del tiempo (es decir, el drama evolutivo, siguiendo las ideas de Hutchinson[41]). La metagenómica nos permite analizar la historia evolutiva de los componentes individuales de las comunidades naturales actuales mediante el análisis evolutivo de los genes contenidos en ella utilizando las herramientas filogenéticas desarrolladas durante el siglo pasado[1,42], así como analizar las interacciones y funcionalidad de las comunidades mediante el análisis de los genes funcionales complementándolo con los enormes avances observados en bioquímica y fisiología.

Así como los naturalistas del siglo XIX necesitaban una sólida formación en la sistemática de plantas y animales, en el estudio de su morfología, en geología, climatología y cartografía, los nuevos naturalistas, esto es, los ecólogos del siglo XXI, necesitan sólidos conocimientos en la nueva sistemática y filogenia de los organismos, sin descuidar a los procariontes y virus, deben tener una buena formación en estadística, genética, biología molecular y bioinformática. Es claro que los nuevos ecólogos no pueden poseer todo el conocimiento de la biología actual como lo tenían los buenos naturalistas del siglo XIX, pero deben de tener una buena formación que les permita interactuar, entender y trabajar

armoniosamente en grandes equipos dentro de proyectos genómicos, de ecología molecular y metagenómicos. Así, tenemos que repensar en cómo formar una nueva generación de biólogos competentes tanto en los métodos de muestreo y diseños experimentales de la Ecología clásica, como en las técnicas de laboratorio de Biología Molecular y en los análisis básicos informáticos de las Secuencias de ADN y de proteínas. Para ello, necesitarán poseer conocimientos robustos en Ecología Teórica, Evolución, Bioquímica, Genética y Bioinformática, aunque ninguno de los planes de estudio actuales cubre un perfil como éste.

Efectivamente, los retos que enfrentamos los nuevos ecólogos son impresionantes, y requieren de una nueva visión y formación y los estudios van a ser complicados y costosos, pero proporcionalmente los riesgos, retos y costos económicos no son mayores que los que enfrentaron y resolvieron los naturalistas del siglo XVIII y XIX, como Linneo, Lamarck, Humboldt o el mismo Darwin. No somos tan listos, pero somos muchos y tenemos una sólida base gracias a sus trabajos pioneros.

Agradecimientos

Muchos colegas, amigos y alumnos han contribuido a nuestra formación, en nuestros proyectos y al desarrollo de estas ideas. Con la Dra. Luisa Falcón y el Dr. René Cerritos iniciamos nuestras meditaciones sobre metagenómica. En el desarrollo de los métodos de ecología molecular debemos mencionar a la Dra. Ana Escalante, a la M. en C. Laura Espinosa y de nuevo a la Dra. Falcón, que han trabajado intensamente muchos años con nosotros, tratando de avanzar en estos complicados problemas técnicos y conceptuales. La Dra. Erika Aguirre revisó el manuscrito. Diferentes proyectos del Conacyt nos han apoyado en nuestros estudios, especialmente el de Metagenómica (Conacyt SEP 57507) y otros (Semarnat-Conacyt 44673Q y Conacyt SEP-2004-C01-46475-Q) y una beca del Conacyt para que Germán Bonilla-Rosso realizara los estudios de doctorado.

Referencias

1. Eguiarte, L.E., Souza, V. & Núñez-Farfán, J. La revolución darwiniana y la evolución molecular. *TIP Revista Especializada en Ciencias Químico-Biológicas* **1(1)**, 8-16 (1998).
2. National Research Council. The role of theory in advancing 21st century biology: Catalyzing transformative research (The National Academies Press, Washington, D.C., 2008).
3. Alcaraz, L.D., *et al.* The genome of *Bacillus coahuilensis* reveals adaptations essential for survival in the relic of an ancient marine environment. *Proceedings of the National Academy of Sciences of the United States of America* **105(15)**, 5803-5808 (2008).
4. Tuskan, G.A., *et al.* The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313(5793)**, 1596-1604 (2006).
5. Warren, W.C., *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453(7192)**, 175-183 (2008).
6. Eguiarte, L.E., Souza, V. & Aguirre, X. Ecología Molecular (INE, SEMARNAT, CONABIO, UNAM, México, D.F., 2007).
7. Kluyver, A.J. & van Niel, C.B. The Microbe's Contribution to Biology (Harvard University Press, Cambride, MA, 1956).
8. Amann, R.I., *et al.* Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Applied and environmental microbiology* **56(6)**, 1919-1925 (1990).
9. Torsvik, V., Øvreas, L. & Thingstad, T.F. Prokaryotic Diversity—Magnitude, Dynamics, and Controlling Factors.*Science* **296(5570)**, 1064-1066 (2002).
10. Methé, B.A., *et al.* Genome of Geobacter sulfurreducens: Metal Reduction in Subsurface Environments .*Science* **302(5652)**, 1967-1969 (2003).
11. Bond, D.R. & Lovley, D.R. Electricity Production by Geobacter sulfurreducens Attached to Electrodes. *Applied and environmental microbiology* **69(3)**, 1548-1555 (2003).
12. Strous, M., *et al.* Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440(7085)**, 790-794 (2006).
13. Pace, N.R., *et al.* Analyzing natural microbial populations by rRNA sequences. *ASM News* **51**, 4-12 (1985).
14. Escalante, A.E. in Ecología Molecular (eds. Eguiarte, L.E., Souza, V. & Aguirre, X.) (INE, SEMARNAT, CONABIO, UNAM, México, D.F., 2007), pp. 393.
15. Auchtung, T.A., Takacs-Vesbach, C.D. & Cavanaugh, C.M. 16S rRNA Phylogenetic Investigation of the Candidate Division "Korarchaeota". *Applied and environmental microbiology* **72(7)**, 5077-5082 (2006).
16. Woese, C.R. Bacterial evolution. *Microbiological reviews* **51(2)**, 221-271 (1987).
17. Woese, C.R., Kandler, O. & Wheelis, M.L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America* **87(12)**, 4576-4579 (1990).
18. Hugenholtz, P., Goebel, B.M. & Pace, N.R. Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. *Journal of Bacteriology* **180(18)**, 4765-4774 (1998).
19. Konstantinidis, K.T. & Tiedje, J.M. Microbial diversity and genomics (John Wiley & Sons, New Jersey, 2004).
20. Handelsman, J., *et al.* Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology* **5(10)**, R245-R249 (1998).
21. Rondon, M.R., *et al.* Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. *Applied and environmental microbiology* **66(6)**, 2541-2547 (2000).
22. Magurran, A.E. Ecological Diversity and Its Measurements (Princeton University Press, Princeton, NJ, 1988).
23. Venter, J.C. *et al.* Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* **304(5667)**, 66-74 (2004).
24. Béjà, O., *et al.* Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea . *Science* **289(5486)**, 1902-1906 (2000).
25. Béjà, O., *et al.* Proteorhodopsin phototrophy in the ocean. *Nature* **411(6839)**, 786-789 (2001).
26. Ram, R.J., *et al.* Community Proteomics of a Natural Microbial Biofilm. *Science* **308(5730)**, 1915-1920 (2005).
27. García Martín, H., *et al.* Metagenomic analysis of two enhanced

biological phosphorus removal (EBPR) sludge communities. *Nature Biotechnology* **24(10)**, 1263-1269 (2006).
28. Rusch, D.B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology* **5(3)**, e77 (2007).
29. Yooseph, S., *et al.* The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS Biology* **5(3)**, e16 (2007).
30. Falcón, L.I., *et al.* Evidence of biogeography in surface ocean bacterioplankton assemblages. *Marine Genomics* (aceptado).
31. Woyke, T., *et al.* Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443(7114)**, 950-955 (2006).
32. Raes, J., Foerstner, K.U. & Bork, P.B. Get the most out of your metagenome: computational analysis of environmental sequence data. *Current opinion in microbiology* **10(5)**, 490-498 (2007).
33. Warnecke, F., *et al.* Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450(7169)**, 560-565 (2007).
34. Minckley, W.L. Cuatro Ciénegas fishes: research review of a local test of diversity versus habit size. *Arizona-Nevada Acad Sci* **19**, 13-21 (1984).
35. Minckley, W.L. A bibliography for natural history of the Cuatro Cienegas basin and environs, Coahuila, Mexico. *Proc Des Fishes Counc* **25**, 47-64 (1994).
36. Souza, V., *et al.* An endangered oasis of aquatic microbial biodiversity in the Chihuahuan desert. *Proceedings of the National Academy of Sciences of the United States of America* **103(17)**, 6565-6570 (2006).
37. Souza V. , Eguiarte, L.E., Siefert J. & Elser J. J. Microbial endemism: does phosphorus limitation enhance speciation? *Nature Reviews Microbiology* **6**, 559-564 (2008).
38. Escalante, A.E., *et al.* Diversity of aquatic prokaryotic communities in the Cuatro Cienegas basin. *FEMS Microbiology Ecology* **65(1)**, 50-60 (2008).
39. Elser, J. J., *et al.* Effects of PO4 enrichment and grazing snails on microbial communities in an ecosystem with living stromatolites. *Freshwater Biology* **50**, 1808-1825 (2005).
40. Desnues, C., *et al.* Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* **452(7185)**, 340 (2008).
41. Hutchinson, G.E. The Ecological Theater and the Evolutionary Play (Yale University Press, New Haven and London, 1965), pp. 139.
42. Eguiarte, L.E., *et al.* El análisis filogenético: Métodos, problemas y perspectivas. *Boletín de la Sociedad Botánica de México* **60**, 169-181 (1997).

CAPÍTULO I:
ANÁLISIS DE LAS METODOLOGÍAS PARA LA MEDICIÓN DE LA DIVERSIDAD TAXONÓMICA

Dada la complejidad intrínseca al concepto de diversidad, la elección de una métrica adecuada para la representación completa de la diversidad taxonómica en una comunidad natural ha sido siempre una tarea complicada porque depende de una definición unitaria de diversidad (Hurlbert 1971, Hill 1973). La primera métrica a considerar es naturalmente el número total de especies o **riqueza específica** como una medida de diversidad. Mientras más compleja sea una comunidad, más complicado será obtener una muestra representativa de ella que contenga todas las especies presentes en la comunidad, de manera que usualmente la riqueza específica mínima de una comunidad se estima a partir de la incidencia y abundancia de nuevas especies (e.g. Chao 1984, Apéndice 1). Sin embargo, es posible que dos comunidades con exactamente el mismo número de especies difieran en la abundancia relativa de los individuos dentro de cada especie, de manera que mientras que en una la mayoría de los individuos pertenecen a una sóla especie (y presenta una mayor **dominancia**), en la otra todas las especies tienen la misma probabilidad de ocurrencia (y presenta una mayor **equitabilidad**). El índice de dominancia de Simpson (Simpson 1949; Apéndice 1) estima la dominancia de una comunidad ponderando las especies más abundantes, y el índice de Berger-Parker considera exculsivamente la proporción de individuos en la comunidad que pertenecen a la especie más abundante (Berger & Parker 1970; Apéndice 1).

Cada una de éstas métricas observa sólo un aspecto de la diversidad, y por lo tanto son métricas unidimensionales que tienden a sobresimplificar la diversidad. El índice de Shannon es una métrica integrativa que utiliza tanto el número de especies como la distribución de sus abundancias para calcular el contenido de información o entropía de una comunidad (Shannon 1948). La elección entre éstas y muchas otras métricas para estimar y comparar la diversidad en comunidades naturales generó polémicas discusiones en la segunda mitad del siglo XX, que en el fondo versaban sobre la definición de diversidad también. Otras métricas más adecuadas pero menos sencillas de calcular e interpretar son las métricas multidimenisonales, familias de índices o perfiles de diversidad, que calculan diversos valores de diversidad en respuesta a un parámetro de escala que permite ponderar diferencialmente el número de especies y sus abundancia relativas (ver Apéndice 2).

Éstas métricas han sido evaluadas y aplicadas a la ecología de comunidades de macroorganismos, en donde los individuos son fáciles de clasificar y cuantificar. En las comunidades microbianas, generalmente se utilizan métodos indirectos para estimar el número de individuos, utilizando marcadores moleculares filogenéticos para cuantificar y clasificar simultáneamente a los individuos dentro de una comunidad. El gen 16SrRNA ha sido utilizado tradicionalmente en consecuencia de su uso extensivo en biología molecular. Sin embargo, dado el número variable de copias en el genoma, su extensa longitud y su heterogeneidad en la variabilidad a lo largo de su extensión, es posible que no sea el mejor marcador molecular para estudios de ecología de comunidades microbianas mediante técnicas metagenómicas. Otros genes conservados que se presentan como alternativa son aquellos que ocurren en copia única en el genoma y de manera ubicua en todos los organismos (Ciccarelli et al. 2006).

La aplicación de las métricas clásicas de diversidad a la ecología de comunidades microbianas utilizando técnicas metagenómicas promete un avance significativo tanto en el conocimiento de la diversidad bacteriana como de la teoría de ensamble y funcionamiento de las comunidades naturales. Sin embargo, su empleo con bases de datos metagenómicas potencialmente puede aportar nuevos sesgos que no se presentan con las matrices ecológicas de abundancia y presencia/ausencia, como la implicación del uso de secuencias de fragmentos de genes para la estimación de las abundancias relativas de los organismos, las diferencias en el tamaño del genoma de diferentes especies y el esfuerzo de secuenciación de cada metagenoma.

En el artículo del presente capítulo, se analizó mediante simulaciones metagenómicas el sesgo y precisión de diferentes métricas de diversidad que han sido frecuentemente exportadas de la ecología de comunidades de macroorganismos al análisis metagenómico (Apéndice 1). Asimismo, gracias al uso de juegos de datos simulados es posible explorar las posibles fuentes del sesgo observado. Se observó que las matrices de abundancia de especies que utilizan los genes codificantes de proteínas (Ciccarelli et al. 2006) son más similares a las comunidades originales reales que las matrices construidas a partir del gen 16SrRNA. Todas las métricas de diversidad estudiadas resultaron estar sesgadas respecto a los valores reales de las comunidades, imposibilitando su aplicación para comparaciones cuantitativas. El sesgo observado tiene su origen principalmente en un muestreo incompleto e insuficiente de las comunidades debido a la baja cobertura en el esfuerzo de secuenciación, y en la composición taxonómica de las comunidades en la muestra, aunque no se resuelve si ésto se debe a las diferencias en la representatividad en las bases de datos de ciertos grupos taxonómicos o en las diferencias en el tamaño de genoma. Finalmente, encontramos que las mejores comparaciones cualitativas ocurren al utilizar métricas multidimensionales y mejoran al incorporar medidas de distancia filogenética, y lo probamos con datos reales de comunidades halófilas en salinas como estudio de caso.

Éste primer capítulo revela que no es posible realizar comparaciones cuantitativas entre comunidades estudiadas mediante aproximaciones metagenómicas, pero que sí es posible realizar comparaciones cualitativas al utilizar métricas multidimensionales y filogenéticas, es decir incorporando la mayor cantidad de información. Asimismo, revela que para bases de datos metagenómicas es más recomendable construir las matrices ecológicas a partir de las abundancias relativas de 31 genes codificantes de proteínas y no del gen 16SrRNA. Ésta información sirvió como punto de partida para los análisis de diversidad presentados en el resto de la tesis.

RESEARCH ARTICLE

Understanding microbial community diversity metrics derived from metagenomes: performance evaluation using simulated data sets

Germán Bonilla-Rosso[1], Luis E. Eguiarte[1], David Romero[2], Michael Travisano[3] & Valeria Souza[1]

[1]Department of Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, México D.F., México; [2]Programa de Ingeniería Genómica, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, México D.F., México; and [3]Departement of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, MN, USA

Correspondence: Valeria Souza, Department Ecología Evolutiva, Instituto de Ecología, Apartado Postal 70-275, Universidad Nacional Autónoma de México, Ciudad Universitaria, México D.F., México.
Tel.: 52 555 622 9006;
fax: 52 555 616 1976;
e-mail: souza@servidor.unam.mx

Received 21 November 2011; revised 25 April 2012; accepted 27 April 2012.
Final version published online 25 May 2012.

DOI: 10.1111/j.1574-6941.2012.01405.x

Editor: Julian Marchesi

Keywords
community structure; diversity; dominance; metagenomics; microbial communities; simulations.

Abstract

Metagenomics holds the promise of greatly advancing the study of diversity in natural communities, but novel theoretical and methodological approaches must first be developed and adjusted for these data sets. We evaluated widely used macroecological metrics of taxonomic diversity on a simulated set of metagenomic samples, using phylogenetically meaningful protein-coding genes as ecological proxies. To our knowledge, this is the first approach of this kind to evaluate taxonomic diversity metrics derived from metagenomic data sets. We demonstrate that abundance matrices derived from protein-coding marker genes reproduce more faithfully the structure of the original community than those derived from SSU-rRNA gene. We also found that the most commonly used diversity metrics are biased estimators of community structure and differ significantly from their corresponding real parameters and that these biases are most likely caused by insufficient sampling and differences in community phylogenetic composition. Our results suggest that the ranking of samples using multidimensional metrics makes a good qualitative alternative for contrasting community structure and that these comparisons can be greatly improved with the incorporation of metrics for both community structure and phylogenetic diversity. These findings will help to achieve a standardized framework for community diversity comparisons derived from metagenomic data sets.

Introduction

In recent years, advances in the metagenomic analysis of microbial communities have been fuelled not only by decreasing sequencing costs, but also by the promise for the identification of general patterns in microbial community ecology. Metagenomics can significantly advance the study of community ecology by a simultaneous access to both functional and taxonomic diversity. It has already been applied to a wide range of environments (Rusch *et al.*, 2007), providing an unprecedented opportunity to identify ecological patterns in the structure and distribution of natural microbial communities (Kemp & Aller, 2004; Lozupone & Knight, 2007; Smith, 2007). Nonetheless, the estimation of taxonomic diversity has long proved to be a difficult task (Hurlbert, 1971; Hill, 1973; Venter *et al.*, 2004; Roesch *et al.*, 2007; Rusch *et al.*, 2007; Bent & Forney, 2008; Quince *et al.*, 2008; Shaw *et al.*, 2008; Sharpton *et al.*, 2011).

Historically, microbial community ecology has relied on SSU-rRNA genotyping as the standard approach, and many studies have estimated species richness directly from SSU-rRNA clone libraries (Roesch *et al.*, 2007; Fulthorpe *et al.*, 2008; Biers *et al.*, 2009). Although SSU-rRNA are powerful phylogenetic markers, the scattered distribution of hypervariable regions across its full length (~1500 bp) makes it very hard to recover comparable, phylogenetically informative fragments that are mutually overlapping (Mills *et al.*, 2006; Kembel

FEMS MICROBIOLOGY ECOLOGY

et al., 2011), and efforts have focused on circumventing this problem through the use of reference alignments and phylogenetic trees (Huson *et al.*, 2007; Rusch *et al.*, 2007; Berger *et al.*, 2011; Sharpton *et al.*, 2011). In addition, concerns have recently been raised against its use to study community structure because variability in gene copy number per genome can lead to biased estimations (Venter *et al.*, 2004; Biers *et al.*, 2009; Kembel *et al.*, 2011; Roux *et al.*, 2011). This is why attention has turned to the use of multiple single-copy, universally conserved protein-coding phylogenetic markers (Ciccarelli *et al.*, 2006; Wu & Eisen, 2008) as ecological proxies of community structure in metagenomic studies (Venter *et al.*, 2004; von Mering *et al.*, 2007; Rusch *et al.*, 2007; Biers *et al.*, 2009; Kembel *et al.*, 2011; Roux *et al.*, 2011).

The large number of microbial sequencing projects is stressing the need to develop new theoretical and methodological approaches to measure diversity across data sets (Rodriguez-Brito *et al.*, 2006; Huson *et al.*, 2009). While a wide range of diversity metrics have been used to compare microbial community richness (Roesch *et al.*, 2007; Schloss & Handelsman, 2008) and ranking (Hughes *et al.*, 2001; Shaw *et al.*, 2008; Youssef & Elshahed, 2009), testing their suitability to be used with microbial communities has received less consideration (Hughes *et al.*, 2001; Curtis *et al.*, 2002; Hill *et al.*, 2003; Quince *et al.*, 2008; Kuczynski *et al.*, 2010). To our knowledge, the applicability of macroecological diversity metrics has been evaluated mostly for SSU-rRNA clone libraries (Hughes *et al.*, 2001; Mills *et al.*, 2006; Bent & Forney, 2008; Shaw *et al.*, 2008; Youssef & Elshahed, 2009; Kuczynski *et al.*, 2010), and only the choice of ecological distances has been explored for metagenomic data sets (Mitra *et al.*, 2010). Furthermore, the use of mathematical models and computer simulated data sets for accurate evaluation of diversity metrics has been scarce (Curtis & Sloan, 2006; Sloan *et al.*, 2006; Green & Plotkin, 2007; Bent & Forney, 2008; Kuczynski *et al.*, 2010), even though it is not possible to test the efficiency of these metrics without knowing the real diversity in natural communities (Shaw *et al.*, 2008). To address this problem, we evaluated the applicability of widely used diversity metrics on a simulated set of metagenomic samples from nine source communities with contrasting structure and proposed a set of considerations for the qualitative comparison of the diversity in metagenomic data sets.

Materials and methods

To evaluate the applicability of macroecological diversity measures to metagenomic data sets, we chose to simulate the sequencing of nine theoretical microbial communities, estimate relative abundances from protein-coding phylogenetic markers and calculate diversity with canonical macroecological metrics. We generated other similar data sets to compare against and evaluate the effect of marker choice, taxonomic composition bias and sampling bias. A summary of the generation of matrices is presented as a flux chart in Fig. 1.

Design of source communities from completely sequenced genomes

As microbial ecology heavily relies on genomic molecular markers, the first step was to design *in silico* a set of theoretical, artificial microbial communities with contrasting diversity that will serve as the known template and starting point for the sequencing simulation. We took advantage of the availability of complete genome sequences from several microbial organisms deposited in public databases and randomly sampled them to construct these source communities (Supporting Information, Table S1). We assume that their relative abundance in the community is equal to the relative abundance of the genome in the community metagenome, so that all species included have only one genomic copy per genome and there is no polyploidy.

To better represent the multidimensional nature of diversity, each source community belonged to one of the three species richness levels (low: 10 species, medium: 100 spp., high: 500 spp.) and one of the three dominance levels. In the low-dominance level, all species had exactly the same number of individuals (a total-evenness scenario). The medium-dominance level was constructed in a way that four species equally contained half of the individuals in the community (one-eighth of the community each), for a scenario of four equally dominant species and a long tail of rare species. The high-dominance level was constructed so that only three species contained half of the individuals of the community, with one species containing one quarter of the community, and the other quarter shared by the other two species. This represents a scenario with one dominant species, two half-dominant and a long tail of rare species. A total of nine source communities were constructed as the result of the cross-product of all three richness levels and all three dominance levels (Fig. S1). The total number of individuals was kept to 1000 for all communities to standardize dominance comparisons, and the abundances were calculated as proportions of the total community, so that the dominance level was conserved across different richness levels. To avoid taxonomic biases, dominance was modified over the same taxa, in a way that community composition at the lower richness levels are a subset of the higher richness levels.

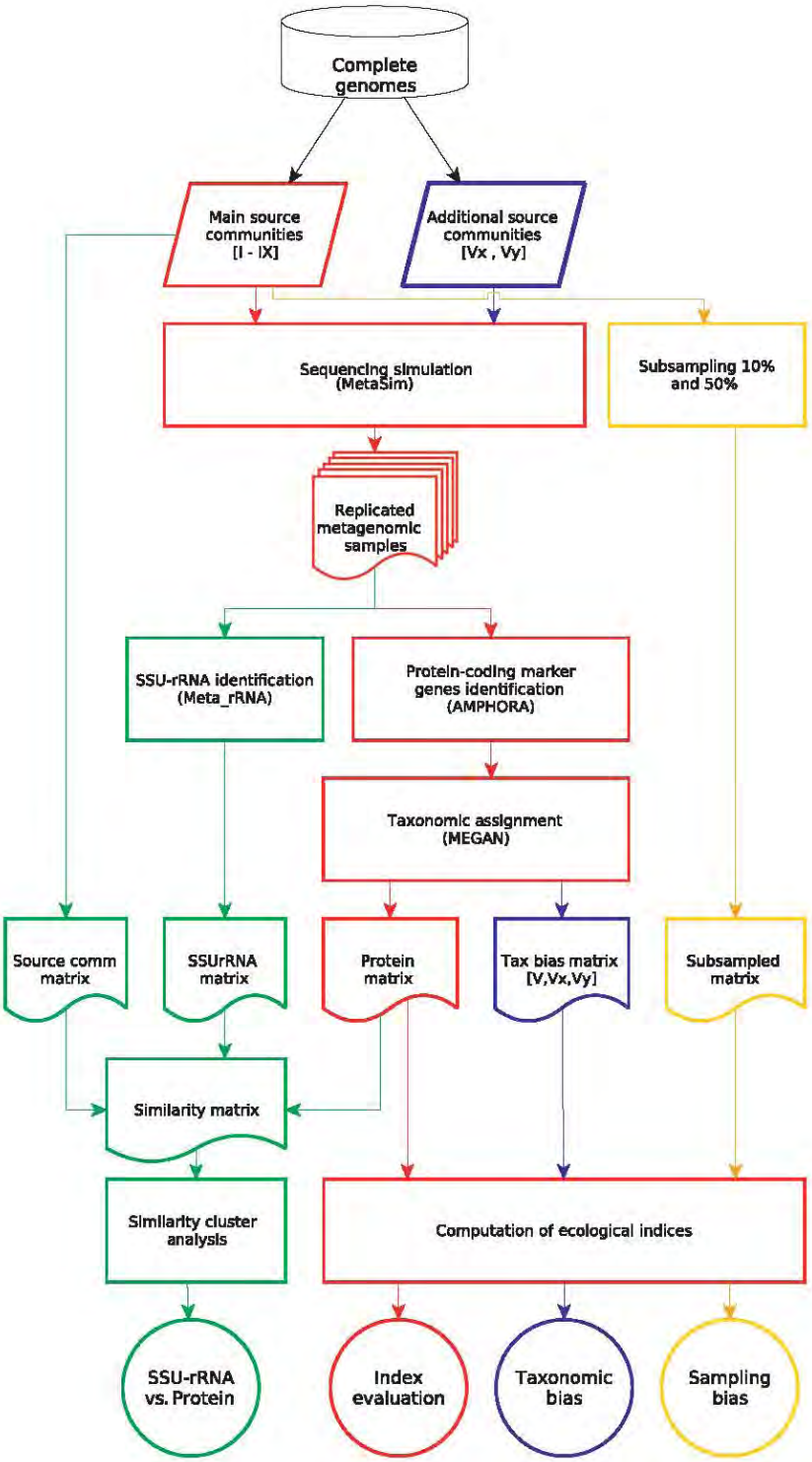


Fig. 1. A flowchart illustrating the main steps in the methodology towards the comparison of diversity metrics calculated from protein-marker matrices derived from simulated metagenomes (red), its contrast against SSU-rRNA derived matrices (-green) and the evaluation of sampling (yellow) and taxonomic (blue) biases.

Metagenomic data sets sequencing simulation

We next simulated the pyrosequencing of each source community with the sequencing simulator software MetaSim (Richter *et al.*, 2008). Briefly, MetaSim generates a set of synthetic sequencing reads (a metagenomic data set) from a species-abundance matrix and a database of the complete genomes, according to the characteristics and error models produced by different sequencing technologies (Richter *et al.*, 2008). We used our source communities as the species-abundance matrices input and simulated five pyrosequencing replicated runs for each source community (450 000 reads each, error model = 454, read length = ~250 bp, distribution mean = 0.23, distribution SD = 0.15, proportionality constant = 0.15, scale SD with square root of mean = true, error clone distribution = normal, error clone mean = 2000, 2nd parameter = 200). Each resulting simulated metagenome replica was corrected for pyrosequencing noise with CDHIT-454 (Li & Godzik, 2006), and ORFs were predicted and translated into proteins with GeneMark (Lukashin & Borodovsky, 1998).

Construction of community matrices

Each translated protein sample replica was scanned for 31 universally conserved, single-copy, protein-coding genes with AMPHORA (*dnaG*, *frr*, *infC*, *nusA*, *pgk*, *pyrG*, *rplA*, *rplB*, *rplC*, *rplD*, *rplE*, *rplF*, *rplK*, *rplL*, *rplM*, *rplN*, *rplP*, *rplS*, *rplT*, *rpmA*, *rpoB*, *rpsB*, *rpsC*, *rpsE*, *rpsI*, *rpsJ*, *rpsK*, *rpsM*, *rpsS*, *smpB*, and *tsf*, Table S2; Wu & Eisen, 2008). These genes are commonly used as taxonomic molecular markers because they are phylogenetically informative, and because they are single-copy in the genomes, we can use them as ecological proxies as an indirect measure of the relative abundance of their species of origin. Each of the identified protein-marker fragment was assigned to a taxonomic category using the last common ancestor (LCA) algorithm implemented in the metagenomic analysis software MEGAN (Huson *et al.*, 2007; Min Support = 1, Min Score = 35, Top Per cent = 3). This software employs the phylogenetic information within the top best BLAST hits of each fragment against the nonredundant protein database and the NCBI taxonomy tree to assign each fragment to a taxonomic category. Once all reads were classified, we used a parsing script to summarize the total number of reads within each taxonomic category in each metagenomic sample in the form of a community or species-abundance matrix.

Diversity metrics calculation

All the canonical macroecological diversity metrics in this work are estimated from ecological distance matrices. We used the protein-marker matrices obtained in the previous section to calculate the Hellinger transformation of ecological distances (Eqn 1), both because the Hellinger distances are more representative of real ecological distance (Legendre & Gallagher, 2001) and because their use with metagenomic data sets has already been evaluated with positive results (Mitra *et al.*, 2010).

Equation (1) Hellinger's Distance

$$D = \sqrt{\sum_{i=1}^{S}\left(\sqrt{\frac{x_i}{\hat{x}}} - \sqrt{\frac{y_i}{\hat{y}}}\right)^2}, \hat{x} = \sum_{i=1}^{S} x_i$$

where x_i = abundante of ith species at site x; y_i = abundante of ith species at site y.

These ecological distance matrices were in turn used to calculate diversity metrics commonly used in macroecology. The richness estimators used were observed richness (S), the nonparametric richness estimator *Chao1* (Chao, 1984), and abundance-based coverage estimator *ACE* (Chao & Lee, 1992). Dominance-based diversity metrics used were Simpson's probability that two randomly sampled individuals belong to the same species (D; Simpson, 1949), and Berger–Parker's proportion of the most abundant species (*BP*; Berger & Parker, 1970). Metrics that incorporate both richness and dominance used in this work are Shannon's diversity index (H; Shannon, 1948), and its derived evenness metrics J and E (Kindt & Kindt, 2008) and Fisher's alpha (α) parameter for a log-series fitted species-abundance curve (Fisher *et al.*, 1943).

All the previous metrics are only point descriptions of diversity (Hurlbert, 1971; Hill, 1973), while parametric diversity families provide a more complete, multidimensional summary of community diversity (Hill, 1973; Patil & Taillie, 1982; Ricotta, 2003). Rényi's entropy profiles (Rényi, 1961) are a generalization of Shannon's informational measure extrapolated to particular moments of the same function with a scale parameter (alpha = 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, infinity) that reflects the partition of abundance between species, constituting the best representation of a continuum of possible diversity measurements (Ricotta, 2003). In consequence, Rényi's metrics span from richness to dominance, across approximations to most of the individual metrics previously mentioned (Fig. S4). All Rényi's profiles were calculated as in Eqn (2) after Tóthmérész (1995). All ecological and statistical analyses were performed in R with packages VEGAN (Oksanen *et al.*, 2007) and BiodiversityR (Kindt & Kindt, 2008).

Equation (2) Rényi's Entropy

$$H\alpha = \frac{\ln\left(\sum_{i=1}^{S} p_i^{\alpha}\right)}{1-\alpha}$$

Where p_i = relative abundante of ith species; α = scale parameter.

Evaluation of individual diversity indices

To assess the performance of each diversity metric relative to the true diversity parameter values from their community of origin, each index was tested against their corresponding value from the source communities for significant differences. The 'real' ecological distance matrices were constructed from the raw source-community species-abundance matrices, and the 'real' diversity metrics of the original communities were calculated from these as described in the previous section. We then tested for statistically significant differences between the estimated diversity metrics (calculated from the replicated metagenomic data sets) and the real diversity values (calculated from the source communities) with a *T*-test for single samples, using the values from the replicated metagenomic data sets as observations and the values from the source communities as the population parameter (Sokal & Rohlf, 1995). Samples were then ranked according to the values of each diversity metric and compared against the ranking obtained from their respective source community (Table 1).

Choice of molecular marker as ecological proxy

To evaluate whether protein-coding genes are superior to SSU-rRNA as ecological proxies of the original community, we scanned each untranslated sample from the replicated metagenomes for SSU-rRNA gene fragments using Meta_RNA, a high-sensitivity algorithm for the detection of ribosomal metagenomic fragments using hidden Markov Models (Huang *et al.*, 2009). To date, there is no consensus on the best methodology and reference database to taxonomically classify complete SSU-rRNA genes, let alone fragmented sequences (McDonald *et al.*, 2011; Sharpton *et al.*, 2011), and their choice can profoundly affect the resulting community matrices. One of the advantages of using simulated metagenomes is that we can track each of the SSU-rRNA fragments back to their genome of origin, allowing us to reconstruct a species-abundance matrix without incorporating the selection of a classifying method as an additional confusion factor. As this results in a highly confident classification of the identified fragments, it gives the SSU-rRNA matrix in this work an advantage over the protein matrix, but we chose this comparison for the sake of simplicity. A matrix of ecological distances was constructed between all the metagenomic SSU-rRNA matrices, protein-marker matrices and the original source-communities matrices (derived from the raw, 'real' data without simulation). The similarities between samples were analysed with a hierarchical cluster analysis by complete linkage as implemented in the CLUSTER package in R (Maechler *et al.*, 2002), and the distances between the SSU-rRNA and protein-marker matrices to their corresponding source communities were tested for statistical differences with a completely randomized block design for ANOVA in R (R Development Core Team, 2006).

Taxonomic composition biases

To analyse the effect of taxonomic composition bias, two additional communities were built with the same community structure as sample 'V' but the dominant species were randomly shifted from the pool of available complete genomes to modify community composition (Table S1). This allowed us to compare three communities with exactly the same diversity but different taxonomic composition. The two resulting source communities (Vx and Vy) were subjected to the exact same procedures as the others as described above, and their diversity metrics compared against sample V.

The mean pairwise phylogenetic distance (MPD; Webb *et al.*, 2002) is a diversity measure that explicitly incorporates differences in community structure, and it was determined between all members in each community following the procedure presented in Kembel *et al.* (2011) using the R package PICANTE (Kembel *et al.*, 2010). Briefly,

Table 1. Rank ordering according to diversity indices values of source communities (S) and metagenomic samples (M)

Rank	Shannon		Simpson		Logalpha		Berger-Parker		Jevenness		Eevenness		Chao1	
1st	**S1**	**M1**	**S1**	**M1**	**S1**	**S1**	S6	M3	S7	M1	S7	M1	S1	M3
2nd	**S4**	**M4**	**S4**	**M4**	S3	S4	S3	M6	S8	M4	S8	M4	**S2**	**M2**
3rd	**S2**	**M2**	**S2**	**M2**	S2	S3	**S9**	**M9**	S4	M7	S4	M7	S3	M1
4th	S3	M5	**S5**	**M5**	S4	S2	**S5**	**M5**	S9	M8	S9	M8	**S4**	**M4**
5th	S5	M3	**S7**	**M7**	S5	S5	S2	M8	S1	M9	S1	M9	**S5**	**M5**
6th	**S6**	**M6**	**S8**	**M8**	S6	S6	S8	M2	**S5**	**M5**	**S5**	**M5**	**S6**	**M6**
7th	**S7**	**M7**	S3	M6	**S7**	**S7**	**S7**	**M7**	S6	M2	S6	M2	S7	M8
8th	**S8**	**M8**	S6	M3	**S9**	**S8**	**S4**	**M4**	S2	M6	S2	M6	S8	M7
9th	**S9**	**M9**	**S9**	**M9**	**S8**	**S9**	**S1**	**M1**	**S3**	**M3**	**S3**	**M3**	**S9**	**M9**

Ranks conserved in both cases are shown in bold.

each protein-marker gene fragment was aligned to a concatenated reference alignment and then placed onto a reference phylogeny using the evolutionary placement of short sequences implemented in RAxML v.7.2.8 (Berger *et al.*, 2011). MPD was then calculated from this phylogenetic tree. The reference phylogeny was calculated via maximum likelihood with a WAG+G model partitioned by gene families from the reference alignment provided in Kembel *et al.* (2011). Statistical differences in MPD were calculated with ANOVA. In addition and because average genome size is deeply affected by the taxonomic community composition, the effective genome size (EGS) was calculated from the protein-marker abundance matrices following the methodology in Raes *et al.* (2007).

Incomplete sampling bias

An important source of bias that is unrelated to the methodology evaluated here is incomplete sampling of the natural community. Because no complex natural community has been sampled to exhaustion (to our knowledge), the effects of these kinds of bias are of the greatest importance. To separate the bias observed because of incomplete sampling from methodological bias, two additional species-abundance matrices were constructed by randomly sampling 10% and 50% of the individuals directly from the source communities, without a sequencing simulation. These samples were processed to obtain Rényi diversity profiles in exactly the same way that has been previously described.

Results and discussion

The comparison of microbial community structure by means of metagenomic data sets relies on the estimation of diversity from abundance matrices. While this comparison is promising for testing ecological hypotheses, in practice, the construction of accurate abundance matrices from metagenomic data sets is challenging and far from being standardized. To address this problem, we designed nine source communities with contrasting structure and simulated the sequencing of five replicas from each. Next, we took the advantage of the fact that we knew the real values of the diversity metrics parameters from the source communities and evaluated the performance of its estimators by contrasting them against the estimated values from the metagenomic samples.

On the type of molecular markers as ecological proxies

The first step towards contrasting communities is the construction of the abundance matrix, and so the choice of molecular markers as ecological proxies for species abundances is fundamental. Previous studies have used SSU-rRNA gene clone libraries and metagenomic fragments (Kemp & Aller, 2004; Edwards *et al.*, 2006; Mills *et al.*, 2006; Roux *et al.*, 2011), conserved protein-marker genes (Kembel *et al.*, 2011; Roux *et al.*, 2011) and even all metagenomic reads (Edwards *et al.*, 2006) as ecological proxies to address community structure and composition. Here, we compared the performance of abundance matrices built from SSU-rRNA fragments or from protein markers recovered from the metagenome sample data sets to reflect the real structure of the source communities.

Overall, the protein-marker matrices were consistently more similar and showed smaller ecological distances (mean = 0.45) to the source communities than the SSU-rRNA matrices (mean = 0.50; Fig. 2). SSU-rRNA were directly classified by their genome of origin, and are free of other common sources of error such as misalignment, misclassification and a lower resolution for detecting taxonomic groups (Roux *et al.*, 2011). This means that even if we had error-free classification methods for SSU-rRNAs, the protein-marker gene matrices would still be more similar to the real source community structure. The exception is sample I, where the SSU-rRNA matrices were more similar to the source communities than the protein matrices. Sample I has the higher richness and evenness, and although this could indicate

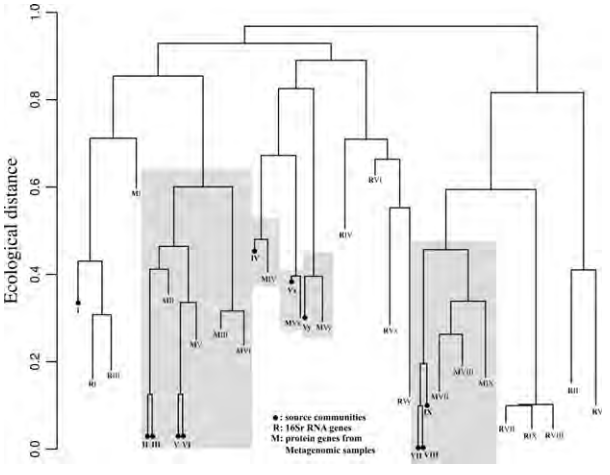


Fig. 2. Dendogram resulting from the complete cluster analysis based on Hellinger distances between source communities (black dots numbered 1–9), rRNA gene abundance matrices (R1–R9) and matrices derived from the 31 protein genes (M1–M9). Agglomerative coefficient = 0.82. The mean distance from protein gene matrices to source communities is 0.45. The mean distance from rRNA gene matrices to source communities is 0.50. This difference in distance is statistically significant (d.f. = 1, $F = 14.055$, $P < 0.01$; $\text{d.f.}_{\text{blocks}} = 10$, $F_{\text{blocks}} = 78.35$, $P_{\text{blocks}} < 0.01$). Source communities are marked with circular tips.

that SSU-rRNA matrices perform better with very complex communities, a most plausible interpretation is that the effect of classification bias on protein matrices is more strongly revealed in complex communities. Misclassification and low resolution of reference databases are prone to modify precisely the relative dominance of closely related clades, affecting samples with large numbers of species and high evenness. We would expect then that sample I would be more strongly affected if our SSU-rRNA matrices were subjected to a classification algorithm. This finding supports the choice of universally conserved, single-copy protein-coding marker genes over SSU-rRNA genes for the estimation of diversity metrics.

Evaluation of diversity metrics

With the 31 protein-marker matrices, we calculated the most commonly reported diversity metrics and Rényi's entropy profiles for each of the metagenomic samples. Because five metagenomic samples were produced by the sequencing simulation replications from each source community, we were able to directly compare the estimated values of each diversity metric (from the sample replicas) against their corresponding known community parameter (from the source community). None of the metrics estimated were statistically similar to their corresponding parameter from the source communities ($P > 0.05$; Table S3), and their results were inconsistent across samples. This means that the particular values for individual diversity metrics from metagenomic data sets differ quantitatively from the ones derived from the real, known community structure. It has been shown that some ecological problems can be approached by qualitative relative measures of diversity like ordering a set of samples according to their diversity rankings relative to each another (Shaw *et al.*, 2008). The ordering and ranking of communities according to individual diversity metrics has already been applied in microbial ecology studies using clone libraries (Hughes *et al.*, 2001; Shaw *et al.*, 2008; Youssef & Elshahed, 2009) and also metagenomes (Biers *et al.*, 2009). However, no individual diversity index recovered the same ranking from their corresponding source community, and the inconsistency of the ranking across different indices prevented us from achieving a consensus ranking (Table 1). This can be attributed to the fact that individual metrics are only point descriptors of particular aspects of diversity, and so a bias in their estimation will result in an erroneous ranking of the samples. Hence, the use of metrics that explores the multidimensional aspects of diversity (Preston, 1948; Hill, 1973) appears as a better option to compare communities. We chose Rényi's entropy profiles (Rényi, 1961) because it clearly depicts diversity graphically (Tóthmérész, 1995), but other possible alternatives are Hill's numbers (Hill, 1973), Patil and Taille's parameter families (Patil & Taillie, 1982), and even a combination of individual metrics that measure richness and different degrees of weight to dominance and richness like the *Chao1* and *BP* indices. Although also biased, the relationship between each pair of source communities Rényi's profiles (Fig. 3a) is faithfully reflected by the relationships of the metagenomic samples (Fig. 3b). Samples are difficult to rank using Rényi's profiles because one sample can be more diverse in one scale and less diverse in other (Tóthmérész, 1995) as the case of samples II and IV in Fig. 3a, but their strength relies on their ability of depicting exactly that complex relationship between the two samples, where sample II has a larger richness than IV, but it has a larger dominance than the even sample IV. An analysis of the Rényi's profiles from our samples reveals that the inconsistencies observed at the ranking with individual indices are caused by real differences in the community structure. Moreover, our results indicate that the relative positions between samples are more faithfully reflected when replicated data sets are pooled together as shown in Fig. 3d. In summary, ranking by single-diversity metrics might not be sufficient to accurately compare the diversity in two communities, and we suggest the use of multidimensional metrics to describe the rankings at different scales of diversity that might be differentially affected during manipulative studies.

Possible sources of estimation bias

There are three factors expected to cause the majority of the estimation bias observed in metagenomic data sets: DNA extraction and sequencing, choice of molecular marker selected as ecological proxy and the effect of an insufficiently sampled community. Biases in DNA extraction are beyond the scope of this work and have been addressed elsewhere (Morgan *et al.*, 2010; Lombard *et al.*, 2011), and because our metagenomic data sets were simulated *in silico,* they are free from this bias. To differentiate biases introduced by the methodology and the effect of subsampling, we constructed abundance matrices by sampling 10% and 50% of the individuals in the source communities directly, without sequencing simulation or taxonomic classification (Fig. 3c). The effect of subsampling is similar to the patterns of general reduction in diversity, and sample aggregation observed in the metagenomic data sets (Figs. 3b and 3c). A much clearer separation among samples is observed when 50% of the source community is sampled (Fig. 3e). Unfortunately, the fraction of the community present in any given sample is very hard to estimate for natural communities, and the definition of the number of sequences required

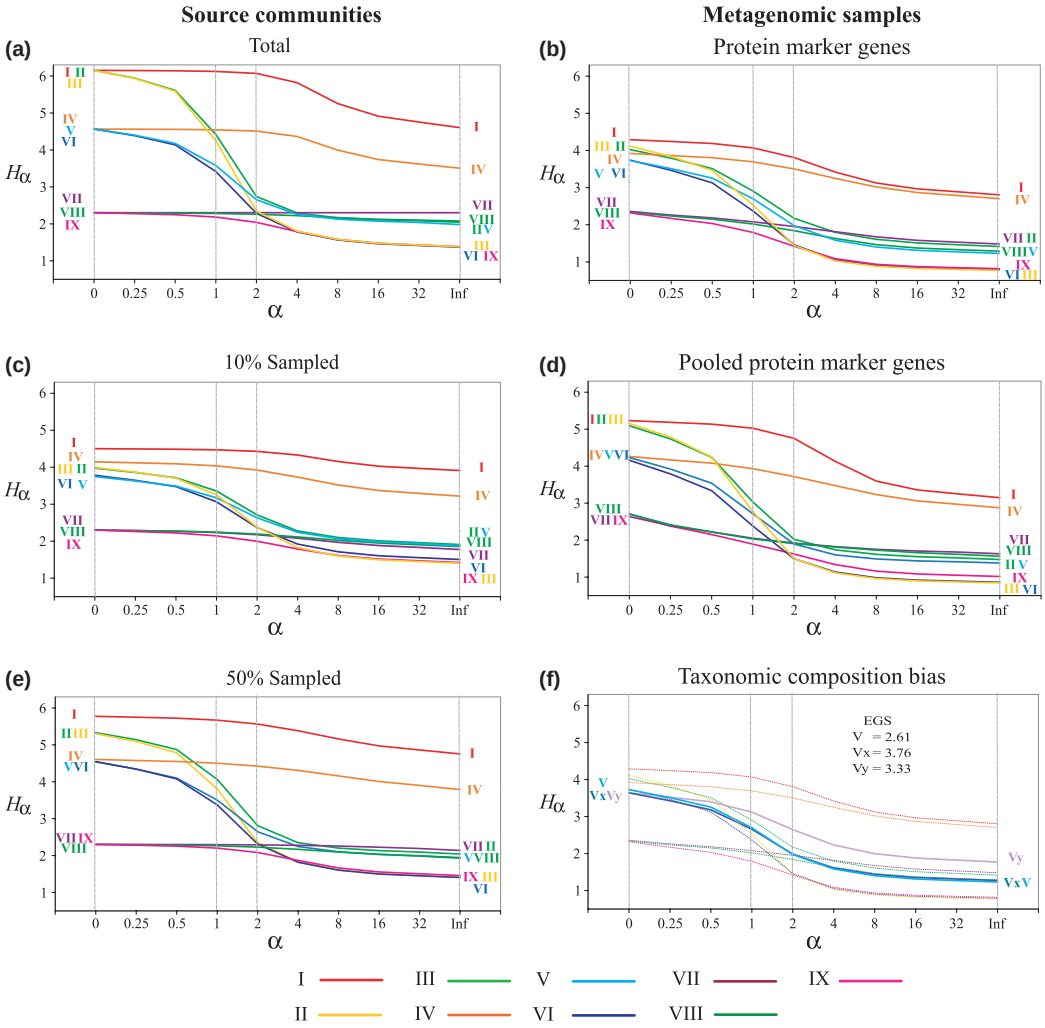


Fig. 3. Rényi's entropy profiles for (a) the source communities; (b) the matrices derived from the averaged 31 protein-marker genes; subgroups of the source communities where only (c) 10% or (e) 50% of the community was sampled; (e) the pooled replicas from the 31 protein-marker gene matrices; and (f) the protein-marker abundances of samples with different taxonomic composition. Samples Vx and Vy in (d) have exactly the same structure as V, but the species showing larger abundances are different. The rest of the samples in (d) are attenuated with dotted lines as they are given only as reference (EGS). The richness, Shannon, Simpson and Berger–Parker indices can be conceived as single moments of the entropy function, and are marked with vertical dashed lines over (α = 0, 1, 2, infinity), respectively.

to obtain a representative data set is one of the major challenges in metagenomic research. Quince *et al.* (2008) suggested that a slight increase in sequencing effort would produce a significant increase in coverage in moderately complex communities. We observed that richness categories can already be differentiated with the pooling of only two samples, each sample being roughly equivalent to the sequencing of one plate in the 454-FLX platform (Fig. S3). Moreover, samples are readily separated by their diversity profiles when the five simulated replicas are pooled together (Fig. 3d). This suggests that most of the confusing factors observed are due to subsampling, which is promising because this is expected to be less of a problem in the future with the decreasing costs of sequencing technologies.

Another potential source of bias for comparing community structure with metagenomics comes from phylogenetic community composition. This arises from the fact that the probability of sequencing any given molecular marker is a factor of the density of that marker in its genome of origin, which in turn depends of the genome size of each particular organism (Raes *et al.*, 2007). This problem is exclusive of metagenomic data sets because other approaches are usually based on the direct observations of species from individual counts. Beszteri *et al.* (2010) proposed that single-copy protein genes suf-

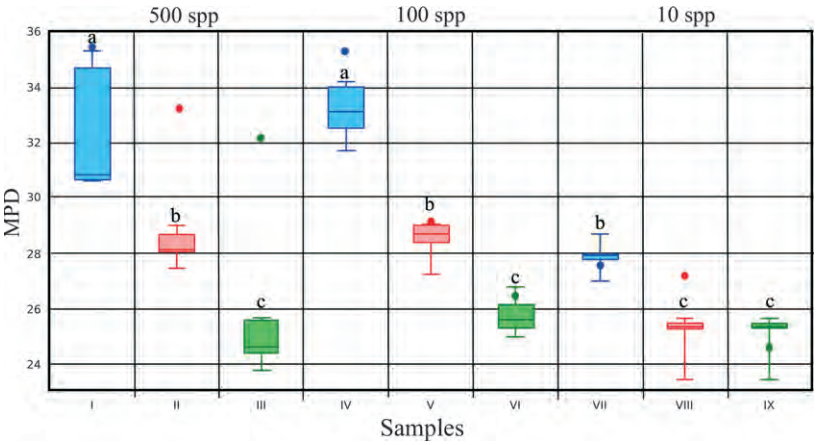


Fig. 4. Box-and-whisker plot comparing the mean pairwise phylogenetic distance values observed for the protein-marker matrices in each sample. Midlines represent the median and box limits represent the first and third quantiles, while whiskers are the maximum and minimum values and bold circles mark the corresponding source-community value. Samples are grouped by their source community richness category in bold black boxes and by their evenness by colours as follows: total evenness, blue; medium evenness, red; high dominance, green. Letters above the boxes denote membership to the statistically significant groups obtained by a *post hoc* Tukey multiple comparison with 95% of confidence.

fer from a reduced sampling probability in metagenomic data sets (as a 'dilution effect'), as they are directly affected by the mean genome size of all individuals in the community (measured as the EGS, Raes *et al.*, 2007). To address this issue, two additional source communities with identical community structure to sample V, but with different taxonomic composition (samples Vx and Vy) were built. Samples V and Vx are more similar than sample Vy, and significant differences were observed between the three samples, with evenness being more profoundly affected than richness (Fig. 3f). The observed pattern is precisely what we would expect from the dilution effect, but the EGS sample ordination does not follow the observed diversity pattern, Vy being the middle value between V and Vx (Fig. 3f). As all the dominant species in these samples belong to different bacterial phyla, this suggests that there are phylogenetic factors other than EGS affecting the estimation of community structure metrics. Our approach is not suited to address these factors, but the variable phylum representation in the reference protein databases is most likely to affect the resolution for classification and relative abundance estimation.

Because we used phylogenetically informative molecular markers as ecological proxies, it seems natural to incorporate that very same phylogenetic information into diversity metrics. Again, we measured the MDP, but other alternatives are available (Cadotte *et al.*, 2010). Differences in evenness were corroborated by variations in MPD; for instance, group *a* (I–IV) had a large MPD that was explained by a large evenness in the Rényi profile (Fig. 4), but these two samples differed in their richness. Samples VII and VIII were undifferentiated by the Rényi profile, but could be separated by the MPD and showed that although the structure was very similar in both samples, a greater clustering was observed in sample VIII. The metagenomic samples are separated by diversity metrics when they are first grouped according to their richness category and then by their MPD category, effectively reflecting the ranking of source communities by their structure (Fig. 4). Although the estimated values were statistically different from that of the source communities, the grouping of samples by MPD reflected the evenness categories from the source communities (Fig. 4). The MPD values from samples in the low richness category (i.e. samples VIII and IX) are equivalent to samples in the medium and high-dominance categories (i.e. samples III and VI), most likely because MPD is also affected by richness (Kembel *et al.*, 2010). These results suggest that measures of phylogenetic diversity can further differentiate communities by their composition and that these values naturally reflect the structure of the community and so can help differentiate samples that have not been differentiated by other multidimensional metrics that do not consider community composition.

It should also be noted that because these simulated metagenomes were constructed using known genomes, these comparisons are a best-case scenario. The diversity metrics resulted in biased estimations even under these optimal conditions, so it is reasonable to expect greater biases with real data sets where the majority of the species are only distantly related to known organisms with sequenced genomes (a case study with real metagenomic data can be found in Fig. S2). Furthermore, misclassifica-

tion errors are expected to be reduced with the advancement of classification algorithms, the availability of sequencing technologies that deliver longer sequencing reads and the phylogenetic expansion of the reference genomes, and these fields have shown significant improvements in recent days (Wu *et al.*, 2009; Ghosh *et al.*, 2010; Meinicke *et al.*, 2011; Parks *et al.*, 2011; Pati *et al.*, 2011). In the meantime, the LCA algorithm allows for reads from organisms that are phylogenetically distant from reference genomes to be only be assigned to high taxonomic ranks, so that a more accurate community structure estimation can be achieved sacrificing phylogenetic resolution. In practice, this means that more representative abundance matrices can be built from metagenomes with phylogenetically uncharacterized members if they are built at genus or family level instead of species level.

A last source of biases that was not addressed here is precisely the combined effect of uncharacterized species in a highly complex community, and we recognize that the behaviour of both diversity metrics and choice of ecological proxy might change at higher complexity. Nevertheless, these differences are difficult to address because no complex community metagenomes have been sequenced to exhaustion. Until then, these problems can be only addressed with the comparison of natural communities where contrasting levels of diversity can be presumed with confidence (Fig. S4).

Conclusion

Modern microbial ecology needs new tools to quantify microbial diversity in a statistically realistic fashion, if we expect to identify general patterns of community structure, composition and assemblage. Moreover, we need to distinguish true patterns from possible artefacts caused by the massive amounts of fragmentary data whose statistical properties are poorly understood and are possibly biased because of genetic, biological and sampling factors. Although it is natural to borrow ecological methods directly from macroecology, microbial ecologists should adjust or develop and evaluate tools and methodological practices, in a way that properly fits the biological and ecological properties of natural microbial communities.

Diversity is a complex community property, and this study illustrates the need to carefully evaluate the behaviour of the metrics used to estimate it using simulated data sets where the real community structure and composition are known. Our results showed that abundance matrices derived from protein-coding marker genes reproduce more faithfully the structure from the original community than those derived from SSU-rRNA genes, even without taking into account the alignment and misclassification biases. We found that, when calculated from metagenomic samples, the most commonly used diversity metrics are biased estimators and differ significantly from their real community parameter counterpart. Our analyses further suggest that these biases are most likely the consequence of insufficient sampling and that, as expected, this problem could be overcome by increasing sequencing coverage depth. We also found that the differences in taxonomic community composition can affect community structure estimation so phylogenetic diversity measures should be incorporated to account for this source of bias. Nevertheless, we show that correct qualitative comparisons can be achieved by the ordering and ranking of samples using a metric that contemplates the multidimensional nature of community structure diversity.

Finally, the incorporation of metrics for both community structure and phylogenetic diversity provides additional understanding of diversity in metagenomic data sets. Although the causes and alternatives to diversity metric bias are to be addressed by mathematical theory, our findings are a first attempt to achieve a standardized framework for community diversity comparisons derived from metagenomic data sets. This will support ongoing work towards the identification of general diversity patterns across geographic space and along environmental gradients.

Acknowledgements

Many thanks to L. Segovia, C. Rooks, F. Reverchon, E. Lopez-Lozano, L. Espinosa-Asuar and E. Aguirre for their suggestions on this project, as well as the comments of two anonymous reviewers that greatly improved this manuscript. Dr Blackburn and his team walked with us through the final versions of the manuscript. Financial support was received from Consejo Nacional de Ciencia y Tecnología – Secretaría de Educación Pública (grant 57507) and Secretaría de Medio Ambiente y Recursos Naturales (grant 2006-C01-23459). All research was carried out at Instituto de Ecología (UNAM), as part of GBR's PhD program at Programa de Doctorado en Ciencias Biomédicas UNAM. G.B.R. was supported with a PhD scholarship from Consejo Nacional de Ciencia y Tecnología (196814). V.S. and L.E.E. worked on this manuscript while in sabbatical at UCI (US) supported by UC-Mexus and DGPA-UNAM, respectively.

References

Bent SJ & Forney LJ (2008) The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. *ISME J* **2**: 689–695.

Berger WH & Parker FL (1970) Diversity of planktonic foraminifera in deep-sea sediments. *Science* **168**: 1345–1347.

Berger SA, Krompass D & Stamatakis A (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol* **60**: 291.

Beszteri B, Temperton B, Frickenhaus S & Giovannoni SJ (2010) Average genome size: a potential source of bias in comparative metagenomics. *ISME J* **4**: 1075–1077.

Biers EJ, Sun S & Howard EC (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl Environ Microbiol* **75**: 2221–2229.

Cadotte MW, Jonathan Davies T, Regetz J, Kembel SW, Cleland E & Oakley TH (2010) Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecol Lett* **13**: 96–105.

Chao A (1984) Nonparametric estimation of the number of classes in a population. *Scand J Stat* **11**: 265–270.

Chao A & Lee S-M (1992) Estimating the number of classes via sample coverage. *J Am Stat Assoc* **87**: 210–217.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B & Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**: 1283–1287.

Curtis TP & Sloan WT (2006) Towards the design of diversity: stochastic models for community assembly in wastewater treatment plants. *Water Sci Technol* **54**: 227.

Curtis TP, Sloan WT & Scannell JW (2002) Estimating prokaryotic diversity and its limits. *Proc Nat Acad Sci* **99**: 10494–10499.

Edwards R, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC & Rohwer F (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**: 57.

Fisher R, Corbet S & Williams C (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *J Anim Ecol* **12**: 42–58.

Fulthorpe RR, Roesch LFW, Riva A & Triplett EW (2008) Distantly sampled soils carry few species in common. *ISME J* **2**: 901–910.

Ghosh TS, M MH & Mande SS (2010) DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformatics* **11**: S14.

Green JL & Plotkin JB (2007) A statistical theory for sampling species abundances. *Ecol Lett* **10**: 1037–1045.

Hill M (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**: 427–432.

Hill TCJ, Walsh KA, Harris JA & Moffett BF (2003) Using ecological diversity measures with bacterial communities. *FEMS Microbiol Ecol* **43**: 1–11.

Huang Y, Gilna P & Li W (2009) Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* **25**: 1338.

Hughes JB, Hellmann JJ, Ricketts TH & Bohannan BJM (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* **67**: 4399–4406.

Hurlbert SH (1971) The nonconcept of species diversity: a critique and alternative parameters. *Ecology* **52**: 577–586.

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

Huson D, Richter D, Mitra S, Auch A & Schuster S (2009) Methods for comparative metagenomics. *BMC Bioinformatics* **10**: S12–S22.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP & Webb CO (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**: 1463.

Kembel SW, Eisen JA, Pollard KS & Green JL (2011) The phylogenetic diversity of metagenomes. *PLoS ONE* **6**: e23214.

Kemp PF & Aller JY (2004) Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. *FEMS Microbiol Ecol* **47**: 161–177.

Kindt R & Coe R (2005) *Tree diversity analysis. A manual and software for common statistical methods for ecological and biodiversity studies*. World of Agroforestry Centre (ICRAF), Nairobi.

Kuczynski J, Liu Z, Lozupone C, McDonald D, Fierer N & Knight R (2010) Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Methods* **7**: 813–819.

Legendre P & Gallagher E (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia* **129**: 271–280.

Li W & Godzik A (2006) Cd-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658.

Lombard N, Prestat E, van Elsas JD & Simonet P (2011) Soil-specific limitations for access and analysis of soil microbial communities by metagenomics. *FEMS Microbiol Ecol* **78**: 31–49.

Lozupone CA & Knight R (2007) Global patterns in bacterial diversity. *P Natl Acad Sci USA* **104**: 11436–11440.

Lukashin AV & Borodovsky M (1998) GeneMark. hmm: new solutions for gene finding. *Nucleic Acids Res* **26**: 1107.

Maechler M, Rousseeuw P, Struyf A, Hubert M & Hornik K (2002) *cluster: Cluster Analysis Basics and Extensions*. R package version 1.14.2.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R & Hugenholtz P (2011) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610–618.

Meinicke P, Aßhauer KP & Lingner T (2011) Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics* **27**: 1618–1624.

Mills DEK, Entry JA, Voss JD, Gillevet PM & Mathee K (2006) An assessment of the hypervariable domains of the 16S rRNA genes for their value in determining microbial community diversity: the paradox of traditional ecological indices. *FEMS Microbiol Ecol* **57**: 496–503.

Mitra S, Gilbert JA, Field D & Huson Daniel H (2010) Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. *ISME J* **4**: 1236–1242.

Morgan JL, Darling AE & Eisen JA (2010) Metagenomic sequencing of an *in vitro*-simulated microbial community. *PLoS ONE* **5**: e10209.

Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P, Stevens MHH & Wagner H (2007) *Vegan: community ecology package*. R package version 1.8-8.

Parks DH, MacDonald NJ & Beiko RG (2011) Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics* **12**: 328.

Pati A, Heath LS, Kyrpides NC & Ivanova N (2011) ClaMS: a classifier for metagenomic sequences. *Stand Genomic Sci* **5**: 248–253.

Patil GP & Taillie C (1982) Diversity as a concept and its measurement. *J Am Stat Assoc* **77**: 548–561.

Preston F (1948) The commonness, and rarity, of species. *Ecology* **29**: 254–283.

Quince C, Curtis TP & Sloan WT (2008) The rational exploration of microbial diversity. *ISME J* **2**: 997–1006.

R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org.

Raes J, Korbel JO, Lercher MJ, von Mering C & Bork P (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol* **8**: R10.

Rényi A (1961) On measures of entropy and information. Fourth Berkeley Symposium on Mathematical Statistics and Probability, pp. 547–561.

Richter DC, Ott F, Auch AF, Schmid R & Huson DH (2008) Metasim—a sequencing simulator for genomics and metagenomics. *PLoS ONE* **3**: e3373.

Ricotta C (2003) On parametric evenness measures. *J Theor Biol* **222**: 189–197.

Rodriguez-Brito B, Rohwer F & Edwards RA (2006) An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**: 162.

Roesch LFW, Fulthorpe RR, Riva A, Casella G & Hadwin A (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**: 283–290.

Roux S, Enault F, Bronner G & Debroas D (2011) Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems. *FEMS Microbiol Ecol* **78**: 617–628.

Rusch DB, Halpern A, Sutton G *et al.* (2007) The sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.

Schloss P & Handelsman J (2008) A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics* **9**: 34.

Shannon CE (1948) A mathematical theory of communication. *AT&T TECH J* **27**: 379–423 and 623–653.

Sharpton TJ, Riesenfeld SJ, Kembel SW, Ladau J, O'Dwyer JP, Green JL, Eisen JA & Pollard KS (2011) PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Comput Biol* **7**: e1001061.

Shaw AK, Halpern AL, Beeson K, Tran B, Venter JC & Martiny JBH (2008) It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol* **10**: 2200–2210.

Simpson EH (1949) Measurement of diversity. *Nature* **163**: 688.

Sloan WT, Woodcock S, Lunn M, Head IM & Curtis TP (2006) Modeling taxa-abundance distributions in microbial communities using environmental sequence data. *Microb Ecol* **53**: 443–455.

Smith VH (2007) Microbial diversity–productivity relationships in aquatic ecosystems. *FEMS Microbiol Ecol* **62**: 181–186.

Sokal RR & Rohlf FJ (1995) *Biometry: The Principles of Statistics in Biological Research*. WH Freeman and Company, New York, NY.

Tóthmérész B (1995) Comparison of different methods for diversity ordering. *J Veg Sci* **6**: 283–290.

Venter JC, Remington K, Heidelberg JF *et al.* (2004) Environmental genome shotgun sequencing of the sargasso sea. *Science* **304**: 66–74.

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N & Bork P (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**: 1126–1130.

Webb CO, Ackerly DD, McPeek MA & Donoghue MJ (2002) Phylogenies and community ecology. *Annu Rev Ecol Syst* **33**: 475–505.

Wu M & Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9**: R151.

Wu D, Hugenholtz P, Mavromatis K *et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**: 1056–1060.

Youssef NH & Elshahed MS (2009) Diversity rankings among bacterial lineages in soil. *ISME J* **3**: 305–313.

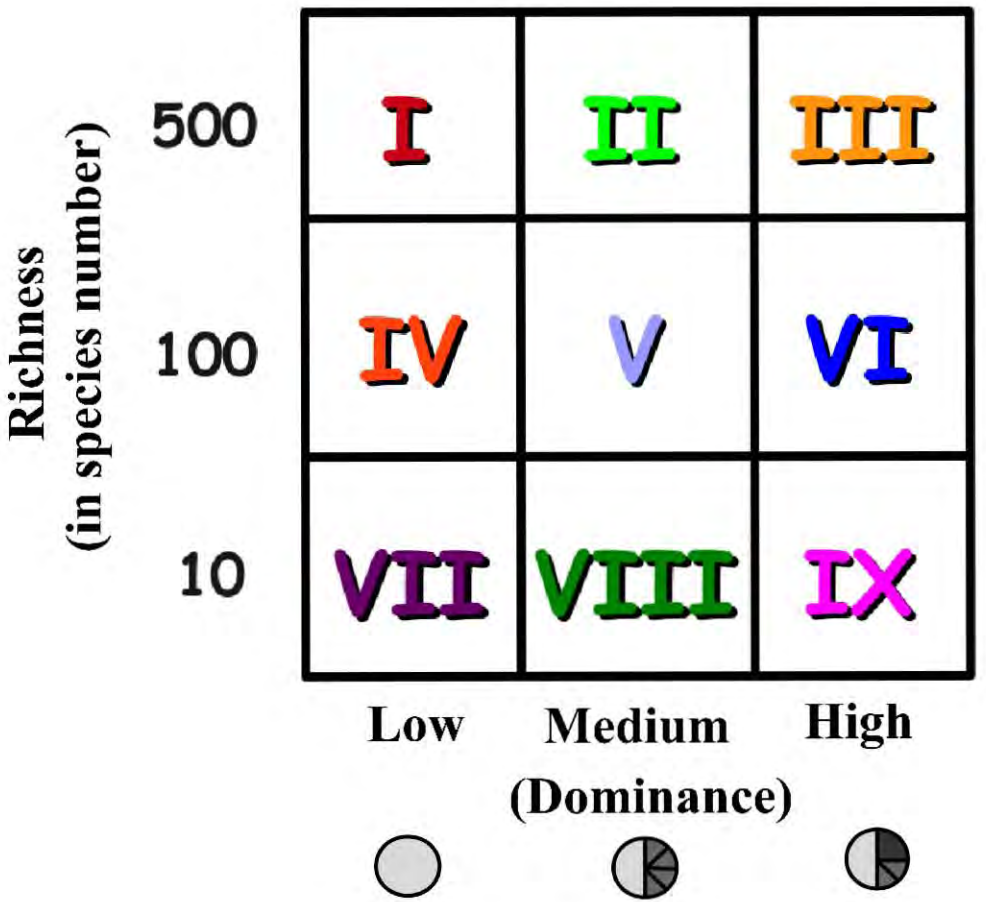
Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Schematic representation of the nine source communities designed as a product of three richness categories (with 10, 100 and 500 species on the *x*-axis) and three dominance/richness categories (total evenness on the extreme right over the *y*-axis, highest dominance at the extreme left).

Fig. S2. Renyi's entropy profiles for the SSU-rRNA derived (a) and the protein-markers derived (b) matrices of three solar saltern ponds with low (squares), medium (triangles) and high (circles) salinity.
Fig. S3. Rarefaction analysis showing the increase in the expected number of species with increasing sequencing depth.
Fig. S4. Quick tutorial for the interpretation of Rényi's entropy profiles.
Table S1. Community profiles describing the structure and genomes of the source communities.
Table S2. Summary of the number of reads found for each one of the 31 protein-coding marker genes.
Table S3. Summary of the resulting *P*-values for the statistical comparisons between the observed diversity values and the values calculated from the original source communities.
Table S4. Summary of the values for all diversity metrics calculated.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

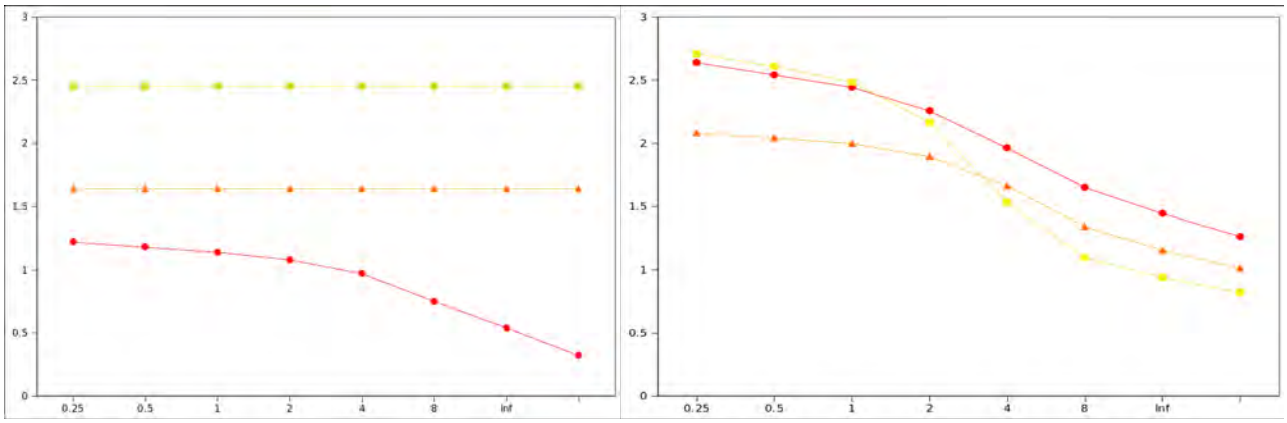

Richness
(in species number)
500
100
10
I
II
III
IV
V
VI
VII
VIII
IX
Low
Medium
High
(Dominance)

Study of San Diego Solar Salterns as Case Study

Finding a natural environmental model to test metagenomic diversity metrics is not an easy task because it would require that we know in advance the true diversity from each natural community. Nevertheless, a good approach is to compare samples along a gradient of extreme environments where diversity would be expected to decrease on one of the extremes (Frontier 1985). Salinity gradients have shown drastic diversity gradients in pond crystallizers (Rodriguez-Valera et al. 1985; Benlloch et al. 2002), although different patterns have been observed in other kinds of salinity gradients (Ben-Dov et al. 2008; Herlemann et al. 2011; Wang et al. 2011). We analysed the aquatic metagenomes from three pond crystallizers at South Bay Saltworks solar salterns (Chula Vista, CA, USA) within a gradient of low (6-8%), medium (12-14%) and high (27-30%) salinity (Beltrán Rodriguez-Brito, unpublished). A decrease in diversity have been shown as salinity increases within this same range in other crystallizers (Legault et al. 2006; Ghai et al. 2011), so our theoretical findings can be tested against these natural communities.

The datasets from these metagenomes are publicly available at the MGRAST portal (http://metagenomics.anl.gov/?page=MetagenomeProject&project=11), with accession numbers 4440437.3, 4440438.3 and 4440425.3. The downloaded metagenomes were subjected to the same pipeline described in the main text for the construction of protein-derived species abundance matrices. We also downloaded the taxonomic classification of SSU-rRNA fragments according to the GreenGenes Taxonomy (DeSantis et al. 2006) that results from the automatic annotation system in MGRAST (Meyer et al. 2008), and constructed a species abundance matrix as well. We chose to work at genus level because these metagenomes are derived from natural communities and the chances of them containing members with sequenced genomes are dim. We calculated the Rényi's entropy profiles for both matrices, and the profiles are plotted in the Supplemenatry Figure 2.

Although much further environmental and metagenomic analyses are needed to formulate hypotheses behind the observed diversity, the patterns depicted in the protein-marker gene profile is in agreement with what would be expected from a salinity gradient. This is, a highly rich and even community in the low salinity sample, a lower richness but similar evenness in the medium-salinity sample, and the lowest salinity and largest dominance in the high-salinity sample. We argue that this is ecologically meaningful because the increasing salinity reduces species richness by limiting growth to halotolerant organisms, and at the highest salinity only a few halophilic organisms will be present and dominant because competitors are excluded by salinity. This is also in agreement with community profiles from other saline pond crystallizers (Ghai et al. 2011). In contrast, the SSU-rRNA derived profile has no clear ecological meaning, with the medium-salinity sample being both the richest and the most dominant sample of the three. Following our discussion in the main text, we are inclined to argue that the observed patterns are a result of the taxa-specific number of rRNA copies per genome.



Supplementary Figure 2. Renyi's entropy profiles for the protein-markers derived (left) and the SSU-rRNA derived (right) matrices of three solar saltern ponds with low (squares), medium (triangles) and high (circles) salinity.

References

Ben-Dov E, Shapiro OH, Gruber R, Brenner A & Kushmaro A (2008) Changes in microbial diversity in industrial wastewater evaporation ponds following artificial salination. *FEMS Microbiology Ecology* 66: 437–446.

Benlloch S, et al. (2002) Prokaryotic genetic diversity throughout the salinity gradient of a coastal solar saltern. *Environmental Microbiology* 4: 349–360.

DeSantis TZ et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology* 72: 5069–5072.

Frontier S (1985) Diversity and structure in aquatic ecosystems. *Oceanography and Marine Biology an Annual Review* 23: 253–312.

Ghai R et al. (2011) New Abundant Microbial Groups in Aquatic Hypersaline Environments. *Scientific Reports* 1(135). doi:10.1038/srep00135

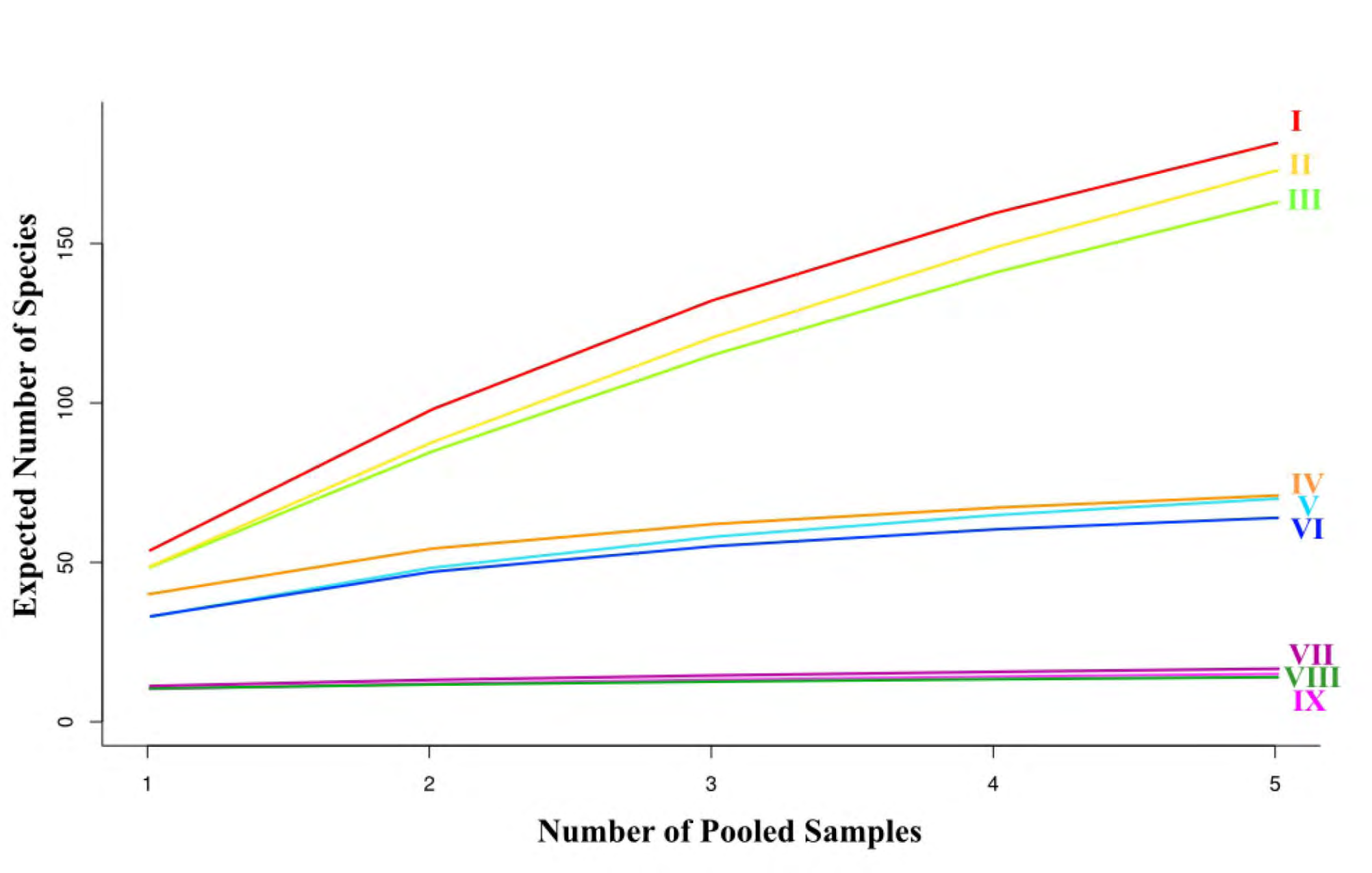
Herlemann DP, Labrenz M, Jorgens K, Bertilsson S, Waniek JJ & Andersson AF (2011) Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME Journal* 5: 1571–1579.

Legault B, Lopez-Lopez A, Alba-Casado J, Doolittle WF, Bolhuis H, Rodriguez-Valera F & Papke RT (2006) Environmental genomics of *Haloquadratum walsbyi*. *BMC genomics* 7: 171.

Meyer F et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* 9: 386.

Rodriguez-Valera F, Ventosa A, Juez G & Imhoff JF (1985) Variation of environmental features and microbial populations with salt concentrations in a multi-pond saltern. *Microbial Ecology* 11: 107–115.

Wang J, Yang D, Zhang Y, Shen J, van der Gast C, Hahn MW & Wu Q (2011) Do Patterns of Bacterial Diversity along Salinity Gradients Differ from Those Observed for Macroorganisms? *PLoS ONE* 6: e27597.


Expected Number of Species
0
50
100
150
1
2
3
4
5
Number of Pooled Samples
I
II
III
IV
V
VI
VII
VIII
IX

CAPÍTULO II:
ANÁLISIS DE LA DIVERSIDAD TAXONÓMICA DE TAPETES MICROBIANOS

En el capítulo anterior, se revisaron diversas métricas para estimar la diversidad taxonómica de una comunidad microbiana a partir de datos metagenómicos. A continuación presento la aplicación de dichas métricas de diversidad taxonómica para la caracterización de la diversidad de dos comunidades naturales, así como demostrar la capacidad de éstas aproximaciones para la identificación de patrones de diversidad. Las comunidades elegidas para su secuenciación metagenómica fueron dos tapetes microbianos provienentes de dos sistemas contrastantes en el Valle de Cuatrociénegas de Carranza, Coahuila (México): uno proviene de una poza de desecación con alta variabilidad ambiental a lo largo del año, mientras que el otro proviene de una poza permanente más estable. La presencia de tapetes microbianos en estos dos sistemas ambientalmente contrastantes sugiere una de dos posibilidades: que existe un conjunto de especies formadoras de tapetes microbianos que han prevalecido a lo largo de la historia geológica del valle y que dada su persistencia y plasticidad son ubícuas en la región, o que las condiciones ambientales particulares del valle promueven el desarrollo de éste tipo de sistemas organosedimentarios y seleccionan las especies.

Éstas dos posibilidades representan dos concepciones sobre el ensamble de comunidades y de tapetes microbianos: en la primera existe un grupo particular de especies especializadas en la formación de los tapetes que debería ser similar en los dos sistemas, y su presencia en el valle sería explicada por la **historia evolutiva** conjunta del valle y las especies. En la segunda, los tapetes microbianos representan sistemas que emergen bajo circunstancias ambientales particulares a partir de la generación de nichos ecológicos específicos, de manera que la presencia de las especies se deberá a un proceso de **filtración ambiental**.

En el artículo presentado a continuación, identificamos dos sistemas constrastantes con dos tipos de diversidad en los tapetes microbianos. El tapete de la poza permanente presenta una enorme diversidad a niveles taxonómicos superiores (orden, clase y phylum), con una elevada equitabilidad y sin ninguna dominancia aparente. El tapete de la poza de desecación presenta un menor número de grupos a niveles superiores a familia, y se encuentra fuertemente dominada por especies del género *Pseudomonas*. Sin embargo, la diversidad al interior de éste grupo es elevada. Al analizar los rasgos funcionales de los grupos taxonómicos, encontramos que los gremios de productividad primaria (fotótrofos, fijadores de nitrógeno, etc.) de diferentes *taxa* son abundantes en la poza permanente, mientras que los organismos degradadores de macromoléculas complejas en la materia orgánica (heterótrofos) son más abundantes en la poza de desecación. Tanto la composición como la estructura son radicalmente distintas entre los dos tapetes, y dado que la tasa de especies compartidas entre los dos sistemas es muy baja, el análisis sugiere que los tapetes se encuentran bajo el efecto de filtración ambiental. Sin embargo, también existe la posibilidad de que los sistemas comparados en realidad sean demasiado diferentes como para permitir su comparación, aunque morfológica y funcionalmente pareciesen similares.

Los metagenomas de los tapetes microbianos se compararon con los metagenomas de dos estromatolitos de la misma región previamente secuenciados (Breitbart et al., 2009), permitiendo la identificación de patrones de diversidad en sistemas organosedimentarios distintos pero comparables. Éstos dos estromatolitos también provienen de dos sistemas contrastantes, uno de una poza permanente y otro de un sistema fluvial. La mayor riqueza y menor dominancia se observa en los metagenomas del tapete microbiano y el estromatolito de las pozas más estables, y en los sistemas variables se observa un incremento en la dominancia. Es decir, que las dos comunidades estables tienen una estructura más similiar entre ellas, a pesar de que pertenecen a sistemas distintos, y que la perturbación parece afectar de manera similar la estructura en los sistemas variables. De igual forma, el número de especies compartidas entre las cuatro comunidades es muy bajo.

Por último, comparé la composición taxonómica de las cuatro comunidades del valle contra el único otro metagenoma de tapete microbiano disponible, en Guerrero Negro, Baja California. Las condiciones ambientales de éste tapete microbiano son radicalmente distintas a Cuatrociénegas, pues se desarrollan en sistemas halófilos costeros. El número de especies compartidas con éste tapete fue, como era de esperarse, muy bajo. Sin embargo, se detectó un patrón similar en la abundancia relativa de los órdenes más abundantes en el tapete microbiano de la poza permanente de Cuatrociénegas, el estromatolito de la poza permanente de Cuatrociénegas, y el tapete microbiano de Guerrero Negro. Ésta sorprendente conservación de un "núcleo" de organismos del mismo orden sugiere una tercera explicación al ensamble de las comunidades en donde estaría definida simultáneamente por la filtración ambiental y la historia evolutiva, pues las condiciones ambientales particulares filtrarían las especies a ocupar el nicho, pero a un nivel taxonómico superior que sugiere la filtración de rasgos taxonómicos filogenéticamente conservados. Ésta idea es compatible con la *teoría de ensamble de comunidades por "lotería"* (Burke et al. 2011), sólo que a diferencia de la propuesta de Burke et al. (2011), los caracteres filtrados serían caracteres conservados filogenéticamente y no genes funcionales individuales. Bajo éste esquema, la filtración de ciertos grupos taxonómicos superiores dados sus rasgos funcionales es un proceso evolutivo determinístico definido por los parámetros ambientales en los nichos ecológicos, mientras que la selección de las especies particulares que colonizan cada nicho es un proceso estocástico afectado por la filtración ambiental y las dinámicas locales de migración y extinción.

En éste capítulo se demuestran los alcances de las métricas de diversidad analizadas en el capítulo anterior aplicadas a datos reales provenientes de comunidades naturales. La caracterización de la diversidad taxonómica en los tapetes microbianos permite explorar la homogeneidad en la composición taxonómica de los tapetes microbianos, así como describir el efecto de los parámetros ambientales sobre la estructura de las comunidades. Finalmente, la comparación de metagenomas de diversos sistemas permite identificar parámetros de diversidad taxonómica asociados a éste tipo de sistemas organosedimentarios.

ASTROBIOLOGY
Volume 12, Number 7, 2012

DOI: 10.1089/ast.2011.0724

Comparative Metagenomics of Two Microbial Mats at Cuatro Ciénegas Basin II: Community Structure and Composition in Oligotrophic Environments

Germán Bonilla-Rosso,[1] Mariana Peimbert,[2] Luis David Alcaraz,[3] Ismael Hernández,[4] Luis E. Eguiarte,[1] Gabriela Olmedo-Alvarez,[4] and Valeria Souza[1]

Abstract

Microbial mats are self-sustained, functionally complex ecosystems that make good models for the understanding of past and present microbial ecosystems as well as putative extraterrestrial ecosystems. Ecological theory suggests that the composition of these communities might be affected by nutrient availability and disturbance frequency. We characterized two microbial mats from two contrasting environments in the oligotrophic Cuatro Ciénegas Basin: a permanent green pool and a red desiccation pond. We analyzed their taxonomic structure and composition by means of 16S rRNA clone libraries and metagenomics and inferred their metabolic role by the analysis of functional traits in the most abundant organisms. Both mats showed a high diversity with metabolically diverse members and strongly differed in structure and composition. The green mat had a higher species richness and evenness than the red mat, which was dominated by a lineage of *Pseudomonas*. Autotrophs were abundant in the green mat, and heterotrophs were abundant in the red mat. When comparing with other mats and stromatolites, we found that taxonomic composition was not shared at species level but at order level, which suggests environmental filtering for phylogenetically conserved functional traits with random selection of particular organisms. The highest diversity and composition similarity was observed among systems from stable environments, which suggests that disturbance regimes might affect diversity more strongly than nutrient availability, since oligotrophy does not appear to prevent the establishment of complex and diverse microbial mat communities. These results are discussed in light of the search for extraterrestrial life. Key Words: Cuatro Ciénegas—Metagenomics—Microbial mats—Oligotrophic—Phosphorus limitation—Stromatolites. Astrobiology 12, 659–673.

1. Introduction

Microbial mats are self-sustained laminated organosedimentary ecosystems that develop on solid/water interfaces and are composed of tightly interacting microorganisms. They form colored multilayered biofilms embedded in a matrix of extrapolymeric substances that bind cells and inorganic substances together, creating steep biogeochemical microgradients as a result of their own metabolism (Jorgensen *et al.,* 1986; van Gemerden, 1993; Stolz, 2000). This compact gradient (2–50 mm) generates a diverse array of microniches in which different functional guilds can thrive and the spatial closeness allows for the development of complex metabolite exchange networks that ensure survival even under extreme conditions (Paerl and Yannarell, 2010).

Although modern microbial mats are geographically widespread (Gerdes, 2010), they are limited in their environmental occurrence to a selected few aquatic environments, both freshwater and marine (Krumbein *et al.,* 1977; Bebout *et al.,* 2002). This contrasts with their former distribution on Earth, as revealed by the abundance of biogenic, laminated reeflike structures in the fossil record from Precambrian shallow marine environments (Awramik, 1984) as early as 3.4 Ga (Allwood *et al.,* 2006). It is believed that both an exclusion of grazing eukaryotes and a lack of competition for light with fast-growing algae are required for a microbial mat to develop, which restricts their current distribution mostly to extreme environments in modern Earth (Cohen, 1989; Bebout *et al.,* 2002). Many current habitats can be considered environmental analogues to those found in

[1]Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Coyoacán, México D.F., México.
[2]Departamento de Ciencias Naturales, Universidad Autónoma Metropolitana, Cuajimalpa, Álvaro Obregón, México D.F., México.
[3]Departamento de Genómica y Salud, Centro Superior de Investigación en Salud Pública, Valencia, España.
[4]Departamento de Ingeniería Genética, Cinvestav, Campus Guanajuato, Irapuato, México.

Precambrian shallow waters (Foster and Mobberley, 2010). As such, microbial mats are the oldest ecosystems known to date and have proven to be successful ecological assemblages, given that they have survived billions of years of environmental change due to their stable, but adaptable, structural properties (Awramik, 1976; Green and Jahnke, 2010).

Since microbial mats are self-sustained heterogeneous ecosystems, they are suitable experimental models that can be used to test how ecosystems respond to rapid environmental disturbances, which cannot be tested with any other ecosystem (Paerl *et al.,* 2003; Yannarell *et al.,* 2007). They have also allowed for the reconstruction of putative metabolisms and the characterization of biosignatures of early life on Earth (Kasting, 2001; Bebout *et al.,* 2004; Foster and Mobberley, 2010). Without doubt, the study of Earth's most common, simple, and pervasive ecosystems at the molecular level will help us better understand the evolution of life on Earth and, as a consequence, will facilitate the search for life elsewhere (Foster and Mobberley, 2010).

Microbial mats can be found in extreme conditions, including cold (0.4–3.4°C; Bottos *et al.,* 2008; Varin *et al.,* 2010) and high concentrations of iron (Emerson and Revsbech, 1994), sulfur (Elshahed *et al.,* 2003), and hydrocarbons (Mills *et al.,* 2005). The vast majority of the literature, however, has focused on halophilic coastal mats (like those in Guerrero Negro, Mexico, or in Shark Bay, Australia) or hyperthermophilic mats (like those in Yellowstone) (Ward *et al.,* 1998; Spear *et al.,* 2003). Mats from coastal hypersaline environments can harbor diverse and complex ecosystems (Ley *et al.,* 2006; Kunin *et al.*, 2008). In contrast, mats that grow at high temperatures (64–82°C) are characterized by a low species diversity (Inskeep *et al.,* 2010; Meyer-Dombard *et al.,* 2011), most likely because temperature sets a limit for photosynthesis (Ward *et al.,* 1998). This suggests that nutrient availability and primary production will determine the complexity of a microbial mat community, as has been previously proposed for plant communities (Grime, 1973; Tilman, 1990). The ecological characterization of communities from low-nutrient (oligotrophic) environments can aid the search of life on other planets by defining a nutrient availability range for the development of life, so that astrobiological targets can be more precisely defined.

Another ecological factor to be considered in the search for life beyond the confines of Earth is environmental disturbance, since communities will not develop where disturbances are too frequent or too intense, while an intermediate disturbance frequency in many cases will maximize complexity by creating several microniches (Grime, 1973; Connell, 1978; Rainey and Travisano, 1998). It is therefore relevant to study how communities are affected by disturbance regimes to be able to predict their possible existence in other planets.

The Cuatro Ciénegas Basin (CCB) is a naturally isolated valley in the Chihuahuan Desert (Coahuila, Mexico) that encompasses hundreds of permanent ponds, marshes, and desiccation pools (Minckley and Cole, 1968) (Fig. 1), which despite their low phosphorus content (Elser *et al.,* 2005; Breitbart *et al.,* 2009, Peimbert *et al.,* 2012 in this issue) harbor varied and diverse microbial communities (Souza *et al.,* 2006; Escalante *et al.,* 2008). The CCB has received the attention of astrobiological research because it is considered a modern

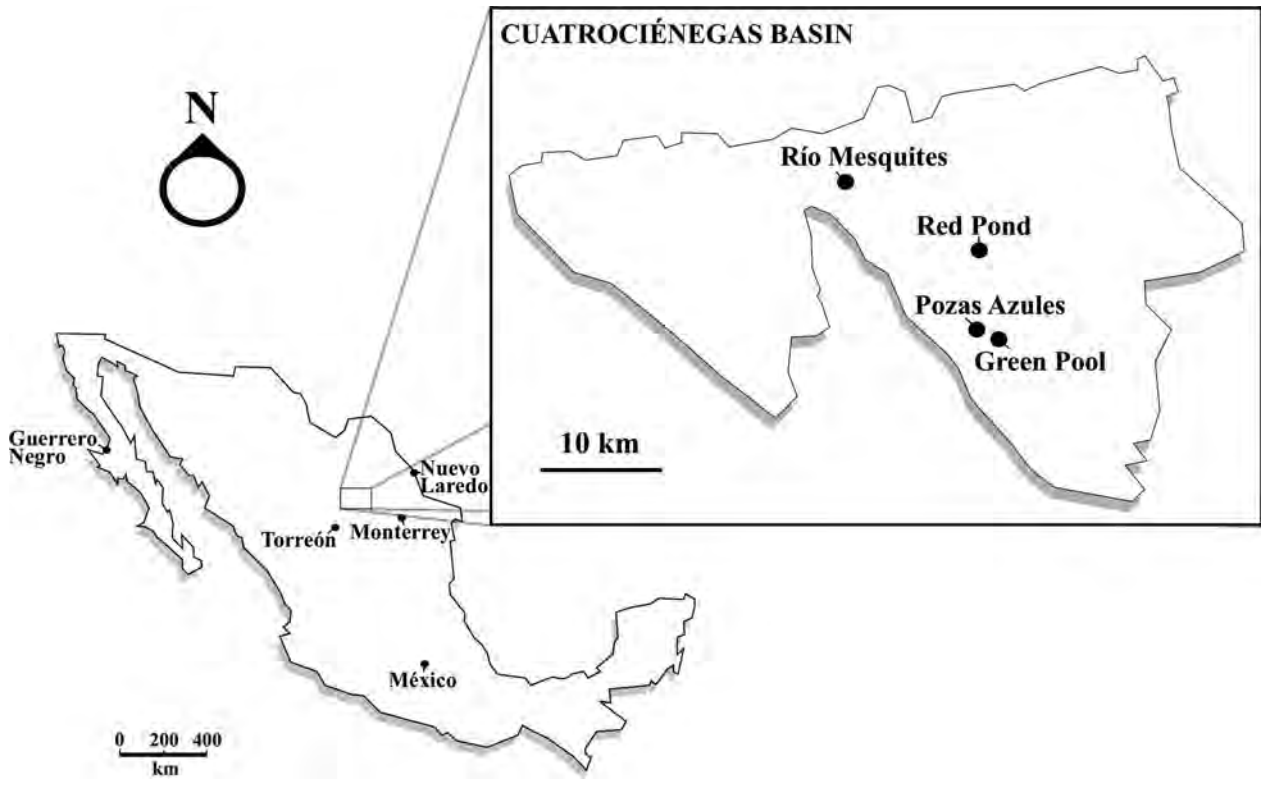


FIG. 1. Schematic map depicting the geographic origin of the metagenomic data sets analyzed in this study.

analogue of the Precambrian oceans (Elser *et al.*, 2006), which have been characterized as having had low levels of circulating phosphorus (Bjerrum and Canfield, 2002; Planavsky *et al.*, 2010) and a reduced deposition of phosphorite rocks (Papineau, 2010). Moreover, complex microbial communities are known to have developed in Precambrian waters, since abundant phosphate-absorptive calcareous stromatolites have been preserved in the geological record (Grotzinger and Knoll, 1999; Gerdes, 2010). Finally, the CCB has also been identified as an analogue of the martian Olympia Undae gypsum sand dunes (Szynkiewicz *et al.*, 2010) and the dynamic fluvial system of the Gale Crater of early Mars, one of the four landing sites targeted for the Mars Science Laboratory mission later this year (Golombek *et al.*, 2011).

All this stresses the need to characterize the microbial communities that inhabit the CCB, while its oligotrophic waters and diversity of aquatic environments offer the opportunity to study the limiting effects of phosphorus availability on microbial community development. In the present study, we taxonomically characterized two aquatic, nonthermophilic, nonhalophilic oligotrophic microbial mats from two environments with contrasting disturbance regimes: a permanent green pool and a red desiccation pond. We analyzed their taxonomic structure and composition by using 16S rRNA gene clone libraries and metagenomics. A low diversity was expected in these oligotrophic environments, with a greater diversity in the mat from the desiccation pond due to intermediate disturbances. We also compared our data with two previously published metagenomes of microbial communities from the CCB (the stromatolites of Río Mesquites and Poza Azul; Breitbart *et al.*, 2009) and with the only other microbial mat metagenome available, the well-studied coastal halophilic microbial mat from the salines of Guerrero Negro (Baja California, Mexico; Spear *et al.*, 2003; Kunin *et al.*, 2008). We expected to find large differences in taxonomic composition, since Guerrero Negro and the CCB are geographically separated by a linear distance of 1200 km.

2. Material and Methods

Two 20 × 20 cm microbial mat cores were collected in July 2008 from a shallow, seasonal red desiccation pond in the Los Hundidos region (red mat, at 26°52′17″N, 102°01′11.3″W) and a permanent green pool in Pozas Azules Ranch (green mat, at 26°49′24.4″N, 102°00′53.2″W) (Fig. 1). Due to the nature of the desiccation ponds, the temperature disturbance frequency was 5 times higher for the red mat than for the green mat (Table 1). Samples were frozen in liquid nitrogen and transported to the laboratory. DNA extraction was performed by Freeze/Thaw, CTAB, phenol-chloroform extraction as described previously (Zhou *et al.*, 1996; Breitbart *et al.*, 2009). The samples were further purified by electrodialysis as described by Rodríguez-Mejía *et al.* (2008). Total DNA was amplified with Genomiphi polymerase (GE Healthcare, Piscataway, NJ, USA) according to the manufacturer's instructions. Ten independent reactions were carried out and later pooled before sequencing to reduce amplification bias. From the total DNA, 400 μL of one red mat (551 ng/μL) and 400 μL of the green mat (890 ng/μL) were independently pyrosequenced with 454 FLX (Roche Diagnostics, IN, USA) at CINVESTAV-LANGEBIO, Irapuato, Mexico. Metagenomic reads were noise-corrected with CD-HIT 454 (Li and Godzik, 2006). RNA sequences were identified and masked with rRNA-HMM and tRNA-scan (Huang *et al.*, 2009). Open reading frames calling was performed with GeneMark (Besemer and Borodovsky, 2005) and annotated with RAMMCAP (Li, 2009). Data sets were uploaded to the online server MG-RAST and subjected to its standard quality control pipeline (Meyer *et al.*, 2008). Data is available via the MG-RAST portal (green mat ID is 4441363.3, and red mat ID is 4442466.3).

Table 1. Physiochemical Properties of the Aquatic Environments from Where the Samples Analyzed in This Study Were Taken

	Tm (°C)	T* (°C)	*pH**	*Conductivity (μS/cm)**	*Stoichiometric N to P ratio***
Green	25 [15–24]	30.27	6.0	2.57	2:1
Red	30 [10–60]	48.26	5.5	117.6	157:1
Red 2	NA	47.58	7.3	83.5	NA

*At time of sampling.
**Numbers of nitrogen atoms per phosphorus atom. For reference, the classic "Redfield" ratio when nutrients are not limiting is 16:1.
Tm = Yearly average temperature mean; minimum and maximum temperatures are shown in brackets.
NA, Data not available.

2.1. 16S rRNA gene clone libraries analyses

A 16S rRNA gene clone library was constructed for each of the two mats and one additional mat from an adjacent red desiccation pool in Los Hundidos (red mat 2). The PCR reaction mixture for amplification of 16S rRNA genes contained 1.5 m*M* $MgCl_2$, 250 m*M* of each nucleotide, 4 m*M* of each primer, 1 U of Taq DNA polymerase (Roche, Mannheim, Germany), and 50 ng of isolated DNA. The bacterial 16S rRNA gene was amplified with the universal primers 27F (5-AGA GTT TGA TCC TGG CTC AG-3) and 1492R (5-GGT TAC CTT GTT ACG ACT T-3). Amplification was performed as follows: a "hot start" (95°C for 5 min) was followed by 25 cycles at 94°C for 40 s, 55°C for 40 s, and 72°C for 90 s with a 10 min extension at 72°C. Amplified products were purified with a Roche PCR purification Kit (Roche, Mannheim, Germany), and PCR products were cloned with a TOPO cloning kit (Invitrogen, Karlsruhe, Germany) following the instructions of the manufacturer. Plasmids were sequenced on one end by using dye terminator chemistry on an automated DNA sequencer (ABI3700, Applied Biosystems).

Raw sequences were chimera-checked with Mallard (Ashelford *et al.*, 2006), trimmed of vector sequences, and checked for errors and low quality with BioEdit (Hall, 1999). Sequences were aligned and analyzed with ARB (Ludwig *et al.*, 2004), assigned to a taxonomic category with the GreenGenes Classifier (DeSantis *et al.*, 2006), and assigned to a phylogenetic position in reference to the SILVA database in ARB. Clustering of operational taxonomic units (OTUs) at 97% identity, alpha-diversity indexes, and rarefaction analyses were performed with MOTHUR (Schloss *et al.*, 2009). Phylogenetic trees were constructed with PhyML software as implemented in ARB (GTR + I model, 1000 bootstraps), and edited with iTOL (Letunic and Bork, 2011).

2.2. Metagenomic protein gene phylogenetic marker analyses

The entire metagenomic data set of stromatolites from Poza Azul (ID 4440067.3) and Río Mesquites (ID 4440060.4) were downloaded from MG-RAST (Meyer *et al.,* 2008) and subjected to the exact same quality control and annotation analyses as described above. The 31 universally conserved, single-copy phylogenetic molecular markers identified with AMPHORA (Wu and Eisen, 2008) were searched for and assigned to taxonomic categories with MEGAN (Huson *et al.,* 2007). Analyses were performed at genus level. Taxa-abundance matrices were built to calculate ecological distance matrices by using the Bray-Curtis distance and were subjected to a principal component analysis. Mantel's test was performed to evaluate correlation between geographical and ecological distance. Community structure for the four data sets was evaluated with Renyi's entropy profiles (Bent and Forney, 2008). All ecological analyses were performed with R (R Development Core Team, 2006). A list of the diversity metrics can be found in Supplementary Material S4b (Supplementary Data are available online at www.liebertonline.com/ast).

2.3. Metagenomic all-reads complement analyses

The resulting metagenomic data sets of the green and red microbial mats were uploaded and annotated with the MG-RAST automatic pipeline, where each single read in a metagenome can be assigned to a taxonomic category by comparing it with the available annotated databases in the SEED platform (Meyer *et al.,* 2008). Hits with an e-value lower than 1×10^{-5} were used to build taxonomic profiles for the metagenomes. The same analyses were performed with the pyrosequenced metagenomes of the Poza Azul and Río Mesquites stromatolites (Breitbart *et al.,* 2009) and the shotgun capillarily sequenced metagenomes from Guerrero Negro (Kunin *et al.,* 2008), which were all downloaded from the MG-RAST portal (identification numbers 4440060.4, 4440067.3, and 4440963.3 to 4440972.3). The profiles of the 10 data sets from Guerrero Negro were pooled together since the referred study analyzed each layer of the microbial mat separately. These profiles allowed us to compare a larger number of reads from each metagenome by assigning them to different taxonomic categories at different taxonomic levels. We considered abundant taxa as those comprising 75% of the total annotated reads in each metagenome and compared these across metagenomes. The fragment recruitment analyses were calculated in MG-RAST with a maximum e-value cutoff of 1e-05. The presence of specific genes in the genomes that recruited the largest number of reads was analyzed in the Integrated Microbial Genomes portal (Markowitz *et al.,* 2009).

3. Results

We evaluated diversity of two microbial mats, using three different approaches: 16S rRNA gene clone library construction, protein-coding phylogenetic marker genes from metagenomes, and all-reads analysis from metagenomes. These approaches are complementary, as each one analyzed different aspects of community structure and composition. For example, the abundance of minor phyla (Armatimonadetes, Caldisericia, Nitrospira, OD1, Lentisphaera, and Deferribacteres) is only apparent with the clone library approach, while the protein marker approach gives a more precise structure estimation and allows the comparison between previously published metagenomes for which no clone libraries are available. The large proportion of cyanobacterial genomes is revealed by comparing the two previous approaches with the all-reads taxonomic composition.

3.1. 16S rRNA gene clone libraries analyses

Highly structured microbial mats were sampled from two different kinds of oligotrophic pools in the CCB: a green mat from a stable permanent pool and two red mats from two desiccation ponds. A total of 516 unique phylotypes were recovered from 724 sequences. Clustering at 97% identity produced 32 OTUs for the red mat and 260 for the green mat, comprising 23 phyla. The red mat had low evenness and was dominated by a few OTUs from the genus *Pseudomonas* (Shannon = 2.003, Simpson = 0.216), while the green mat showed a very high evenness with no dominant OTU (Shannon = 5.389, Simpson = 0.003), though the phyla Cyanobacteria, Clostridia, Gammaproteobacteria, Epsilonproteobacteria, and Deltaproteobacteria were abundant (Fig. 2). The *Pseudomonas* sequences belonged to the *P. fluorescens* (*P. fluorescens* and *P. corrugata/P. brassicacearum*) subgroups and the *P. pachastrellae* lineages (Mulet *et al.*, 2010). The nonparametric richness estimator Chao estimated 1221 OTUs (C.I. = 904–1704) for the green mat and 80 OTUs (C.I. = 42–221) for the red mat. A total regional richness of 1198 OTUs (C.I. = 902–1643) is expected in microbial mats of the CCB, calculated from a composite pool of the individual libraries. The two microbial mat communities were phylogenetically and statistically different, both in terms of composition (Sorensen = 0.095, UniFrac U = 0.78) and structure (Morisita-Horn = 0.042, Weighted UniFrac W = 0.91). Both communities shared only 8 OTUs, with a total of 21 OTUs expected to be shared by the Chao richness estimator.

The 16S rRNA gene clone library of a second red mat (red mat 2) from a second desiccation pool was produced for comparison. Forty-two operational taxonomic units were recovered at 97% identity from 218 sequences. This red mat 2 was also dominated by *Pseudomonas,* although it was slightly more diverse than the first red mat (Shannon = 3.546, Simpson = 0.022, Chao = 122 [73-246]). The red mat 2 was more similar to the first red mat (Sorensen = 0.253, Morisita-Horn =0.193) and shared 15 OTUs, with a total of 77 shared OTUs expected (Fig. 3).

3.2. Metagenomic protein gene phylogenetic marker analyses

The metagenomes of both mats were sequenced, which generated a total of 150,381,320 bp and 709,799 reads. The green mat metagenome consisted of 427,366 reads with an average length of 202.54 bp and a maximum length of 390 bp. The average GC content was 39.7%. The red mat metagenome consisted of 282,433 reads with an average length of 225.98 bp and a maximum length of 366 bp. The average GC content was 52.8%.

A total of 1066 sequences from 31 protein marker genes were retrieved from the data sets with AMPHORA (Wu and Eisen, 2008): 831 from the green mat and 235 from the

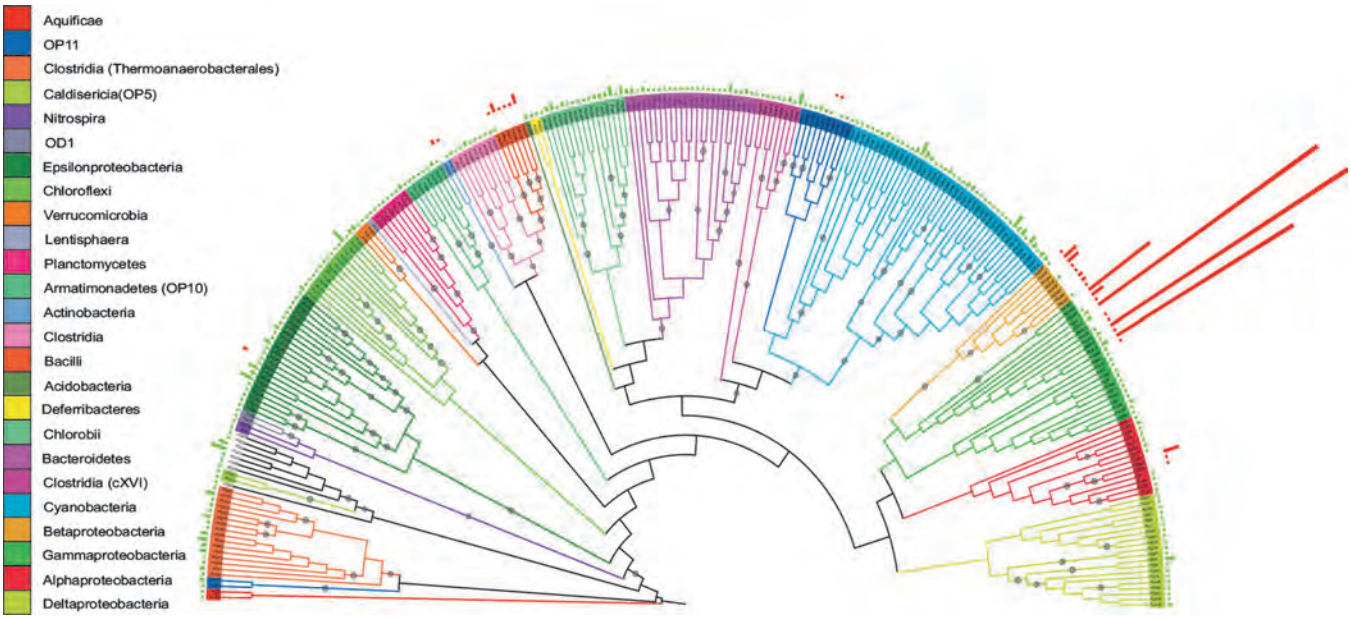


FIG. 2. Consensus phylogenetic tree of 16S rRNA gene clone libraries from the green and red microbial mats reconstructed by Maximum Likelihood with 1000 bootstraps. The outer rings represent the relative abundance of each OTU clustered at 97% of the green (green bars) and red (red bars) mats. Black points on the branches indicate clades present in >80% of the phylogenies. Color key for bacterial phyla are given clockwise from top to bottom.

red mat. In agreement with the clone library, the red mat was less diverse (Shannon = 1.315, Simpson = 0.464, Chao = 23.5 ± 31.108 SE) and dominated by *Pseudomonas*. Genera *Sphingopixis* (Sphingomonadales), *Chitinophaga* (Sphingobacteria), and *Microcoleus* (Cyanobacteria) were also abundant. In contrast, the green mat was highly diverse (Shannon = 4.108, Simpson = 0.019, Chao = 148 ± 33.618), with no dominant genus. Among the abundant genera were the phototrophic *Synechococcus* (Cyanobacteria) and *Chloroherpeton* (Chlorobi); the heterotrophic *Legionella* (Gammaproteobacteria), *Algoriphagus* and *Bacteroides* (Bacteroidetes), and the deltaproteobacterial *Desulfococcus* and *Synthrophobacter* (Fig. 4).

The microbial mat metagenomes were compared with two other metagenomes from the CCB: a stromatolite from the permanent pool Pozas Azules and a riverine oncolite from Río Mesquites (Breitbart *et al.*, 2009). The analysis of the same 31 marker genes retrieved 785 sequences from Río Mesquites and 238 from Pozas Azules. The most abundant genera in Pozas Azules all belonged to the superphylum Verrucomicrobia–Planctomycetes (*Chthoniobacter, Verrucomicrobium, Rhodopirellula, Akkermansia, Gemmata, Planctomyces*) and exhibited a high diversity (Shannon = 3.230, Simpson = 0.053, Chao = 145.2 ± 127.552). Río Mesquites showed far less diversity (Shannon = 0.614, Simpson = 0.824, Chao = 132 ± 147.344), with most sequences belonging to the oligotrophic genus *Pelagibacter* (formerly known as the SAR11 cluster, Giovannoni *et al.*, 2005). These findings are in agreement with recent clone libraries from DNA obtained from similar structures in the same system (Nitti *et al.*, 2012).

The principal component analysis (Fig. 4) showed a larger similarity between the green mat and the Pozas Azules stromatolite because these samples aggregated together, though independently from the red mat and Río Mesquites stromatolite. Ecological indexes also showed a larger similarity between the green mat and the Pozas Azules stromatolite (Sorensen = 0.194, Morisita-Horn = 0.2693) than between any other pair of samples (< 0.05).

Renyi's community profiles revealed that the green mat was the most diverse sample, while the red mat was the least rich and Río Mesquites showed the largest dominance. Although the two stromatolite samples showed a similar level of richness, they differed in dominance (Fig. 5). Mantel's test was not significant and showed no correlation between ecological and geographical distances ($r = 0.2821$, $p = 0.704$). A guide to the interpretation of Renyi's profiles can be found in Supplementary Material S4c.

3.3. Metagenomic all-reads complement analyses

The green mat and the Pozas Azules stromatolite samples displayed the highest evenness: 28 different orders composed 75% of the classified reads in the green mat and 22 orders in Pozas Azules. Among the most abundant taxa in the green mat were photosynthetic taxa (Cyanobacteria, Chloroflexales, Chlorobiales, Chromatiales, Rhodobacterales) and known heterotrophic taxa (Clostridiales, Bacillales, Burkholderiales). In Pozas Azules, the most abundant taxa were the Planctomycetes/Verrucomicrobia complex (Planctomycetales, Verrucomicrobiales, Spartobacteria) and the Cyanobacteria (Chroococcales, Nostocales, Oscillatoriales). The red mat was dominated by reads from heterotrophic orders, with Pseudomonadales, Burkholderiales, and Bacillales comprising 50% of the total reads. Photosynthetic orders represented 15% of total reads. The Río Mesquites stromatolite was dominated by reads from Cyanobacteria, with Nostocales, Chroococcales, and Oscillatoriales comprising 78% of the total reads.

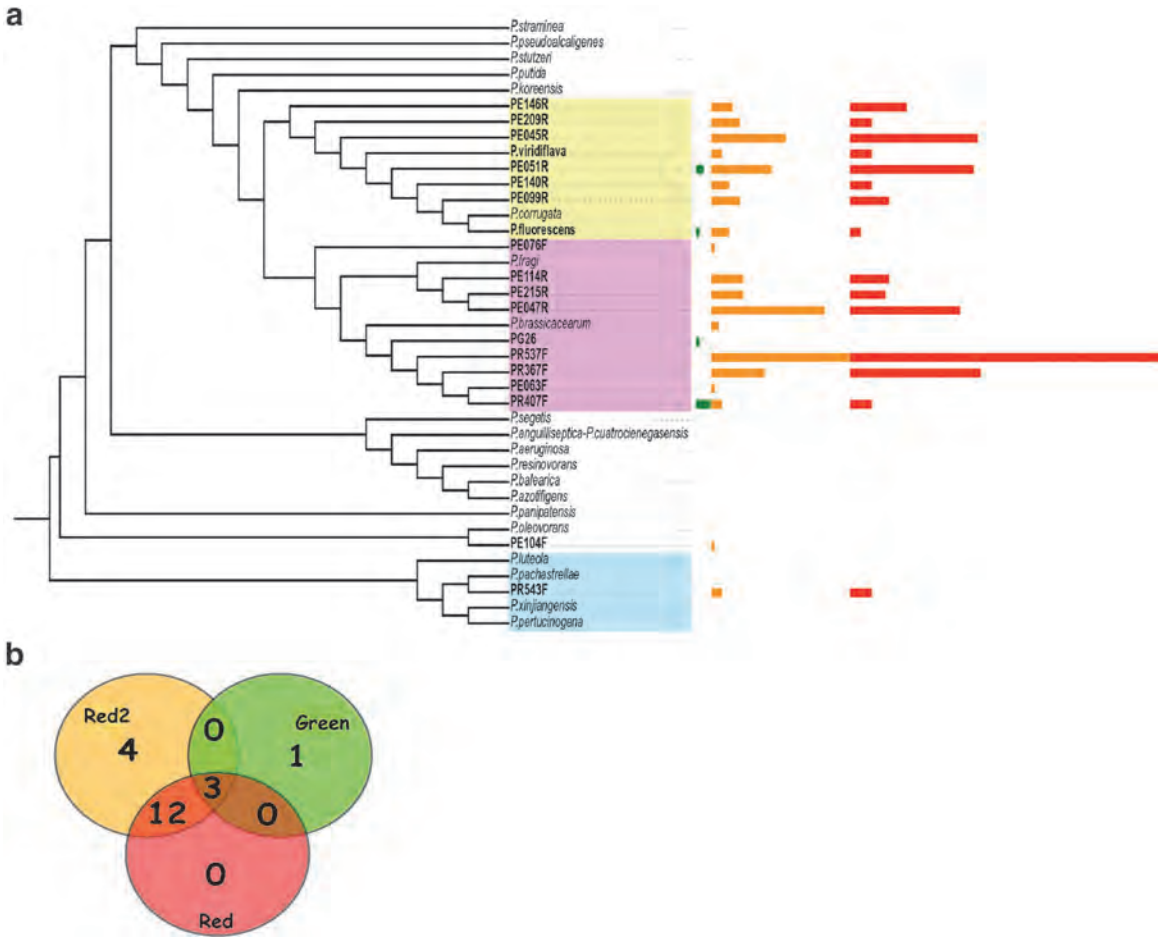


FIG. 3. (**a**) Phylogenetic tree of the *Pseudomonas* subclade reconstructed by maximum likelihood. OTUs containing sequences from this study are typed in boldface, and reference sequences are in italics. The abundance bars represent the number of sequences contained in each OTU at 98% identity. Bars are colored as follows: green mat (green), red mat (red), red mat 2 (orange). (**b**) Venn diagram depicting the number of shared and unique OTUs between libraries. Color images available online at www.liebertpub.com/ast

To include another nonthermophilic microbial mat in the study, we analyzed the pooled taxonomic sequence profile from the coastal hypersaline mat metagenome from Guerrero Negro (Kunin *et al.,* 2008). It also displayed a high evenness, with 75% of the reads distributed among 27 orders. Abundant taxa were Rhodobacterales, Rhizobiales, Clostridiales, Chroococcales, Actinomycetales, and Bacteroidales. We found a remarkable conservation in the relative proportion of the analyzed taxa across the green mat, Pozas Azules, and Guerrero Negro (Fig. 6).

Fragment recruitment revealed that reference genomes from members of the Nostocales recruited most reads from the green mat (1140 recruited by *Anabaena variabilis* and 820 from *Nostoc punctiforme*), but they did so with low similarity values (mode > 1e-10). The genomes of unicellular diazotrophic *Cyanothece* species (Chroococcales: Aphanotecoideae) recruited a large amount of reads from the green mat data set (1017 reads), while reference genomes from different strains of *Pseudomonas fluorescens* recruited most reads from the red mat (3996–5626 reads) with high similarity values (mode < 1e-30). The cyanobacteria *Microcoleus chthonoplastes* was the next reference genome that recruited most reads from the red mat (4847 reads). Given the cosmopolitan distribution and high similarity between *M. chthonoplastes* populations (Garcia-Pichel *et al.,* 1996) and the fact that the recruited fragments from the red mat had very high similarity values (mode < 1e-30), the organisms present in the CCB red mat are very likely non-heterocystous filamentous cyanobacteria from *M. chthonoplastes.* Fragment recruitment plots can be found in Supplementary Material S3.

4. Discussion

4.1. On the dominant organisms in microbial mats

Cyanobacteria are important and often dominant components of microbial mats (Ward *et al.,* 1998), and are responsible for the formation of the mat tissue with interlaced filaments. This study is no exception, since they were abundant in both mats, and order Chroococcales was the most abundant and diverse cyanobacterial order in both mats. Although no single OTU was dominant within Chroococcales, several species from genus *Synechococcus* were found. Organisms from this order would have had no role in the formation of the mat tissue, since it comprises only unicellular species. Order Oscillatoriales contains mat-forming species, and although it was among the most abundant groups in the green mat, no

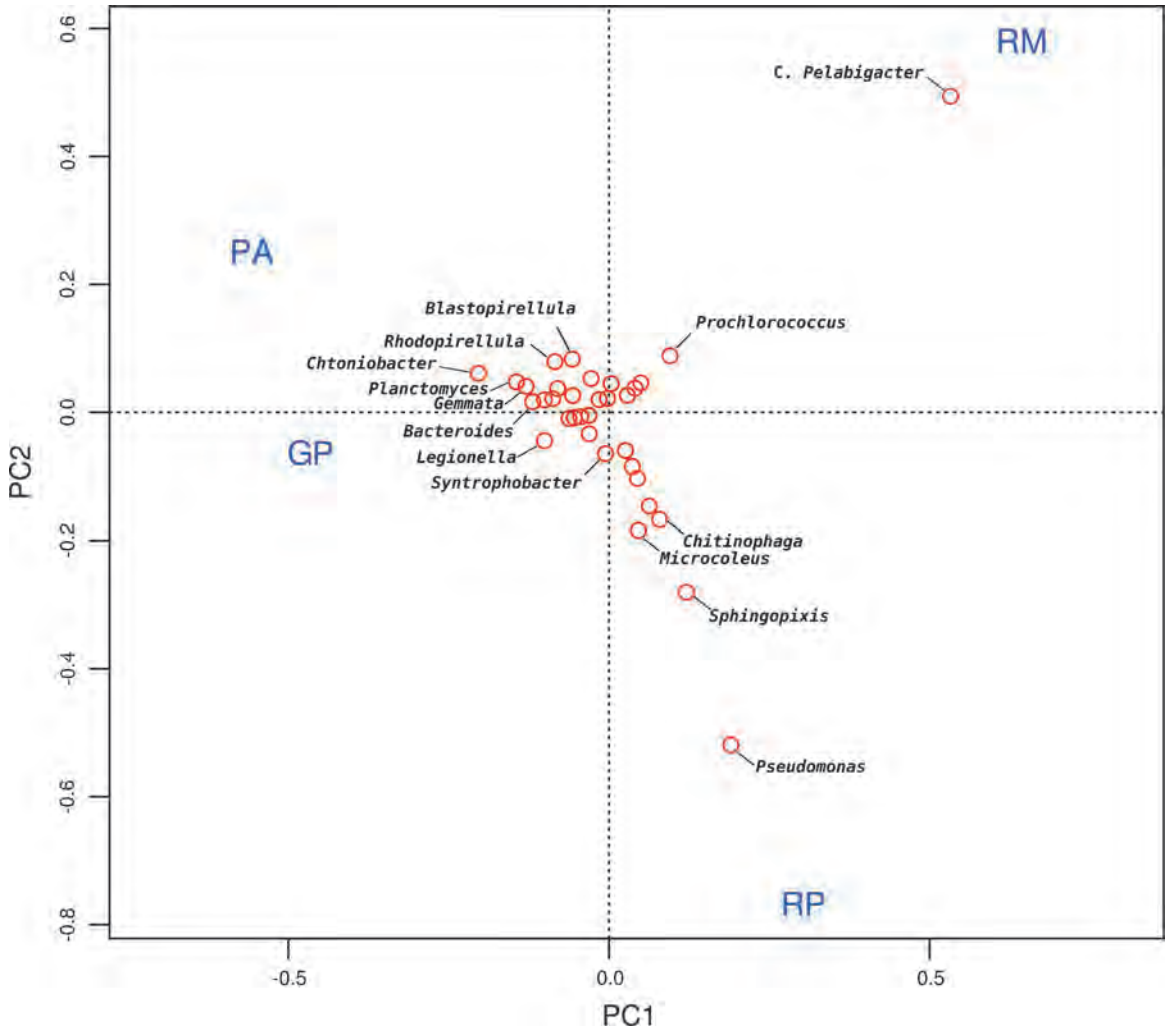


FIG. 4. Principal component analysis ordination plot of the Cuatro Ciénegas metagenomes using the taxon-abundance matrix constructed with deduced protein markers in the green (GP) and red (RP) microbial mats and Pozas Azules (PA) and Río Mesquites (RM) stromatolites. The names of the top 14 genera most greatly contributing to the principal components are indicated. Color images available online at www.liebertonline.com/ast

sequences from the well-known mat-forming cosmopolitan *Microcoleus chthonoplastes* (Garcia-Pichel *et al.,* 1996) were found in GP. Order Nostocales was the most abundant group that contained known mat-forming Cyanobacteria. Abundant filamentous Nostocales from family Rivulariaceae (*Calothrix* and *Tolypothrix*) have been observed in other CCB microbial mats and stromatolites (Garcia-Pichel *et al.,* 2002; Domínguez-Escobar *et al.,* 2011), and this group has also been reported to grow better under low-phosphorus conditions (Berrendero *et al.,* 2008). Unfortunately, family Rivulariaceae cannot be accurately determined from protein sequences, since no reference genomes are available yet, which makes cultured members of this group good targets for future sequencing projects. In consequence, the mat-forming filamentous cyanobacteria in the green mat could be uncharacterized species from order Oscillatoriales, Nostocales, or both.

In contrast, most sequences from Oscillatoriales in the red mat show a remarkable similarity to *M. chthonoplastes* (BLAST hits below 1e-20), despite the fact that they comprise only 4% of the total metagenome. Nevertheless, only one 16S rRNA sequence from the clone library and only two protein-coding markers from the red mat displayed affinity to this organism. This is what would be expected from organisms with large genomes (*M. chthonoplastes* has a genome of 7.36 Mb), and this discrepancy has also been noted in other microbial mats (Sørensen *et al.,* 2005; Dillon *et al.,* 2009). All approaches agreed in that the red mat was dominated by *Pseudomonas*. Although unexpected, this is not surprising in that the existence of mats dominated by noncyanobacterial organisms like *Thioploca* (Oschmann, 2000), *Beggiatoa* (Mussmann *et al.,* 2007), and even fungi mycelia (Verrecchia, 2000) have also been reported. Moreover, *Pseudomonas* species are recognized for their metabolic versatility, and many are acknowledged to form biofilms (Silby *et al.,* 2009). More remarkable is the finding of several closely related but distinct lineages with affinity to *P. fluorescens.* Since the metagenomes described here represent only a snapshot of the red mat community evolution in time, we can infer that the composition observed in the red mats is actually a radiative burst of *Pseudomonas* and other heterotrophs after a desiccation disturbance. Even though it has already been shown that environmental heterogeneity promotes diversification in

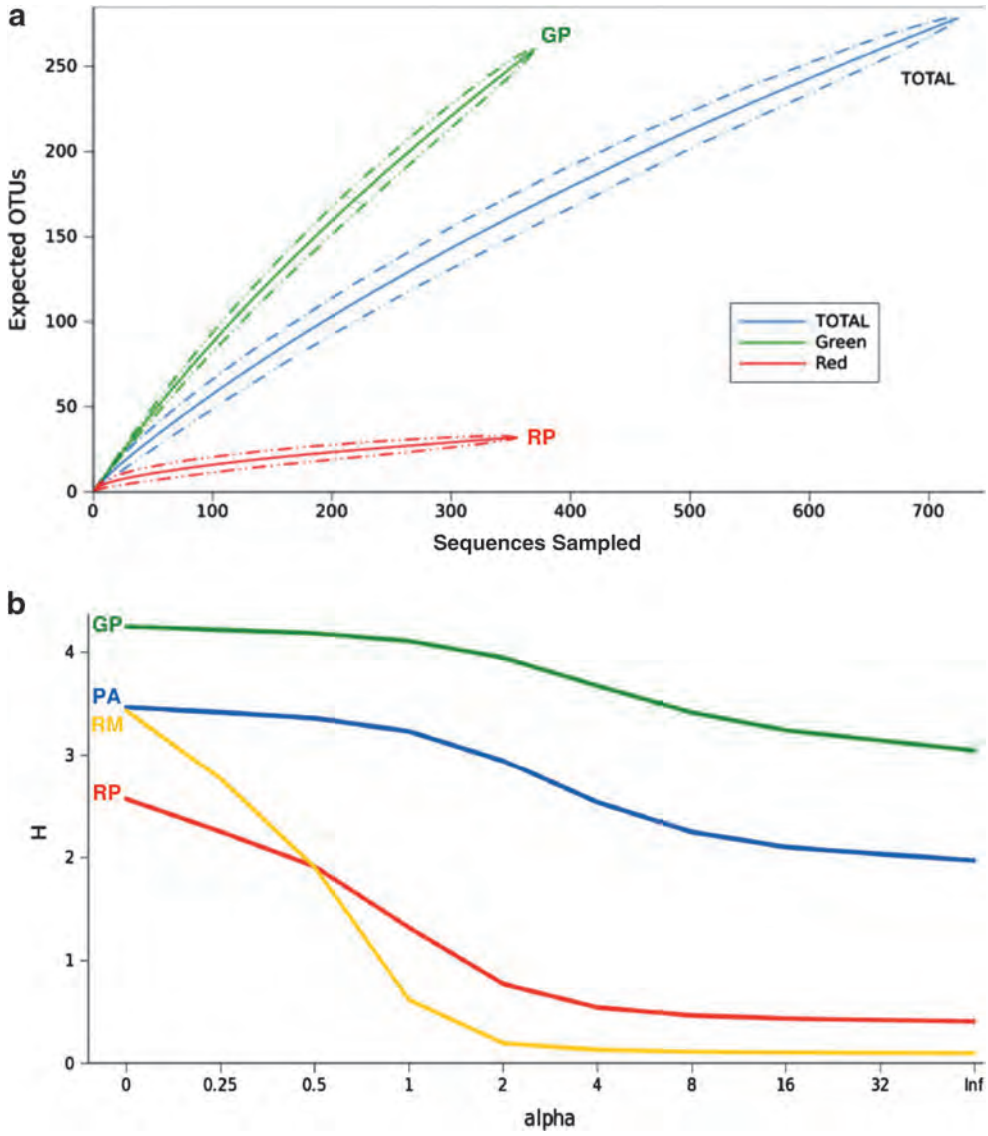


FIG. 5. (**a**) Estimated rarefaction curves for the 16S rRNA gene clone libraries, with OTUs clustered at 97% identity. (**b**) Renyi's entropy profile for the four Cuatro Ciénegas metagenomes. Profiles were calculated with the taxon-abundance matrix constructed with protein marker analyses. Green mat (GP), red mat (RP), Río Mesquites (RM), and Pozas Azules (PA). Color images available online at www.liebertonline.com/ast

P. fluorescens populations (Rainey and Travisano, 1998), the evolutionary consequences of these continuous disturbances are unknown. Our results suggest that this could lead to an adaptive radiation burst over the long term, but the observed pattern could also be caused simply by the metabolic plasticity of *Pseudomonas.* The similar composition of a second desiccation pond (red mat 2) suggests that the blooming of *Pseudomonas* could be an effect of the seasonal cycles common to all desiccation ponds in this area.

4.2. Taxonomic composition and functional trait analysis of the most abundant organisms

Although most organisms in these mats are uncharacterized, we can rely on the abundant literature on related species and genome projects to infer the general metabolic traits of the organisms that are abundant, according to the three methods, and have known and well-characterized relatives. In the green mat, despite the large amount of organisms represented, the main primary producers are likely to be Cyanobacteria, more specifically the abundant picooxygenic *Synechococcus* (Chroococcales). Moreover, the sequences found reveal high affinity to species OS-A and OS-B' found in thermophilic mats in Yellowstone (Ward *et al.*, 1998). Interestingly, these are the only *Synechococcus* genomes to carry nitrogen-fixing genes, as well as other genes that allow them to grow in oligotrophic environments such as genes for urea cycle, sulfolipid production, nitrate, phosphate and phosphonate utilization, and the biosynthesis and degradation of cyanophycin, a reserve polymer rich in nitrogen (Bhaya *et al.*, 2007). This suggests that the most abundant phototrophs in the mat are not filamentous bacteria but efficient unicellular oligotrophic cyanobacteria and potentially important contributors to primary

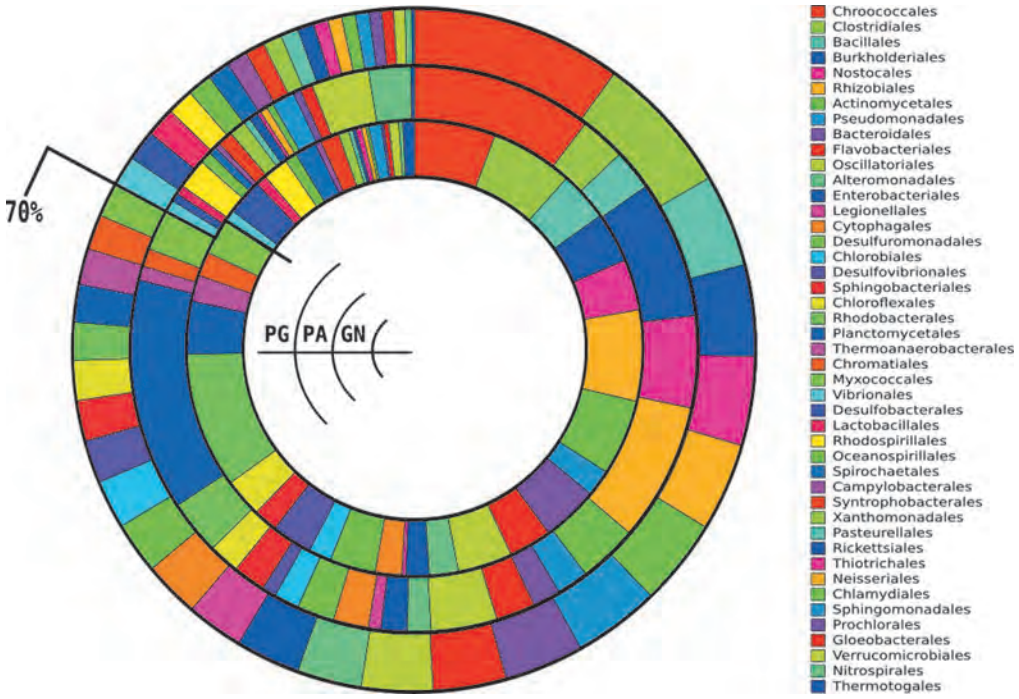


FIG. 6. Relative abundance distributions of the most abundant orders by means of the all-reads metagenomic complement approach. Abundant taxa are considered as those comprising 75% of the total annotated reads in each metagenome. Green mat, outer circle; Pozas Azules, middle circle; Guerrero Negro, inner circle. The black line indicates where the cumulative proportion of the three samples reaches 70%, just after Myxococcales.

productivity and inorganic nutrient incorporation. Nevertheless, large filamentous cyanobacteria might still be the dominant organisms when considering biomass contribution (S.J. Green *et al.,* 2008).

The organic compounds synthesized by these primary producers in the green mat are likely recycled by the abundant and efficient oxygenic heterotrophs like *Legionella,* which are common in oligotrophic water bodies, even though they are auxotrophic for seven amino acids, including L-cysteine, and require iron salts for growth (Declerck, 2010). They satisfy their nutrient requirements from living organisms (such as amoeboid hosts and microbial biofilms) or decaying organic matter (Declerck, 2010). Therefore, the green microbial mat appears to provide physicochemical protection from the environment, while it concentrates large amounts of nutrients. *Algoriphagus* species are aerobic heterotrophs characterized by the degradation of high-molecular-weight polysaccharides, with *Algoriphagus* sp. PR1 genome coding for 145 polysaccharide degradation enzymes (Alegado *et al.,* 2011). In the anoxic part of microbial mats, *Desulfococcus* species are able to completely degrade short- and long-chain fatty acids and aromatic compounds coupled to the reduction of sulfate to hydrogen sulfide (Muyzer and Stams, 2008) and can fix CO_2 via the Wood-Ljungdahl pathway (Platen *et al.,* 1990). In contrast, the green-sulfur bacteria *Chloroherpethon* utilizes sulfide ions as electron donors for carbon assimilation (Bryant and Frigaard, 2006), while it contributes with carbon and nitrogen fixation to primary productivity and the production of phosphate granules via polyphosphate kinases. Even though Chlorobii are not properly classified as oligotrophs, they specialize in low light use and have been shown to dominate in environments with low organic matter input (Gonzalez *et al.,* 2011).

The red mat, in contrast, is dominated by *Pseudomonas,* a metabolically versatile, ubiquitous heterotroph with broad catabolic and transport capabilities (Moore *et al.,* 2006). We corroborated the *Pseudomonas* dominance with another clone library built from a second red mat, located in a similar desiccation pond that was adjacent to the first red mat pond (Fig. 3). *Pseudomonas fluorescens* species are aerobes; but some strains can use nitrate as electron acceptor instead of oxygen, and their genomes code for a large number of high-specificity nutrient transporters, as well as a large array of efflux systems for metal, organic solvent, and antibiotic detoxification (Silby *et al.,* 2009). Their large genomes, high-affinity transporters, and broad sensing capabilities position *Pseudomonas* as copiotrophs (organisms adapted to grow in nutrient-rich environments) (Lauro *et al.,* 2009). Nevertheless, *P. fluorescens* can use several carbon sources at very low concentrations, with a marked preference for amino acids (van der Kooij *et al.,* 1982), and can form biofilms under starvation conditions to maximize exposure to diluted nutrients (Kroukamp *et al.,* 2010), which suggests that they are at least tolerant to nutrient depletion. More relevant to the red mat environment, several *P. fluorescens* strains can efficiently solubilize mineralized inorganic phosphates (Fankem *et al.,* 2008; Woo *et al.,* 2010). Acidification increases solubilization by the production of carboxylic acids that have a high affinity for phosphate-bound ions (Khan *et al.,* 2009) and reaches an optimum between a pH of 4.5 and 5.5 (Fankem *et al.,* 2008). Phosphate solubilization is maximized with citrate and malate (Fankem *et al.,* 2008), which are two of the most abundant carboxylic acids produced by *P. fluorescens*

(Vyas and Gulati, 2009). None of the OTUs from the mats belonged to previously reported abundant species in the CCB (*P. mendocina, P. otitidis*, or *P. cuatrocienegasensis*, Escalante *et al.*, 2009).

The aerobic heterotroph *Chitinophaga* is also abundant in the red mat. It is also a copiotroph, with a large genome of 9.1 Mb that contains a large and diverse collection of enzymes for the degradation of sugars (169 glycosyl hydrolases; Del Rio *et al.*, 2010), most notably the degradation of chitin and casein (Sangkhobol and Skerman, 1981). In contrast, other abundant members in the red mat include the oligotrophic ultramicrobia (cell volume $< 0.1\ \mu m^3$) *Sphingopixis alaskensis* (Lauro *et al.*, 2009) and *Janthinobacterium* sp. Marseille (*Minibacterium massiliensis*, Audic *et al.*, 2007), whose common traits include very small cells and genomes, preference for amino acids over sugars, active iron scavenging, and a broad array of high-affinity but low-specificity transport systems. In addition, *S. alaskensis* exhibits degradation of high-energy yielding fatty acids and has a large number of genes for secondary metabolite catabolism and detoxification (Lauro *et al.*, 2009). Surprisingly, the phosphorus metabolism in *S. alaskensis* is reduced to the nonspecific alkaline phosphatases and a single ATP-dependent transporter, which is the only ABC transporter in the genome (Williams *et al.*, 2009). In contrast, the genome of *Janthinobacterium* sp. contains ABC transporters for sulfate, sulfonate, thiosulfate, nitrate, most amino acids, phosphate, and phosphonates (Audic *et al.*, 2007).

While *Microcoleus chthonoplastes*, which is also represented in the red mat, has a large genome (7.36 Mb) and is capable of performing both oxygenic and anoxygenic photosynthesis in the presence of sulfide (Stal, 1991), it cannot utilize nitrate or fix nitrogen (Zimmermann, 1989). Mats dominated by *M. chthonoplastes* show very low nitrogen fixation (Camacho and de Wit, 2003), and most strains analyzed to date lack the nitrogenase gene (Bolhuis *et al.*, 2009). Moreover, anoxygenic photosynthesis and nitrogen fixation in *M. chthonoplastes* mats are strongly dependent on phosphate abundance (Zimmermann, 1989; Camacho and de Wit, 2003), so it is not surprising that its genome contains genes for low- and high-affinity transporters (*pitA* and *pstSBCA*) and for the synthesis of the reservoir compound polyphosphate (*ppk*).

4.3. Food or stability? Oligotrophy and disturbance as limiting factors for the development of complex microbial mat communities

The functional trait analysis showed that, in the green mat, oligotrophic organisms are abundant primary producers that can fix nitrogen and optimize phosphorus utilization. Moreover, at least three of the five most abundant organisms are capable of CO_2 fixation. The fixed organic matter would then be recycled by non-oligotrophic heterotrophs. In contrast, oligotrophic organisms in the red mat are heterotrophs, while the dominant organisms are versatile copiotrophs. This reveals two very different strategies to cope with an oligotrophic environment, one mainly based on autotrophic primary production and the other on very efficient heterotrophic recycling. The results suggest that oligotrophy is not a limiting factor for the development of complex and functionally diverse microbial communities.

The analysis of diversity revealed that the green mat was by far the most diverse community and that communities in the more stable environments (the green mat and Pozas Azules) harbor higher richness and evenness than those in the more variable environments (the red mat and Río Mesquites). This suggests that disturbance frequency is a determinant factor of community structure in microbial mats, which is consistent with the exclusion of microbial mats by disturbing eukaryotes (Cohen, 1989; Bebout *et al.*, 2002). In the red mat, the disturbance regime is apparently too frequent and may drive the reduction in abundance of several species, most notably autotrophs. These systems, as those growing in hyperthermophilic environments, demonstrate that microbial mats can exist with a simplified diversity, as long as a chemical gradient exists where primary producers are present and their nutrients are recycled by heterotrophs.

4.4. A common high-rank taxonomic composition as a result of trait conservation in microbial mats

Experimental studies on Guerrero Negro microbial mats have reported that, although salinity seems to affect community structure, cyanobacterial communities are only modestly affected by changes in sulfate and salinity concentrations (S.J. Green *et al.*, 2008). Mathematical models of nutrient and population dynamics in microbial mats where complexity is reduced to a few functional groups show a large resemblance to natural mats (Decker *et al.*, 2005). Moreover, functionally similar mats develop under varying oxygen concentrations because oxygenic phototrophs create a similar oxygen gradient in the upper layers (Herman and Kump, 2005). The role of oxygenic phototrophs is stressed not only as a source of nutrient incorporation but also as the generators of this gradient themselves. The potential metabolic differences between the two mats are mainly a result of differences between the proportion of photoautotrophs in the mats, which suggests that primary productivity is determinant to the community structure complexity. Since an unexpected diversity has been found in systems that produce a similar layered pattern with steep biogeochemical gradient (Jorgensen *et al.*, 1986; Stolz, 2000; Kunin *et al.*, 2008), it would appear that microbial mats are completely independent of the taxonomic composition of its conforming species and that similar systems may arise in any place where these gradient conditions are met. Hence, we compared our microbial mat communities with similar structures developed under different environmental conditions, the stromatolites from the CCB, and the hypersaline mat from Guerrero Negro. The recently published literature on microbialite community diversity comprising microbial mats (Ley *et al.*, 2006; Abed *et al.*, 2007; Allen *et al.*, 2009), stromatolites (Burns *et al.*, 2004; Papineau *et al.*, 2005; Baumgartner *et al.*, 2009), and endoevaporites (Sahl *et al.*, 2008) reveals that microbial mat communities have a far larger expected richness than stromatolites and endoevaporites (Table S2b in Supplementary Material). As expected, we found very few shared organisms between the two CCB mats, and no species were shared between the four systems from the CCB. Moreover, only two genomes (*M. chthonoplastes* and *Desulfococcus oleovorans*) showed high recruitment in both Guerrero Negro and CCB metagenomes. However, a larger similarity at higher taxonomic levels was observed between the Guerrero Negro

microbial mat, the green mat, and the Pozas Azules stromatolite even though these two mats were geographically separated by a linear distance of *ca.* 1200 km. The pattern appears unexpected, since other clone library–based investigations in which both stromatolites and microbial mats from the same sampling site were analyzed produced very different community compositions (for example Burns *et al.,* 2004; Allen *et al.,* 2009).

The finding of a common phylogenetic pattern at high taxonomic ranks suggests that layered microbial communities at the water-sediment interface assemble in biogeochemical gradients that fill defined ecological niches (photosynthesis, sulfate reduction, heterotrophy) according to their functional traits and independently of their phylogeny. It also supports the theory of the assembly of bacterial communities by functional genes rather than species (Burke *et al.,* 2011) and the existence of general functional traits shared by organisms at deep phylogenetic nodes (J.L. Green *et al.,* 2008; Philippot *et al.,* 2010). Under this model, phylogenetically unrelated species are able to colonize the same niche in an ecosystem as long as they are ecologically equivalent (same trophic level or metabolic function), just as similar ecosystems will have communities with common functional "guilds," but species within these guilds will be selected at random (Burke *et al.,* 2011). This explains the success of microbial matlike communities because they benefit from a simple set of environmental requirements, namely, a water-sediment interface and a steep physicochemical gradient. Hence, complex communities are likely to be found wherever these conditions are met, with metabolically diverse functional guilds benefiting from the generation of diverse microniches and from the environmental buffering the mat structure provides. This would suggest that the prevalence of remarkably similar, but taxonomically distinct, mat structures in stable environments is a natural consequence of the undisturbed association of metabolically diverse organisms that exploit a locally microdiverse niche. The identification of a conserved taxonomic group core also provides a base system with which to test the effects of different kinds of disturbances on complex microbial ecosystems, while it also sets a start-up for the design of synthetic artificial model microbial communities in mesocosms.

Microbial mats appear to be the most ancient and pervasive ecosystems because they conform biogeochemical structures that are likely to be found wherever a gradient occurs across a sediment-water interface, and they comprise physicochemical structures that protect organisms from environmental extremes (S.J. Green *et al.,* 2008) while concentrating nutrients. This could explain their success even under the low-phosphorus environments of the Precambrian oceans (Papineau, 2010). The fact that their structure depends more on the ecological traits than the taxonomic component of their members suggests that analogues of these kinds of ecosystems are also the most likely structures to be found beyond the confines of Earth. Their organosedimentary nature makes them ideal targets for the detection of evidence of former life in the geological record by means of the identification of stable isotopic signatures and molecules of biological origin.

Currently, a wide array of biosignatures has been proposed with which to narrow the number of astrological bodies that could host life, such as photosynthetic pigments (Seager *et al.,* 2005), sulfur gases (Domagal-Goldman *et al.,* 2011), methane (Hoehler *et al.,* 2001), and ammonia (Des Marais *et al.,* 2002). Theoretically, it would seem more likely to detect a complex community of unknown organisms than a simple one, since a complex community would contain and exude a larger variety of metabolic products, which would increase the probability of detecting one of these as biosignatures. Our results suggest that complex communities can be found in environments where nutrient concentrations are very low, which underscores the importance of considering microhabitats in otherwise nonviable environments. However, an environment with a frequent disturbance regime might significantly reduce diversity, while environments with intense disturbances might not host life at all. This suggests that more complex and, hence, more detectable communities are likely to be found in more stable environments, so that efforts in the search for life should incorporate disturbance intensity and frequency when narrowing the list of planets where search efforts are to be directed.

5. Conclusion

The CCB displays astounding microbial diversity despite being a mixed oligotrophic environment, which makes its ponds desirable astrobiological experimental models. The microbial mats from this study revealed two extremes of diversity, with the green mat showing a large diversity above phylum level and the red mat exhibiting diversity below genus level. Our results widen the ranges within which life may be found, since they show that complex communities can develop in environments with nutrient concentrations below detection levels, even though phosphate is readily mineralized and made biologically unavailable. However, our results also stress the need to incorporate disturbance intensity and frequency models into a more precise definition of astrobiological targets, since high disturbance regimes can lower diversity (with a concomitant lowering of measurable biosignatures) and even prevent the development of life.

Acknowledgments

We thank Rodrigo Gonzalez Chauvet for technical logistics and field assistance; E. Lopez, A. Islas, V. Lopez, F. Reverchon, E. Rebollar, M. Avitia, and A. Gutierrez for assistance in sample collection and DNA isolation; and Laura Espinosa from IE/UNAM for laboratory and technical assistance. We thank three anonymous reviewers, whose insightful comments greatly improved this manuscript. We also thank O. Rodriguez and C. Bixler for their inertiatic propulsion. This work was supported by grants CONACyT 057507, SEMARNAT 2006-C01-23459, and WWF-Alianza Carlos Slim L039 to V.S., and CINVESTAV-multidisciplinario to G.O.A. The manuscript was written while V.S. and L.E.E. were on sabbatical at UCI with support from DGAPA to V.S. and UC-Mexus to L.E.E. G.B.R. was supported with Ph.D. scholarship CONACYT 196814 and Programa de Posgrado en Ciencias Biomédicas UNAM.

Abbreviations

CCB, Cuatro Ciénegas Basin; OTUs, operational taxonomic units.

References

Abed, R.M., Kohls, K., and de Beer, D. (2007) Effect of salinity changes on the bacterial diversity, photosynthesis and oxygen consumption of cyanobacterial mats from an intertidal flat of the Arabian Gulf. *Environ Microbiol* 9:1384–1392.

Alegado, R.A., Ferriera, S., Nusbaum, C., Young, S.K., Zeng, Q., Imamovic, A., Fairclough, S.R., and King, N. (2011) Complete genome sequence of *Algoriphagus* sp. PR1, bacterial prey of a colony-forming choanoflagellate. *J Bacteriol* 193:1485–1486.

Allen, M.A., Goh, F., Burns, B.P., and Neilan, B.A. (2009) Bacterial, archaeal and eukaryotic diversity of smooth and pustular microbial mat communities in the hypersaline lagoon of Shark Bay. *Geobiology* 7:82–96.

Allwood, A.C., Walter, M.R., Kamber, B.S., Marshall, C.P., and Burch, I.W. (2006) Stromatolite reef from the Early Archaean era of Australia. *Nature* 441:714–718.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J., and Weightman, A.J. (2006) New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl Environ Microbiol* 72:5734–5741.

Audic, S., Robert, C., Campagna, B., Parinello, H., Claverie, J.M., Raoult, D., and Drancourt, M. (2007) Genome analysis of *Minibacterium massiliensis* highlights the convergent evolution of water-living bacteria. *PLoS Genet* 3:e138.

Awramik, S.A. (1984) Ancient stromatolites and microbial mats. In *Microbial Mats: Stromatolites,* edited by Y. Cohen, R.W. Castenholz, and H.O. Halvorson, Alan R. Liss, New York, pp 1–22.

Awramik, W.S. (1976) Gunflint stromatolites: microfossil distribution in relation to stromatolite morphology. In *Developments in Sedimentology,* Vol. 20, edited by M.R. Walter, Elsevier, Amsterdam, pp 311–320.

Baumgartner, L.K., Spear, J.R., Buckley, D.H., Pace, N.R., Reid, R.P., Dupraz, C., and Visscher, P.T. (2009) Microbial diversity in modern marine stromatolites, Highborne Cay, Bahamas. *Environ Microbiol* 11:2710–2719.

Bebout, B.M., Carpenter, S.P., Des Marais, D.J., Discipulo, M., Embaye, T., Garcia-Pichel, F., Hoehler, T.M., Hogan, M., Jahnke, L.L., Keller, R.M., Miller, S.R., Prufert-Bebout, L.E., Raleigh, C., Rothrock, M., and Turk, K. (2002) Long-term manipulations of intact microbial mat communities in a greenhouse collaboratory: simulating Earth's present and past field environments. *Astrobiology* 2:383–402.

Bebout, B.M., Hoehler, T.M., Thamdrup, B., Albert, D., Carpenter, S.P., Hogan, M., Turk, K., and Des Marais, D.J. (2004) Methane production by microbial mats under low sulphate concentrations. *Geobiology* 2:87–96.

Bent, S.J. and Forney, L.J. (2008) The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. *ISME J* 2:689–695.

Berrendero, E., Perona, E., and Mateo, P. (2008) Genetic and morphological characterization of *Rivularia* and *Calothrix* (Nostocales, Cyanobacteria) from running water. *Int J Syst Evol Microbiol* 58:447–460.

Besemer, J. and Borodovsky, M. (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* 33:W451–W454.

Bhaya, D., Grossman, A.R., Steunou, A.S., Khuri, N., Cohan, F.M., Hamamura, N., Melendrez, M.C., Bateson, M.M., Ward, D.M., and Heidelberg, J.F. (2007) Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J* 1:703–713.

Bjerrum, C.J. and Canfield, D.E. (2002) Ocean productivity before about 1.9 Gyr ago limited by phosphorus adsorption onto iron oxides. *Nature* 417:159–162.

Bolhuis, H., Severin, I., Confurius-Guns, V., Wollenzien, U.I.A., and Stal, L.J. (2009) Horizontal transfer of the nitrogen fixation gene cluster in the cyanobacterium *Microcoleus chthonoplastes*. *ISME J* 4:121–130.

Bottos, E.M., Vincent, W.F., Greer, C.W., and Whyte, L.G. (2008) Prokaryotic diversity of arctic ice shelf microbial mats. *Environ Microbiol* 10:950–966.

Breitbart, M., Hoare, A., Nitti, A., Siefert, J., Haynes, M., Dinsdale, E., Edwards, R., Souza, V., Rohwer, F., and Hollander, D. (2009) Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Ciénegas, Mexico. *Environ Microbiol* 11:16–34.

Bryant, D.A. and Frigaard, N.U. (2006) Prokaryotic photosynthesis and phototrophy illuminated. *Trends Microbiol* 14:488–496.

Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S., and Thomas, T. (2011) Bacterial community assembly based on functional genes rather than species. *Proc Natl Acad Sci USA* 108:14288–14293.

Burns, B.P., Goh, F., Allen, M., and Neilan, B.A. (2004) Microbial diversity of extant stromatolites in the hypersaline marine environment of Shark Bay, Australia. *Environ Microbiol* 6: 1096–1101.

Camacho, A. and de Wit, R. (2003) Effect of nitrogen and phosphorus additions on a benthic microbial mat from a hypersaline lake. *Aquat Microb Ecol* 32:261–273.

Cohen, Y. (1989) Photosynthesis in cyanobacterial mats and its relation to the sulfur cycle: a model for microbial sulfur interactions. In *Microbial Mats: Physiological Ecology of Benthic Microbial Communities*, edited by Y. Cohen and E. Rosenberg, American Society for Microbiology, Washington DC, pp 22–36.

Connell, J.H. (1978) Diversity in tropical rain forests and coral reefs. *Science* 199:1302–1310.

Decker, K., Potter, C., Bebout, B.M., Des Marais, D.J., Carpenter, S., Discipulo, M., Hoehler, T.M., Miller, S.R., Thamdrup, B., Turk, K.A., and Visscher, P.T. (2005) Mathematical simulation of the diel O, S, and C biogeochemistry of a hypersaline microbial mat. *FEMS Microbiol Ecol* 52:377–395.

Declerck, P. (2010) Biofilms: the environmental playground of *Legionella pneumophila*. *Environ Microbiol* 12: 557–566.

Del Rio, T.G., Abt, B., Spring, S., Lapidus, A., Nolan, M., Tice, H., Copeland, A., Cheng, J., Chen, F., Bruce, D., Goodwin, L., Pitluck, S., Ivanova, N., Mavromatis, K., Mikhailova, N., Pati, A., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y., Jeffries, C., Chain, P., Saunders, E., Detter, J., Brettin, T., Rohde, M., Göker, M., Bristow, J., Eisen, J., Markowitz, V., Hugenholtz, P., Kyrpides, N., Klenk, H., and Lucas, S. (2010) Complete genome sequence of *Chitinophaga pinensis* type strain (UQM 2034T). *Stand Genomic Sci* 2:87.

Des Marais, D.J., Harwit, M.O., Jucks, K.W., Kasting, J.F., Lin, D.N.C., Lunine, J.I., Schneider, J., Seager, S., Traub, W.A., and Woolf, N.J. (2002) Remote sensing of planetary properties and biosignatures on extrasolar terrestrial planets. *Astrobiology* 2:153–181.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72:5069–5072.

Dillon, J.G., Miller, S., Bebout, B., Hullar, M., Pinel, N., and Stahl, D.A. (2009) Spatial and temporal variability in a stratified

hypersaline microbial mat community. *FEMS Microbiol Ecol* 68:46–58.

Domagal-Goldman, S.D., Meadows, V.S., Claire, M.W., and Kasting, J.F. (2011) Using biogenic sulfur gases as remotely detectable biosignatures on anoxic planets. *Astrobiology* 11: 419–441.

Domínguez-Escobar, J., Beltrán, Y., Bergman, B., Díez, B., Ininbergs, K., Souza, V., and Falcón, L.I. (2011) Phylogenetic and molecular clock inferences of cyanobacterial strains within Rivulariaceae from distant environments. *FEMS Microbiol Lett* 316:90–99.

Elser, J.J., Schampel, J.H., Garcia-Pichel, F., Wade, B.D., Souza, V., Eguiarte, L., Escalante, A., and Farmer, J.D. (2005) Effects of phosphorus enrichment and grazing snails on modern stromatolitic microbial communities. *Freshw Biol* 50:1808–1825.

Elser, J.J., Watts, J., Schampel, J.H., and Farmer, J. (2006) Early Cambrian food webs on a trophic knife-edge? A hypothesis and preliminary data from a modern stromatolite-based ecosystem. *Ecological Letters* 9:295–303.

Elshahed, M.S., Senko, J.M., Najar, F.Z., Kenton, S.M., Roe, B.A., Dewers, T.A., Spear, J.R., and Krumholz, L.R. (2003) Bacterial diversity and sulfur cycling in a mesophilic sulfide-rich spring. *Appl Environ Microbiol* 69:5609–5621.

Emerson, D. and Revsbech, N.P. (1994) Investigation of an iron-oxidizing microbial mat community located near Aarhus, Denmark: field studies. *Appl Environ Microbiol* 60:4022–4031.

Escalante, A.E., Eguiarte, L.E., Espinosa-Asuar, L., Forney, L.J., Noguez, A.M., and Souza-Saldivar, V. (2008) Diversity of aquatic prokaryotic communities in the Cuatro Cienegas basin. *FEMS Microbiol Ecol* 65:50–60.

Escalante, A.E., Caballero-Mellado, J., Martínez-Aguilar, L., Rodríguez-Verdugo, A., González-González, A., Toribio-Jiménez, J., and Souza, V. (2009) *Pseudomonas cuatrocienegasensis* sp. nov., isolated from an evaporating lagoon in the Cuatro Ciénegas valley in Coahuila, Mexico. *Int J Syst Evol Microbiol* 59:1416–1420.

Fankem, H., Ngo Nkot, L., Deubel, A., Quinn, J., Merbach, W., Etoa, F.-X., and Nwaga, D. (2008) Solubilization of inorganic phosphates and plant growth promotion by strains of *Pseudomonas fluorescens* isolated from acidic soils of Cameroon. *Afr J Microbiol Res* 2:171–178.

Foster, J.S. and Mobberley, J.M. (2010) Past, present, and future: microbial mats as models for astrobiological research. In *Microbial Mats: Modern and Ancient Microorganisms in Stratified Systems,* Springer, Dordrecht, pp 563–582.

Garcia-Pichel, F., Prufert-Bebout, L., and Muyzer, G. (1996) Phenotypic and phylogenetic analyses show *Microcoleus chthonoplastes* to be a cosmopolitan cyanobacterium. *Appl Environ Microbiol* 62:3284–3291.

Garcia-Pichel, F., Wade, B.D., and Farmer J.D. (2002) Jet-suspended, calcite-ballasted cyanobacterial waterwarts in a desert spring. *J Phycol* 38:420–428.

Gerdes, G. (2010) What are microbial mats? In *Microbial Mats: Modern and Ancient Microorganisms in Stratified Systems*, edited by J. Seckbach and A. Oren, Springer, Dordrecht, pp 3–25.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., Rappé, M.S., Short, J.M., Carrington, J.C., and Mathur, E.J. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309:1242–1245.

Golombek, M., Grant, J., Vasavada, A.R., Grotzinger, J., Watkins, M., Kipp, D., Noe Dobrea, E., Griffes, J., and Parker, T. (2011) Final four landing sites for the Mars Science Laboratory [abstract 1520]. In *42nd Lunar and Planetary Science Conference,* Lunar and Planetary Institute, Houston.

Gonzalez, B.C., Iliffe, T.M., Macalady, J.L., Schaperdoth, I., and Kakuk, B. (2011) Microbial hotspots in anchialine blue holes: initial discoveries from the Bahamas. *Hydrobiologia* 677: 149–156.

Green, J.L., Bohannan, B.J.M., and Whitaker, R.J. (2008) Microbial biogeography: from taxonomy to traits. *Science* 320: 1039–1043.

Green, S.J. and Jahnke, L.L. (2010) Molecular investigations and experimental manipulations of microbial mats: a view to paleomicrobial ecosystems. In *Microbial Mats: Modern and Ancient Microorganisms in Stratified Systems,* edited by J. Seckbach and A. Oren, Springer, Dordrecht, pp 185–208.

Green, S.J., Blackford, C., Bucki, P., Jahnke, L.L., and Prufert-Bebout, L. (2008) A salinity and sulfate manipulation of hypersaline microbial mats reveals stasis in the cyanobacterial community structure. *ISME J* 2:457–470.

Grime, J.P. (1973) Competitive exclusion in herbaceous vegetation. *Nature* 242:344–347.

Grotzinger, J.P. and Knoll, A.H. (1999) Stromatolites in Precambrian carbonates: evolutionary mileposts or environmental dipsticks? *Annu Rev Earth Planet Sci* 27:313–358.

Hall, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* 41:95–98.

Herman, E. and Kump, L. (2005) Biogeochemistry of microbial mats under Precambrian environmental conditions: a modelling study. *Geobiology* 3:77–92.

Hoehler, T.M., Bebout, B.M., and Des Marais, D.J. (2001) The role of microbial mats in the production of reduced gases on the early Earth. *Nature* 412:324–327.

Huang, Y., Gilna, P., and Li, W. (2009) Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* 25:1338–1340.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386.

Inskeep, W.P., Rusch, D.B., Jay, Z.J., Herrgard, M.J., Kozubal, M.A., Richardson, T.H., Macur, R.E., Hamamura, N., Jennings, R.D., Fouke, B.W., Reysenbach, A.-L., Roberto, F., Young, M., Schwartz, A., Boyd, E.S., Badger, J.H., Mathur, E.J., Ortmann, A.C., Bateson, M., Geesey, G., Frazier, M., and Rodriguez-Valera, F. (2010) Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* 5:e9773.

Jorgensen, B.B., Cohen, Y., and Revsbech, N.P. (1986) Transition from anoxygenic to oxygenic photosynthesis in a *Microcoleus chthonoplastes* cyanobacterial mat. *Appl Environ Microbiol* 51: 408–417.

Kasting, J.F. (2001) Earth history. The rise of atmospheric oxygen. *Science* 293:819–820.

Khan, A., Jilani, G., Akhtar, M.S., Naqvi, S.M.S., and Rasheed, M. (2009) Phosphorus solubilizing bacteria: occurrence, mechanisms and their role in crop production. *J Agric Biol Sci* 1:48–58.

Kroukamp, O., Dumitrache, R.G., and Wolfaardt, G.M. (2010) Pronounced effect of the nature of the inoculum on biofilm development in flow systems. *Appl Environ Microbiol* 76:6025–6031.

Krumbein, W.E., Cohen, Y., and Shilo, M. (1977) Solar Lake (Sinai). 4. Stromatolitic cyanobacterial mats. *Limnol Oceanogr* 22:635–656.

Kunin, V., Raes, J., Harris, J.K., Spear, J.R., Walker, J.J., Ivanova, N., von Mering, C., Bebout, B.M., Pace, N.R., Bork, P., and

Hugenholtz, P. (2008) Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol Syst Biol* 4:198.

Lauro, F.M., McDougald, D., Thomas, T., Williams, T.J., Egan, S., Rice, S., DeMaere, M.Z., Ting, L., Ertan, H., Johnson, J., Ferriera, S., Lapidus, A., Anderson, I., Kyrpides, N., Munk, A.C., Detter, C., Han, C.S., Brown, M.V., Robb, F.T., Kjelleberg, S., and Cavicchioli, R. (2009) The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* 106:15527–15533.

Letunic, I. and Bork, P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 39:W475–W478.

Ley, R.E., Harris, J.K., Wilcox, J., Spear, J.R, Miller, S.R., Bebout, B.M., Maresca, J.A., Bryant, D.A., and Sogin, M.L. (2006) Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. *Appl Environ Microbiol* 72:3685–3695.

Li, W. (2009) Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics* 10:359.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüssmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., and Schleifer, K.H. (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* 32:1363–1371.

Markowitz, V.M., Chen, I.-M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Anderson, I., Lykidis, A., Mavromatis, K., Ivanova, N.N., and Kyrpides, N.C. (2009) The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res* 38:D382–D390.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R.A. (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386.

Meyer-Dombard, D.R., Swingley, W., Raymond, J., Havig, J., Shock, E.L., and Summons, R.E. (2011) Hydrothermal ecotones and streamer biofilm communities in the Lower Geyser Basin, Yellowstone National Park. *Environ Microbiol* 13:2216–2231.

Mills, H.J., Martinez, R.J., Story, S., and Sobecky, P.A. (2005) Characterization of microbial community structure in Gulf of Mexico gas hydrates: comparative analysis of DNA- and RNA-derived clone libraries. *Appl Environ Microbiol* 71:3235–3247.

Minckley, W.L. and Cole, G.A. (1968) Preliminary limnologic information on waters of the Cuatro Cienegas basin, Coahuila, Mexico. *Southwest Nat* 13:421–431.

Moore, E.R.B., Tindall, B.J., Martins Dos Santos, V., Pieper, D.H., Ramos, J.L., and Palleroni, N.J. (2006) Nonmedical pseudomonas. In *The Prokaryotes*, Vol. 6, edited by M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, and E. Stackebrandt, Springer, New York, pp 646–703.

Mulet, M., Lalucat, J., and García-Valdés, E. (2010) DNA sequence-based analysis of the *Pseudomonas* species. *Environ Microbiol* 12:1513–1530.

Mussmann, M., Hu, F.Z., Richter, M., de Beer, D., Preisler, A., Jørgensen, B.B., Huntemann, M., Glöckner, F.O., Amann, R., Koopman, W.J., Lasken, R.S., Janto, B., Hogg, J., Stoodley, P., Boissy, R., and Ehrlich, G.D. (2007) Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. *PLoS Biol* 5:e230.

Muyzer, G. and Stams, A.J.M. (2008) The ecology and biotechnology of sulphate-reducing bacteria. *Nat Rev Microbiol* 6:441–454.

Nitti, A., Daniels, C.A., Siefert, J., Souza, V., Hollander, D., and Breitbart, M. (2012) Spatially resolved genomic, stable isotopic, and lipid analyses of a modern freshwater microbialite from Cuatro Ciénegas, Mexico. *Astrobiology* 12:685–698.

Oschmann, W. (2000) Microbes and black shales. In *Microbial Sediments*, edited by R.E. Riding and S.M. Awramik, Springer Verlag, Berlin, pp 137–148.

Paerl, H.W. and Yannarell, A.C. (2010) Environmental dynamics, community structure and function in a hypersaline microbial mat. In *Microbial Mats: Modern and Ancient Microorganisms in Stratified Systems,* edited by J. Seckbach and A. Oren, Springer, Dordrecht, pp 423–444.

Paerl, H.W., Steppe, T.F., Buchan, K.C., and Potts, M. (2003) Hypersaline cyanobacterial mats as indicators of elevated tropical hurricane activity and associated climate change. *AMBIO: A Journal of the Human Environment* 32:87–90.

Papineau, D. (2010) Global biogeochemical changes at both ends of the proterozoic: insights from phosphorites. *Astrobiology* 10:165–181.

Papineau, D., Walker, J.J., Mojzsis, S.J., and Pace, N.R. (2005) Composition and structure of microbial communities from stromatolites of Hamelin Pool in Shark Bay, Western Australia. *Appl Environ Microbiol* 71:4822–4832.

Peimbert, M., Alcaraz, L.D., Bonilla-Rosso, G., Olmedo-Alvarez, G., García-Oliva, F., Segovia, L., Eguiarte, L.E., and Souza, V. (2012) Comparative metagenomics of two microbial mats at Cuatro Ciénegas Basin I: ancient lessons on how to cope with an environment under severe nutrient stress. *Astrobiology* 12:648–658.

Philippot, L., Andersson, S.G.E., Battin, T.J., Prosser, J.I., Schimel, J.P., Whitman, W.B., and Hallin, S. (2010) The ecological coherence of high bacterial taxonomic ranks. *Nat Rev Microbiol* 8:523–529.

Planavsky, N.J., Rouxel, O.J., Bekker, A., Lalonde, S.V., Konhauser, K.O., Reinhard, C.T., and Lyons, T.W. (2010) The evolution of the marine phosphate reservoir. *Nature* 467:1088–1090.

Platen, H., Temmes, A., and Schink, B. (1990) Anaerobic degradation of acetone by *Desulfococcus biacutus* spec. nov. *Arch Microbiol* 154:355–361.

R Development Core Team. (2006) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at http://www.r-project.org.

Rainey, P.B. and Travisano, M. (1998) Adaptive radiation in a heterogeneous environment. *Nature* 394:69–72.

Rodríguez-Mejía, J.L., Martinez-Anaya, C., Folch-Mallol, J.L., and Dantán-González, E. (2008) A two-step electrodialysis method for DNA purification from polluted metallic environmental samples. *Electrophoresis* 29:3239–3244.

Sahl, J.W., Pace, N.R., and Spear, J.R. (2008) Comparative molecular analysis of endoevaporitic microbial communities. *Appl Environ Microbiol* 74:6444–6446.

Sangkhobol, V. and Skerman, V.B.D. (1981) *Chitinophaga,* a new genus of chitinolytic myxobacteria. *Int J Syst Bacteriol* 31:285–293.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H.,

Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., and Weber, C.F. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75:7537–7541.

Seager, S., Turner, E.L., Schafer, J., and Ford, E.B. (2005) Vegetation's red edge: a possible spectroscopic biosignature of extraterrestrial plants. *Astrobiology* 5:372–390.

Silby, M.W., Cerdeño-Tárraga, A.M., Vernikos, G.S., Giddens, S.R., Jackson, R.W., Preston, G.M., Zhang, X.X., Moon, C.D., Gehrig, S.M., Godfrey, S.A., Knight, C.G., Malone, J.G., Robinson, Z., Spiers, A.J., Harris, S., Challis, G.L., Yaxley, A.M., Harris, D., Seeger, K., Murphy, L., Rutter, S., Squares, R., Quail, M.A., Saunders, E., Mavromatis, K., Brettin, T.S., Bentley, S.D., Hothersall, J., Stephens, E., Thomas, C.M., Parkhill, J., Levy, S.B., Rainey, P.B., and Thomson, N.R. (2009) Genomic and genetic analyses of diversity and plant interactions of *Pseudomonas fluorescens*. *Genome Biol* 10:R51.

Sørensen, K.B., Canfield, D.E., Teske, A.P., and Oren, A. (2005) Community composition of a hypersaline endoevaporitic microbial mat. *Appl Environ Microbiol* 71:7352–7365.

Souza, V., Espinosa-Asuar, L., Escalante, A.E., Eguiarte, L.E., Farmer, J., Forney, L., Lloret, L., Rodríguez-Martínez, J.M., Soberón, X., Dirzo, R., and Elser, J.J. (2006) An endangered oasis of aquatic microbial biodiversity in the Chihuahuan desert. *Proc Natl Acad Sci USA* 103:6565–6570.

Spear, J.R., Ley, R.E., Berger, A.B., and Pace, N.R. (2003) Complexity in natural microbial ecosystems: the Guerrero Negro experience. *Biol Bull* 204:168–173.

Stal, L.J. (1991) The metabolic versatility of the mat-building cyanobacteria *Microcoleus chthonoplastes* and *Oscillatoria limosa* and its ecological significance. *Algological Studies/Archiv für Hydrobiologie* 64:453–467.

Stolz, J.F. (2000) Structure of microbial mats and biofilms. In *Microbial Sediments,* edited by R.E. Riding and S.M. Awramik, Springer-Verlag, Berlin, pp 1–8.

Szynkiewicz, A., Ewing, R.C., Moore, C.H., Glamoclija, M., Bustos, D., and Pratt, L.M. (2010) Origin of terrestrial gypsum dunes—implications for martian gypsum-rich dunes of Olympia Undae. *Geomorphology* 121:69–83.

Tilman, D. (1990) Constraints and tradeoffs: toward a predictive theory of competition and succession. *Oikos* 58:3–15.

van der Kooij, D., Oranje, J., and Hijnen, W. (1982) Growth of *Pseudomonas aeruginosa* in tap water in relation to utilization of substrates at concentrations of a few micrograms per liter. *Appl Environ Microbiol* 44:1086–1095.

van Gemerden, H. (1993) Microbial mats: A joint venture. *Mar Geol* 113:3–25.

Varin, T., Lovejoy, C., Jungblut, A.D., and Vincent, W.F. (2010) Metagenomic profiling of Arctic microbial mat communities as nutrient scavenging and recycling systems. *Limnol Oceanogr* 55:1901–1911.

Verrecchia, E.P. (2000) Fungi and sediments. In *Microbial Sediments*, edited by R.E. Riding and S.M. Awramik, Springer-Verlag, Berlin, pp 69–75.

Vyas, P. and Gulati, A. (2009) Organic acid production *in vitro* and plant growth promotion in maize under controlled environment by phosphate-solubilizing fluorescent *Pseudomonas*. *BMC Microbiol* 9:174.

Ward, D.M., Ferris, M.J., Nold, S.C., and Bateson, M.M. (1998) A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. *Microbiol Mol Biol Rev* 62: 1353–1370.

Williams, T.J., Ertan, H., Ting, L., and Cavicchioli, R. (2009) Carbon and nitrogen substrate utilization in the marine bacterium *Sphingopyxis alaskensis* strain RB2256. *ISME J* 3:1036–1052.

Woo, S.M., Lee, M.K., Hong, I.S., Poonguzhali, S., and Sa, T.M. (2010) Isolation and characterization of phosphate solubilizing bacteria from Chinese cabbage. In *Soil Solutions for a Changing World,* Brisbane, Australia, pp 56–59.

Wu, M. and Eisen, J.A. (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* 9:R151.

Yannarell, A.C., Steppe, T.F., and Paerl, H.W. (2007) Disturbance and recovery of microbial community structure and function following Hurricane Frances. *Environ Microbiol* 9: 576–583.

Zhou, J., Bruns, M.A., and Tiedje, J.M. (1996) DNA recovery from soils of diverse composition. *Appl Environ Microbiol* 62: 316–322.

Zimmermann, C.F. (1989) Nitrogen and phosphorus uptake and release by the blue-green alga *Microcoleus lyngbyaceus*. *Journal of Aquatic Plant Management* 27:49–51.

Address correspondence to:
Valeria Souza
Departamento de Ecología Evolutiva
Instituto de Ecología
Universidad Nacional Autónoma de México
Apartado Postal 70-275
Ciudad Universitaria
Coyoacán 04510
México D.F.
México

E-mail: souza@servidor.unam.mx

Submitted 5 September 2011
Accepted 7 April 2012

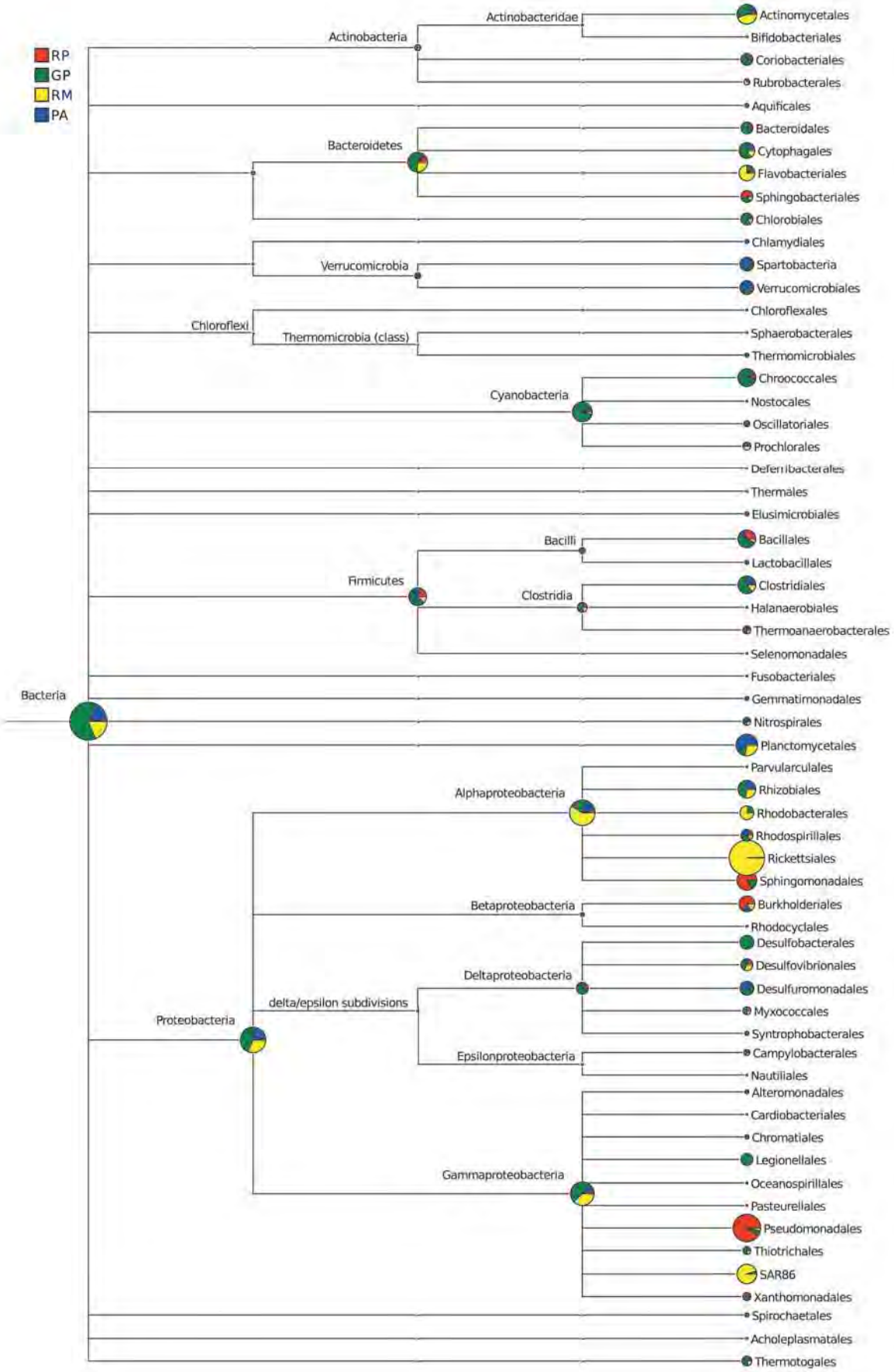


FIG. S1. Phylogram depicting the relative abundances at order level of the four CCB metagenomes according to the protein-coding gene abundance matrices.

16S rRNA gene clone library microbial diversity of several organosedimentary microbial communities including recently published studies on microbial mats, stromatolites and evaporites.

Name	System	Origin	ID %	Clones	OTUs	Chao	Reference
AbuDhabi-L	Intertidal mat	Sabkha, Arabian Gulf	97	105	51	NA	*Abed et al., 2007*
AbuDhabi-M	Intertidal mat	Sabkha, Arabian Gulf	97	116	39	NA	*Abed et al., 2007*
AbuDhabi-U	Intertidal mat	Sabkha, Arabian Gulf	97	112	58	NA	*Abed et al., 2007*
SharkBay-S	Hypersaline mat	Shark Bay, Australia	97	111	109	6216	*Allen et al., 2009*
SharkBay-S	Hypersaline mat	Shark Bay, Australia	99	111	111	1999	*Allen et al., 2009*
SharkBay-P	Hypersaline mat	Shark Bay, Australia	97	111	110	3053	*Allen et al., 2009*
SharkBay-P	Hypersaline mat	Shark Bay, Australia	99	111	110	3053	*Allen et al., 2009*
Arctic-W	Ice Shelf mat	Canadian High Arctic	98	128	52	106	*Bottos et al., 2008*
Arctic-M	Ice Shelf mat	Canadian High Arctic	98	189	105	243	*Bottos et al., 2008*
GuerreroNegro	Hypersaline mat	Guerrero Negro, México	99	1586	752	1000	*Ley et al., 2006*
CCC-PG	Oligotrophic mat	Cuatrociénegas, México	99	354	342	4034	*This work*
CCC-PG	Oligotrophic mat	Cuatrociénegas, México	97	354	287	1337	*This work*
CCC-PR	Oligotrophic mat	Cuatrociénegas, México	99	371	89	176	*This work*
CCC-PR	Oligotrophic mat	Cuatrociénegas, México	97	371	40	65	*This work*
CCC-PE	Oligotrophic mat	Cuatrociénegas, México	99	154	66	97	*This work*
CCC-PE	Oligotrophic mat	Cuatrociénegas, México	97	154	45	73	*This work*
SharkBay-I	Intertidal Stromatolite	Shark Bay, Australia	99	35	33	46	*Burns et al., 2004*
HamelinPool-DS	Intertidal Stromatolite	Hameling Pool, Australia	99	192	71	178	*Papineau et al., 2005*
HamelinPool-DS	Intertidal Stromatolite	Hameling Pool, Australia	97	192	61	117	*Papineau et al., 2005*
HamelinPool-DI	Intertidal Stromatolite	Hameling Pool, Australia	99	192	124	505	*Papineau et al., 2005*
HamelinPool-DI	Intertidal Stromatolite	Hameling Pool, Australia	97	192	111	314	*Papineau et al., 2005*
HamelinPool-R	Intertidal Stromatolite	Hameling Pool, Australia	99	192	90	566	*Papineau et al., 2005*
HamelinPool-R	Intertidal Stromatolite	Hameling Pool, Australia	97	192	66	288	*Papineau et al., 2005*
HighborneCay-1	Subtidal Stromatolites	Exumas, Bahamas	97	251	128	229	*Baumgartner et al., 2009*
HighborneCay-2	Subtidal Stromatolites	Exumas, Bahamas	97	251	133	274	*Baumgartner et al., 2009*
HighborneCay-3	Subtidal Stromatolites	Exumas, Bahamas	97	350	181	398	*Baumgartner et al., 2009*
HighborneCay-N	Intertidal Stromatolite	Exumas, Bahamas	97	NA	172	NA	*Havemann et al., 2009*
GuerreroEv-06	Endoevaporitic	Guerrero Negro, México	97	442	189	1240	*Sahl et al., 2008*
GuerreroEv-05	Endoevaporitic	Guerrero Negro, México	97	277	158	911	*Sahl et al., 2008*
LindseyLake03	Endoevaporitic	Lindsey Lake, NM	97	328	110	413	*Sahl et al., 2008*

FIG. S2. Supplementary tables of (**a**) species richness according to 16S rRNA clone libraries in previously published microbial mat studies; (**b**) relative frequency distributions of reads assigned to most-abundant orders according to the all-read metagenomic content.

Relative frequency distributions of reads assigned to most abundant orders by means of the all-reads metagenomic complement approach for the Green Mat (G), Red Mat (R), Guerrero Negro mat (GN) and Pozas Azules Stromatolite (PA). Different shades of gray are used to remark high frequency (>0.1, dark gray), medium frequency (>0.03, medium gray), and very low frequency (<0.003, light gray) of particular taxonomic groups.

	G	GN	PA	R
Chroococcales	0.084	0.048	0.088	0.043
Clostridiales	0.056	0.049	0.022	0.016
Bacillales	0.037	0.029	0.020	0.051
Burkholderiales	0.037	0.030	0.067	0.126
Nostocales	0.036	0.030	0.045	0.030
Rhizobiales	0.035	0.054	0.068	0.037
Actinomycetales	0.035	0.045	0.029	0.020
Pseudomonadales	0.034	0.016	0.015	0.326
Bacteroidales	0.032	0.034	0.013	0.013
Flavobacteriales	0.029	0.025	0.019	0.022
Oscillatoriales	0.028	0.033	0.032	0.038
Alteromonadales	0.026	0.017	0.012	0.014
Enterobacteriales	0.026	0.014	0.013	0.027
Legionellales	0.023	0.002	0.006	0.001
Cytophagales	0.022	0.016	0.017	0.011
Desulfuromonadales	0.021	0.027	0.018	0.005
Chlorobiales	0.019	0.014	0.012	0.004
Desulfovibrionales	0.017	0.024	0.007	0.005
Sphingobacteriales	0.016	0.014	0.016	0.011
Chloroflexales	0.016	0.023	0.015	0.005
Rhodobacterales	0.015	0.081	0.026	0.015
Planctomycetales	0.015	0.031	0.112	0.003
Thermoanaerobacterales	0.014	0.015	0.007	0.004
Chromatiales	0.014	0.014	0.009	0.004
Myxococcales	0.013	0.021	0.020	0.004
Vibrionales	0.012	0.007	0.006	0.006
Desulfobacterales	0.012	0.021	0.004	0.004
Lactobacillales	0.012	0.006	0.004	0.003
Rhodospirillales	0.011	0.015	0.015	0.008
Oceanospirillales	0.010	0.007	0.006	0.007
Spirochaetales	0.009	0.013	0.004	0.002
Campylobacterales	0.008	0.003	0.003	0.002
Syntrophobacterales	0.008	0.011	0.007	0.002
Xanthomonadales	0.007	0.006	0.008	0.010
Pasteurellales	0.007	0.003	0.003	0.002
Rickettsiales	0.007	0.002	0.004	0.001
Thiotrichales	0.006	0.003	0.002	0.001
Neisseriales	0.006	0.003	0.004	0.004
Chlamydiales	0.006	0.002	0.003	0.000
Sphingomonadales	0.005	0.007	0.011	0.047
Prochlorales	0.005	0.001	0.004	0.001
Gloeobacterales	0.005	0.003	0.006	0.001
Verrucomicrobiales	0.005	0.006	0.028	0.001
Nitrospirales	0.003	0.002	0.020	0.000
Thermotogales	0.001	0.007	0.002	0.001

FIG. S2. (Continued).

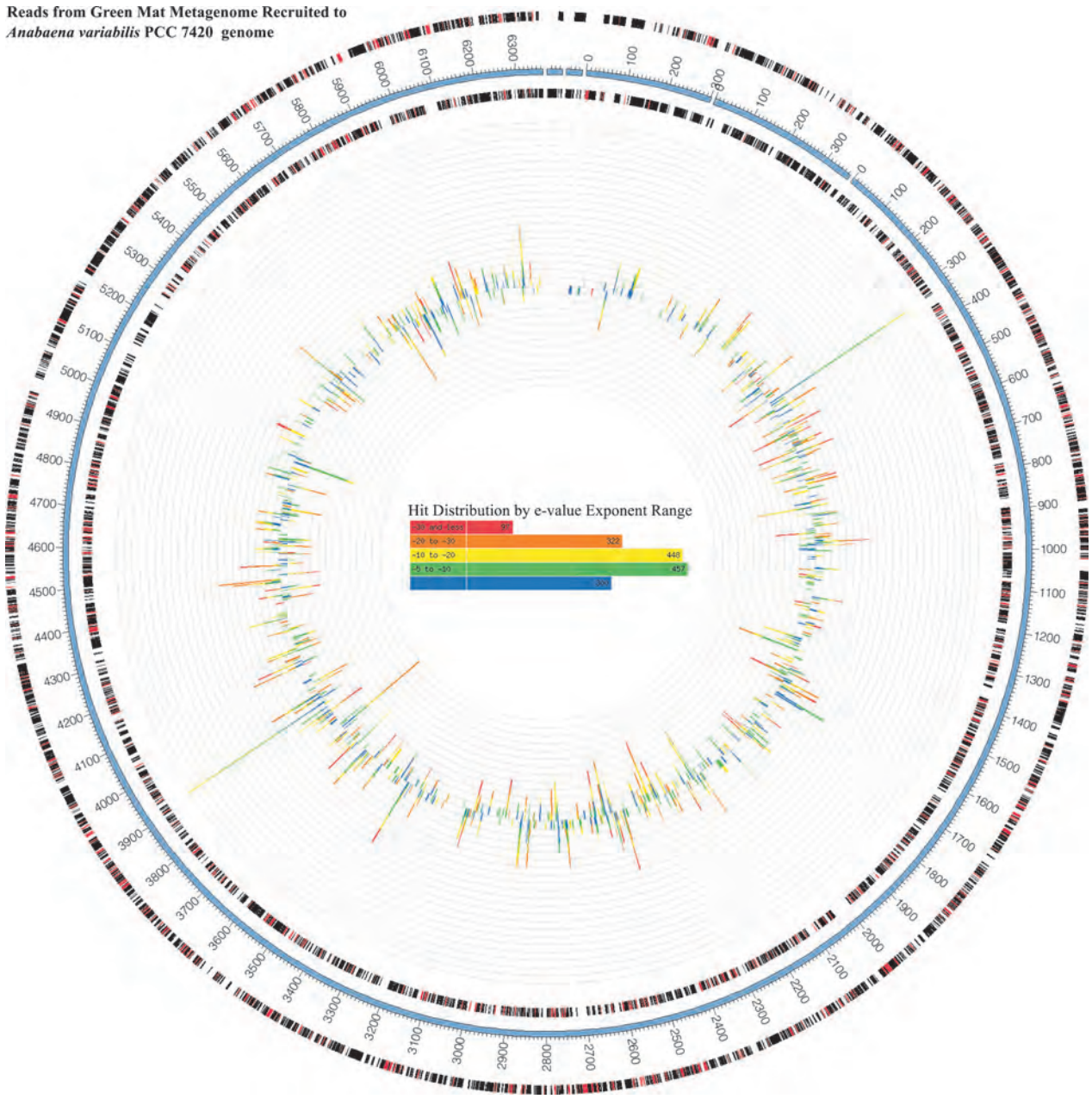


FIG. S3. Fragment recruitment diagrams of the reference genomes recruiting the highest amount of reads from the green and red mats' metagenomes.

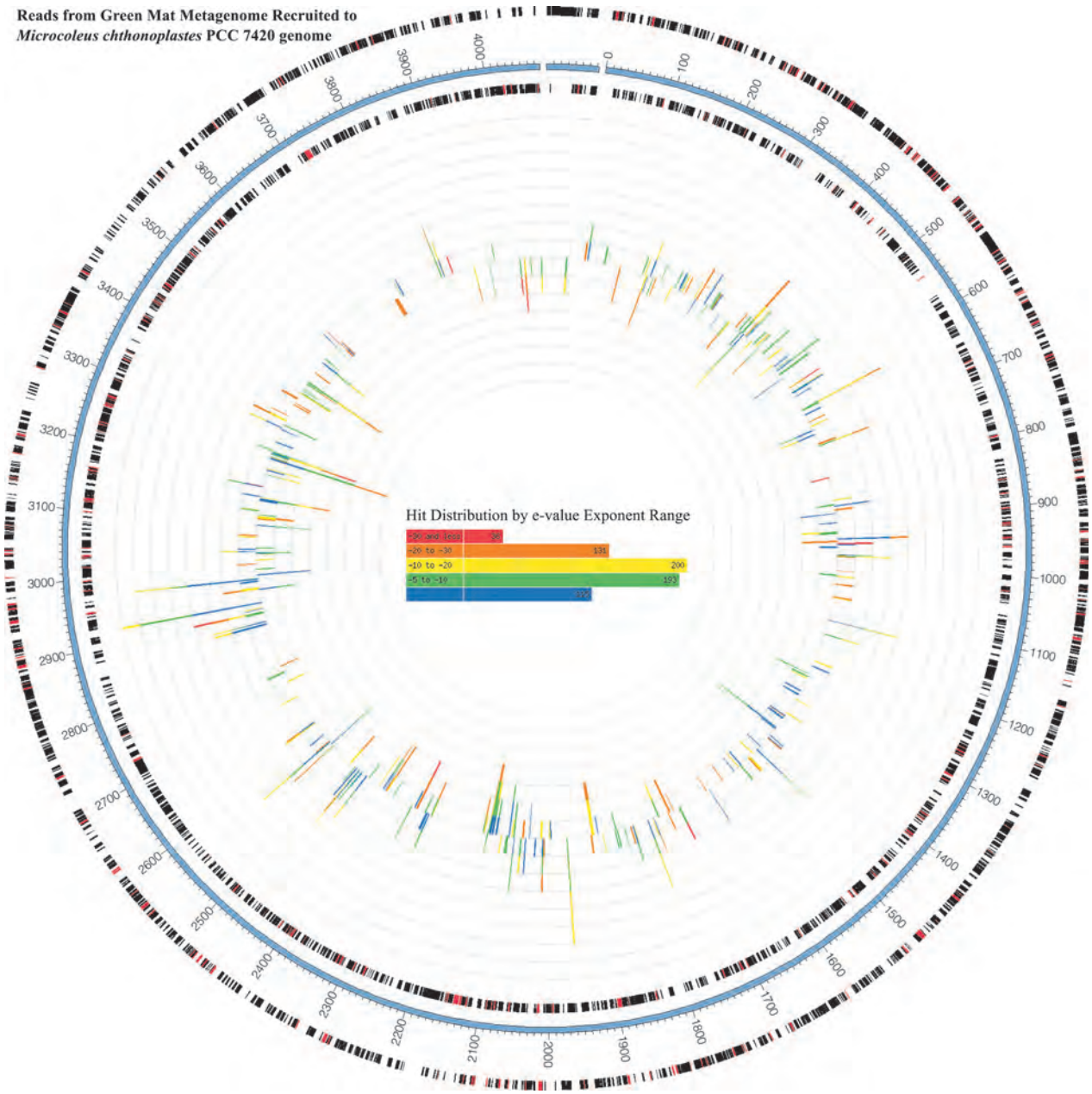


FIG. S3. (Continued).

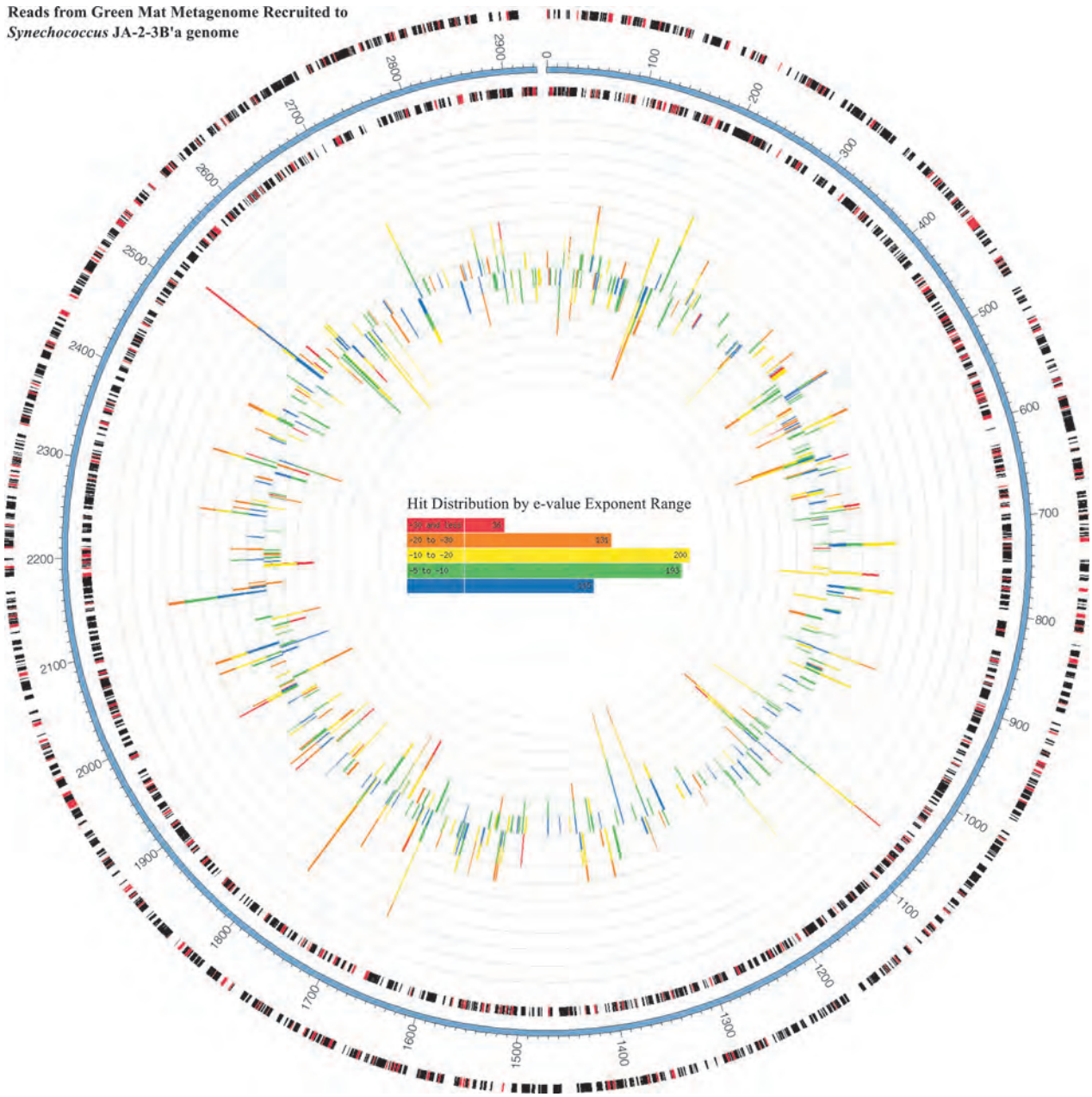


FIG. S3. (Continued).

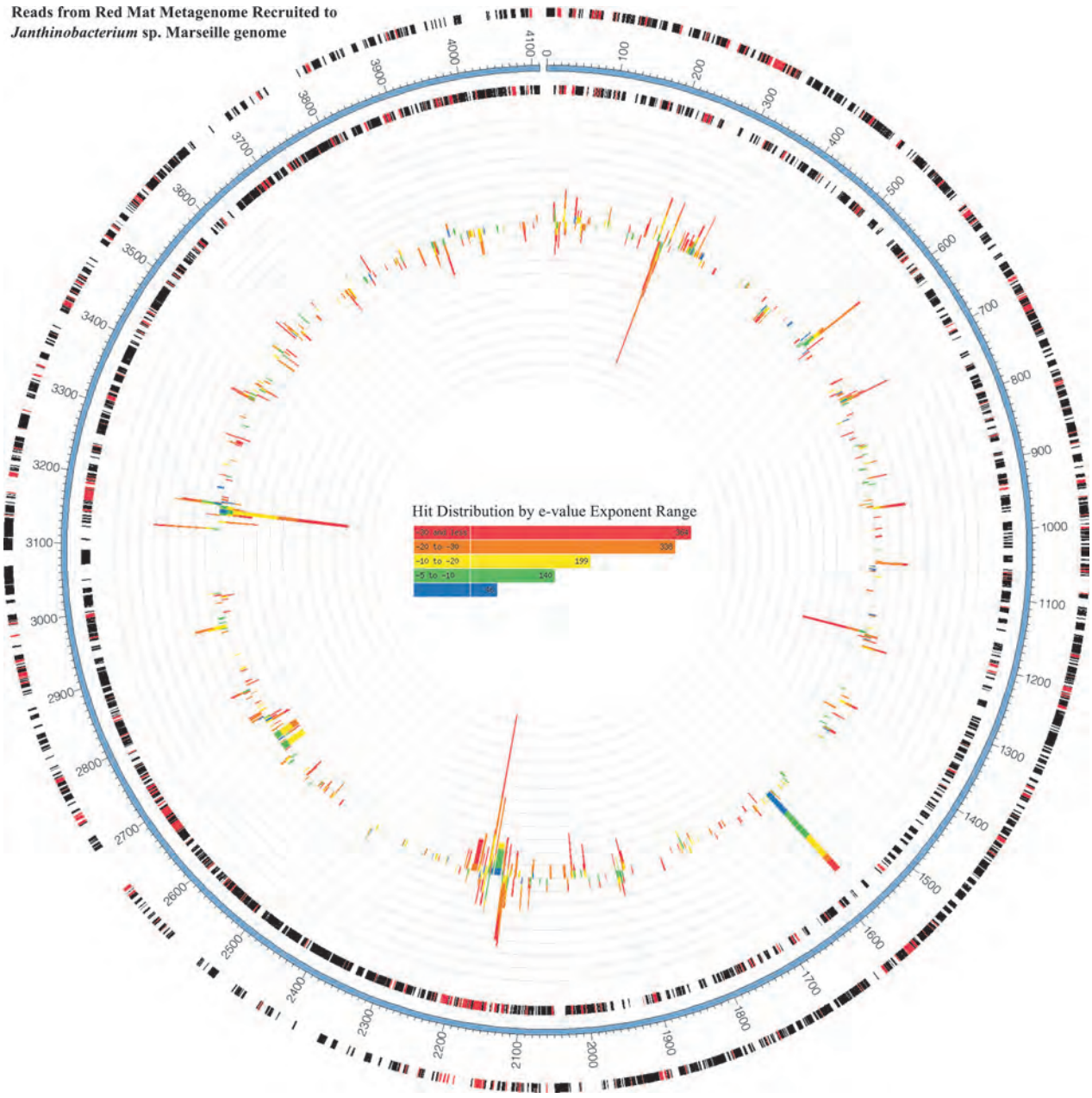


FIG. S3. (Continued).

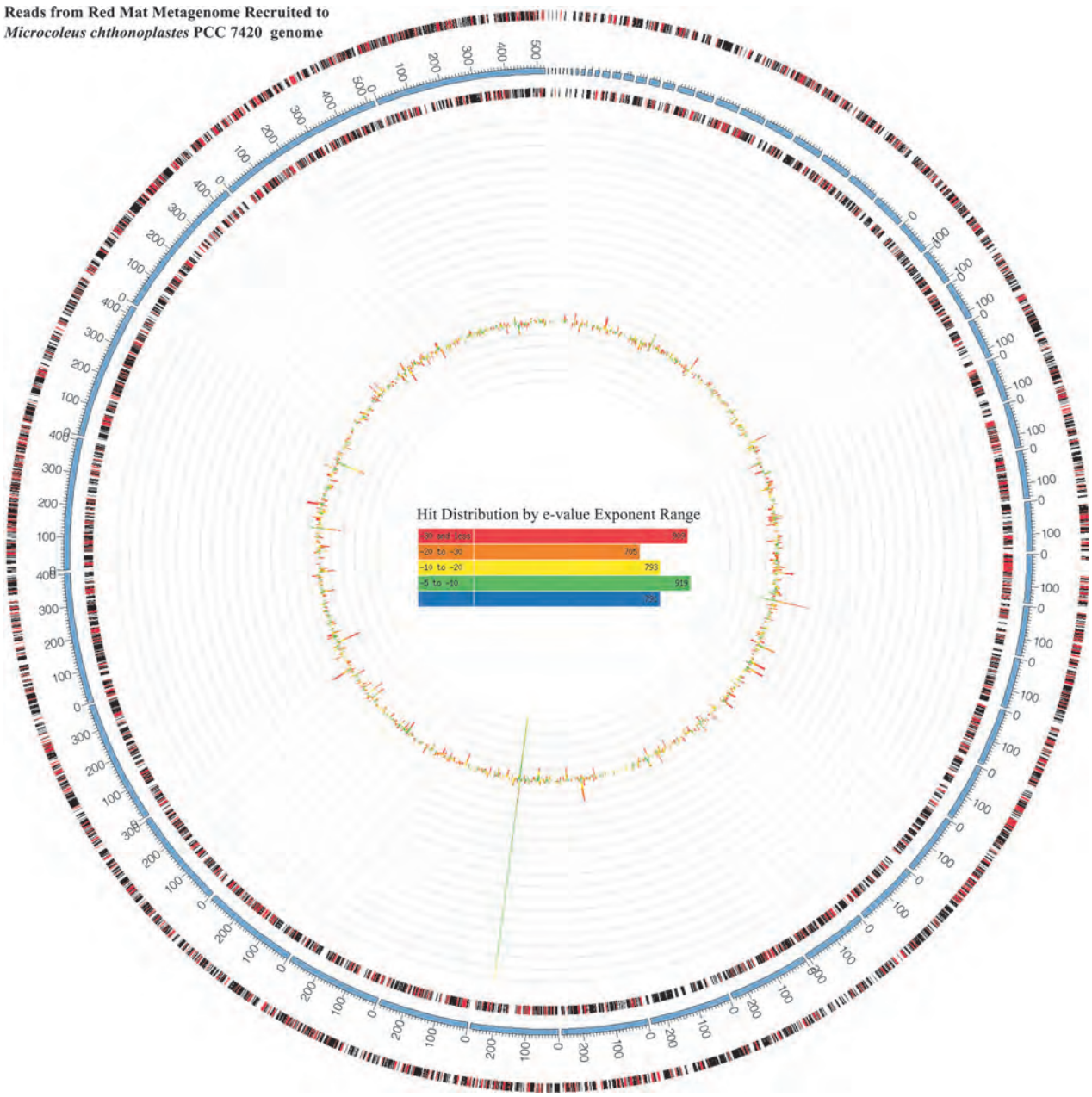


FIG. S3. (Continued).

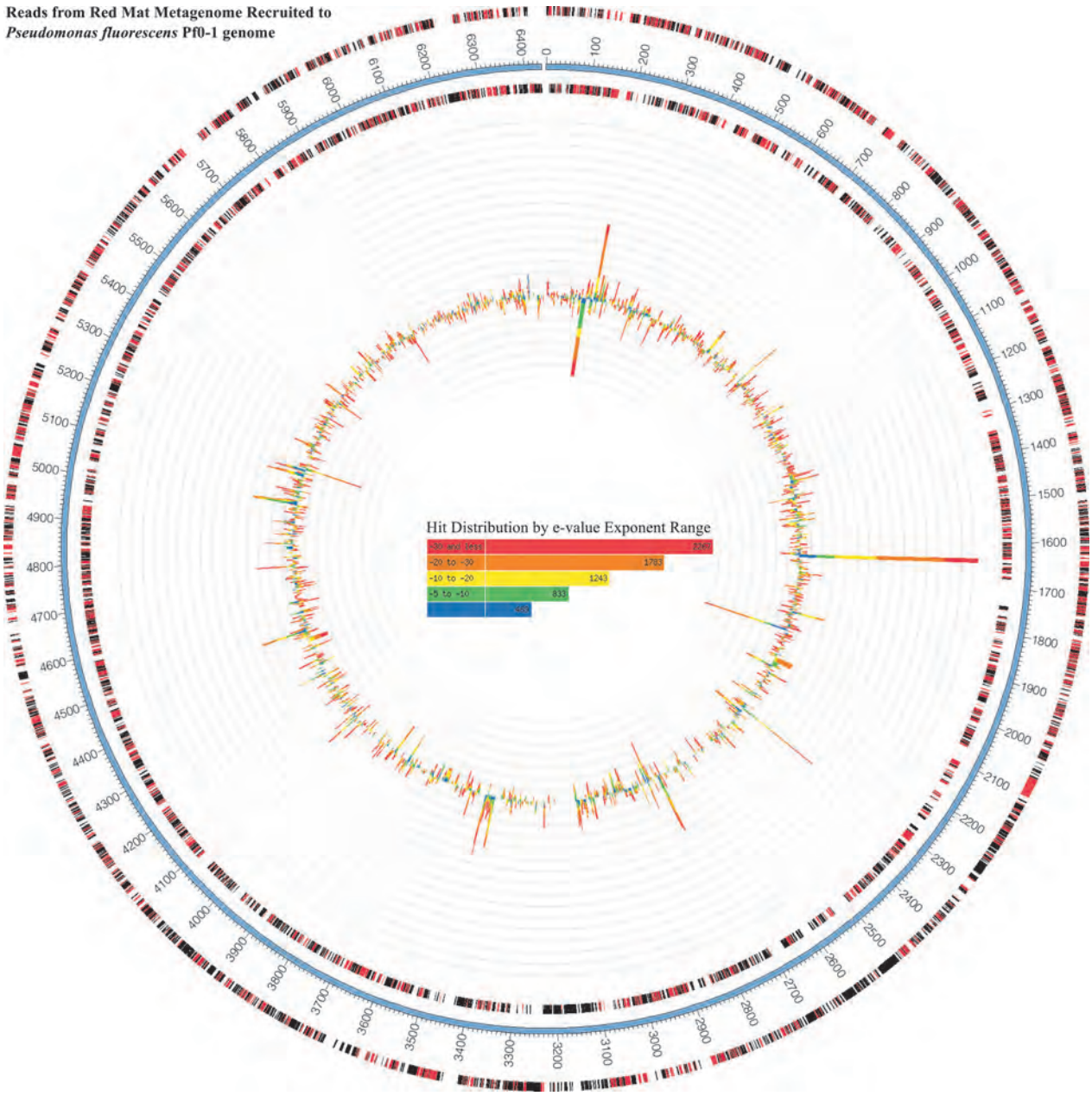


FIG. S3. (Continued).

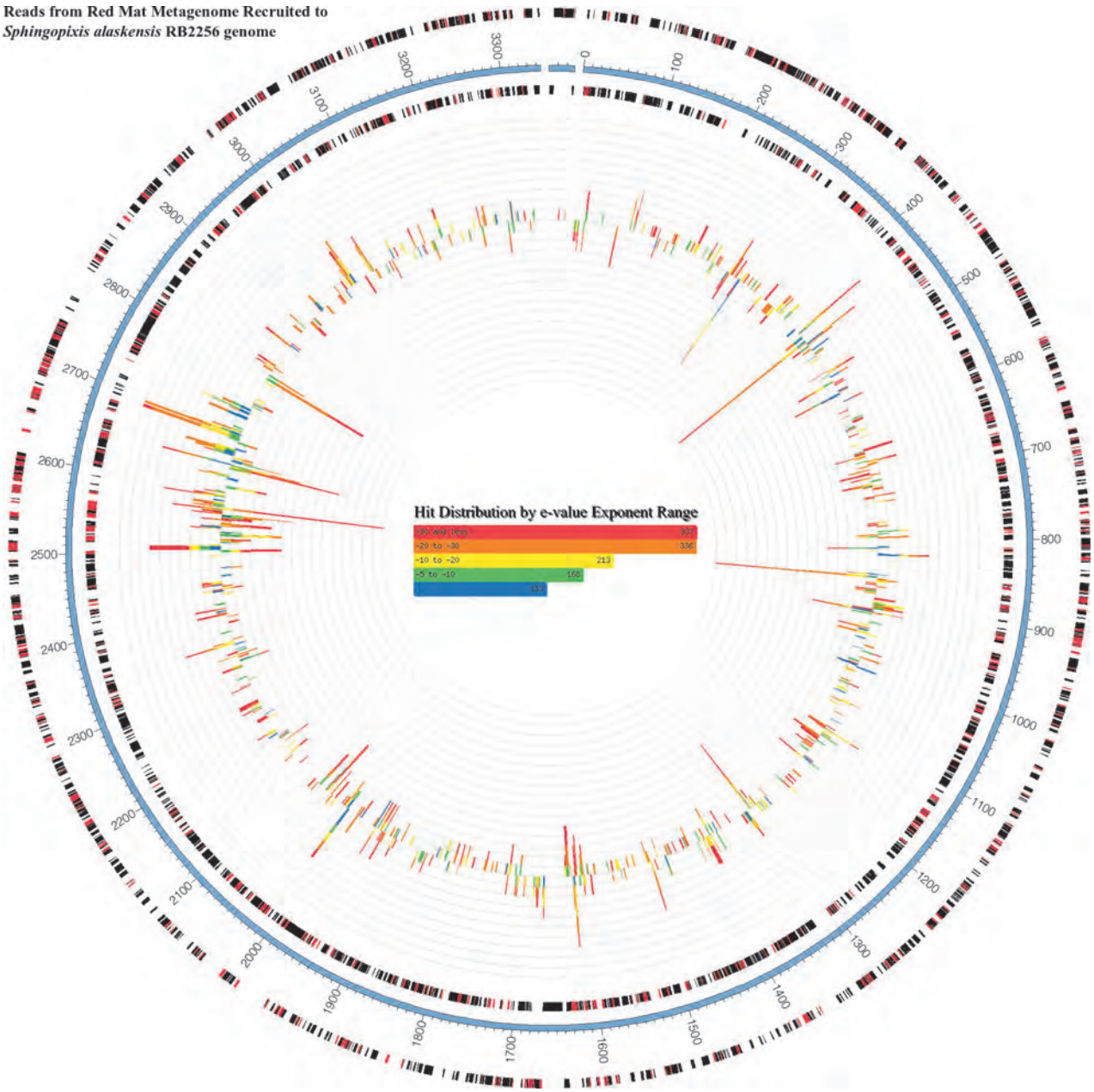


FIG. S3. (Continued).

CAPÍTULO III:
ANÁLISIS DE LA DIVERSIDAD FUNCIONAL DE TAPETES MICROBIANOS

En el capítulo anterior, se caracterizó la diversidad taxonómica de los tapetes microbianos, y mediante la comparación contra otros sistemas organosedimentarios, se identificaron patrones de diversidad comunes a éstos sistemas afectados simultáneamente por procesos de filtración ambiental y de rasgos funcionales filogenéticamente conservados. Las bases de datos metagenómicas permiten el análisis simultáneo del componente funcional y filogenético de una comunidad, de manera que el siguiente paso natural es el análisis del componente funcional de la diversidad de los tapetes microbianos del Valle de Cuatrociénegas.

Introducción

Es de gran importancia profundizar en el conocimiento sobre la estructura y composición de los tapetes microbianos, dado que conforman sistemas ecológicos modelo para analizar el funcionamiento y ciclaje de nutrientes en los ecosistemas (Foster y Mobberly, 2010). Los tapetes microbianos constituyen modelos idóneos para manipulación experimental espacial y temporal (Paerl et al. 2003; Yannarell et al. 2007); adicionalmente reproducen la estructura y funciones de otros ecosistemas, pero a una escala ordenes de magnitud menor, reducida de cientos de kilómetros a centímetros (Foster y Mobberley 2010).

Por las razones anteriores, el funcionamiento fisiológico de los tapetes microbianos ha sido estudiado desde hace más de tres décadas por los ecólogos microbianos, revelando por ejemplo un marcado microgradiente en las concentraciones de oxígeno, pH y ácido sulfídrico (Revsbech et al. 1983; Canfield y Des Marais 1993) y en las tasas de respiración y reducción de sulfatos (Jorgensen y Des Marais 1990; Visscher et al. 1992) correlacionado con la profundidad milimétrica del tapete. Este gradiente se debe principalmente a que el tapete se encuentra tan finamente entretejido que se vuelve rápidamente anóxico a pocos milímetros de la superficie, y la estructuración de diversos gremios funcionales bacterianos corresponde también a la laminación sedimentaria. Sin embargo, sólo con el desarrollo reciente de técnicas de secuenciación fue posible identificar la gran diversidad taxonómica contenida en comunidades microbianas organosedimentarias (Ley et al. 2006, Capítulo III) y a la fecha existen pocos estudios sobre el complemento genómico funcional de éstas comunidades (Kunin et al. 2008; Breitbart et al. 2009).

La caracterización de la estructura y composición de la comunidad de dos tapetes microbianos en dos ambientes oligotróficos del Valle de Cuatrociénegas (CCC) fue presentada en el Capítulo anterior. Estos tapetes contienen una gran diversidad taxonómica, y además son muy diferentes entre sí, a pesar de que las pozas de CCC son los cuerpos de aguas continentales con menor concentración de fósforo reportados a la fecha (Elser et al. 2005). Por otra parte, se han reportado diversas estrategias para la utilización de nutrientes en los genomas bacterianos provenientes del valle de CCC, como lo son la substitución de fosfolípidos por sulfolípidos en la membrana plasmática en *Bacillus coahuilensis* (Alcaraz et al. 2008) y la utilización de fosfonatos en *Bacillus sp.* M3-13 (Alcaraz et al. 2010). A pesar de que se espera que el funcionamiento global de los tapetes de CCC siga el mismo modelo determinado por el microgradiente de concentración de oxígeno (Kunin et al. 2008), no existe a la fecha un modelo claro que explique la alta diversidad de los tapetes microbianos en ambientes de oligotrofía extrema, ni datos sobre las estrategias seguidas por comunidades enteras para sobrevivir bajo éstas condiciones. En este Capítulo presentamos el análisis funcional comparativo de los metagenomas de los dos tapetes microbianos cuyo complemento taxonómico fue presentado en el capítulo anterior.

Métodos

Como ya mencionamos en el Capítulo II, los tapetes microbianos estudiados provienen de dos sistemas acuáticos contrastantes: una laguna permanente (PG) y una poza de desecación (PR). La caracterización geoquímica de éstas dos pozas se presenta en Peimbert et al. (2012) al final de éste capítulo. Dada la naturaleza evaporítica de la poza, el tapete PR está sujeto a variaciones ambientales mayores, como una reducción gradual y constante en la cantidad de agua en la poza y un aumento concomitante en el diferencial de temperatura y en las concentraciones de los compuestos en solución. La poza permanente de PG es, en contraste, significativamente más estable.

El DNA total metagenómico fue pirosecuenciado con la plataforma FLX-454 (Life Sciences), y el juego de datos resultante fue sujeto a los controles de calidad y la tubería de análisis y anotación descritas en Bonilla-Rosso et al. (2012a) (Capítulo II), y Peimbert et al. (2012), del que soy coautor, al final de éste capítulo. Los perfiles funcionales correspondientes a las categorías funcionales de COGs (Clusters of Orthologous Genes, (Tatusov et al. 2003), KEGG (Kanehisa et al. 2008) y el SEED (Overbeek et al. 2005) fueron comparados entre los cuatro metagenomas de Cuatrociénegas, obtenidos de la anotación por RAMMCAP (Li 2009) con un valor de corte menor a 1e-5 y una similitud mayor a 60%.

Uno de los principales problemas en el análisis funcional de metagenomas radica en obtener información útil y verosímil de las diferencias reales en las abundancias de categorías funcionales estudiadas. En éste caso, los perfiles funcionales fueron normalizados respecto al tamaño del metagenoma, considerado como el número de secuencias finales en cada juego de datos tras la revisión de calidad y redundancia. Las diferencias entre metagenomas fueron identificadas mediante el programa STAMP (STatistical Analysis of Metagenomic Profiles, (Parks y Beiko 2010), el cual permite realizar pruebas estadísticas robustas y diferenciar entre categorías funcionales sobrerrepresentadas mediante efectos de tamaño. Se estimó la tasa de falsos descubrimientos de Storey (False Discovery Rate, Storey et al. 2004) como corrección para pruebas múltiples, las diferencias estadísticas fueron evaluadas mediante la prueba exacta de Fisher y los intervalos de confianza se calcularon mediante el método de Newcombe-Wilson (Newcombe 1998). Es importante notar que en las ciencias biológicas es frecuente encontrar diferencias que a pesar de ser estadísticamente significativas, son tan pequeñas que carecen de significado biológico. El uso de los tamaños del efecto ayuda a identificar diferencias biológicamente significativas (Nakagawa y Cuthill 2007).

También se realizó el análisis comparativo genecéntrico de los ciclos biogeoquímicos del Carbono (C), Nitrógeno (N), Fósforo (P) y Azufre (S). En éste caso, se buscaron los genes correspondientes a enzimas clave involucradas en las reacciones principales de los ciclos (Tablas A-C) en cada uno de los metagenomas estudiados mediante BLAST. Se consideró como un hit positivo las secuencias mayores a 75 bp de longitud con un valor de corte menor a 1e-5 y una similitud mayor a 40% , y las diferencias fueron normalizadas respecto al tamaño del metagenoma.

Resultados y Discusión

La caracterización fisicoquímica de las pozas y la descripción metabólica general de los metagenomas se presentan en Peimbert et al., 2012. A continuación se presenta un análisis más detallado del significado biológico de las diferencias en la sobrerrepresentación del análisis genecéntrico y por categorías funcionales, basados en la comparación estadística entre las categorías funcionales de los subsistemas del SEED.

Análisis por Categorías Funcionales

Las comparaciones pareadas significativas con sus respectivos tamaños de efecto e intervalos de confianza entre los perfiles funcionales de los cuatro metagenomas de Cuatrociénegas (Figs. A-D). En la primera columna de izquierda a derecha se encuentran los nombres de cada categoría. La segunda columna consiste en una gráfica de barras que compara el número de secuencias asignadas a cada categoría en cada metagenoma. La tercera columna corresponde a la diferencia entre las proporciones de abundancia de cada categoría, donde cada punto se presenta con su intervalo de confianza al 95%. Sólo se presentan las diferencias estadísticamente significativas (p<0.01).

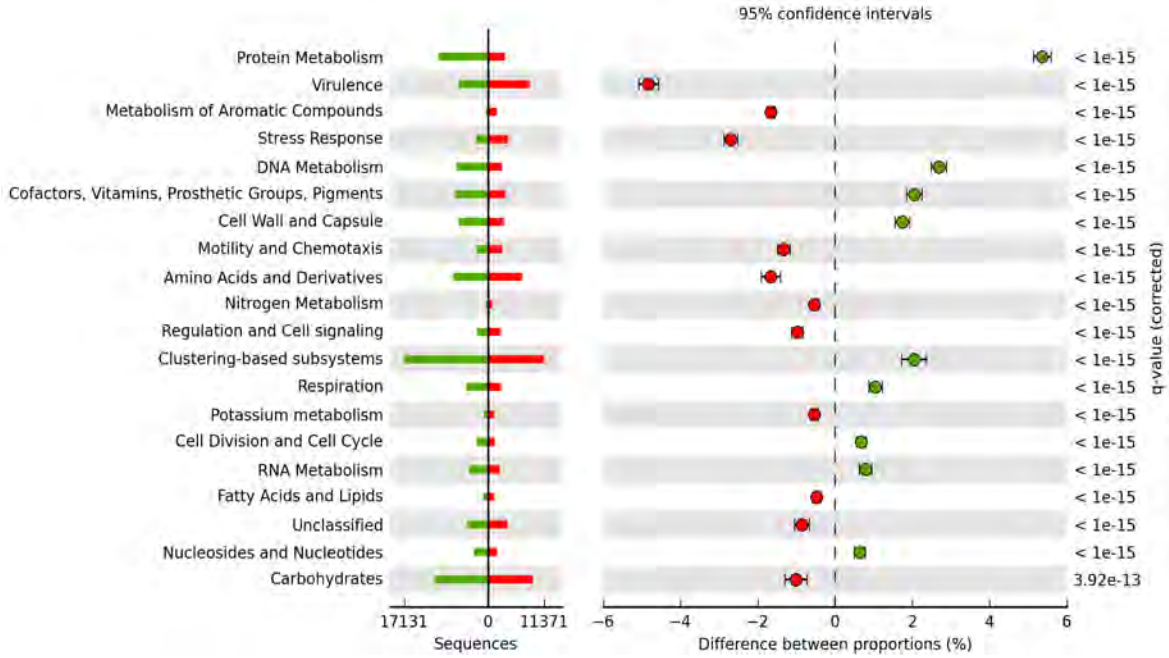


Fig A. Comparaciones pareadas significativas entre las abundancias relativas estandarizadas de la jerarquía más generales de los sistemas del SEED en los tapetes PR (rojo) y PG (verde).

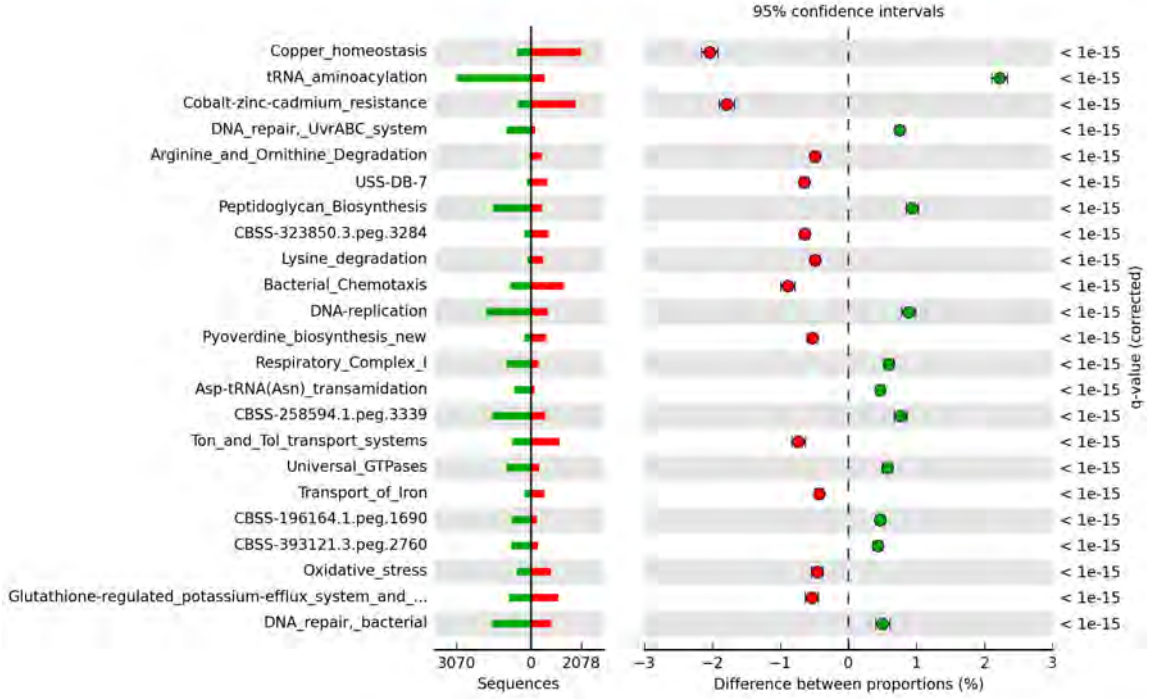


Fig B. Comparaciones pareadas significativas entre las abundancias relativas estandarizadas de los subsistemas del SEED en los tapetes PR (rojo) y PG (verde).

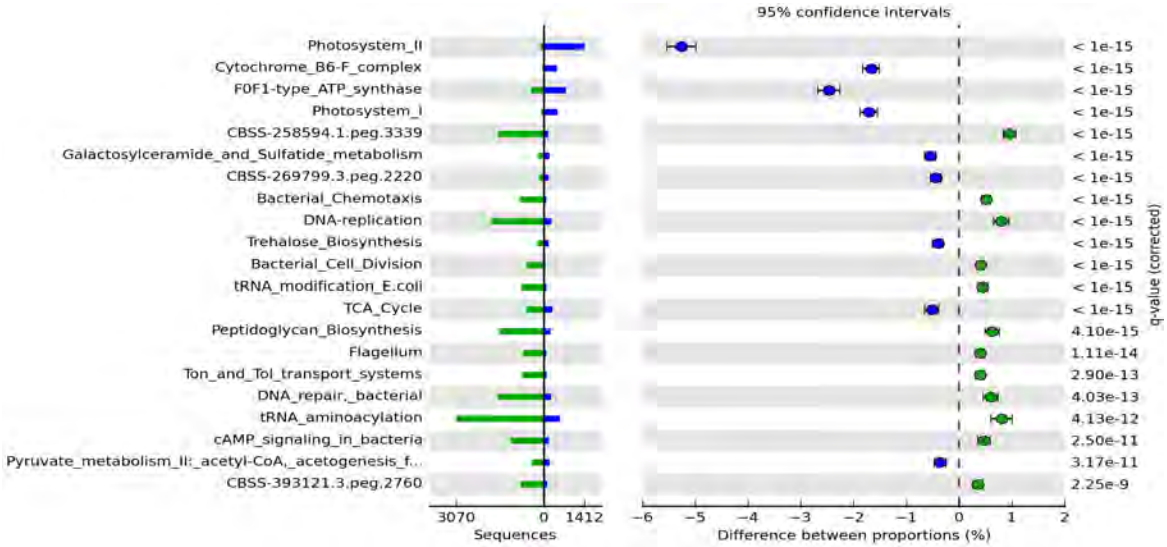


Fig C. Comparaciones pareadas significativas entre las abundancias relativas estandarizadas de los subsistemas del SEED en el estromatolito PA (azul) y el tapete PG (verde).

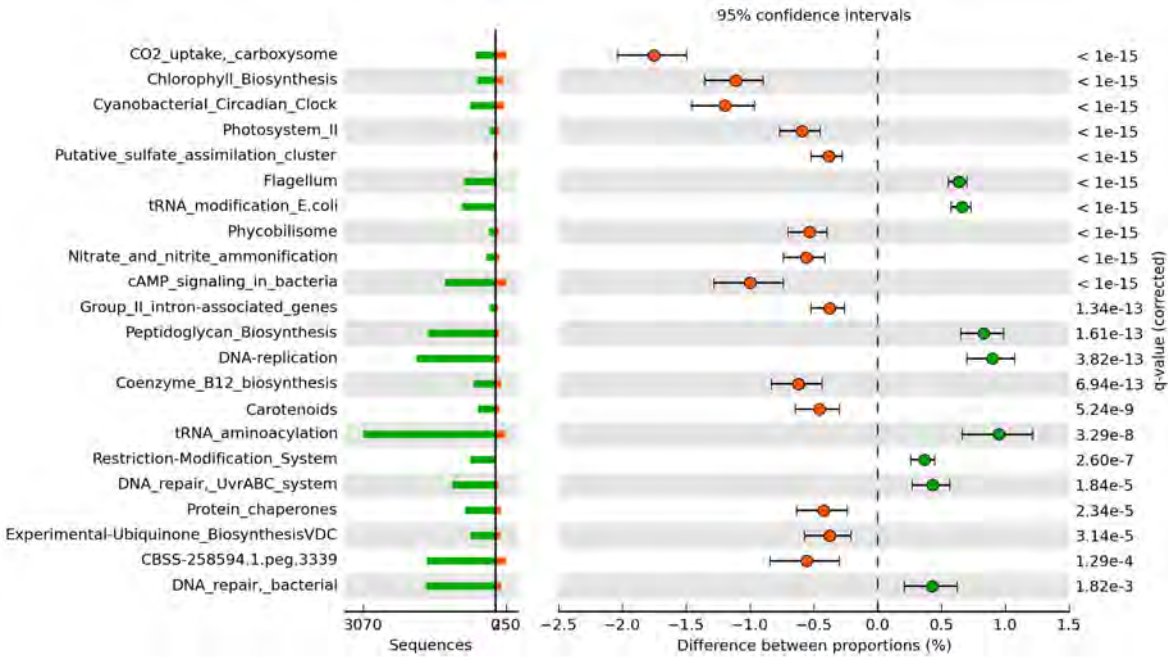


Fig D. Comparaciones pareadas significativas entre las abundancias relativas estandarizadas de los subsistemas del SEED en el estromatolito RM (naranja) y el tapete PG (verde).

Los dos tapetes microbianos son funcionalmente diversos, y cada uno contiene elementos de prácticamente todas las rutas metabólicas representadas en los mapas del KEGG (Peimbert et al. 2012). Notablemente, las dos categorías del metabolismo de carbohidratos más representadas en ambos tapetes son las mismas: la de utilización de carbohidratos complejos como el glucógeno (SEED: Maltose utilization: Glycogen phosphorylase) y la del empleo de compuestos de un sólo carbono (SEED: Serine-glyoxylate cycle). Una cantidad importante de secuencias fueron clasificadas dentro de categorías de función desconocida (COG S: function unknown; SEED: Clustering-based subsystems). Ésto es relativamente común en metagenomas provenientes de ambientes poco estudiados, y resalta la existencia de familias enteras de proteínas de las cuales no conocemos sus funciones metabólicas ni enzimáticas, y en ciertos casos familias presentes en la naturaleza pero de las cuales no tenemos representantes en las bases de datos genómicas.

El tapete PG tiene una sobrerrepresentación de funciones relacionadas con producción de energía, fotosíntesis y funciones celulares basales. Las sobrerrepresentación de las categorías relacionadas con metabolismo basal (COGs D, J, L, M, O) en éste metagenoma, que tiene también una alta diversidad de especies y no presenta dominancia de grupos específicos, es explicable como un artefacto de la clasificación en sistemas complejos con alta diversidad metabólica, pues son éstas categorías las únicas comúnes a todos los microorganismos presentes en la muestra. Es por eso que las categorías representadas corresponden a las tRNA aminoacil transferasas (SEED: tRNA aminoacylation) que son diversas y genómicamente abundantes (alrededor de una por cada aminoácido), las largas y conservadas proteínas de reparación de DNA por escisión de nucleótidos (SEED: DNA repair UvrABC subsystem) y la maquinaria de replicación del DNA (SEED: DNA replication) y de la síntesis de pared celular (SEED: Peptidoglycan Biosynthesis).

Se calculó el tamaño efectivo de genoma (EGS, Peimbert et al., 2012), una medida indirecta del tamaño promedio de los genomas presentes en un metagenoma a partir del número de genes de copia única, como lo reportan Raes et al. (2007). Encontramos que el EGS en el tapete de PG es muy pequeño (1.27 Mb), mientras que el EGS del tapete PR se encuentra en un tamaño medio (3.69 Mb). Dado que las categorías funcionales de metabolismo basal se encuentran conservadas en todos los genomas, su abundancia en una comunidad incrementará linealmente con cada genoma incorporado, independientemente del tamaño de éste. Sin embargo, la abundancia relativa de éstas categorías será mayor en comunidades cuyos miembros tienen genomas más pequeños, y se verán sobrerrepresentadas en comunidades con una abundancia de miembros con genomas pequeños respecto a otras donde los miembros tienen genomas grandes. En contraste, el número de funciones metabólicas crecerá exponencialmente con cada genoma añadido, y el crecimiento será más acelerado en comunidades con genomas más grandes. Los tamaños pequeños de EGS son característicos de genomas de organismos oligótrofos, mientras que los tamaños grandes de EGS son característicos de genomas de organismos copiótrofos, o sea que viven en lugares ricos en nutrientes (Lauro et al. 2009).

La sobrerrepresentación de funciones de biosíntesis y generación de energía en PG sugieren que esta comunidad es principalmente autótrofa y fuertemente basada en la producción primaria, pues las categorías de biosíntesis de porfirinas y clorofila [PATH:ko00860], oxidación fosforilativa [PATH:ko00190], fijación de carbono en organismos fotosintéticos [PATH:ko00710] y fotosíntesis [PATH:ko00195]. Más áun, las categorías más abundantes dentro del metabolismo de carbohidratos en PG corresponden al metabolismo central aeróbico (SEED: TCA Cycle, Pyruvate Metabolism: anaplerotic reactions) y a la biosíntesis de carbohidratos (SEED: Sucrose metabolism, Pentose Phosphate Pathway, Glycogen metabolism, CO2 uptake and carboxysome, Pyruvate metabolism II: acetyl-CoA and acetogenesis).

En contraste, las características más notables del tapete PR (compartidas por los tres sistemas de categorización funcional) son una sobrerrepresentación de funciones relacionadas con metabolismo secundario y degradación de xenobióticos y de respuesta a cambios ambientales y estrés (COGs E, N, P, T). La mayor parte de las funciones relacionadas con el procesamiento de la información ambiental corresponden a motilidad celular (COG N, síntesis de flagelo [PATH:ko02040] y quimiotaxis [PATH:ko02030]) y la transducción de señales de dos componentes [PATH:ko02020] y el transporte trans-membranal mediante sistemas ABC [PATH:ko02010]. En cuanto a estrés y resistencia, las categorías más abundantes en PR corresponden a resistencia a la intoxicación por cobre (SEED: Copper-translocating P-type ATPase; Multicopper oxidase; Copper resistance and tolerance proteins), cobalto/zinc/cadmio (SEED: Co-Zn-Cd resistance and efflux proteins), arsénico (SEED: Arsenate reductase, Arsenical pump ATPase, Arsenic efflux pump, Arsenate resistance proteins), bombas de expulsión multidrogas (SEED: Acriflavin resistance, RND efflux systems, MATE family of multi antimocrobial extrusion protein) y mecanismos de regulación de la homeostasis de potasio (SEED: Gluthathione-regulated potassium efflux system *KefABC*, Potassium-transporting ATPase, Potassium uptake protein *Trk*, Potassium channel *Kdp*).

La sobrerepresentación de éstas categorías sugiere que los organismos que componen el tapete PR se encuentran bien adaptados para resistir altas concentraciones de elementos tóxicos, como se esperaría en una poza de desecación en donde la evaporación incrementa gradualmente la concentración de sales, metales y otros compuestos tóxicos que se encuentran en solución, especialmente Cobre, del cual CCB es reconocido por contener cobre de alto grado (http://www.santafemetals.com/index.php?page=cuatro_cienegas).

Entre las categorías de obtención de energía en PR que son más abundantes que en PG, se encuentran los metabolismos de metano [PATH:ko00680], nitrógeno [PATH:ko00910] y azufre [PATH:ko00920]. Los sistemas más abundantes involucrados en metabolismo de carbohidratos corresponden a metabolismos para la fermentación de carbohidratos complejos (SEED: Glycolysis, Acetyl-CoA Fermentation to Butyrate; Acetone Butanol Ethanol Synthesis, Butanol Biosynthesis) y la utilización anabólica de carbohidratos simples como fuentes de carbono (SEED: Glyoxylate Cycle).

Las categorías sobrerrepresentadas en PR son catabólicas y heterótrofas y evidencian una comunidad que busca activamente sus nutrimentos y posee la habilidad de obtener su energía de diversas fuentes pero no por fotoautotrofía, y puede utilizar compuestos tanto simples como complejos como fuentes de carbono. Estas características concuerdan con la dominancia de *Pseudomonas* en este sitio. Los miembros del género *Pseudomonas* poseen una plasticidad metabólica sorprendente (Shen et al. 2006; Silby et al. 2009; Mulet et al. 2010), que aparentemente les da una ventaja competitiva sobre otras especies. Vale la pena notar que las rutas metabólicas presentes en el tapete PR no son explicables por la presencia de una sóla especie de *Pseudomonas*, lo que sugiere la existencia de linajes de este género metabólicamente distintos, especializados en la degradación de conjuntos diferentes de complejos macromoleculares.

Análisis Genecéntrico

La segunda aproximación consiste en un análisis genecéntrico de la frecuencia relativa de los genes involucrados en los ciclos de Carbono, Nitrógeno, Azufre y Fósforo (Tablas A-C).

Tabla A. Genes del ciclo del Fósforo utilizados en el análisis genecéntrico.

phnA	phosphonoacetate hydrolase
phnD	phosphonate-binding transporter
phnH	phosphonate metabolism protein
phnW	2-aminoethylphosphonate:pyruvate transaminase
phnX	Phosphonoacetaldehyde hydrolase
pitA	low-affinity inorganic phosphate transporter
ppA	inorganic pyrophosphatase
ppK	polyphosphate kinase
ppX	exopolyphosphatase
pstA	phosphate transport permease
pstS	phosphate-binding protein
dedA	alkaline phosphatase
phoA	alkaline phosphatase
phoE2	uncharacterized phosphatase
phoX	alkaline phosphatase X
ptxB	phosphite/phosphonate transport system
ptxD	phosphite dehydrogenase
htxA	hypophosphite dehydrogenase

Tabla B. Genes del ciclo del Azufre utilizados en el análisis genecéntrico.

cysT	Sulfate transport system permease
cysA	Sulfate and thiosulfate import ATP-binding protein CysA (EC 3.6.3.25)
cysP	Sulfate and thiosulfate binding protein CysP
tcyP	L-cystine uptake protein TcyP
ssuF	Organosulfate transporter
aslA	Arylsulfatase (EC 3.1.6.1)
dmsA	Dimehtylsulfoxide reductase
sreA2	DMSO reductase
ddhA	Dimehtylsulfide dehydrogenase
dmoA	Dimethylsulfide monooxigenase
dsoABC	Dimethylsulfide oxygenase
sfnG	Dimethylsulfone reductase
ssuD	Alkanesulfonate monooxygenase (EC 1.14.14.5)
ssuE	Ferredoxin--NADP(+) reductase (EC 1.18.1.2)
apt	Adenylyl-phosphate transferase (APAT)
dsrA	Sulfite reductase, dissimilatory-type gamma subunit (EC 1.8.99.3)
dsrB	Sulfite reductase, dissimilatory-type gamma subunit (EC 1.8.99.3)
dsrO	Sulfite reduction-associated complex DsrMKJOP iron-sulfur protein DsrO (=HmeA)
dsrK	Sulfite reduction-associated complex DsrMKJOP iron-sulfur protein DsrO (=HmeA)
dsrP	Sulfite reduction-associated complex DsrMKJOP iron-sulfur protein DsrO (=HmeA)
soxC	Sulfite oxidase
aprA	Adenylylsulfate reductase (EC 1.8.99.2)
satA	Sulfate adenylyltransferase, dissimilatory-type [ATP sulfurylase/adenylyl/sulfate transferase](EC 2.7.7.4)
trxB	Thioredoxin reductase (EC 1.8.1.9)
tpx/bcp	Thiol peroxidase (EC 1.11.1.15)
cysC	Adenylylsulfate kinase (EC 2.7.1.25)
cysN	Sulfate adenylyltransferase subunit 1 [ATP sulfurylase] (EC 2.7.7.4)
cysI	Sulfite reductase [NADPH] hemoprotein beta-component (EC 1.8.1.2)
cysH	Phosphoadenylyl-sulfate reductase [thioredoxin] (PAPS) (EC 1.8.4.8)

Tabla C. Genes de los ciclos de Carbono (izquierda) y Nitrógeno (derecha) utilizados en el análisis genecéntrico.

aclB	citrate lyase
ccsA	citryl-CoA synthetase
ccsI	citryl-CoA lyase
acs	acetyl-CoA synthase
cooS	carbon monoxide dehydrogenase
mcm	methylmalonyl-CoA mutase
cbbL	rubisco
ccbM	rubisco
hpaA1	4-hydroxyphenylacetate-3-hydroxylase
pgk	phosphoglycerate kinase

amoA	ammonia oxygenase
narG	nitrate reductase
napA	nitrate reductase
narB	assimilative nitrate reductase
nirK	nitrite reductase
nirS	nitrite reductase
nirA	nitrite reductase
norB	nitric oxide reductase
nosZ	nitrous oxide reductase

El ciclo del nitrógeno está mejor representado en el tapete de PR que en PG, con una preferencia hacia la desnitrificación disimilativa, que reduce el nitrato completamente hasta dinitrógeno y sin la presencia de genes involucrados en la nitrificación. De las dos nitrito reductasas existentes (*nirK* que utiliza cobre como cofactor y *nirS* que posee un citoctomo *cd1*, pero son funcionalmente equivalentes), sólo se detectó *nirS* en ambos tapetes, a pesar de que *nirK* es dominante en comunidades de suelos y sedimentos y *nirS* es dominante en comunidades marinas (Jones & Hallin 2010). Los genes de ésta ruta respiratoria que utiliza nitrógeno como aceptor final de electrones (*narG*, *napA, nirS, norB, nosZ*), han sido reportados previamente como presentes en *Pseudomonas* (Philippot et al. 2001; Kimbrel et al. 2010). También se detectaron genes para la reducción asimilativa del nitrato, que incorpora nitrógeno a materia orgánica, de las cuales el más abundante fue el gen de la nitrato reductasa cianobacteriana *narB*, detectada inicialmente en cianobacterias unicelulares del género *Synechococcus,* las cuales son abundantes en los tapetes, y aunque en menor proporción, también se encontró la nitrito reductasa *nirAB*.

Aunque previamente se ha reportado la desnitrificación asimilativa en *Pseudomonas* (Betlach et al. 1981), cabe resaltar que la asimilación de nitrito es mayor en PG que en PR. Asimismo, el gen involucrado en la fijación de nitrógeno, la nitrogenasa *nifH*, sólo se detectó en el tapete PG. Éstos resultados indican que el ciclo del nitrógeno en el tapete PR es predominantemente respiratorio, es decir que la comunidad en PR preferentemente utiliza el nitrógeno mineral en rutas de obtención de energía, mientras que en PG el ciclo es predominantemente asimilativo y el nitrógeno inorgánico es incorporado a materia orgánica.

Respecto al ciclo del azufre, el tapete PG presenta un ciclo inorgánico del azufre cerrado, en donde el sulfito (SO_3) es oxidado a sulfato (SO_4) por *soxC*, el SO_4 es en turno convertido a adenosina-5'-fosfosulfato (APS) por *sat* y el APS regresa a SO_3 mediante *dsrA*. En la parte orgánica, PG presenta una abundancia de arilsulfatasas (*aslA*) y la degradación de dimetilsulfóxido (DMSO) está orientada a cubrir la demanada de SO_3 por medio de *ssuD* y *dsrA*. En contraste, la parte inorgánica en el tapete PR está dominada por la reducción asimilativa de SO_4 a SO_3 por la ruta de APS y fosfoadenosina-5'-fosfosulfato (PAPS) mediante los genes *cysNC*, para ser completamente reducido a ácido sulfídrico (H_2S) mediante *cysIJ*. La degradación de compuestos orgánicos tienen el mismo fin, pues la dimetilsulfona es convertida en SO_3 por *sfnG* y *ssuD* y el DMS es convertido en H_2S mediante *dmoAB* y *ssuD*.

Cabe resaltar que ninguno de los tapetes presentan los genes para la transformación de H_2S a azufre elemental citoplásmico (S0) *fccAB* y *sqr*, por lo tanto, se esperaría que la comunidad del tapete PR libere ácido sulfídrico a la atmósfera. Igualmente, los tapetes no presentan genes involucrados en la degradación del DMS vía la oxidación de DMSO (*dsoAB*). Ambos tapetes presentan una preferencia por el transporte transmembranal de compuestos inorgánicos del azufre (sulfato y tiosulfato) que por compuestos organosulfatados (DMSO, DMS, y L-cisteína), y el tapete PR es ligeramente más rico en éstos transportadores que PG. Finalmente, ambos tapetes presentan una abundancia en thioredoxinreductasas.

Cabe resaltar que el ciclo del azufre presenta un patrón contrario al observado en el ciclo del nitrógeno en el tapete PR, pues en éste caso la mayoría de los genes hallados codifican para proteínas involucradas en procesos de asimilación. Ésto sugiere que en la comunidad de PG hay un mayor balance entre las diferentes partes de un ciclo cerrado del azufre, mientras que PR parece estar más limitado por azufre, dado que existe una sobrerrepresentación de las rutas asimilatorias.

A pesar de que se ha predicho una preferencia por la utilización de fosfonatos en ambientes acuáticos (DyhrmaN et al. 2007; Gilbert et al. 2009; Martinez et al. 2010), los tapetes microbianos presentan una mayor preferencia por la utilización de fosfatos, ambos presentan transportadores de alta (*pstS*) y baja afinidad (*pitA*), con PR mostrando una mayor preferencia por *pitA*. La producción de compuestos macromoleculares de reserva es de gran importancia en éstos tapetes, pues el gen más abundante en ambos corresponde a la polifosfato cinasa (*ppK*), que lleva a cabo la síntesis y utilización de polifosfatos, y también en gran abundancia está presente la exopolyfosfatasa *ppX*. Se encontró una mayor preferencia por la fosfatasa alcalina *dedA* en PG, mientras que PR tiene una mayoría de la fosfatasa alcalina *phoX*, que hasta ahora había sido sólo identificada en ambientes marinos (Martínez et al. 2010), y en el caso de PG, la segunda categoría más abundante corresponde a la utilización de pirofosfato (*ppA*), cuya reacción produce dos fosfatos. En contraste, la segunda categoría más abundante en PR corresponde a la utilización de fosfito (*ptxBD*) e hipofosfito (*htxAB*), lo que considero es un efecto de la dominancia de *Pseudomonas,* dado que éstos operones fueron caracterizados en *P. stutzeri,* en donde son activados en condiciones limitantes de fosfato como fuente alternativa de fósforo (White & Metcalf 2004). Los genes involucrados en la utilización de fosfonatos fueron los menos abundantes en ambos tapetes, aunque el tapete PG es ligeramente más diverso en genes de utilización de fosfonatos.

Estudios anteriores detectaron en un genoma bacteriano de la región los genes que codifican para la sulfoquinovosa sintasa (*sqdB, sqdX*), involucrados en la síntesis de sulfolípidos de membrana que substituyen a los fosfolípidos en condiciones de limitación por fósforo (Alcaraz et al. 2008). Éstos genes fueron detectados en baja abundancia en relación con el resto de los genes de utilización de fosfato, pero fueron más comunes en el tapete PG. En resumen, en ambos tapetes parece haber una preferencia por la utilización de fosfatos sobre fosfonatos y fosfitos, y los miembros de *Pseudomonas* se ven beneficiados de su capacidad de utilizar otras fuentes alternativas de fosfato.

Respecto a la utilización de carbono, el tapete PG tiene significativamente una mayor proporción de genes involucrados en la fijación de carbono inorgánico, en concordancia con una mayor abundancia de organismos autotróficos. Las categorías más abundantes corresponden al Ciclo de Calvin (*ccbML, pgk*) y al ciclo del 3-hidroxipropionato/4-hidroxibutirato (3HP/4HB), a pesar de que éste ciclo ha sido hallado solamente en miembros del phylum *Archaea* y que éste phylum se encuentra en muy bajas proporciones en la comunidad.

En contraste, el metabolismo de carbono en el tapete PR está dominado por rutas metabólicas fermentativas de compuestos complejos y por rutas de utilización de monosacáridos y compuestos de un sólo carbono. Esto refuerza la idea de que el tapete en PG tiene una mayor producción primaria por fijación de carbono que PR, mientras que PR se especializa en la utilización de compuestos simples y en la degradación de carbohidratos macromoleculares como substancias de reserva.

Conclusiones

En éste capítulo exploramos dos aproximaciones para el análisis funcional de metagenomas. La aproximación más simple y sencilla es la genecéntrica, en donde se compara la abundancia relativa de ciertos genes conocidos involucrados en funciones muy particulares, en éste caso de importancia biogeoquímica. Ésta aproximación resulta muy informativa cuando se conocen las secuencias de los genes de interés y cuando las funciones en cuestión son llevadas a cabo por pocos genes. Sin embargo, utiliza un porcentaje muy pequeño de la información total del metagenoma, y es sumamente sensible a errores causados por la existencia de genes no homólogos que lleven a cabo la misma función (*alternólogos*, Hernández-Montes et al. 2008).

La segunda aproximación integra el conocimiento bioquímico y enzimático relacionado con las secuencias homólogas de referencia en bases de datos y lo posiciona en un contexto metabólico funcional. Ésto permite comparar no sólo genes individuales sino rutas metabólicas contextuales enteras, robusteciendo las conclusiones realizadas sobre las diferencias metabólicas de dos metagenomas, y tiene la ventaja de que todas las rutas metabólicas análogas (y en consecuencia todos su alternólogos) serán agrupadas dentro de la misma categoría funcional.

En resumen, el metabolismo del tapete PR es preferentemente heterótrofo, con una gran diversidad de rutas involucradas en la utilización de compuestos orgánicos simples y complejos, y las rutas metabólicas de nutrimentos son preferentemente disimilatorias, utilizando carbono y nitrógeno como aceptor final de electrones en la respiración. La excepción es el metabolismo del azufre, en donde los compuestos inorgánicos son utilizados principalmente como componentes de materia orgánica. El tapete PR se beneficia de tener genes involucrados en la utilización de fosfatos y fosfitos en proporciones equivalentes. Por lo tanto, la capacidad de asimilar fuentes alternativas de fósforo, de degradar una gran variedad de compuestos orgánicos y de tolerar toxinas y metales le brinda a las especies de *Pseudomonas* una ventaja competitiva en las condiciones permanentemente cambiantes de las pozas de desecación.

La comunidad en el tapete PG, presenta un metabolismo preferentemente autotrófico, con una gran cantidad de rutas destinadas a la asimilación de compuestos inorgánicos y por lo tanto a la producción primaria de nueva materia orgánica. La abundancia relativa de las rutas es también mucho menos dispar, como se espera de una comunidad taxonómicamente equitativa. Una mayor diversidad en las rutas metabólicas sugiere la existencia de ciclos biogeoquímicos cerrados con un mejor aprovechamiento y reciclaje de los elementos fijados en la materia orgánica, como sucede con el caso del ciclo del azufre.

Es posible construir un escenario teórico en donde los tapetes microbianos desarrollan una mayor complejidad en las aguas oligotróficas más estables (que es lo que observamos por ejemplo en PG). En condiciones de perturbación por desecación, gran parte de las especies que componen el tapete no son capaces de sobrevivir las concentraciones crecientes de sales, metales u otras toxinas, ni los cambios en los régimenes diurnos de temperatura o el incremento en la temperatura promedio.

En este escenario, la materia orgánica en pie queda expuesta a la degradación por organismos de rápido crecimiento con gran versatilidad en la descomposición en compuestos moleculares más simples. Las *Pseudomonas* son organismos que presentan todas estas características y probablemente se vean competitivamente beneficiadas bajo éstas condiciones (que es lo que observamos en la PR).

Las comunidades de tapetes microbianos analizadas en ésta sección son muy diferentes tanto en su composición y estructura taxonómica como en su complemento funcional. Ésto revela una enorme versatilidad de estrategias para la sobrevivencia en ambientes pobres en nutrientes que van desde la incorporación *de novo* de compuestos inorgánicos a la biomasa hasta el eficiente reciclaje de materia orgánica en pie. Sin embargo, la búsqueda del uso de estrategias comúnes y nuevas rutas metabólicas específicas para la sobrevivencia en éste tipo de ambientes requiere análisis más específicos que contemplen la comparación de ambientes fisicoquímicamente equivalentes y bajo regímenes de perturbación similares pero con composiciones taxonómicas diferentes, de manera que sea posible eliminar los factores de confusión incoporados por la diferencia de ambientes. Alternativamente, el análisis de comunidades a lo largo de gradientes en los regímenes de perturbación podrían revelar el enriquecimiento gradual de ciertos genes o rutas metabólicas.

ASTROBIOLOGY
Volume 12, Number 7, 2012

DOI: 10.1089/ast.2011.0694

Comparative Metagenomics of Two Microbial Mats at Cuatro Ciénegas Basin I: Ancient Lessons on How to Cope with an Environment Under Severe Nutrient Stress

Mariana Peimbert,[1,2] Luis David Alcaraz,[3,4] Germán Bonilla-Rosso,[1] Gabriela Olmedo-Alvarez,[3] Felipe García-Oliva,[5] Lorenzo Segovia,[6] Luis E. Eguiarte,[1] and Valeria Souza[1]

Abstract

The Cuatro Ciénegas Basin (CCB) is an oasis in the desert of Mexico characterized by low phosphorus availability and by its great diversity of microbial mats. We compared the metagenomes of two aquatic microbial mats from the CCB with different nutrient limitations. We observed that the red mat was P-limited and dominated by *Pseudomonas,* while the green mat was N-limited and had higher species richness, with Proteobacteria and Cyanobacteria as the most abundant phyla. From their gene content, we deduced that both mats were very metabolically diverse despite their use of different strategies to cope with their respective environments. The red mat was found to be mostly heterotrophic, while the green mat was more autotrophic. The red mat had a higher number of transporters in general, including transporters of cellobiose and osmoprotectants. We suggest that generalists with plastic genomes dominate the red mat, while specialists with minimal genomes dominate the green mat. Nutrient limitation was a common scenario on the early planet; despite this, biogeochemical cycles were performed, and as a result the planet changed. The metagenomes of microbial mats from the CCB show the different strategies a community can use to cope with oligotrophy and persist. Key Words: Microbial mats—Metagenomics—Metabolism. Astrobiology 12, 648–658.

1. Introduction

Microbial mats are self-sustaining communities with the capacity to perform all major biogeochemical cycles. Microbial mats are characterized by stratification of the microbial populations into distinct layers and are thought to be the earliest biological communities on Earth, as suggested by fossil stromatolites dated to 3.4 billion years ago (Des Marais, 1990; Tice and Lowe, 2004). Mats have inhabited Earth for many years and are found in many different environments, which highlights their great plasticity in adapting to different conditions (Paerl *et al.,* 2000; Kunin *et al.,* 2008; Lau *et al.,* 2009). Their wide range of metabolic capabilities (see Bender and Phillips, 2004) makes them systems of particular biotechnological importance. The geographical distribution of modern microbial mats is currently restricted to only a few aquatic systems, one of which is the Cuatro Ciénegas Basin (CCB), an oasis in the Chihuahuan Desert of Mexico.

The CCB is composed of a system of springs, pools, and streams that form an inverse archipelago in which each pool is an island (Souza *et al.,* 2008). The pools of the CCB exhibit the lowest phosphorous concentration reported in continental waters (Elser *et al.,* 2005). Geological data show that the valley and its hydrological systems have a unique ancient history (Souza *et al.,* 2006; Szynkiewicz *et al.,* 2009). It has been proposed that aquatic communities in the CCB diverged from their marine ancestors in the Jurassic period, when the CCB was under the ocean (Souza *et al.,* 2006; Moreno-Letelier *et al.,* 2011). Paleo-pollen data show that no

[1]Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, México D.F., México.
[2]Departamento de Ciencias Naturales, Universidad Autónoma Metropolitana, Cuajimalpa, México D.F., México.
[3]Departamento de Ingeniería Genética, Cinvestav, Campus Guanajuato, Irapuato, México.
[4]Departamento de Genómica y Salud, Centro Superior de Investigación en Salud Pública, Valencia, España.
[5]Centro de Investigaciones en Ecosistemas, Universidad Nacional Autónoma de México, Morelia, México.
[6]Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, México.

new soil has settled in the basin floor, and as a consequence no stores of nutrients have been built up (Minckley and Jackson, 2007). New phosphorus (P) arrives by wind transport and dust deposition, and is quickly occluded by the abundant calcium associated with limestone parent material. This long history of nutrient deprivation has resulted in adaptations, such as bacteria with reduced genomes or cell membranes where sulfolipids substitute for phospholipids (Elser *et al.,* 2005; Alcaraz *et al.,* 2008; Desnues *et al.,* 2008; Souza *et al.,* 2008). Paradoxically, this extremely unbalanced ecosystem has a large diversity of microbes, endemic fishes, snails, and zooplankton, and represents one of the most biodiverse sites by area in North America and the most important hot spot of endemicity at all taxa levels (Tatusov *et al.,* 2001; Souza *et al.,* 2006; Carson *et al.,* 2008; Cerritos *et al.,* 2008; Desnues *et al.,* 2008; Escalante *et al.,* 2008, 2009; Breitbart *et al.,* 2009; The Nature Conservancy, 2010; Wilson and Sherman, 2010).

It is still unknown how a nutrient-deprived ecosystem sustained solely by microbial mats can host such complex and diverse communities. To understand the ecological particularities of CCB microbial communities, we describe and compare in the present study the metagenomes of two aquatic microbial mats that live under different conditions within the oasis. A red mat was sampled from a shallow, highly variable desiccation pond with a very low C:N:P ratio (15820:157:1). The second mat, which we will call the green mat, was sampled from a permanent pool with a different C:N:P ratio (51:2:1) and a constant temperature. The two metagenomes show markedly different community structures and display different strategies for contending with oligotrophy.

2. Materials and Methods

2.1. Sampling site

Microbial mats and water samples were collected in July 2008 in the CCB. Sampling site coordinates were 26°52′17″N, 102°01′11.3″W for the red pond in the ejido Los Venados and 26°49′24.4″N, 102°00′53.2″W for the green pool in the Pozas Azules Ranch (PRONATURA). For each microbial mat, four temperature measurements were recorded daily for more than 6 months with a UA-002-64 data logger (Onset, MA, USA).

2.2. Physicochemical analysis

Two hundred milliliters of water were filtered through a 0.45 μm Millipore filter. All carbon (C) forms were determined with a total carbon analyzer (UIC Mod. CM5012; Chicago, IL, USA), and nitrogen (N) and phosphorus (P) forms were determined by colorimetric methods with use of a Bran-Luebbe Auto Analyzer III (Norderstedt, Germany). Total C and inorganic C were determined by combustion and coulometric detection (Huffman, 1977), respectively. Total organic C was calculated as the difference between total C and inorganic C. Total N and P were determined after acid digestion. P was determined by a molybdate-based colorimetric method after reduction with ascorbic acid (Murphy and Riley, 1962), and N was assayed by a macro-Kjeldahl method with colorimetric determination (Bremner and Mulvaney, 1982). The stoichiometric C:N:P ratio was calculated based on the mass (mg/L) of total organic carbon, total organic nitrogen, and total organic phosphorous. Water analyses were performed in CIECO/UNAM.

2.3. DNA isolation and sequencing

The microbial mat was collected with sterile equipment. DNA extraction was performed as described previously (Zhou *et al.,* 1997; Breitbart *et al.,* 2009) by using freeze/thaw, CTAB, and phenol-chloroform extraction. Samples were then further purified by electrodialysis as described by Rodriguez-Mejia *et al.* (2008). Total DNA was amplified with Genomiphi polymerase (GE Healthcare, Piscataway, NJ, USA) according to the manufacturer's instructions; random hexamers were used as primers. Ten independent 4 h reactions were carried out and then pooled before sequencing to reduce amplification bias. The resulting DNA was purified on silica columns (Qiagen) and concentrated by ethanol precipitation. Approximately 10 μg DNA was sequenced with pyrosequencing technology (454 FLX Roche Diagnostics, IN, USA) at Cinvestav-LANGEBIO, Irapuato, Mexico.

2.4. Data analysis

Each read from the data set was annotated with BlastN and BlastX (Altschul *et al.,* 1990) with a cutoff e-value of 10^{-5} with the following databases and annotation systems: Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.,* 2008), the SEED (Overbeek *et al.,* 2005), NCBI's NR, and the Clusters of Orthologous Groups (COGs) (Szynkiewicz *et al.,* 2009).

The metagenomes of the microbial mats were uploaded and compared against other metagenomes with MG-RAST (Meyer *et al.,* 2008). Whole metagenome data sets were used for functional comparisons; we used a cutoff value of 10^{-5} and the SEED Subsystems with a minimum identity of 60% in the overall alignment. The data sets employed for the comparisons are publicly available in MG-RAST and have the following accession numbers or names: 4442466.3, 4441363.3, 4440060.4, 4440964.3, 4440963.3, 4440965.3, 4440966.3, 4440967.3, 4440969.3, 4440970.3, 4440968.3, 4440971.3, 4440972.3, 4441576.3, 4441587.3, 4440067.3, 4441124.3, 4441125.3, 4441126.3, 4441127.3, 4441128.3, 4441129.3, 4441130.3, 4441131.3, 4441143.3, 4441144.3, 4441148.3, 4441152.3, 4441153.3, 4441579.3, 4441580.3, 4441582.3, 4441658.3, 4441584.3, 4441590.3, GS001a, GS002, GS003, GS004, GS006, GS009, GS010, GS011, GS012, GS017, GS020, GS038, GS040, GS041, GS042, GS043, GS044, GS045, GS046, GS117b, Guerrero Negro (all samples), PASstromBahamasMic20050722, Río Mesquites Bacteria, and the two metagenomes presented in this work. All abundance data have been normalized to values between 0 and 1. For the statistical analysis, we made use of R version 2.12.1 (x86_64) (R Development Core Team, 2008), and principal component analysis (PCA) was done with the FactoMineR library (Lê *et al.,* 2008). We also conducted a χ^2 test for independence across all samples. For the PCA, a normalized matrix, via multiple sample scaling, that contained the abundance of SEED Level 1 functional groups was used as input, as well as a qualitative supplementary variable indicating the sample's origin. PCA clustering was used to determine group formation along with the sample's origin. We

chose to cluster the following groups: CC, Guerrero Negro mats, GOS open ocean, and GOS coastal and estuarine samples. One-way ANOVA was performed with the PCA-derived groups, and the significance of this grouping across the gene's functional roles was assessed. A false discovery rate (FDR) (White *et al.*, 2009) p-value was calculated when comparing individual features with 1000 permutations and a significance threshold of 0.05.

3. Results

3.1. Pools

The unstable desiccation red pond (the name of the pond is due to the color of its water) from which the red mat was obtained is characterized by a very low P content (0.6 mg/L total organic phosphorus). The pond's water temperature fluctuates between 10°C and 60°C, with maximum daily fluctuations of 15°C during winter and 45°C in the summer. The water is slightly acidic (pH 5.5), while its conductivity (117.6 μS/cm) reveals a low-salt freshwater similar to that found in the Ice Lake (Minnesota, USA) or the oligotrophic Lakes Superior and Tahoe. This pond also shows an extremely oligotrophic C:N:P ratio of 15820:157:1 (for comparison, the Redfield ratio of 150:15:1 has been found in most aquatic ecosystems).

The permanent green pool is more stable (the water of this pool is green). The temperature was 25°C (±4 standard deviation) across the 6 months in which measurements were taken; the water is also slightly acidic (pH 6.0), and the conductivity is 2.57 μS/cm, which is very low for lentic soft-water bodies and an order of magnitude less than that of the red pond. Even though there are 2 orders of magnitude more P in the green mat than in the red mat, the proportion of N is very low (C:N:P ratio of 51:1.8:1). The green pool is also poor in other elements such as magnesium, sodium, potassium, chloride, and sulfate (the detailed chemical composition of the pools is available as Supplementary Material; Supplementary Data are available online at www.liebertonline.com/ast).

3.2. Metagenomes

The 454 pyrosequencing of the DNA obtained from the red mat yielded 347,728 reads that resulted in a total of 64 Mb of sequence data. The average read length was 226 bp, and the GC content mode was 60%. We identified 991 different ribosomal gene sequences, each of which was assigned to a genus in the SILVA database (Pruesse *et al.*, 2007). The red mat was found to be composed almost exclusively of bacteria (Archaea 0.26%, Eukaryota 1.78%), with the phylum Proteobacteria being predominant, followed by Firmicutes and Cyanobacteria (Table 1). Of the 347,728 total reads, 105,549 were assigned to taxa through the MG-RAST server and use of the RefSeq database (Pruitt *et al.*, 2005; Meyer *et al.*, 2008). Of these assigned reads, 55% belonged to *Pseudomonas*, which reveals that this gammaproteobacteria was dominant in the mat.

The green mat metagenome yielded 427,366 reads with an average size of 202 bp, which resulted in 86 Mb with a GC content mode of 35%. Of these, 94,009 reads were assigned to taxa through MG-RAST (Meyer *et al.*, 2008). In addition, 603 ribosomal genes were identified and assigned to a genus by using the SILVA database (Pruesse *et al.*, 2007). This mat was also composed almost exclusively of bacteria (Archaea 2.06%, Eukaryota 2.79%). In this case, the phyla Proteobacteria and Cyanobacteria were dominant, but no clear dominant taxon was identified (Table 1).

Table 1. Apparent Taxonomic Distribution of Metagenome Sequences

	Red mat all reads[a]	*Red mat rRNA*[b]	*Green mat all reads*[a]	*Green mat rRNA*[b]
Bacteria	97.96	93.95	95.15	92.87
Proteobacteria	76.52	70.33	36.27	36.48
Cyanobacteria	11.24	1.61	18.19	7.30
Firmicutes	4.31	10.9	12.91	19.24
Bacteroidetes	3.86	3.53	9.99	7.30
Actinobacteria	0.65	1.51	2.59	1.33
Chloroflexi	0.45	0.1	3.75	2.16
Planctomycetes	0.12	0	1.28	1.66
Verrucomicrobia	0.11	0	1.75	0.33
Chlorobi	0.1	0	1.99	0.17
Archaea	0.26	0.00	2.06	0.66
Eukaryota	1.78	6.05	2.79	6.47
Number of reads	105,549	991	94,009	603

[a]Metagenome affiliation was obtained by using the Metagenome RAST server (Meyer *et al.*, 2008). Only the most common phyla are shown.
[b]Reads identified as LSU and SSU rRNA genes.

3.3. Red mat metabolic analysis

Functional assignment of the reads was carried out by performing BLAST searches against the KEGG and COGs databases. The COGs annotation (Szynkiewicz *et al.*, 2009) matched 53,485 sequences. These results indicate that 37% of the identified sequences encoded proteins with known metabolic functions. A further 29% were similar to genes involved in cellular processes, while 18% were similar to genes involved in replication, transcription, and translation (Table 2).

By using the KEGG database (Kanehisa *et al.*, 2008), 43,112 sequences were assigned. The sequences were found to correspond to 3228 KEGG unique features (KO, KEGG Orthology) and 211 metabolic pathways. The most abundant pathways were the ABC transporters (8.83%), purine metabolism (6.43%), and the two-component systems (4.59%) (Table 3). *Pseudomonas* genomes have 111–123 KEGG pathways represented per genome. All the metabolic pathways present in the completely sequenced *Pseudomonas* genomes were found in the red mat metagenome, in addition to many other pathways such as photosynthesis, glycan biosynthesis, and linoleic acid metabolism. The red mat metagenome also contained pathways for the synthesis and degradation of secondary metabolites that are not present in any known *Pseudomonas* genome.

Similarly to *Pseudomonas*, the red mat metagenome also exhibited many genes involved in metabolic pathways for the degradation of organic compounds. We found genes involved in the degradation of toxic compounds such as toluene, xylene, ethylbenzene, naphthalene, styrene, DDT (insecticide), atrazine (herbicide), dichlorobenzene, hexachlorocyclohexane, tetrachloroethane, dichloroethane, chloroacrylate fluorine, and fluorobenzoate. We also found pathways for producing

Table 2. Percentage of Metagenome Sequences Similar to Major Metabolism

COG category[a]	*Red mat*[b]	*Green mat*[b]	*P*[c]
Information storage and processing			
Translation, ribosomal structure, and biogenesis	3.89 ± 3.02	10.05 ± 1.94	2e-300
DNA replication, recombination, and repair	8.99 ± 3.02	10 ± 2.88	2e-14
Transcription	4.77 ± 1.67	2.82 ± 2.14	1e-92
	17.65	**22.87**	
Metabolism			
Amino acid transport and metabolism	10.66 ± 2.85	8.78 ± 3.1	1e-41
Energy production and conversion	8.26 ± 2.62	7.33 ± 2.77	1e-14
Carbohydrate transport and metabolism	5.41 ± 2.31	5.6 ± 2.28	0.968
Coenzyme metabolism	3.88 ± 2.14	4.74 ± 1.94	2e-20
Nucleotide transport and metabolism	2.43 ± 1.81	3.33 ± 1.55	4e-31
Lipid metabolism	4.52 ± 1.79	3.28 ± 2.09	2e-42
Secondary metabolites biosynthesis	1.84 ± 1.07	1.14 ± 1.35	2e-35
	37	**34.21**	
Cellular processes			
Cell envelope biogenesis	5.26 ± 2.8	8.48 ± 2.25	2e-130
Post-translational modification	3.56 ± 2.27	5.36 ± 1.86	1e-71
Inorganic ion transport and metabolism	7.99 ± 2.03	4.26 ± 2.73	7e-174
Signal transduction mechanisms	5.85 ± 2.02	4.23 ± 2.36	2e-54
Defense	2.4 ± 1.59	2.56 ± 1.54	0.989
Cell division and segregation	0.73 ± 1.28	1.65 ± 0.86	2e-68
Cell motility	2.11 ± 1.25	1.56 ± 1.44	2e-19
Secretion	0.97 ± 1.15	1.33 ± 0.98	5e-14
	28.86	**29.41**	
Poorly characterized			
general function	9.48 ± 2.81	8.51 ± 2.95	6e-14
unknown	6.94 ± 2.17	4.89 ± 2.56	3e-71
	16.42	**13.41**	

[a]Szynkiewicz *et al.,* 2009.
[b]The given error is the standard error from a FDR analysis.
[c]The P-values were calculated with a two-tailed *t* test.

antibiotics such as tetracycline, penicillin, streptomycin, novomycin, ansamycin, and vancomycin, as well as genes such as those coding for beta-lactamases and metallo-lactamases that indicate ways to resist those antibiotics.

3.4. Green mat metabolic analysis

The green mat metagenome was analyzed in the same way as the red mat, and 63,364 reads were assigned and classified by using the COGs database (Szynkiewicz *et al.,* 2009). Of these reads, 34% corresponded to genes involved in metabolism, 29% to genes involved in cellular processes, and 23% to genes involved in information storage and processing (Table 2). By using the KEGG database (Kanehisa *et al.,* 2008), 106,841 reads were classified by sequence analysis. These corresponded to 228 pathways, 19 of which are not represented in the red mat. Among the pathways absent in the red mat are those involved in signaling pathways and posttranslational modifications in eukaryotes. The most frequent pathways in the green mat are those involved in purine metabolism (6.54%), ABC transport (5.12%), and aminoacyl-tRNA biosynthesis (4.34%) (Table 4).

3.5. Metagenome comparison

By analyzing the frequency distributions of the COGs categories, we observed that categories involved in translation and cellular envelope biogenesis are more abundant in the green mat, while genes involved in transcriptional regulation and metabolism of inorganic ions are more frequent in the red mat (Table 2). We explored whether gene function differences correlated with genome size. To this end, the effective genome sizes were calculated as described by Raes *et al.* (2007), and we obtained a genome average size of 3.69 Mb for the red mat and 1.27 Mb for the green mat.

Specific genes were searched for as proxies or indicators of particular biogeochemical cycles. We looked for genes involved in phosphonate utilization (*phnD, phnH, htxB,* and *ptxB*), polyphosphate metabolism (*ppA, ppK,* and *ppX*), and phosphate recycling *(phoA, phoX,* and *pstS*). All these genes were found in both mats, with the exception being genes for phosphite and phosphonate utilization (*ptxB* and *phnH*) that were not found in the green mat. Sulfolipid biosynthesis genes *sqdB* and *sqdX,* in contrast, were more notable in the green mat. The genes *rbcL, rbcS, codH,* and *aclY,* used as markers for major pathways of carbon fixation, were all present in both mats.

Nitrogen analysis was performed by examining the genes *nifH, nrfA, narG, napA, narB, nirS, norB,* and *nosZ.* Nitrogen assimilatory pathways were observed to be more abundant in the green mat, with a marked preference for assimilatory nitrate reductases (*narB*) over their respiratory counterparts (*narG, napA, nirS*). There is a remarkable preference for iron-containing cytochrome cd1 nitrate reductase (*nirS*) because no copper-containing dissimilatory nitrate reductase (*nirK*) was found in either mat. Nitrogen fixation genes were detected only in the green mat (Fig. 1).

Table 3. Percentage of Red Mat Sequences Showing Homology to Genes Associated with KEGG Pathways (Kanehisa *et al.,* 2008)

KEGG pathway [path id]	*%*
ABC transporters [PATH:ko02010]	8.83
Purine metabolism [PATH:ko00230]	6.43
Two-component system [PATH:ko02020]	4.59
Oxidative phosphorylation [PATH:ko00190]	3.83
Glycolysis/Gluconeogenesis [PATH:ko00010]	3.40
Glycine, serine, and threonine metabolism [PATH:ko00260]	2.75
Porphyrin and chlorophyll metabolism [PATH:ko00860]	2.59
Alanine, aspartate, and glutamate metabolism [PATH:ko00250]	2.30
Arginine and proline metabolism [PATH:ko00330]	2.29
Aminoacyl-tRNA biosynthesis [PATH:ko00970]	2.16
Pyrimidine metabolism [PATH:ko00240]	2.07
Fatty acid metabolism [PATH:ko00071]	2.07
Cysteine and methionine metabolism [PATH:ko00270]	1.97
Fructose and mannose metabolism [PATH:ko00051]	1.87
Citrate cycle (TCA cycle) [PATH:ko00020]	1.85
Ribosome [PATH:ko03010]	1.81
Fatty acid biosynthesis [PATH:ko00061]	1.64
Valine, leucine, and isoleucine biosynthesis [PATH:ko00290]	1.52
Pentose phosphate pathway [PATH:ko00030]	1.51
Starch and sucrose metabolism [PATH:ko00500]	1.39

The most frequent paths are shown. The full table is available as Supplementary Data, Table S1 (www.liebertonline.com/ast).

Table 4. Percentage of Green Mat Sequences Showing Homology to Genes Associated with KEGG Pathways (Kanehisa *et al.,* 2008)

KEGG pathway [path id]	*%*
Purine metabolism [PATH:ko00230]	6.54
ABC transporters [PATH:ko02010]	5.12
Aminoacyl-tRNA biosynthesis [PATH:ko00970]	4.34
Oxidative phosphorylation [PATH:ko00190]	3.89
Glycolysis/Gluconeogenesis [PATH:ko00010]	3.31
Nucleotide excision repair [PATH:ko03420]	2.43
Two-component system [PATH:ko02020]	2.43
Valine, leucine, and isoleucine biosynthesis [PATH:ko00290]	2.35
Porphyrin and chlorophyll metabolism [PATH:ko00860]	2.34
Pyrimidine metabolism [PATH:ko00240]	2.26
Alanine, aspartate, and glutamate metabolism [PATH:ko00250]	2.12
Ribosome [PATH:ko03010]	1.95
Glycine, serine, and threonine metabolism [PATH:ko00260]	1.83
Fructose and mannose metabolism [PATH:ko00051]	1.72
Peptidoglycan biosynthesis [PATH:ko00550]	1.61
Protein export [PATH:ko03060]	1.60
Citrate cycle (TCA cycle) [PATH:ko00020]	1.58
Homologous recombination [PATH:ko03440]	1.50
Pentose phosphate pathway [PATH:ko00030]	1.24
Phenylalanine, tyrosine, and tryptophan biosynthesis [PATH:ko00400]	1.22

The most frequent paths are shown. The full table is available as Supplementary Data, Table S1 (www.liebertonline.com/ast).

We next compared these two CCB mats against other aquatic microbiomes whose metagenomes are known. We used metagenomes from two stromatolites from the CCB (also isolated from phosphorus-deprived oligotrophic conditions), the hypersaline microbial mat of Guerrero Negro—for which each layer was studied independently—and the Global Ocean Sampling (GOS) Expedition (Rusch *et al.,* 2007; Kunin *et al.,* 2008; Breitbart *et al.,* 2009). We used the SEED subsystems to compare the annotations for each of the analyzed metagenomes by PCA. Three large clusters were formed: (1) GOS open ocean, (2) Guerrero Negro, and (3) GOS coastal and estuarine. The red mat clustered together with the GOS coastal water samples, and the green mat was also close to this cluster. The nearest neighbor of stromatolite PA was stromatolite RM, although these two clustered apart from the rest of the samples. Interestingly, when analyzing differences in gene functions by clustering the groups from the CCB, marine environments, and layers of the Guerrero Negro hypersaline mat, the only significant difference found by PCA (SEED level 1) was in genes of the photosynthesis category (FDR $p = 0.0374$). Within this category, the most differences were found in genes involved in electron transport and photophosphorylation that were more abundant (FDR $p = 0.0096$) in the metagenomes from coastal waters and the CCB than in the metagenomes from the open ocean. Light-harvesting complex-related genes were also more abundant (FDR $p = 0.0111$) in the basal group of the PCA (Fig. 2).

4. Discussion

In this work, we compared the metagenomes of two oligotrophic microbial mats, one a P-limited red mat from a desiccation pond and the other a N-limited green mat from a permanent pool at the CCB. The two mats were collected from locations that are less than 10 km apart, and their pools are immersed in a calcareous environment. Oligotrophic environments are defined by their low nutrient availability; oligotrophic continental water bodies are limited by P (Correll, 1999), while oceans are limited by P, N, and Fe (Mills *et al.,* 2004). In the latter, primary producers such as cyanobacteria can fix atmospheric nitrogen, using nitrogenase. This is not the case with phosphorus because it lacks volatile atmospheric compounds and thus enters into aquatic

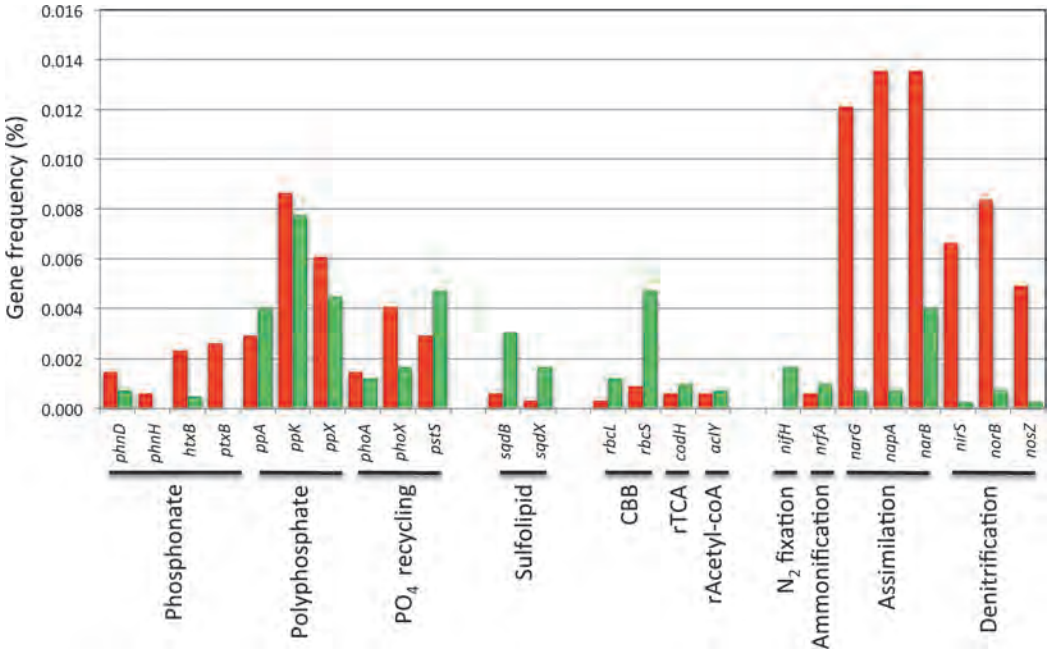


FIG. 1. Frequency of reads corresponding to gene markers for biochemical cycles. Red mat is shown in red bars while green mat is shown in green bars. Genes correspond to the following proteins: *phnD* and *ptxB* (phosphonate transporters), *phnH* and *htxB* (C-P lyase), *ppA* (pyrophosphatase), *ppK* (polyphosphatase kinase), *ppX* (exopolyphosphatase), *phoA* and *phoX* (alkaline phosphatases), *pstS* (phosphate transporter), *rbcL* and *rbcS* (RuBisCO), *codH* (CO dehydrogenase), *aclY* (citrate lyase), *nifH* (nitrogenase reductase), *nrfA* (nitrite reductase), *narG* (nitrate reductase), *napA* (nitrate reductase), *narB* (nitrate reductase), *nirS* (nitrite reductase), *norB* (nitric oxide reductase), and *nosZ* (nitrous oxide reductase). Color images available online at www.liebertpub.com/ast

systems mainly through deposition on surface waters. The C:N:P ratio in both pools is well below the Redfield ratio (106:16:1) required in the ocean to allow phytoplankton growth and sustain ecosystems (Redfield, 1934; Souza *et al.,* 2007).

We found that microbial communities in both the red and green mats are dominated by known heterotrophic taxa, while other previously reported phosphorus-limited microbial communities, such as the Mediterranean Ocean and the Río Mesquites oncolite (stromatolite RM), are dominated by autotrophs (Krom *et al.,* 1991; Breitbart *et al.,* 2009). The oceanic N:P ratio is regulated by nitrogen-fixing organisms. When the N:P ratio decreases, diazotrophs acquire an adaptive advantage due to their ability to obtain new nitrogen. As the ratio increases, rapidly growing heterotrophic organisms displace the high-energy-requiring diazotrophs (Tyrrell, 1999). This fine homeostatic mechanism seems to be disrupted in the CCB, where phosphorus is rapidly

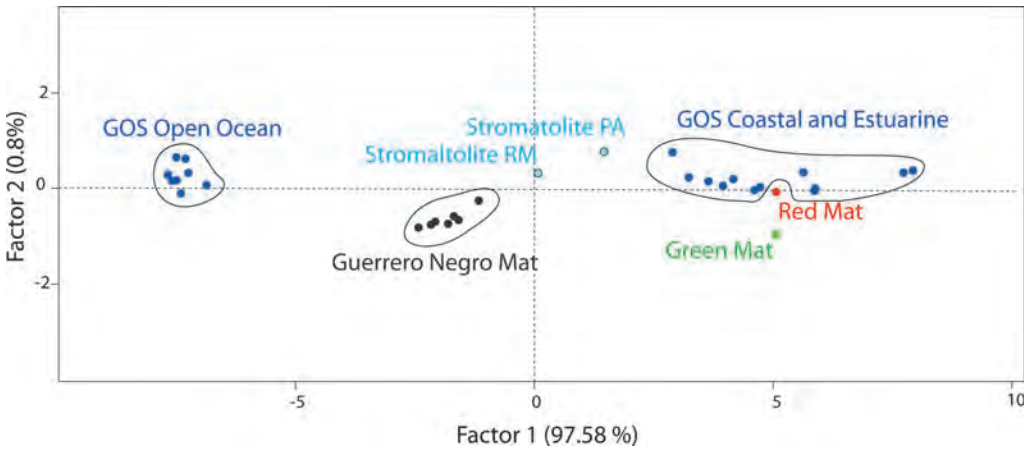


FIG. 2. Functional comparison of metagenomes from aquatic microbiomes by PCA analysis of SEED Subsystems level 1 (Meyer *et al.,* 2008). Stromatolites PA and RM (Breitbart *et al.,* 2009), Global Ocean Sampling (GOS) (Rusch *et al.,* 2007), Guerrero Negro hypersaline mat (Kunin *et al.,* 2008). The correlation of each variable with factor 1 is available as Supplementary Material, Table S2. Color images available online at www.liebertonline.com/ast

mineralized and most of the available phosphorus-containing compounds exist within standing organic biomass. This sequestration of phosphorus explains the abundance of heterotrophic taxa.

As previously mentioned, the pond where the red mat was sampled is small and has large fluctuations in temperature and water volume, with consequent changes in conductivity. The red mat is strongly dominated by the genus *Pseudomonas,* as 55% of the red mat sequence reads reveal identity to this genus. *Pseudomonas* are characterized by a large genomic and metabolic plasticity that allows them to survive in many environmental conditions (Mathee *et al.,* 2008; Kümmerli *et al.,* 2009; Klockgether *et al.,* 2011). In contrast, the pool from which the green mat was harvested has much lower temperature fluctuations, an extreme N limitation instead of a P limitation, and a large microbial diversity. When assigning a genus to each read, we observed a greater richness in the green mat (801 genera in the green mat vs. 698 in the red mat). Likewise, we observe a greater evenness in the green mat; *Cyanothece* is the most abundant genus in the green mat, with a representation of 4.5%. Both richness and evenness indicate that the diversity of the green mat is much greater than the diversity of the red carpet, as is confirmed by the Simpson's index (*D*), which is 0.0087 for the green mat and 0.2823 for the red mat.

Surprisingly, despite the differences in ecology, nutrient limitation, and species richness and diversity, both mats can perform a wide array of metabolic functions and have almost the same diversity of COGs (*i.e.,* 3125 COGs in the red mat versus 3025 COGs in the green mat). In this work, we studied gene presence, and it should be noted that further gene expression analysis is needed for validation of the metabolic capacities of the mats. The number of metabolic pathways detected in each mat is quite similar (211 for the red mat and 228 for the green mat) (Supplementary Fig. S1). Nevertheless, large differences were observed within the relative frequencies in which different pathways are represented. In the red mat, we observed a higher frequency of ABC transporters and two-component system pathways (with red-mat-to-green-mat ratios of 1.7 and 1.9, respectively; Tables 3 and 4). ABC transporters and two-component system genes are particularly important for organisms to be able to survive in a variety of environmental conditions.

Biogeochemical cycles are of special interest because these pathways indicate how a community incorporates essential elements for later reuse by various mat components. Microbial mats are known to display complete biogeochemical cycles (Canfield and Des Marais, 1993). Here, we determined whether a given biogeochemical cycle was present by analyzing the data for the presence of key proxy or marker genes within the cycle's various pathways.

Many different mechanisms exist to cope with low phosphorus availability, such as using alternative phosphorus sources, using polyphosphates as storage compounds, or employing a highly effective phosphate-recycling mechanism. The use of alternative phosphorus sources (phosphonates, phosphites, and hypophosphites) is revealed by the presence of the high affinity transporters *phnD* and *ptxB* as well as by C-P lyase genes *phnH* and *htxB* (White and Metcalf, 2004). The genes *ppA, ppK,* and *ppX* are also induced under phosphate-limiting conditions; these genes are involved in polyphosphate metabolism and code for pyrophosphatase, polyphosphatase kinase, and exopolyphosphatase, respectively. Polyphosphate acts as a reservoir of intracellular phosphate, a strategy that seems to be particularly important for motility and biofilms (Brown and Kornberg, 2004). Extracellular phosphates are recycled by the overexpression of alkaline phosphatases *phoA* and *phoX* as well as by the high-affinity phosphate transporter *pstS* (Scanlan *et al.,* 1993; Suzuki *et al.,* 2004; Zaheer *et al.,* 2009). All three of the strategies described above are utilized by both mats; however, the use of alternative sources of phosphates appeared to be more important for the red mat because *ptxB* and *phnH* were not found in the green mat (Fig. 1). Also exclusive to the red mat was the presence of coding genes for phosphate-binding DING proteins, which may be another resource more typically used by *Pseudomonas* to deal with phosphate starvation (Berna *et al.,* 2009). We detected in both mats the *sqdB* and *sqdX* genes involved in sulfolipid biosynthesis, which is another mechanism to contend with limited phosphate (van Mooy *et al.,* 2004: Alcaraz *et al.,* 2008). Interestingly, despite the low N:P ratio of the green mat, *seqB* and *sqdX* abundance is 6 times higher than in the red mat, which suggests that sulfolipid biosynthesis is not induced only by P limitation as expected (Fig. 1).

Autotrophic primary production appears to be more important in the green mat, because the gene frequency of carbon-fixation pathways in the green mat is greater than in the red mat. This is especially true for RuBisCO, which has a frequency 5 times greater in the green mat than in the red mat (Fig. 1). This is consistent with a higher proportion of Cyanobacteria detected in the green mat (Table 1). However, both mats also contain genes for the reductive acetyl-CoA pathway and the reductive tricarboxylic acid (rTCA) cycle, two alternative carbon fixation pathways that are present in a variety of microorganisms (Hugler *et al.,* 2011). The reductive acetyl-CoA pathway is twofold more common in the green mat, as indicated by the frequency of the CO dehydrogenase gene *codH.* This pathway is present in both archaea and bacteria under reductive and anaerobic conditions (Berg *et al.,* 2010). Citrate lyase is one of the few enzymes unique to the rTCA cycle (Wahlund and Tabita, 1997), a pathway exclusive to anaerobic and microaerophilic bacteria that also uses many of the enzymes involved in the TCA cycle. Citrate lyase (*aclY*) was observed to be 1.5 times more frequent in the green than in the red mat. The other three known pathways for carbon fixation were not detected in any of the mats. It is noteworthy that the distribution of these routes is much more restricted: the 3-hydroxypropionate (3-HP) bicycle is exclusive to Chloroflexaceae, while the 3-hydroxypropionate/4-hydroxybutyrate (3-HP/4-HB) cycle and the dicarboxylate/4-hydroxybutyrate (DC/4-HB) cycle occur only in Crenarchaeota (Hugler *et al.,* 2011).

Nitrogen metabolism was found to be very different between the two mats. Nitrogen fixation genes were only detected in the green mat, where we also observed a low GC content (GC mode of 35%). This is consistent with the extreme N limitation of the site, as a high GC content requires more nitrogen. The low GC content represents an adaptive advantage in environments with low nitrogen availability (Biers *et al.,* 2009). In contrast, nitrogen cycle genes in the red mat are predominantly involved with nitrate (NO_3) assimilation and respiration in agreement with a much higher GC content, which peaks at 60%. These observations also suggest

that nitrogen is not a limiting nutrient in the red mat environment.

Transmembrane transporter analysis revealed other relationships between the organisms and their environment (Patel *et al.,* 2010). A larger amount of transporters were found in the red mat, notably the ATP-hydrolyzing ABC transporter family (8.83% in the red mat vs. 5.12% in the green mat). The red mat also has cellobiose transporters, which suggests that it has the capacity to degrade cellulose as a carbon source. In the red mat, the frequency of genes coding for transporters of the osmoprotectants choline and betaine is more than 10 times that of the green mat. This reinforces the observation that the red mat experiences stressful conditions during periods of desiccation. In contrast, there are 8 times more transporters involved in the assembly of Fe-S clusters in the green mat, which confirms the importance of photosynthesis and nitrogen fixation for this mat.

We observed yet more contrasts between the two mats through the analysis of those genes with greater differences in their relative frequencies of appearance (Table 2). In the green mat, there are several genes involved in synthesis and degradation of the cell wall. In particular, carboxypepetidase genes are twice as frequent in the green mat as in the red mat. In contrast, the red mat has 3 times more outer membrane-related genes, such as porin genes. The high frequency of outer membrane proteins in the red mat is consistent with the particularities of *Pseudomonas,* as they have several of these proteins that help them respond to environmental changes (Remans *et al.*, 2010).

The nucleotide excision repair and mismatch repair (MMR) pathways also showed significant differences, as there are more than twice as many genes from these pathways in the green mat than in the red. Nucleotide excision repair is associated with DNA repair following UV damage (Goosen and Moolenaar, 2008). Both pools receive similar high levels of solar radiation, but the green pool is considerably more translucent than the red pool. MMR is usually involved with DNA repair after replication errors; however, the MMR pathway also participates in repairing damage caused by different types of stress (Kunkel and Erie, 2005). The above data suggest that maintenance of genome stability is more important in the green mat than in the red mat, where the plasticity of the *Pseudomonas* genomes may allow the community to survive the fluctuating environment.

We estimated the average size of the genomes in these metagenomes, using a method that relies on the relative frequency of 35 genes that are found as a single copy in bacterial genomes (Raes *et al.,* 2007). The estimated size of the genomes in the green mat was 3 times smaller (~1.27 Mb) than that estimated for the genomes in the red mat (~3.69 Mb), which suggests very different environmental strategies. The differences in genome size are linked to the overrepresentation of some functional categories, such as translation and replication in the green mat. This phenomenon occurs because different functional categories can be lost during genome streamlining in different organisms, while informational pathways are essential in all organisms and hence will be common to all genomes in the sample. Small genomes, such as those inferred to occur in the green mat, have been reported to be the consequence of genome streamlining as a response to low nitrogen and phosphate availability in some oceanic environments (Giovannoni *et al.,* 2005; Lauro *et al.,* 2009). Small genomes suggest an abundance of specialized bacteria that can only survive in very specific microniches, because they lack the required plasticity to survive in other environments.

In contrast, in the red mat we observed larger genomes with an overrepresentation of COGs involved in energy-dependent transport systems, cell motility, and transcriptional regulation, as well as genes involved in signal transduction that are crucial to sense and respond to changing environmental conditions (Konstantinidis and Tiedje, 2004). All these genomic features are characteristic of copiotrophic lifestyles (Lauro *et al.,* 2009), in which organisms are not dependent on carbon and nitrogen fixation and heterotrophic organisms predominate. These features also suggest that, despite the extremely biased Redfield ratio, the microbial community from the red mat is not nutrient limited because it is slowly consuming the already fixed nutrient-rich standing biomass. In contrast, in the oligotrophic green mat community, N limitation results in smaller genomes, an abundance of carbon- and nitrogen-fixing pathways, a high abundance of COGs involved in DNA repair, and a lower abundance of COGs involved in energy-dependent transport systems.

5. Conclusions

Thus we can conclude that, although both mat communities exist under the enormous environmental pressure that is nutrient deprivation, they seem to cope with it in very different ways, which suggests the existence of a wide array of strategies to survive in low-nutrient environments. The green mat represents a highly structured and fractioned niche inhabited by highly specialized bacteria (Diamond, 1975), while the red mat follows a "Red Queen" model, in which the plastic *Pseudomonas* must continuously change their strategy to maintain their dominance in an ever-changing environment (Table 5).

Finally, if we compare the functions inferred for CCB metagenomes with those from the GOS (Rusch *et al.,* 2007) and Guerrero Negro (Kunin *et al.,* 2008) by PCA, we observe that metabolic differences in such dissimilar microbiomes can be

Table 5. Summary of the Differences between the Red and Green Mat

	Red mat	*Green mat*
Pool	Shallow pond	Permanent pool
Nutrient limitation	Phosphorus	Nitrogen
Richness	698 genera	801 genera
Evenness (most abundant genus)	*Pseudomonas* 55%	*Cyanothece* 4.5%
Simpson index (*D*)	0.2823	0.0087
Gene presence	More transporters	More photosynthesis
	More transcriptional regulators	More DNA repair
		Nitrogen fixation
Average genome size	3.69 Mb	1.27 Mb
Strategy	Generalist bacteria	Specialized bacteria

explained almost entirely by the photosynthesis category. This suggests that the functional differences between aquatic microbiomes are subtle and should be studied in detail rather than by large functional categories. Moreover, this relatively low functional diversity invites us to think that, in these ancient communities, the function is what is being selected for rather than the species composition. Detailed community ecology and experimental evolution studies are needed in order to explain how the CCB microbial communities, which are dissimilar species assemblies analogous to ancient life on Earth, perform the same functions.

Acknowledgments

We thank Rodrigo González Chauvet for extraordinary technical logistics and field assistance. We also thank Africa Islas, Varinia Lopez, Frederique Reverchon, Eria Rebollar, Morena Avitia, Ana Gutierrez for assistance in DNA isolation; Celeste Martinez-Piedragil from CIECO/UNAM for assistance in the chemical analyses; Laura Espinosa from IE/UNAM for laboratory and technical assistance. This work was done with grants CONACyT 057507 and SEMARNAT 2006-C01-23459 to V.S. The manuscript was done while V.S. and L.E.E. were on sabbatical in UCI with support of DGAPA, UNAM, and UC-Mexus-CONACYT. G.B.R. was supported by CONACYT scholarship 196814 and Programa de Posgrado en Ciencias Biomedicas UNAM, and M.P. had a postdoctoral salary from CONACyT 057507.

Author Disclosure Statement

No competing financial interests exist.

Abbreviations

CCB, Cuatro Ciénegas Basin; COGs, Clusters of Orthologous Groups; FDR, false discovery rate; GOS, Global Ocean Sampling; KEGG, Kyoto Encyclopedia of Genes and Genomes; MMR, mismatch repair; PCA, principal component analysis.

References

Alcaraz, L.D., Olmedo, G., Bonilla, G., Cerritos, R., Hernandez, G., Cruz, A., Ramirez, E., Putonti, C., Jimenez, B., Martinez, E., Lopez, V., Arvizu, J.L., Ayala, F., Razo, F., Caballero, J., Siefert, J., Eguiarte, L., Vielle, J.P., Martinez, O., Souza, V., Herrera-Estrella, A., and Herrera-Estrella, L. (2008) The genome of *Bacillus coahuilensis* reveals adaptations essential for survival in the relic of an ancient marine environment. *Proc Natl Acad Sci USA* 105:5803–5808.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.

Bender, J. and Phillips, P. (2004) Microbial mats for multiple applications in aquaculture and bioremediation. *Bioresour Technol* 94:229–238.

Berg, I.A., Ramos-Vera, W.H., Petri, A., Huber, H., and Fuchs, G. (2010) Study of the distribution of autotrophic CO_2 fixation cycles in Crenarchaeota. *Microbiology* 156:256–269.

Berna, A., Scott, K., Chabriere, E., and Bernier, F. (2009) The DING family of proteins: ubiquitous in eukaryotes, but where are the genes? *Bioessays* 31:570–580.

Biers, E.J., Sun, S.L., and Howard, E.C. (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl Environ Microbiol* 75:2221–2229.

Breitbart, M., Hoare, A., Nitti, A., Siefert, J., Haynes, M., Dinsdale, E., Edwards, R., Souza, V., Rohwer, F., and Hollander, D. (2009) Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Cienegas, Mexico. *Environ Microbiol* 11:16–34.

Bremner, J.M. and Mulvaney, C.S. (1982) Nitrogen-total. In *Methods of Soil Analysis: Chemical and Microbiological Properties,* edited by A.L. Page, R.H. Miller, and D.R. Keeney, American Society of Agronomy and Soil Science Society of America, Madison, WI, pp 595–624.

Brown, M.R.W. and Kornberg, A. (2004) Inorganic polyphosphate in the origin and survival of species. *Proc Natl Acad Sci USA* 101:16085–16087.

Canfield, D.E. and Des Marais, D.J. (1993) Biogeochemical cycles of carbon, sulfur, and free oxygen in a microbial mat. *Geochim Cosmochim Acta* 57:3971–3984.

Carson, E.W., Elser, J.J., and Dowling, T.E. (2008) Importance of exogenous selection in a fish hybrid zone: insights from reciprocal transplant experiments. *Copeia* 4:794–800.

Cerritos, R., Vinuesa, P., Eguiarte, L.E., Herrera-Estrella, L., Alcaraz-Peraza, L.D., Arvizu-Gomez, J.L., Olmedo, G., Ramirez, E., Siefert, J.L., and Souza, V. (2008) *Bacillus coahuilensis* sp. nov., a moderately halophilic species from a desiccation lagoon in the Cuatro Cienegas Valley in Coahuila, Mexico. *Int J Syst Evol Microbiol* 58:919–923.

Correll, D.L. (1999) Phosphorus: a rate limiting nutrient in surface waters. *Poult Sci* 78:674–682.

Des Marais, D.J. (1990) Microbial mats and the early evolution of life. *Trends Ecol Evol* 5:140–144.

Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., Liu, H., Furlan, M., Wegley, L., Chau, B., Ruan, Y., Hall, D., Angly, F.E., Edwards, R.A., Li, L., Thurber, R.V., Reid, R.P., Siefert, J., Souza, V., Valentine, D.L., Swan, B.K., Breitbart, M., and Rohwer, F. (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* 452:340–343.

Diamond, J.M. (1975) Assembly of species communities. In *Ecology and Evolution of Communities,* edited by M.L. Cody and J.M. Diamond, The Belknap Press of Harvard University Press, Cambridge, MA, pp 342–444.

Elser, J.J., Schampel, J.H., Garcia-Pichel, F., Wade, B.D., Souza, V., Eguiarte, L., Escalante, A.E., and Farmer, J.D. (2005) Effects of phosphorus enrichment and grazing snails on modern stromatolitic microbial communities. *Freshw Biol* 50:1808–1825.

Escalante, A.E., Eguiarte, L.E., Espinosa-Asuar, L., Forney, L.J., Noguez, A.M., and Souza Saldivar, V. (2008) Diversity of aquatic prokaryotic communities in the Cuatro Cienegas basin. *FEMS Microbiol Ecol* 65:50–60.

Escalante, A.E., Caballero-Mellado, J., Martinez-Aguilar, L., Rodriguez-Verdugo, A., Gonzalez-Gonzalez, A., Toribio-Jimenez, J., and Souza, V. (2009) *Pseudomonas cuatrocienegasensis* sp. nov., isolated from an evaporating lagoon in the Cuatro Cienegas valley in Coahuila, Mexico. *Int J Syst Evol Microbiol* 59:1416–1420.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., Rappe, M.S., Short, J.M., Carrington, J.C., and Mathur, E.J. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309:1242–1245.

Goosen, N. and Moolenaar, G.F. (2008) Repair of UV damage in bacteria. *DNA Repair (Amst)* 7:353–379.

Huffman, E.W.D. (1977) Performance of a new automatic carbon dioxide coulometer. *Microchem J* 22:567–573.

Hugler, M., Petersen, J.M., Dubilier, N., Imhoff, J.F., and Sievert, S.M. (2011) Pathways of carbon and energy metabolism of the epibiotic community associated with the deep-sea hydrothermal vent shrimp *Rimicaris exoculata*. *PLoS One* 6:e1601810.1371.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* 36:D480–D484.

Klockgether, J., Cramer, N., Wiehlmann, L., Davenport, C.F., and Tummler, B. (2011) *Pseudomonas aeruginosa* genomic structure and diversity. *Front Microbiol* 2, doi:10.3389/fmicb.2011.00150.

Konstantinidis, K.T. and Tiedje, J.M. (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci USA* 101:3160–3165.

Krom, M.D., Kress, N., and Brenner, S. (1991) Phosphorus limitation of primary productivity in the E. Mediterranean sea. *Limnol Oceanogr* 36:424–432.

Kümmerli, R., Jiricny, N., Clarke, L.S., West, S.A., and Griffin, A.S. (2009) Phenotypic plasticity of a cooperative behaviour in bacteria. *J Evol Biol* 22:589–598.

Kunin, V., Raes, J., Harris, J.K., Spear, J.R., Walker, J.J., Ivanova, N., von Mering, C., Bebout, B.M., Pace, N.R., Bork, P., and Hugenholtz, P. (2008) Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol Syst Biol* 4, doi:10.1038/msb.2008.35.

Kunkel, T.A. and Erie, D.A. (2005) DNA mismatch repair. *Annu Rev Biochem* 74:681–710.

Lau, M.C., Aitchison, J.C., and Pointing, S.B. (2009) Bacterial community composition in thermophilic microbial mats from five hot springs in central Tibet. *Extremophiles* 13:139–149.

Lauro, F.M., McDougald, D., Thomas, T., Williams, T.J., Egan, S., Rice, S., DeMaere, M.Z., Ting, L., Ertan, H., Johnson, J., Ferriera, S., Lapidus, A., Anderson, I., Kyrpides, N., Munk, A.C., Detter, C., Han, C.S., Brown, M.V., Robb, F.T., Kjelleberg, S., and Cavicchioli, R. (2009) The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* 106:15527–15533.

Lê, S., Josse, J., and Husson, F. (2008) FactoMineR: an R package for multivariate analysis. *J Stat Softw* 25:1–18.

Letunic, I., Yamada, T., Kanehisa, M., and Bork, P. (2008) iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci* 33:101–103.

Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J.M., Koehrsen, M., Rokas, A., Yandava, C.N., Engels, R., Zeng, E., Olavarietta, R., Doud, M., Smith, R.S., Montgomery, P., White, J.R., Godfrey, P.A., Kodira, C., Birren, B., Galagan, J.E., and Lory, S. (2008) Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proc Natl Acad Sci USA* 105:3100–3105.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R.A. (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386.

Mills, M.M., Ridame, C., Davey, M., La Roche, J., and Geider, R.J. (2004) Iron and phosphorus co-limit nitrogen fixation in the eastern tropical North Atlantic. *Nature* 429:292–294.

Minckley, T. and Jackson, S. (2007) Ecological stability in a changing world? Reassessment of the palaeo-environmental history of Cuatrociénegas, Mexico. *J Biogeogr* 35:188–190.

Moreno-Letelier, A., Olmedo, G., Eguiarte, L.E., Martinez-Castilla, L., and Souza, V. (2011) Parallel evolution and horizontal gene transfer of the *pst* operon in *Firmicutes* from oligotrophic environments. *Int J Evol Biol* 2011, doi:10.4061/2011/781642.

Murphy, J. and Riley, J.P. (1962) A modified single solution method for the determination of phosphorus in natural water. *Anal Chim Acta* 27:31–36.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., and Vonstein, V. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33:5691–5702.

Paerl, H.W., Pinckney, J.L., and Steppe, T.F. (2000) Cyanobacterial-bacterial mat consortia: examining the functional unit of microbial survival and growth in extreme environments. *Environ Microbiol* 2:11–26.

Patel, P.V., Gianoulis, T.A., Bjornson, R.D., Yip, K.Y., Engelman, D.M., and Gerstein, M.B. (2010) Analysis of membrane proteins in metagenomics: networks of correlated environmental features and protein families. *Genome Res* 20:960–971.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glockner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35:7188–7196.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 33:D501–D504.

R Development Core Team. (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at http://www.r-project.org.

Raes, J., Korbel, J.O., Lercher, M.J., von Mering, C., and Bork, P. (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol* 8, doi:10.1186/gb-2007-8-1-r10.

Redfield, A.C. (1934) On the proportions of organic derivations in sea water and their relation to the composition of plankton. In *James Johnstone Memorial Volume,* edited by R.J. Daniel, University Press of Liverpool, Liverpool, pp 177–192.

Remans, K., Vercammen, K., Bodilis, J., and Cornelis, P. (2010) Genome-wide analysis and literature-based survey of lipoproteins in *Pseudomonas aeruginosa*. *Microbiology* 156:2597–2607.

Rodriguez-Mejia, J.L., Martinez-Anaya, C., Folch-Mallol, J.L., and Dantan-Gonzalez, E. (2008) A two-step electrodialysis method for DNA purification from polluted metallic environmental samples. *Electrophoresis* 29:3239–3244.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F., Utterback, T., Rogers, Y.H., Falcon, L.I., Souza, V., Bonilla-Rosso, G., Eguiarte, L.E., Karl, D.M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R., Strausberg, R.L., Nealson, K., Friedman, R., Frazier, M., and Venter, J.C. (2007) The Sorcerer II Global Ocean Sampling

expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5:e77.

Scanlan, D.J., Mann, N.H., and Carr, N.G. (1993) The response of the picoplanktonic marine cyanobacterium *Synechococcus* species Wh7803 to phosphate starvation involves a protein homologous to the periplasmic phosphate-binding protein of *Escherichia coli*. *Mol Microbiol* 10:181–191.

Souza, V., Espinosa-Asuar, L., Escalante, A.E., Eguiarte, L.E., Farmer, J., Forney, L., Lloret, L., Rodriguez-Martinez, J.M., Soberon, X., Dirzo, R., and Elser, J.J. (2006) An endangered oasis of aquatic microbial biodiversity in the Chihuahuan Desert. *Proc Natl Acad Sci USA* 103:6565–6570.

Souza, V., Falcón, L.I., Elser, J.J., and Eguiarte, L.E. (2007) Protecting a window into the ancient Earth: towards a Precambrian park at Cuatro Cienegas, Mexico. *The Citizen's Page, Evolutionary Ecology Research.* Available online at http://www.evolutionary-ecology.com/citizen/citizen.html.

Souza, V., Eguiarte, L.E., Siefert, J., and Elser, J.J. (2008) Microbial endemism: does phosphorus limitation enhance speciation? *Nat Rev Microbiol* 6:559–564.

Suzuki, S., Ferjani, A., Suzuki, I., and Murata, N. (2004) The SphS-SphR two component system is the exclusive sensor for the induction of gene expression in response to phosphate limitation in synechocystis. *J Biol Chem* 279:13234–13240.

Szynkiewicz, A., Ewing, R.C., Moore, C.H., Glamoclija, M., Bustos, D., and Pratt, L.M. (2009) Origin of terrestrial gypsum dunes—implications for martian gypsum-rich dunes of Olympia Undae. *Geomorphology* 121:69–83.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* 29:22–28.

The Nature Conservancy. (2010) The Nature Conservancy. Available online at http://www.nature.org.

Tice, M.M. and Lowe, D.R. (2004) Photosynthetic microbial mats in the 3,416-Myr-old ocean. *Nature* 431:549–552.

Tyrrell, T. (1999). The relative influences of nitrogen and phosphorus on oceanic primary production. *Nature* 400:525–531.

van Mooy, B.A.S., Devol, A.H., and Keil, R.G. (2004) Quantifying H-3-thymidine incorporation rates by a phylogenetically defined group of marine planktonic bacteria (Bacteriodetes phylum). *Environ Microbiol* 6:1061–1069.

Wahlund, T.M. and Tabita, F.R. (1997) The reductive tricarboxylic acid cycle of carbon dioxide assimilation: initial studies and purification of ATP-citrate lyase from the green sulfur bacterium *Chlorobium tepidum*. *J Bacteriol* 179: 4859–4867.

White, A.K. and Metcalf, W.W. (2004) Two C-P lyase operons in *Pseudomonas stutzeri* and their roles in the oxidation of phosphonates, phosphite, and hypophosphite. *J Bacteriol* 186: 4730–4739.

White, J.R., Nagarajan, N., and Pop, M. (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 5:e1000352.

Wilson, C.G. and Sherman, P.W. (2010) Anciently asexual bdelloid rotifers escape lethal fungal parasites by drying up and blowing away. *Science* 327:574–576.

Zaheer, R., Morton, R., Proudfoot, M., Yakunin, A., and Finan, T.M. (2009) Genetic and biochemical properties of an alkaline phosphatase PhoX family protein found in many bacteria. *Environ Microbiol* 11:1572–1587.

Zhou, J., Davey, M.E., Figueras, J.B., Rivkina, E., Gilichinsky, D., and Tiedje, J.M. (1997) Phylogenetic diversity of a bacterial community determined from Siberian tundra soil DNA. *Microbiology* 143:3913–3919.

Address correspondence to:
Valeria Souza
Departamento de Ecología Evolutiva
Instituto de Ecología
Universidad Nacional Autónoma de México
Apartado Postal 70-275
Ciudad Universitaria
Coyoacán 04510
México D.F.
México

E-mail: souza@servidor.unam.mx

Submitted 14 June 2011
Accepted 3 October 2011

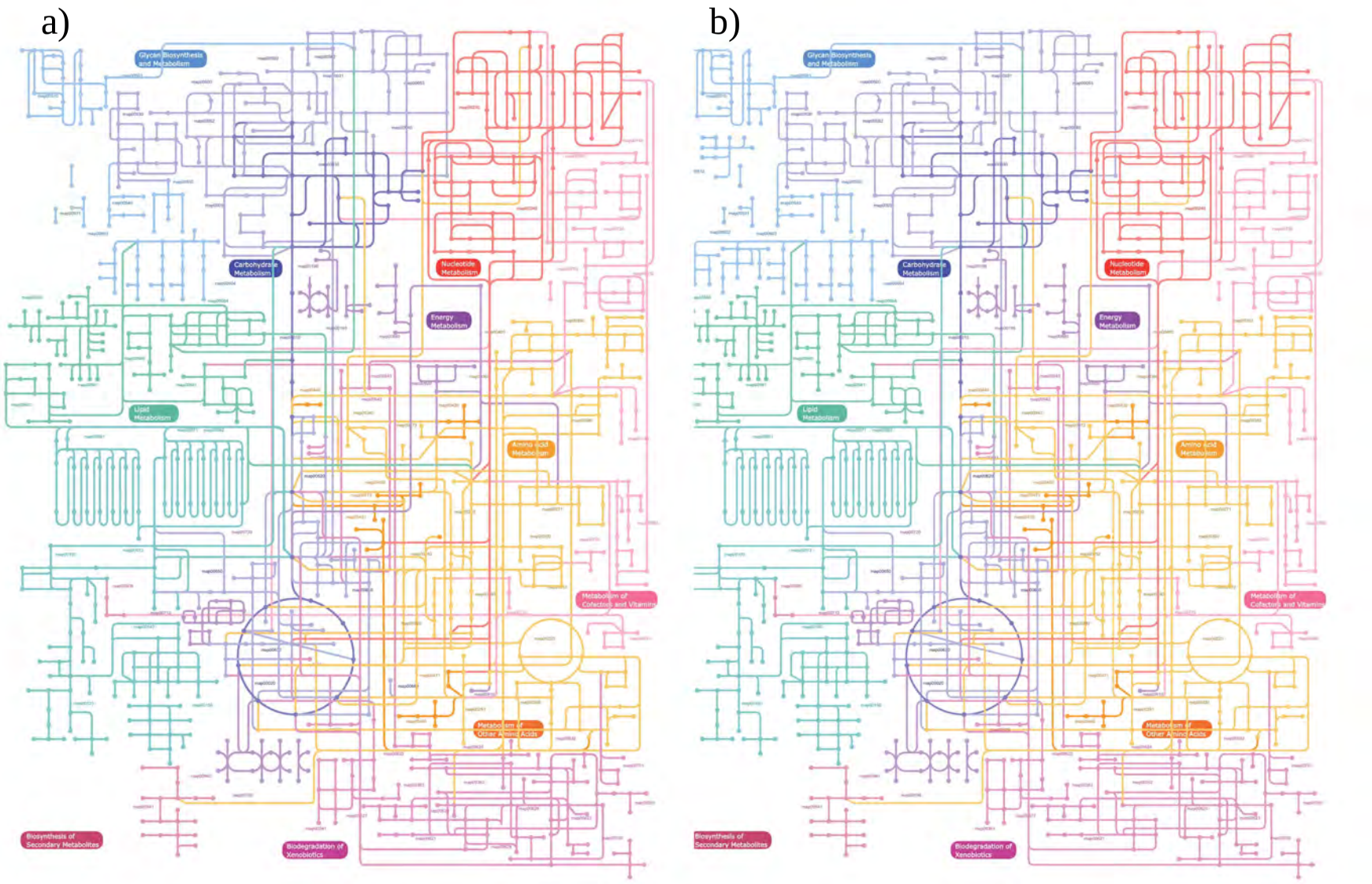


Figure S1. Graphic overview of the metabolism identified in the mats. Nodes represent compounds while edges represent enzymes. a) Pathways observed in Red mat; b) Pathways observed in Green mat. The underlying global pathways map is constructed using 123 KEGG pathways (Kanehisa 2008). This figure was made using iPath (Letunic 2008).

CAPÍTULO IV:
INTEGRACIÓN DE LOS PATRONES DE DIVERSIDAD TAXONÓMICA Y FUNCIONAL

Introducción

De manera intuitiva, en ecología se asume la existencia de una relación entre la diversidad taxonómica y la diversidad funcional (Tilman et al. 1997). Esto se debe a que a pesar de que ciertas especies pueden ocupar nichos funcionales equivalentes, el conjunto de todas las características fisiológicas y en el desarrollo de cada especie define la manera en que se particiona la energía en un ecosistema, y ésto a su vez determina la estructura de las comunidades (Brown 1995; Blackburn y Gaston 1998; Maurer 1999), de manera que cada especie es ecológicamente única. Siguiendo ésta lógica, asumimos también que en general una comunidad con una mayor diversidad de especies contiene también una mayor diversidad de funciones.

Sin embargo, mientras que el número de especies puede ser teóricamente infinito, el número de nichos ecológicos tiene que ser más limitado. Por ejemplo, se ha encontrado una alta redundancia en el número de gremios funcionales encontrados en comunidades vegetales naturales (Naeem et al. 1999; Díaz y Cabido 2001; Fonseca y Ganade 2001) y ecosistemas microbianos (Naeem & Li 1997; Wohl et al. 2004; Allison y Martiny 2008), y en ambos casos se demostró que ésta redundancia afecta el funcionamiento ecosistémico. Lo anterior ha llevado a la proposición de teorías de "ensamble de comunidades" basadas en la existencia de un número finito de funciones en un ecosistema; de manera que en una comunidad donde todos los nichos ya están ocupados, la incorporación de nuevas especies se traducirá en redundancia funcional (Sale 1977; Cheeson y Warner 1981; Munday 2004). Éstas teorías encajan en la teoría neutral unificada (Hubbell 2001) en donde se considera que las especies en un mismo nivel trófico son funcionalmente indistintas, por lo que la riqueza y abundancia de cada especie en particular depende exclusivamente de procesos estocásticos gobernados por la deriva y las tasas de migración y extinción. Es necesario tener en cuenta, sin embargo, que la teoría neutral ha recibido muchas críticas teóricas (Nee y Stone 2003; Ricklefs 2006) y ha fallado en describir comunidades complejas (Wootton 2005; Dornelas et al. 2006). En contraste, también existe evidencia que apunta hacia procesos determinísticos como la naturaleza del nicho ecológico para la determinación de la estructura de comunidades microbianas (Yang et al. 2005; Dumbrell et al. 2010).

En macroorganismos eucariontes la divergencia evolutiva conlleva también una diferenciación, generalmente morfológica y a veces fisiológica y funcional, mediada por la conducta y la misma morfología, pero con una diversidad metabólica limitada (Pace 1997). En las comunidades bacterianas el panorama es todavía más complejo, pues a pesar de que la morfología y la conducta parecen jugar papeles menos importantes, hay una gran cantidad de rutas metabólicas diferentes especializadas en llevar a cabo el mismo proceso enzimático de manera más eficiente en distintas condiciones ambientales, complicando las definiciones de redundancia funcional y diversidad metabólica (Pace 1997). Una complicación adicional en las comunidades bacterianas es que la transferencia horizontal puede homogeneizar funcionalmente una población taxonómicamente distinta, resultando en que organismos que provienen de distintos linajes llevan a cabo la misma función, e incluso que organismos que no se conoce que lleven a cabo una función de improviso la adquieran por transferencia horizontal de otros miembros en la comunidad (Ochman et al. 2000; Hacker y Carniel 2001; Dobrindt et al. 2004).

Las propuestas recientes en ensamble de comunidades tratan de reconciliar las perspectivas de nicho y neutrales (Leibold y McPeek 2006), y particularmente en comunidades bacterianas se ha retomado la "*teoría del ensamble por lotería*" (Burke et al. 2011), la cual incorpora aspectos neutrales y de nicho al proponer que los nichos ecológicos son colonizados aleatoriamente a partir de una poza de especies ecológica y funcionalmente similares que les permite coexistir dentro del mismo nicho (Sale 1977; Munday 2004). Bajo esta concepción, el nicho ecológico determina el número y tipo de gremios funcionales que es posible que existan dentro de un ecosistema, pero las dinámicas de sustitución dentro de cada gremio funcional son puramente aleatorias. Es decir, que la presencia de las funciones es determinística, pero la identidad de la especie particular que lleve a cabo una función es estocástica. Esto no quiere decir que todas las especies dentro de un mismo gremio sean equivalentes e indistintamente intercambiables, sino que el proceso determinístico que determina su presencia en una comunidad opera sobre funciones generales, permitiendo que cada especie mantenga su identidad funcional única mientras cumpla la función del gremio. Ésta identidad única también permite que especies ecológicamente similares ocupen el mismo nicho, pues las características particulares de cada especie permite una explotación diferencial del nicho.

Para explorar la relación entre la diversidad taxonómica y la funcional bajo el esquema teórico expuesto anteriormente, en éste Capítulo se contrastan los patrones de diversidad de cuatro metagenomas de Cuatrociénegas, que incluye a los tapetes microbianos PR y PG (Bonilla-Rosso et al 2012; Peimbert et al. 2012), y los estromatolitos PA y RM (Breitbart et al. 2009).

Metodología

La diversidad taxonómica de los cuatro metagenomas de Cuatrociénegas se analizó mediante perfiles de entropía de Rényi calculados a partir de las matrices de abundancia de especies obtenidas utilizando 31 genes codificantes de proteínas como indicadores ecológicos, como se describe en el Capítulo I (Bonilla-Rosso et al., 2012).

Análisis Funcional Global de las categorías del SEED

Los perfiles de Rényi proveen información sobre la comparación de la distribución de elementos en categorías (Rényi 1961, Tóthmérész 1995), por lo tanto es posible utilizarlos no sólo para comparar la diversidad taxonómica sino también las categorías funcionales de las comunidades. En un primer análisis, se consideraron como categorías funcionales cada uno de los subsistemas del SEED. Un subsistema del SEED es un juego de roles funcionales que en conjunto implementan un proceso biológico específico de manera análoga a una ruta metabólica pero incluyendo también complejos estructurales, y los sistemas y subsistemas son organizados y curados por expertos en cada tema (Overbeek et al. 2005).

Durante la anotación funcional de un metagenoma, las secuencias de cada metagenoma son asignadas a un subsistema, de manera que éstos representan las categorías y el número de secuencias la abundacia relativa de cada una de ellas, permitiendo directamente el cálculo de perfiles funcionales de la entropía de Rényi (Rényi 1961, Tóthmérész 1995). De ésta manera se obtuvieron perfiles de Rényi comparables que presentan información a diferentes escalas sobre la diversidad de especies y categorías funcionales.

Análisis de Diversidad de Gremios Funcionales

Utilizando los algoritmos de clasificación filogenética por último ancestro común implementados en MEGAN (Huson et al. 2007) es posible asignar una clasificación filogenética a cada secuencia dentro de cada categoría funcional del SEED. De esta manera, es posible definir "gremios funcionales" correspondientes a las

categorías funcionales y obtener medidas de diversidad taxonómica dentro de cada gremio también. En consecuencia se calcularon los perfiles funcionales de Rényi de los gremios Respiratorio, Fermentador y Fotosintético,y también dentro de los gremios funcionales que llevan a cabo los ciclos del Azufre, Nitrógeno y Fósforo, considerando como categorías los subsistemas individuales dentro de cada gremio, y como abundancia relativa el número de secuencias asignadas a cada subsistema. Las secuencias dentro de cada gremio fueron asignadas a una categoría taxonómica mediante MEGAN (Huson et al. 2007), y los perfiles de Rényi de diversidad taxonómica fueron calculados para cada gremio también. Ésto permite comparar a escalas más manejables la relación entre la diversidad taxonómica y funcional dentro de gremios funcionales particulares.

Análisis de Diversidad de Familias de Secuencias

Las categorizaciones funcionales arbitrarias del SEED presentan un problema que no es trivial para el análisis de patrones generales de diversidad. En lo que al análisis de la diversidad funcional existen dos aproximaciones alternativas: la primera consiste en analizar la diversidad taxonómica sólo dentro de las categorías funcionales bien conocidas, en analogía a los análisis de gremios funcionales o niveles tróficos de la ecología de macroorganismos. Esta aproximación permitiría abordar preguntas particulares sobre la diversidad taxonómica de organismos que realizan una misma función conocida. Sin embargo, es muy posible que los patrones generales de diversidad funcional o la relación general entre la diversidad funcional y taxonómica no sean evidentes con análisis a ésta escala, perdiendo información central para la identificación de similitudes y diferencias entre comunidades ecológicas.

Para poder aproximarnos a la comprensión de la diversidad y función de una comunidad microbiana, se analizaron las secuencias metagenómicas clasificadas en bases de datos existentes, analizando la diversidad de familias de secuencias. Se consideró la anotación de las proteínas traducidas de los metagenomas respecto a dos bases de datos de grupos de proteínas: la base de datos Pfam (Finn et al. 2008) y la base de datos del COG (Tatusov et al., 2003). El Pfam es una base de datos de familias de proteínas curada manualmente basada en la similitud de secuencias a nivel primario, mientras que el COG (Clusters of Orthologous Genes) define aglomeraciones de genes filogenéticamente relacionados por ortología. La diferencia fundamental radica en que mientras que dentro de las familias del Pfam se encuentran proteínas similares a nivel de secuencia, con alineaciones curadas manualmente, las aglomeraciones del COG se definen como proteínas que en un análisis pareado de genomas son más similares entre sí que hacia otras proteínas dentro del propio genoma. Para este análisis, se utilizó la anotación generada mediante la tubería de proceso RAMMCAP descrita en el Capítulo I, que asocia cada secuencia del metagenoma a una familia del Pfam y a un COG. Dado que ambas bases de datos comprenden categorías funcionales discretas y que las secuencias de los metagenomas son asignadas a éstas categorías, es posible calcular los perfiles entropía de Rényi para las categorías funcionales de COGs y Pfams también.

Análisis de Diversidad de Conglomerados Genómicos

Una segunda aproximación consiste en utilizar sistemas de clasificación de secuencias menos subjetivos, como lo son los sistemas de clasificación de proteínas basados en relaciones filogenéticas o similitud de secuencias (Li y Godzik 2006 ; Yooseph et al. 2007). Esta alternativa parte de que las funciones metabólicas son llevadas a cabo por proteínas, y que la similitud en estructura primaria de las proteínas indica también la similitud en la función que llevan a cabo. La consecuencia lógica es que una mayor diversidad funcional está respaldada por una mayor diversidad de proteínas. El problema es que la aproximación es insensible a las diferencias funcionales de secuencias muy similares y por lo tanto refleja

una diversidad funcional a niveles generales del mecanismo de acción enzimática. En el presente trabajo se eligió explorar ambas alternativas. En éste caso, se realizó una aglomeración jerárquica *de novo* de las secuencias metagenómicas de los cuatro metagenomas mediante el programa CD-HIT (Li & Godzik 2006). Es decir, se formaron conglomerados de secuencias relacionadas por la similitud de sus secuencias. Ésta es una medida indirecta de la diversidad funcional que parte del hecho que dado que una comunidad funcionalmente diversa debe poseer una mayor diversidad genómica total, debe existir un mayor número de conglomerados de secuencias diferentes a un determinado nivel de similitud. La aglomeración jerárquica permite analizar cómo cambia la agrupación de secuencias a diferentes niveles de similitud, al tiempo que agiliza el proceso de aglomeración. Más aún, la aglomeración jerárquica de las secuencias de DNA permite incluir en el análisis otros elementos de la diversidad genómica como lo son los RNAs, secuencias no codificantes y elementos de regulación. Para explorar las similitudes y diferencias de la diversidad funcional entre comunidades, se analizó el número de conglomerados compartidos entre los cuatro metagenomas, tanto en conglomerados que contienen secuencias altamente similares que reflejan las mismas funciones (95%) como conglomerados laxos e incluyentes que reflejan familias de secuencias relacionadas distantemente (45%). Finalmente, se calcularon los perfiles de entropía de Rényi para cada metagenoma considerando como categorías los conglomerados y el número de secuencias como abundancia relativa.

Por último, se compararon los perfiles de Rényi de diversidad funcional por conglomerados contra los perfiles de Rényi de diversidad taxonómica por marcadores filogenéticos. Para cada escala de entropía, se correlacionaron los valores taxonómicos y funcionales de cada metagenoma mediante una correlación de Pearson.

Resultados y Discusión

Análisis Funcional Global de las categorías SEED

En total, se detectaron 136,066 categorías del SEED para PG, 106,317 para PR, 46,014 para PA y 21,026 para RM, lo que refleja el esfuerzo de secuenciación de cada metagenoma (PG: 78.288 Mb; PR: 50.857 Mb; PA:12,202 Mb; RM:31,103 Mb) y se traduce en una densidad de rutas por megabase secuenciada de 1738 para PG, 2090 para PR, 1480 para PA y 1723 para RM. Éste patrón resulta extraño dado que la diversidad taxonómica de los metagenomas en orden decreciente es PG > PR > PA/RM (Capítulo II, Bonilla-Rosso et al. 2012), y sugiere que PR, aunque está dominada por *Pseudomonas* y tiene la menor riqueza de especies, es la que presenta la mayor diversidad funcional.

La comparación de la diversidad funcional según los perfiles de Rényi de las categorías del SEED (Fig. A) revela que las comunidades más diversas corresponden a los tapetes microbianos PG y PR, mientras que RM es la más equitativa y PA la más dominante.

Taxonómicamente, RM y PA son las menos diversas con RM siendo la más dominante de todas y PG y PA son más diversas que PR/RM tanto en riqueza como en equitabilidad (Capítulo II, Bonilla-Rosso et al. 2012. Los resultados podrían sugerir una interpretación que soporte la ausencia de una relación entre la diversidad funcional y la taxonómica. Sin embargo, es necesario tener en cuenta que los subsistemas del SEED son categorizaciones arbitrarias de rutas metabólicas, separadas y organizadas de manera que facilite su análisis y clasificación, y que una gran parte del metabolismo bacteriano se escapa al conocimiento humano contemporáneo, evidenciado por el elevado número de secuencias de metagenomas dentro de categorías marcadas como desconocidas.

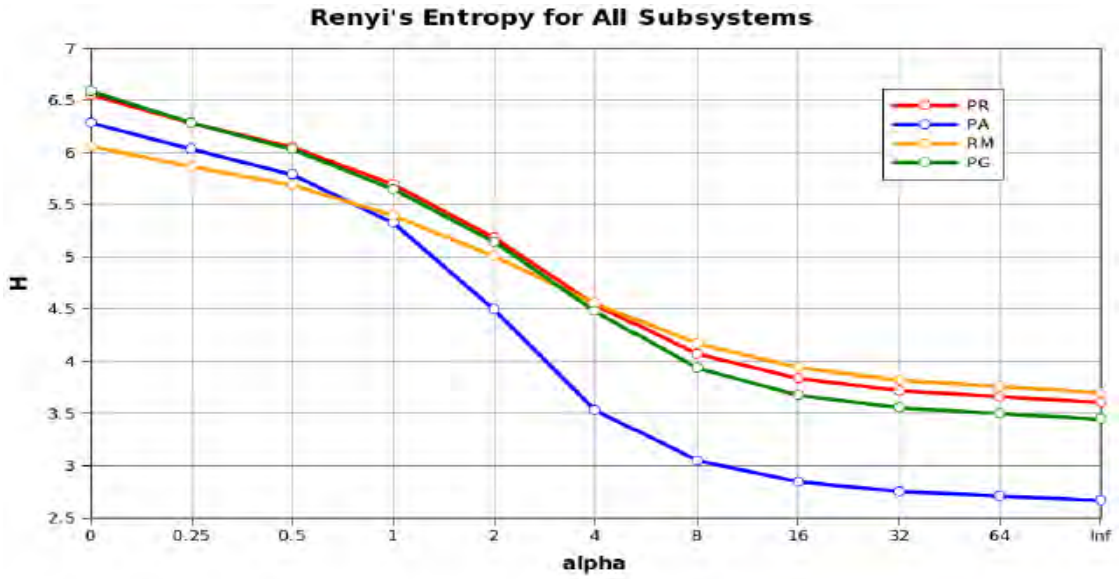


Fig. A. Perfiles de entropía de Rényi para los sistemas del SEED en PA (azul), PR (rojo), PG (verde) y RM (amarillo).

Si consideramos particularmente que PR (taxonómicamente la menos diversa) está dominada por *Pseudomonas*, y que éste género está profundamente estudiado y por lo tanto su metabolismo está bien representado en las bases de datos de referencia, esperaríamos observar una diversidad funcional alta en el perfil de Rényi. Sin embargo, los perfiles de diversidad funcional sugieren que PR es en total más diversa funcionalmente que PA, la cual es taxonómicamente más diversa y compleja que PR, pero presenta una gran proporción de organismos provenientes de grupos taxonómicos muy poco estudiados como el superphylum Planctomycetes-Verrucomicrobia, y por lo tanto los metabolismos de ésta muestra no están representados en las bases de datos de referencia y no son considerados por el SEED.

Análisis de Diversidad de Gremios Funcionales

Adicionalmente, analicé la diversidad exclusivamente dentro de funciones metabólicas conocidas a profundidad (respiración, fermentación, fotosíntesis y ciclos de nitrógeno, fósforo y azufre), en analogía al análisis de comunidades por gremio funcional o nivel trófico de macroorganismos. Esto permite contrastar la diversidad particular de un subconjunto de la comunidad relacionado ecológicamente.

Los patrones de diversidad funcional fueron visualizados mediante la construcción de perfiles de Rényi, en donde todas las secuencias categorizadas dentro de un subsistema fueron re-categorizadas dentro de los sistemas contenidos en el subsistema. Para el análisis taxonómico, las mismas secuencias de cada subsistema fueron clasificadas taxonómicamente mediante MEGAN (Fig. B).

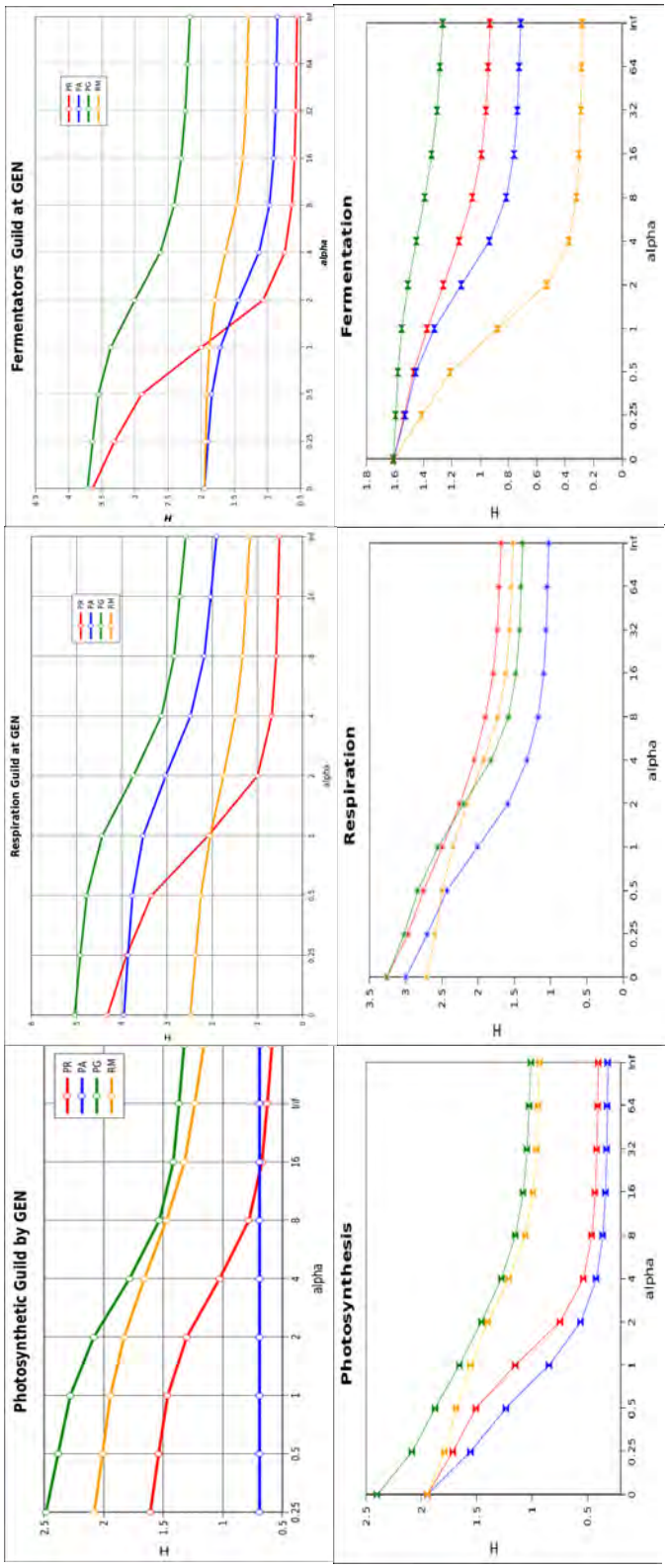


Fig. B. Perfiles de entropía de Rényi para el gremios funcional fotosintético (izquierda), respiratorio (centro) y fermentativo (derecha) para estimar diversidad taxonómica (renglón superior) y funcional (renglón inferior).

En un sistema complejo con una gran variedad de micro-nichos y una gran diversidad bacteriana, los patrones de diversidad se esperan diferentes en cada gremio funcional. El tapete PG resultó ser la comunidad con mayor diversidad taxonómica, independientemente del gremio funcional analizado (Fig. B). Sin embargo, ésta no se tradujo en una mayor diversidad funcional en todos los gremios, sino solamente en el gremio fotosintético y en el fermentativo, y en el respiratorio sólo presenta una mayor riqueza pero no menor dominancia.

Es interesante observar que el gremio fotosintético resultó ser el único en el que el patrón de diversidad funcional refleja el de diversidad taxonómica (FIG. B), o sea que tienen el mismo orden de mayor a menor diversidad (PG>RM>PR>PA). Por otra parte, el análisis revela una mayor diversidad funcional en el gremio de fermentadores en los tapetes (PG y PR) contra lo hallado en estromatolitos (RM y PA), a pesar de que PR tiene una marcada dominancia dentro de éste gremio. Ésto quiere decir que a pesar de que PR tiene una gran variedad de rutas para la fermentación, éstas se encuentran concentradas en pocas especies dominantes. Dicho de otra forma, las *Pseudomonas* dominantes presentan una gran diversidad fermentadora.

Asimismo, se observa que la dominancia taxonómica global observada en RM (principalmente *Alphaproteobacteria* y *Cyanobacteria*; Capítulo III, Bonilla-Rosso et al., 2012) (Fig. B) tiene un efecto sobre la diversidad funcional de todos los gremios menos el fotosintético, en el cual es la segunda comunidad más diversa tanto funcional como taxonómicamente. Dicho de otra forma, a pesar de que RM presenta una alta dominancia, el gremio de organismos fotosintéticos es taxonómica y funcionalmente diverso.

Los tapetes microbianos también son funcionalmente más diversos en cuanto al ciclo del azufre que los estromatolitos (Fig. C). Esto probablemente se debe no sólo a que CCB está dominado por suelos de sulfato de magnesio ($MgSO_4$; (Minckley 1969), y que las substancias exopoliméricas que recubren los tapetes contienen azufre (Braissant et al. 2007; Braissant et al. 2009), sino también porque los sulfonatos de bajo peso molecular son importantes fuentes de carbono y azufre en tapetes microbianos (Visscher et al. 1999).

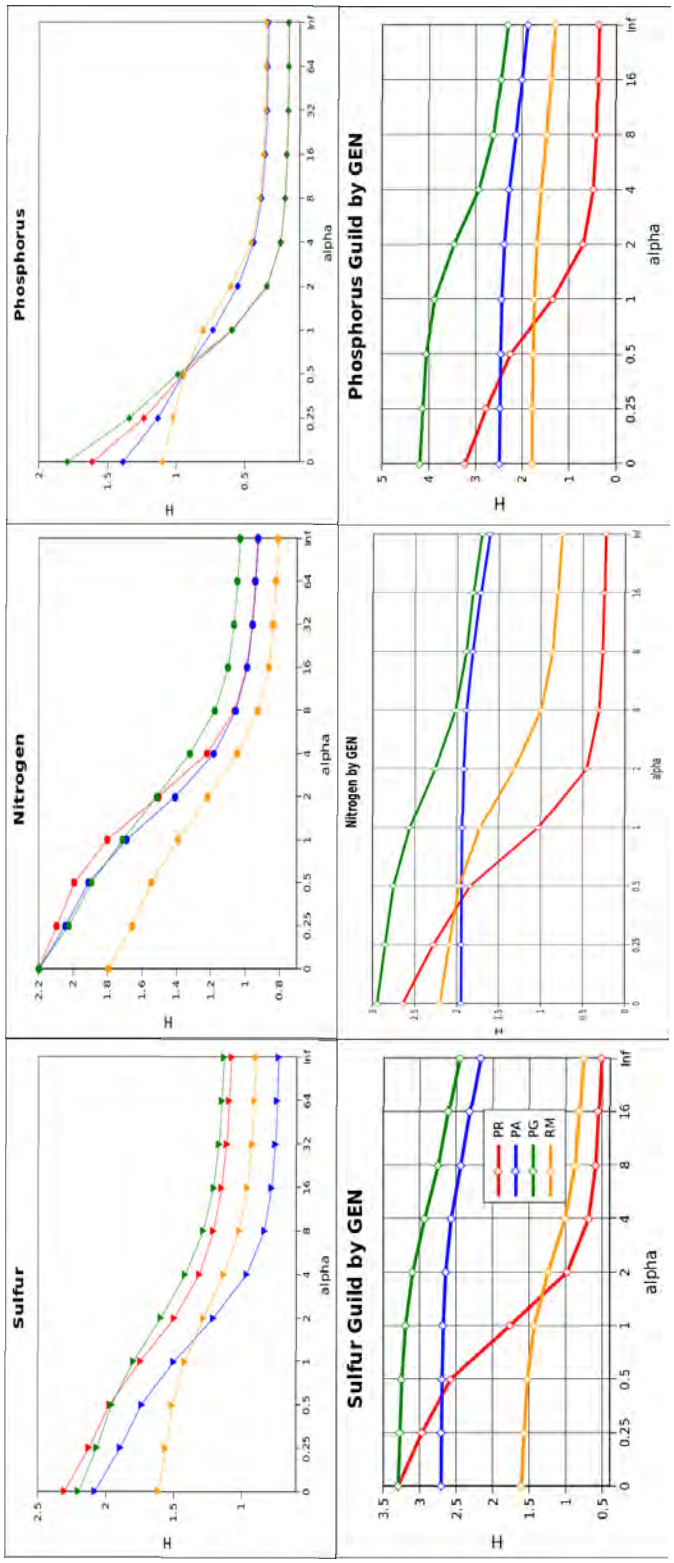


Fig. C. Perfiles de entropía de Rényi para el gremios funcional del ciclo del azufre(izquierda), nitrógeno(centro) y fósforo(derecha) para estimar diversidad taxonómica (renglón superior) y funcional (renglón inferior).

En cuanto al gremio que lleva a cabo el ciclaje de nitrógeno (Fig. C), todas las comunidades tienen un patrón de diversidad funcional relativamente similar, con el estromatolito RM como el menos diverso funcionalmente. A pesar de que la diversidad de categorías funcionales relacionadas con el ciclo del nitrógeno es similar las comunidades PG, PA y PR, las funciones están repartidas por una mayor diversidad de especies en PG que es la que presenta mayor diversidad taxonómica. En PR, el éste gremio funcional presenta una gran dominancia, correspondiente a la versatilidad metabólica del género *Pseudomonas* para llevar a cabo funciones de éste ciclo como se presenta en el Capítulo IV (Peimbert et al., 2012).

En cuanto al ciclo del fósforo, aún cuando se conocen pocas rutas de utilización de fósforo, es notable que el gremio funcional en PG es taxonómicamente más diverso, lo que sugiere que el aprovechamiento de fósforo en la comunidad se encuentra repartido en diferentes especies y no depende de un sólo grupo como en el caso de PR, en donde prácticamente todas las secuencias encontradas están afiliadas al género *Pseudomonas*. Éste tipo de análisis, incorporado a lo que se conoce sobre las características ambientales de las pozas, permiten identificar especies claves en éstas comunidades, pues aparentemente en PR la mayor parte de la asimilación de fósforo depende de un sólo género. Asimismo, ésto sugiere que en comunidades más complejas como PG ésta función está repartida entre más especies y por lo tanto le brinda a la comunidad una mayor resiliencia.

Análisis de Diversidad de Familias de Secuencias Pfam y COG.

El número de familias y COGs observados en cada metagenoma, así como el número total calculado mediante el estimador no paramétrico de Chao se presentan en la Tabla A.

Con las matrices de abundancias de COG y Pfam, se calcularon los perfiles de diversidad de Rényi para analizar los patrones de diversidad de conjuntos de proteínas (Fig. D). En ellos podemos observar que los tapetes PG y PR contienen un mayor número de categorías que los estromatolitos PA y RM, y que en ambos casos la riqueza es mayor en PA que en RM. En cuanto a la dominancia, PG y RM son más equitativos que PA y PR, siendo éste último el que presenta la mayor dominancia en ambos casos. Es interesante que los dos perfiles reconstruidos a partir de sistemas independientes de clasificación produzcan resultados similares a gran escala, y ésto se debe probablemente en que los dos métodos de clasificación se basan en la similitud a nivel de estructura primaria. Las principales diferencias se observan en la dominancia de PR y PA, lo cual cobra sentido si consideramos que las familias de Pfam pueden permitir la presencia de parálogos y por lo tanto una sóla familia de Pfam puede contener varios COGs. Ésto sugiere que cualquiera de los perfiles de la Fig. D son una buena representación de la diversidad de aglomeraciones de proteínas.

Tabla A. Número de familias en Pfam y COGs encontrados

en cada uno de los metagenomas analizados.

	PG	**PR**	**PA**	**RM**
COG	2165	2373	1182	1121
Pfam	2782	2887	1728	1436

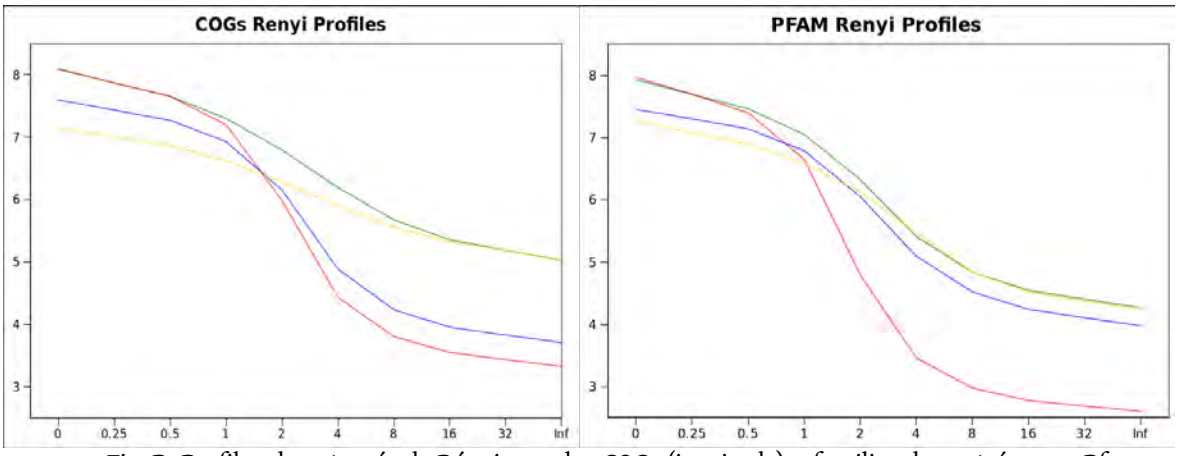


Fig. D. Perfiles de entropía de Rényi para los COGs (izquierda) y familias de proteínas en Pfam (derecha) encontrados en PA (azul), PR (rojo), PG (verde) y RM (amarillo).

Sin embargo, éstos perfiles de diversidad funcional no son totalmente coherentes con los perfiles de diversidad taxonómica presentados en el Capítulo III. Es decir, el tapete PG es el que presenta la mayor riqueza y menor dominancia taxonómica y funcional, lo que sugiere que las comunidades complejas de alta diversidad taxonómica presentan también una alta diversidad funcional. El tapete PR presenta la menor riqueza taxonómica, pero a pesar de que tiene la mayor dominancia funcional, presenta una riqueza funcional equiparable con la de PG. En el caso de RM presenta la mayor dominancia taxonómica, pero tiene una dominancia funcional equiparable con la de PG.

Lo observado es un indicador de que la diversidad funcional dependerá más de la composición taxonómica de la comunidad que de la riqueza o la dominancia. Es decir, el estromatolito RM está fuertemente dominado por un grupo taxonómico, pero no existe dominancia entre las diferentes funciones realizadas por los miembros de dicho grupo. El caso del tapete PR es más claro, pues la dominancia de *Pseudomonas* se traduce en una dominancia metabólica común, pero la plasticidad metabólica inherente al género incrementa enormemente la riqueza funcional con la inclusión de diversas y variadas enzimas que llevan a cabo funciones particulares, probablemente relacionadas con metabolismo secundario (Cap. III). Lo anterior sugiere que no hay una relación general entre la diversidad taxonómica y la funcional, y que dicha relación deberá ser estudiada caso por caso, por que dependerá más de la composición taxonómica de la comunidad y la plasticidad metabólica y funcional de los organismos que la componen.

Análisis de Conglomerados de Familias de Secuencias

En general, el tapete PG resultó consistentemente con el mayor número de conglomerados a todos los niveles de corte, y el tapete PR resultó con el menor número de conglomerados a todos los niveles exceptuando el de 99%. (Tabla B). Dado que los conglomerados tienen a aglutinarse a niveles de corte más bajos (55%), se observa una reducción en el número de conglomerados conforme el nivel de corte se reduce. El estromatolito PA presenta un número más constante de conglomerados a través de todos los niveles de corte, lo que indica que los conglomerados identificados pertenecen a familias funcionales distintas desde el 99% (Tabla B y Fig. E).

Tabla B. Número de conglomerados en cada metagenoma formados a diferentes niveles de corte.

	Total	99%	95%	90%	85%	75%	65%	55%
PG	313104	294410	288567	285551	282667	278219	271164	250499
PR	211271	144685	135580	131396	127489	122845	118542	111675
PA	153046	142126	140786	139857	138761	137206	135271	130250
RM	179143	161840	156261	151408	146092	138479	131054	120164

Al normalizar el número de conglomerados respecto al número total de secuencias iniciales de cada metagenoma (Fig. E), se observa que el tapete PR presenta la menor diversidad de conglomerados, y que un 30% de las secuencias en ese metagenoma se agrupan en conglomerados con valores de corte muy altos (99%). Ésto quiere decir que una gran parte de las secuencias en PR son repeticiones o redundancias genómicas, como se esperaría en un metagenoma dominado por especies de un sólo género.

En contraste, menos de un 10% del total de secuencias en los otros metagenomas se agrupan en conglomerados al mismo nivel de corte. El estromatolito PA y el tapete PG, ambos provenientes de la misma región, presentan el menor cambio entre el número de conglomerados a diferentes niveles de corte, lo que revela que las familias de secuencias que los componen son familias que no están relacionadas entre ellas. Al nivel más bajo de similitud (55%), PA presenta un mayor número de conglomerados relativo al tamaño total que PG, lo que sugiere una mayor diversidad de superfamilias.

Se calcularon los perfiles de entropía de Rényi para evaluar la diversidad de conglomerados en cada uno de los metagenomas a tres niveles de corte representativos (45, 75 y 95%) (Fig. F) . El tapete PG resultó consistentemente el más diverso de todos a los tres niveles de corte, y el tapete PR resultó ser el menos diverso a los tres niveles de corte. El estromatolito RM reveló una mayor riqueza y dominancia que PA al nivel de corte de 95%, pero ésta relación se invirtió al nivel de corte de 45%.

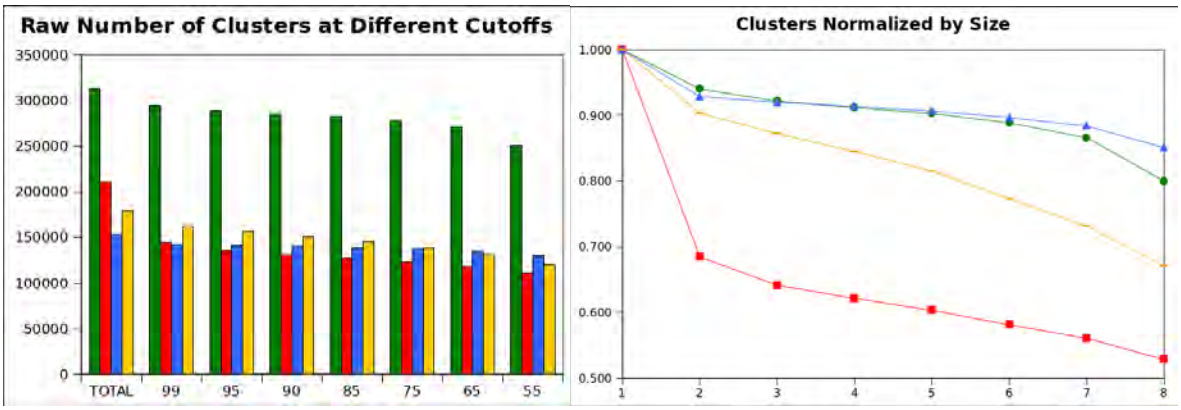


Fig. E. Conglomerados conformados a diferentes niveles de corte en bruto (izquierda) y normalizados (derecha) en PG (verde), PR (rojo), PA (azul) y RM (amarillo).

El perfil de Rényi de los conglomerados a 95% presenta el patrón más similar al perfil de diversidad taxonómica, con PG como el más diverso, seguido por PA que tiene un patrón similar y finalmente RM y PR. Sin embargo, en éste perfil PR tiene una mayor dominancia funcional y una menor riqueza que RM. En contraste, a un corte de 45%, la ordenación de las muestras en escalas de diversidad funcional más relacionadas con la riqueza (alpha = 0-1) corresponden a la ordenación de las muestras en escalas de diversidad taxonómica también relacionadas con riqueza.

Esto sugiere que el número de grupos funcionales genómico se encuentra en parte determinado por el número de especies en la comunidad, pero que las abundancias relativas de las especies de la comunidad no son determinantes de las abundancias relativas de los grupos funcionales. Lo anterior puede ser causado por una redundancia funcional de grupos filogenéticamente distantes o por la diversificación funcional de grupos filogenéticamente coherentes. Por lo tanto, tampoco parece existir una relación directa entre la diversidad taxonómica y la funcional, y probablemente dependa de nuevo de la composición taxonómica de cada comunidad.

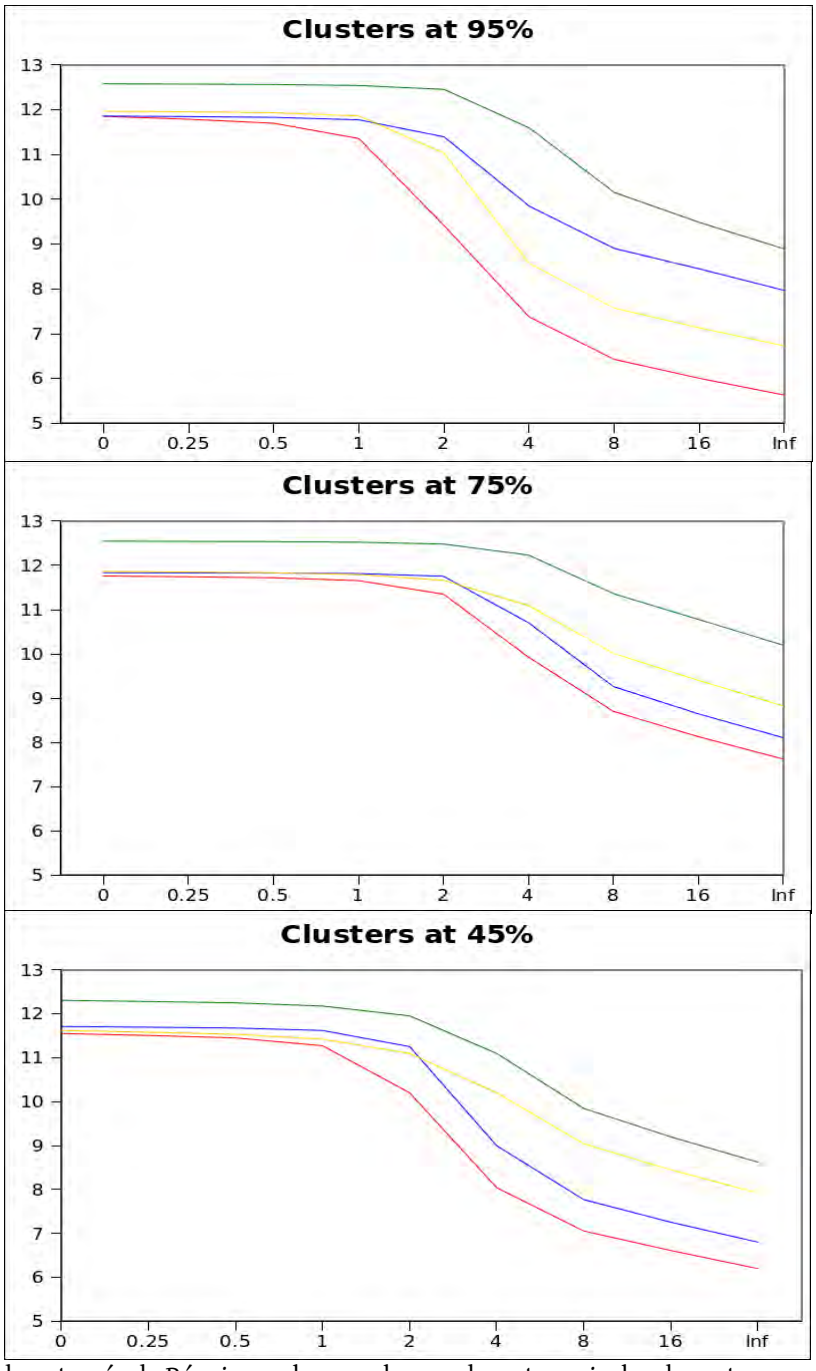


Fig. F. Perfiles de entropía de Rényi para los conglomerados a tres niveles de corte representativos en bruto (izquierda) y normalizados (derecha) en PG (verde), PR (rojo), PA (azul) y RM (amarillo).

Patrones de Similitud de Conglomerados y Diversidad Taxonómica

La comunidad con mayor número de conglomerados únicos o no-compartidos (que no contienen secuencias de ninguna de las otras tres comunidades) fue el tapete microbiano PG, seguido por el estromatolito PA, el estromatolito RM y por último el tapete PR, como la muestra con menor número de conglomerados únicos (Fig. G).

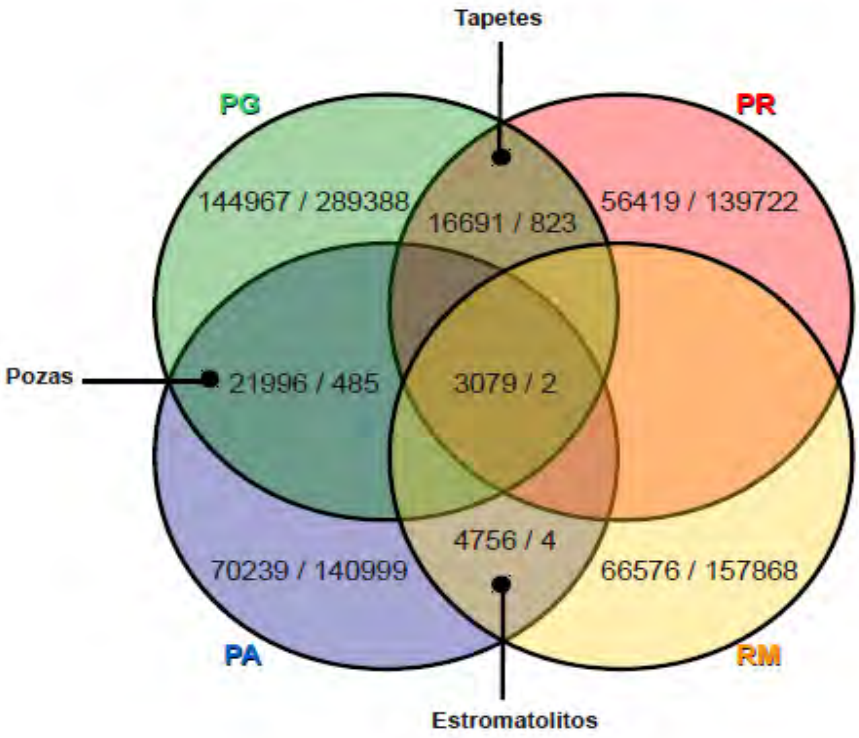
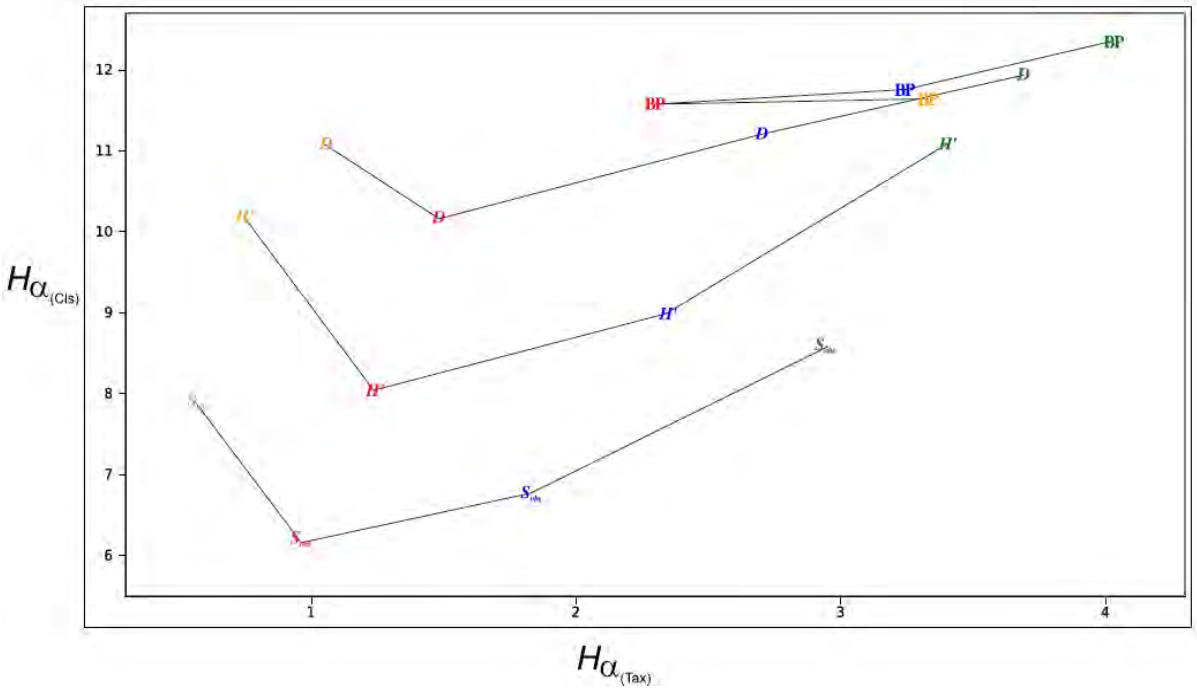


Fig. G. Diagrama de Venn que representa el número de conglomerados únicos y compartidos entre los cuatro metagenomas a 45% (izquierda) y 95% (derecha)

Una proporción muy baja de conglomerados resultó compartida por las cuatro comunidades, y muchos de ellos corresponden a grupos de secuencias de baja complejidad o de elementos móviles relacionados con transposasas. Esto indica una alta diversidad funcional beta, con una gran cantidad de conglomerados presentes, pero la mayoría únicos a cada comunidad local y muy pocos compartidos entre todas, semejante al patrón de diversidad reportado para las comunidades en columna de agua de CCB (Escalante et al. 2008).

En consecuencia, muy pocos conglomerados son compartidos entre pares de secuencias a un nivel de corte de 95%. Los dos tapetes microbianos (PG Y PR) son los que comparten un mayor número de conglomerados a éste nivel, lo que sugiere que comparten pocos organismos estrechamente relacionados entre sí. En contraste, a un nivel de corte de 45% las comunidades que comparten un mayor número de conglomerados son las dos comunidades que se desarrollaron en los ambientes estables de las pozas permanentes: el tapete microbiano PG y el estromatolito PA. Los dos estromatolitos comparten un número comparativamente reducido de conglomerados a los dos niveles de corte.



No existe una correlación estadísticamente significativa (p<0.05) entre la diversidad taxonómica y la funcional, analizada como la correlación entre la entropía de Rényi del mismo orden del número de especies y del número de conglomerados. Ésto probablemente se deba por una parte a que la relación entre éstos dos tipos de diversidad no parece ser lineal (Fig) y a que el número de muestras analizadas sigue siendo muy bajo como para dar resultados confiables. A pesar de ello parece existir una tendencia general a que la diversidad funcional aumente con la diversidad taxonómica. Ésto es evidente en la Fig para los metagenomas PR, PA y PG en cuanto a la riqueza (Sobs), y los índices de Shannon (H'), Simpson (D) y Berger-Parker (BP). Lo anterior sugeriría que el aumento en el número de especies microbianas se traduce en un incremento en el número de funciones codificadas en el genoma. También revelaría que la abundancia relativa de éstas funciones refleja la abundancia relativa de las especies, pues éste patrón se observa a lo largo de todos los órdenes de la entropía de Rényi.

Sin embargo, al incluir el metagenoma de RM éste patrón desaparece, pues RM presenta una mayor diversidad funcional que PR a pesar de que posee una menor diversidad taxonómica. Dado que RM se encuentra dominado por cianobacterias, éste patrón revela también que la relación entre diversidad taxonómica y funcional es también dependiente de la composición particular de cada comunidad.

El contraste de los patrones de diversidad taxonómica y funcional dentro de gremios funcionales conocidos ayuda a la identificación de linajes clave que tienen la exclusividad de ciertas funciones ecológicas, al tiempo que permite observar más finamente el efecto que tienen las diferencias o cambios en la composición filogenética de las comunidades sobre la diversidad funcional de las mismas. Sin embargo, su eficacia está restringida al conocimiento que se tenga sobre los gremios funcionales particulares. La reducción en la diversidad taxonómica total parece afectar diferencialmente a cada gremio funcional, sugiriendo la ausencia de una relación general entre la diversidad taxonómica y la funcional.

Los patrones de similitud entre conglomerados reflejan los patrones de similitud entre comunidades a nivel taxonómico. Es decir, las dos comunidades provenientes de ambientes estables (PG y PA) son más similares funcional y taxonómicamente entre sí que respecto a las demás. Éste patrón resulta de particular interés porque no se observó una mayor similitud en el tipo de sistema, dado que PG es un tapete microbiano y PA un estromatolito.

Éste patrón de similitud puede deberse a muchas causas, como lo son una mayor cercanía geográfica, migración o una mayor similitud en las variables fisicoquímicas en los ambientes de las pozas (temperatura, pH, salinidad, etc.). Desafortunadamente, los datos de éstas variables ambientales en el momento del muestreo de PA y RM no se encuentran disponibles (Breitbart et al., 2009), lo que imposibilita el análisis del componente abiótico. Podemos suponer que la migración se encuentra estrechamente relacionada con la distancia geográfica, aunque es posible que ésto se vea afectado por la compleja interconectividad de los manantiales de cuatrociénegas (Souza et al. 2006). De manera que el factor que podemos probar es la similitud por una menor distancia geográfica. La prueba de Mantel aplicada en el Capítulo I rechazó el aislamiento por distancia en las muestras, sugiriendo que el patrón observado no es un producto de la distancia geográfica entre las muestras. Aunque tampoco es posible descartar sesgos en la amplificación y secuenciación, o el efecto de un diferencial en la longitud promedio de los *reads* secuenciados en las muestras que pueda afectar la clasificación y anotación, la diferencia en la composición taxonómica de al menos PG y PR es apoyada por la observación independiente de las librerías de clonas (Capítulo III, Bonilla-Rosso et al. 2012). El análisis detallado de la composición funcional y taxonómica, junto con la estimación de los rasgos funcionales de los grupos dominantes en cada muestra, apoyan también el patrón observado. Adicionalmente, dado que las pozas que albergan a PG y PA están en el mismo sistema hidrológico (Minckley y Cole, 1968), una mayor similitud en las variables ambientales es altamente probable.

El patrón de similitud observado sugiere que la aparición de los sistemas organosedimentarios laminares de consistencia suave y no-calcificados identificados, o sea, los tapetes microbianos, no depende de la composición taxonómica de la comunidad que lo conforma, y que la diferencia entre los tapetes microbianos y los estromatolitos calcificados se debe más bien a una diferencia en las condiciones fisicoquímicas locales del ambiente que se desarrolla, como lo son la disponibilidad de carbonatos en solución y su solubilidad al pH observado, como proponen Breitbart et al. (2009). Por lo tanto, se espera que los sistemas organosedimentarios aparezcan en interfases agua-sedimento que presenten un gradiente fisicoquímico pronunciado en ambientes que excluyan la presencia de organismos generadores de disturbios, pero la calcificación dependerá de las variables fisicoquímicas mencionadas. Es interesante encontrar una baja proporción de especies compartidas entre las comunidades, lo que sugiere que la emergencia de tapetes microbianos depende de la conjunción de variables ambientales y no de "especies formadoras de tapetes". Sin embargo, encontramos una similitud en la abundancia relativa de grupos taxonómicos profundos, lo que sugiere la conservación de rasgos funcionales filogenéticamente conservados (Philippot et al. 2009).

La semejanza entre los patrones de similitud taxonómico y funcional sugiere también una relación causal entre la composición taxonómica y la funcional, pues revela que las especies similares entre comunidades tienen también un complemento funcional similar. En Cuatrociénegas las comunidades más diversas en cuanto a especies (PG y PA) son las que poseen también una mayor diversidad genómica medida por el número de conglomerados.

Esto es una consecuencia lógica del número y diversidad de genomas representados en la comunidad, pues un mayor número de genomas diferentes se traducirá en un mayor número de genes, proteínas y por lo tanto funciones diferentes. Sin embargo, cada gremio funcional de la comunidad se ve afectado diferencialmente, de manera que la importancia de la diversidad taxonómica para la funcionalidad ecosistémica dependerá de la importancia relativa de cada gremio funcional y la sinergia de éstos en la comunidad.

Por otra parte, la riqueza específica está correlacionada con el número de conglomerados funcionales pero no con la abundancia relativa de éstos, y las comunidades taxonómicamente más similares comparten también un mayor número de conglomerados funcionales. Éstos patrones podría encajar con la "hipótesis de la lotería" (Sale 1977; Munday 2004). Ésta teoría asume que en las comunidades el espacio es un recurso limitante, y que en éste pueden coexistir especies con patrones de uso de recursos similares mediante un proceso de recolonización de espacio disponible por el primer recluta disponible (análogo a ganarse la lotería). Una extensión a ésta teoría ha sido aplicada a comunidades microbianas asociadas a algas, en donde se muestra que los nichos ecológicos son ocupados aleatoriamente a partir de una poza de especies funcional y ecológicamente equivalentes (Burke et al. 2011). En nuestro caso, el proceso de ensamble incluye también rasgos funcionales filogenéticamente conservados y el modelo sería de la siguiente manera:

Un conjunto de nichos ecológicos son abiertos a lo largo del gradiente fisicoquímico que aparece en las interfase sedimento-agua. El gradiente cambiará y se diversificará conforme nuevas especies se establezcan y modifiquen éste gradiente, resultando en la generación e incremento en la complejidad de los nichos disponibles. De esta manera, cada nicho es colonizado aleatoriamente por especies que provienen de una poza de especies capaces de colonizar un mismo nicho ecológico, y que son por lo tanto básicamente ecológicamente equivalentes y funcionalmente similares. A pesar de que a nivel fino, el metabolismo particular de casa especie es único, los miembros de un mismo linaje filogenético comparten rasgos funcionales generales y por lo tanto la poza de especies capaz de colonizar un nicho ecológico constituye un gremio funcional compuesto por un número reducido de linajes filogenéticos con rasgos funcionales conservados. El resultado es que las comunidades que surjan en ambientes fisicoquímicamente análogos y bajo un régimen de perturbación equivalente tendrán pocas especies compartidas (porque la presencia de cada especie particular en cada nicho es producto de un proceso estocástico) pero tendrán gremios funcionales compartidos (porque la capacidad de colonización está determinada por el mismo nicho) y por lo tanto serán más similares en rangos taxonómicos superiores porque los rasgos funcionales fundamentales están conservados a ese nivel. El incremento en diversidad aumenta también la complejidad en las interacciones dentro de la comunidad (Becker et al., 2012), de manera que los nichos abiertos se tornan más específicos, con la conservación de rasgos funcionales compartidos por linajes cohesivos como última consecuencia.

En el caso de los tapetes microbianos, la interfase agua-sedimento representa un nicho compuesto a su vez por muchos micronichos a lo largo del gradiente fisicoquímico característico (Paerl y Yannarell 2010). El establecimiento del tapete dependerá de que el nicho sea ocupado por grupos formadores de tejidos y biopelículas, como lo son las bacterias filamentosas (Cyanobacteria, Thiotrichales) o las generadoras de mucílagos (como *Pseudomonas*). Una vez establecida la estructura física del tapete, el metabolismo inherente a cada miembro de la comunidad modificará y explotará diferencialmente el gradiente, abriendo una multitud de micronichos nuevos. De esta manera, el nicho de fotoautótrofos podrá ser ocupado por ejemplo por la cianobacteria filamentosa *Microcoleus chthonoplastes* como en PR, por cianobacterias unicelulares como en PG o por miembros de Alphaproteobacteria como en RM.

Cabe señalar que los altos niveles de riqueza encontrados dentro de los tapetes microbianos en general, y en gremios funcionales específicos en particular, esbozan la existencia de una coexistencia fuertemente competitiva (Becker et al., 2012). Ésto resulta opuesto al modelo presentado por la hipótesis de la lotería, que asume que una vez colonizado el nicho, la especie no será desplazada competitivamente y por lo tanto favorece la existencia de un efecto de prioridad (Munday 2004). En el caso de los tapetes microbianos puede haber dos explicaciones a la alta diversidad:

La primera es que la conjunción de varios gradientes fisicoquímicos (luz, oxígeno, ácido sulfídrico, pH, materia orgánica; Ley et al. 2006), genera tal variedad de micronichos que difieren apenas ligeramente entre ellos que la diversidad observada dentro de un mismo gremio funcional en realidad es producto de una gran diversidad de organismos, ocupando nichos ecológicamente diferentes en la práctica y por lo tanto los organismos no están en competencia directa entre ellos. La segunda explicación surge de la consideración de la comunidad como un sistema en estado transitivo más que en un estado estable, en donde los gradientes están en cambio constante y éste cambio constante contribuye a la apertura del nicho a nuevas especies aún cuando permanece "ocupado". Ésta explicación cobra validez al considerar que los gradientes fisicoquímicos tienen una fuerte variación diurna (Ley et al., 2006). Bajo éste modelo, la fuerza que define el ensamble sigue siendo la apertura de nuevos nichos, más que la competencia entre las especies que lo ocupan. La última explicación consiste aceptar la existencia de una fuerte exclusión competitiva dentro de cada gremio funcional, y que el resultado de ésta competencia es una especialización y diversificación en el uso particular de recursos, de manera que son las especies las que se abrirían a sí mismas nuevos nichos. Ésta explicación, sin embargo, cobra sentido sólo a una escala evolutiva y no a una escala ecológica. De cualquier forma, nuevos estudios experimentales son necesarios para decidir entre cualquiera de éstas explicaciones.

En resumen, en éste capítulo se exploraron diferentes métodos para medir la diversidad funcional en muestras metagenómicas, y se trató de incorporarlas al análisis de diversidad taxonómica presentado en el Capítulo III (Bonilla-Rosso et al., 2012). El principal problema para hacer éste tipo de comparaciones es definir metodológicamente la diversidad funcional. En éste caso se concluye que para muestras metagenómicas es teóricamente más apropiado utilizar conglomerados producto de una aglomeración jerárquica basada en la similitud en la estructura primaria de las secuencias, porque de ésta forma se evitan sesgos en el conocimiento del metabolismo microbiano.

De cualquier manera, todas las aproximaciones exploradas sugieren una relación entre la diversidad taxonómica y funcional, en donde el incremento en la riqueza y la disminución en la dominancia taxonómica va aunado a un incremento en la diversidad funcional. Ésta es una relación compleja y no lineal, lo que sugiere que en futuros trabajos será necesario desarrollar modelos matemáticos contra los cuales se puedan contrastar los datos. Adicionalmente, la divergencia en el patrón observado en RM sugiere que la relación entre la diversidad taxonómica y funcional es también dependiente de la composición taxonómica de la comunidad. Esto subraya la importancia de incorporar análisis de gradientes en comunidades más similares que permitan reducir los factores de confusión que afectan el cambio en la composición taxonómica y la estructura de las comunidades para obtener conclusiones sólidas sobre su relación con la diversidad funcional.

PERSPECTIVAS Y CONCLUSIONES

Conclusiones

Las técnicas metagenómicas proveen una oportunidad sin precedente para el estudio de patrones globales de diversidad en comunidades complejas, pero su análisis indudablemente presenta formidables retos tanto teóricos como tecnológicos moleculares e informáticos. En esta tesis se exploró la relación entre la diversidad en términos de especies y de funciones de comunidades microbianas a partir de datos metagenómicos de dos tapetes mircobianos de ambientes oligotróficos en el valle de Cuatrociénegas, y se compararon contra otros metagenomas de tapetes microbianos y estromatolitos mexicanos.

En el Capítulo I se exploraron los métodos de estimación de diversidad taxonómica a partir de metagenomas, y se encontró que no es posible hacer comparaciones cuantitativas válidas con las técnicas disponibles en la actualidad, pero que es posible hacer comparaciones cualitativas mediante la incorporación de medidas multidimensionales de diversidad. El sesgo observado es explicado por un efecto de muestreo incompleto, más que por efecto de la metodología empleada en la secuenciación y proceso de los marcadores. También se halló que el uso de marcadores moleculares de proteínas codificantes como indicadores ecológicos en la construcción de matrices de abundancia refleja más fielmente la estructura comunitaria que con SSU-rRNAs.

En el Capítulo II se aplican los métodos analizados en el capítulo anterior al análisis de diversidad taxonómica en los metagenomas de dos tapetes microbianos secuenciados *de novo* y de dos estromatolitos, todos provenientes de ambientes oligotróficos del valle de Cuatrociénegas. Se encontró una marcada diferencia en la composición y estructura, en donde los ambientes más estables poseen las comunidades más diversas y complejas. También se identificó un grupo conservado de organismos en metagenomas de sistemas organosedimentarios, patrón coherente con un modelo de ensamble por lotería (Mundlay 20004; Burke et al. 2011) y con la conservación de rasgos funcionales en niveles taxonómicos profundos.

En los Capítulos III y IV se abordó la caracterización y análisis de la diversidad funcional en éstos metagenomas. Hasta ahora, el conocimiento existente sobre el metabolismo bacteriano está limitado a lo cultivable, limitando nuestra capacidad de realizar análisis comparativos, porque se encontrarán sesgados hacia la representación de nuestro conocimiento en las muestras. Las tres alternativas exploradas en ésta tesis abarcan diferentes escalas de la diversidad funcional: el análisis genecéntrico y de gremios conocidos, ayuda a entender a las comunidades en sus componentes mejor conocidos, por otra parte, el análisis de familias de proteínas abarca un espectro más amplio de las funciones conocidas, y finalmente el análisis de conglomerados permite una comparación del complemento genómico total de una comunidad. Los resultados de las tres aproximaciones son coherentes entre sí y complementarios para el análisis integral de la diversidad funcional.

Los resultados indican que el tapete PG, que está desarrollado en una poza permanente fisicoquímicamente estable, posee una mayor complejidad y diversidad taxonómica y funcional, caracterizada por la abundancia de mecanismos de producción primaria y metabolismos autotróficos aeróbicos. El tapete PR se desarrolla en una poza de desecación fisicoquímicamente variable, y se encuentra fuertemente dominado por varias especies del género *Pseudomonas*. Su metabolismo es predominantemente heterotrófico, y está caracterizado por la presencia de rutas relacionadas con la detoxificación y tolerancia a altas concentraciones de toxinas, iones y metales. Los resultados también

indican que los ambientes oligotróficos con bajas concentraciones de fósforo y nitrógeno pueden sostener comunidades complejas y diversas, y que la estructura del tapete presenta una solución para la concentración de nutrientes y protección de los miembros de la comunidad *per se*.

Al comparar los tapetes microbianos contra los metagenomas de estromatolitos de la misma región, se encontró una mayor similitud entre los sistemas con regímenes de perturbación similares que entre tipos de sistema o proximidad geográfica. Es decir que la mayor similitud observada no ocurre entre los dos metagenomas de tapetes microbianos (PG y PR) o de estromatolitos (PA y RM), sino entre los dos sistemas desarrollados en los ambientes estables de las pozas permanentes (PG y PA). Éstas dos comunidades son las más diversas taxonómica y funcionalmente, es decir con composiciones más complejas, mayor número de *phyla* y especies estimadas, un mayor número de proteínas y de funciones y con genomas más pequeños.

En contraste, aunque PR y RM no son similares entre sí, presentan un menor número de especies y una mayor dominancia, un menor número de proteínas y funciones, una mayor redundancia funcional y sus genomas son más grandes. Éstos resultados sugieren que los ambientes más estables, COMO PG Y PA permiten la diversificación de organismos con genomas pequeños especializados en la explotación de micronichos diversos, al tiempo que los ambientes más variables, PR Y RM promueven la diversificación de organismos generalistas con genomas grandes que les permite una mayor plasticidad y versatilidad metabólica.

Por otra parte, se identificó un patrón de similitud en la composición taxonómica de comunidades tan distanciadas geográfica y ecológicamente como los tapetes microbianos halófilos marinos de Guerrero Negro, Baja California, y los tapetes y estromatolitos de las pozas permanentes oligotróficas continentales de Cuatrociénegas, Coahuila. Curiosamente, la similitud no es a nivel de especies, sino en rangos taxonómicos más altos. Ésto, aunado a la similitud en la composición de conglomerados, es explicado por un modelo de teoría de lotería competitiva para el ensamble de comunidades (Sale 1977; Munday 2004; Burke et al. 2011) combinado con la existencia de rasgos funcionales (Philippot et al. 2009).

Perspectivas

Éste trabajo representa una demostración de concepto de la utilidad de comunidades microbianas para el estudio de patrones de diversidad taxonómica y funcional mediante técnicas metagenómicas y abre la posibilidad de formular diseños experimentales precisos y detallados para el estudio de cada uno de los problemas teóricos en la ecología de comunidades.

El principal resultado subraya la necesidad de explorar los sesgos en las métricas de diversidad estimadas a partir de datos metagenómicos. Aunque se espera que el costo de la secuenciación continúe disminuyendo en los próximos años, resulta esperable que el aumento en el poder de secuenciación se concentre en la multiplicación en el número de muestras de cada estudio más que en aumentar la cobertura individual de cada muestra. Es por eso que es necesario un mayor desarrollo de la teoría detrás de la aplicación de las métricas de diversidad y estudiar a detalle el efecto de los sesgos causados por un muestreo incompleto. El problema de muestreos incompletos y estimación diversidad no es exclusivo de las bases de datos metagenómicas, y ha sido ampliamente estudiado en las matrices ecológicas de comunidades de macroorganismos, por lo que posiblemente sea necesario el desarrollo de nuevos algoritmos específicos para bases de datos metagenómicas.

En el Capítulo I se exploraron los sesgos relacionados con un sólo tipo de plataforma de secuenciación, el pirosecuenciador GS-FLX 454, sin embargo a la fecha ésta tecnología ha sido significativamente mejorada en cuanto a la tasa de errores y la longitud promedio de los reads de secuenciación resultantes. Más aún, nuevas tecnologías de secuenciación han proliferado o nuevas versiones han sido desarrolladas (p.ej. GS-FLX+, Illumina, SOLiD, IonTorrent) y otras más se encuentran en desarrollo (p. ej. PacBio, NanoPore).

Cada versión de plataforma de secuenciación, así como cada procedimiento para evaluar calidad de los reads y llamar los ORFs, genera un modelo de error diferente, y por lo tanto necesita la simulación de nuevos juegos de datos para su evaluación, no sólo para evaluar la tasa de errores y cobertura, sino también para evaluar la fidelidad en la representación de la estructura y composición de las comunidades de donde provienen. Hasta que entendamos bien los sesgos provenientes de las tecnologías, las comparaciones de diversidad entre comunidades y las implicaciones e interpretaciones obtenidas continuarán siendo meramente anecdóticas.

En éste mismo sentido, los análisis cualitativos comparativos de patrones de diversidad no resultaron con un respaldo teórico válido, especialmente cuando se comparan juegos de datos obtenidos con diferentes esfuerzos de colecta, diferentes esfuerzos de secuenciación y plataformas de secuenciación. De acuerdo con los resultados de la simulación, sólo es posible realizar comparaciones cualitativas siempre y cuando las metodologías de obtención de datos sean equiparables. Ésto resulta desalentador para el área de ecología de comunidades bacterianas, pues raras veces se encuentran trabajos cuya metodología sea tan similar que permita la comparación entre ellos, y por lo tanto se recomienda no utilizar ningún tipo de metaanálisis.

Afortunadamente, las comparaciones cualitativas permiten el diseño de experimentos más sutiles, precisos y elegantes para responder preguntas específicas. Por lo tanto, se esperaría un cambio en el paradigma de la aplicación de técnicas metagenómicas. Actualmente, éste ha consistido en la descripción y la caracterización de comunidades naturales, y sus diseños experimentales se han enfocado más en el desarrollo técnico y metodológico de la secuenciación y anotación que en responder hipótesis. Este paradigma debe cambiar hacia la secuenciación de tratamientos con sus respectivas réplicas para responder hipótesis de trabajo particulares, y ésto es posible mediante comparaciones cualitativas entre muestras de un mismo estudio.

Por otra parte, la mayor parte de las métricas de diversidad utilizadas sobre bases de datos metagenómicas son métricas de estimación de diversidad taxonómica, y la diversidad funcional ha sido significativamente menos explorada. Esta tesis explora tres aproximaciones complementarias para medir diversidad funcional (genes, familias y conglomerados), todas ellas basadas en similitud a nivel de secuencia primaria, y extrapolamos las métricas de diversidad de especies que consideran categorías y abundancias relativas porque son fácilmente aplicables a las categorías funcionales que fueron estudiadas.

Sin embargo, el concepto de diversidad funcional ha sido estudiado ampliamente en la ecología de comunidades de macroorganismos, definiendo la diversidad funcional como el rango y distribución de lo que los organismos son capaces de hacer en comunidades y ecosistemas, incorporando implícitamente el concepto de redundancia y complementaridad funcional que no es abordado en éste trabajo (Schleuter et al. 2010).

La extrapolación de éstos índices macroecológicos de diversidad, sin embargo, no es sencilla, puesto que están basados en el estudio de rasgos funcionales que todavía no es posible medir en metagenomas. La línea de investigación consecuencia natural del presente estudio es entonces la exploración de índices de diversidad explícitamente desarrollados para comunidades microbianas, que necesariamente se definirán a partir de bases de datos metagenómicas y por lo tanto deberán estar basadas en el potencial funcional genómico. La ventaja de ésta aproximación es sin duda la incorporación del componente filogenético intrínseco a los datos de secuencias. Los análisis cuantitativos de expresión (transcriptómica) y de desempeño metabólico (metabolómica) prometen ser un excelente complemento para continuar con ésta línea.

Finalmente, dado que ésta tesis propone nuevas posibilidades para el desarrollo de estudios teóricos sobre la diversidad en comunidades microbianas naturales, abre la posibilidad de utilizar las herramientas metagenómicas para el avance de la ecología teórica de comunidades en áreas como el ensamble de comunidades y pruebas sobre la teoría neutral de ecología de comunidades. La perspectiva más interesante es la posibilidad de explorar la liga entre la teoría evolutiva durante el siglo pasado (y de evolución molecular en particular), desarrollada principalmente a nivel de poblaciones ecológicas de una sóla especie o la interacción de algunas pocas especies, y la ecología de comunidades.

Ésta conexión ha sido poco explorada principalmente por la complejidad metodológica y tecnológica en el diseño experimental y análisis de datos necesarios para abordar ésta pregunta. Dado que las bases de datos metagenómicas proveen al mismo tiempo información filogenética y funcional, e incluyen datos sobre las poblaciones y las comunidades, presentan una oportunidad sin precedentes para analizar las dinámicas poblacionales dentro de una comunidad. El objetivo final de éste tipo de estudios es sin duda desarrollar la capacidad predictiva de las dinámicas poblacionales, la riqueza taxonómica y la consecuencia funcional de éstas sobre el desempeño, resistencia y resiliencia de comunidades y ecosistemas.

Literatura Citada

Alcaraz, L.D. et al., 2010. Understanding the evolutionary relationships and major traits of Bacillus through comparative genomics. *BMC Genomics*, 11(1), p.332.

Alcaraz, L.D. et al., 2008. The genome of Bacillus coahuilensis reveals adaptations essential for survival in the relic of an ancient marine environment. *Proceedings of the National Academy of Sciences of the United States of America*, 105(15), p.5803–8.

Allison, S.D. & Martiny, J.B.H., 2008. Colloquium Paper: Resistance, resilience, and redundancy in microbial communities. *Proceedings of the National Academy of Sciences*, 105(Supplement 1), p.11512–11519

Allwood, A.C. et al., 2006. Stromatolite reef from the Early Archaean era of Australia. *Nature*, 441(7094), p714–8

Becker, J. et al., 2012. Increasing antagonistic interactions cause bacterial communities to collapse at high diversity. *Ecology Letters*, 15(5), p.468–474.

Begon, M., Townsend, C.R. & Harper, J.L., 2006. *Ecology: from individuals to ecosystems*, Wiley-Blackwell.

Benson, D.A. et al., 2005. GenBank. *Nucleic Acids Research*, 33(Database Issue), p.D34–D38.

Berger, W.H. & Parker, F.L. 1970. Diversity of planktonic foraminifera in deep-sea sediments. *Science* 168:1345-1347

Betlach, M.R., Tiedje, J.M. & Firestone, R.B., 1981. Assimilatory nitrate uptake in Pseudomonas fluorescens studied using nitrogen-13. *Archives of Microbiology*, 129(2), p.135–140.

Blackburn, T.M. & Gaston, K.J., 1998. Some methodological issues in macroecology. *The American Naturalist*, 151(1), p.68–83.

Bonilla-Rosso G., Souza V. y Eguiarte L.E. 2008. Metagenómica, Genómica y Ecología Molecular: La Nueva Ecología en el Bicentenario de Darwin. *TIP Revista Especializada en Ciencias Químico-Biológicas* 11(1):41-51.

Bonilla-Rosso G., Eguiarte L.E., Romero D.R., Travisano M. y Souza V. 2012a. Understanding microbial community diversity: use of simulated datasets to evaluate the performance of diversity metrics derived from metagenomes. *FEMS Microbiology Ecology* 82(1):37-49

Bonilla-Rosso G., Peimbert M., Alcaraz L.D., Hernandez I., Eguarte L.E., Olmedo G. y Souza V. 2012b. Comparative metagenomics of two microbial mats at Cuatro Ciénegas Basin II: Community Structure and Composition in Oligotrophic Environments. *Astrobiology* 12 (7), 659-67

Braissant, O. et al., 2009. Characteristics and turnover of exopolymeric substances in a hypersaline microbial mat. *FEMS Microbiology Ecology*, 67(2), p.293–307.

Braissant, O. et al., 2007. Exopolymeric substances of sulfate-reducing bacteria: Interactions with calcium at alkaline pH and implication for formation of carbonate minerals. *Geobiology*, 5(4), p.401–411.

Breitbart, M. et al., 2009. Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Ciénegas, Mexico. *Environmental microbiology*, 11(1), p.16–34.

Brown, J.H., 1995. Macroecology, University of Chicago Press. *Chicago, Illinois, USA*.

Burke, C. et al., 2011. Bacterial community assembly based on functional genes rather than species. *Proceedings of the National Academy of Sciences*, 108(34), p.14288–14293.

Cadotte, M.W. et al., 2010. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecology letters*, 13(1), p.96–105.

Calegari, V.R., 1997. *Environmental perceptions and local conservation efforts in Cuatro Ciénegas, Coahuila, Mexico*. University of Texas at Austin.

Canfield, D.E. & Des Marais, D.J., 1993. Biogeochemical cycles of carbon, sulfur, and free oxygen in a microbial mat. *Geochimica et cosmochimica acta*, 57(16), p.3971–3984.

Chao A. 1984. Non-parametric estimation of the number of classes in a population. *Scand J Stat* 11:265-270

Chao, A. & Lee SM. 1992. Estimating the number of classes via sample coverage. *J Am Stat Assoc.* 87:210-217.

Chesson, P.L. & Warner, R.R., 1981. Environmental variability promotes coexistence in lottery competitive systems. *American Naturalist*, p.923–943.

Chodorowski, A., 1959. Ecological differentiation of turbellarians in Harsz-Lake. *Polsk. Arch. Hydrobiol*, 6(19), p.33–75.

Ciccarelli, F.D., et al. 2006. Towards Automatic Reconstruction of a Highly Resolved Tree of Life. *Science* 311(5765:1283-1287

Cohen, Y., 1989. Photosynthesis in cyanobacterial mats and its relation to the sulfur cycle: A model for microbial sulfur interactions. In Y. Cohen & E. Rosenberg, eds. *Microbial Mats. Physiological Ecology of Benthic Microbial Communities*. American Society for Microbiology, pp. 22–36.

Díaz, S. & Cabido, M., 2001. Vive la différence: plant functional diversity matters to ecosystem processes. Trends in Ecology & Evolution, 16(11), p.646–655.

Dobrindt, U. et al., 2004. Genomic islands in pathogenic and environmental microorganisms. Nature Reviews Microbiology, 2(5), p.414–424.

Dornelas, M., Connolly, S.R. & Hughes, T.P., 2006. Coral reef diversity refutes the neutral theory of biodiversity. Nature, 440(7080), p.80–82.

Dumbrell, A.J. et al., 2009. Relative roles of niche and neutral processes in structuring a soil microbial community. The ISME journal, 4(3), p.337–345.

Dyhrman, S.N.., Ammerman, J.W. & Van Mooy, B.A.., 2007. Microbes and the marine phosphorus cycle.

Elser, J.J. et al., 2005. Effects of phosphorus enrichment and grazing snails on modern stromatolitic microbial communities. *Freshwater Biology*, 50(11), p.1808–1825.

Escalante, A.E. et al., 2008. Diversity of aquatic prokaryotic communities in the Cuatro Cienegas basin. *FEMS microbiology ecology*, 65(1), p.50–60.

Faith, D.P., 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1), p.1–10.

Finn, R.D. et al., 2008. The Pfam protein families database. *Nucleic Acids Research*, 36(Database issue), p.D281–D288.

Fonseca, C.R. & Ganade, G., 2001. Species functional redundancy, random extinctions and the stability of ecosystems. *Journal of Ecology*, 89(1), p.118–125.

Foster, J.S. & Mobberley, J.M., 2010. Past, present, and future: microbial mats as models for astrobiological research. In *Microbial Mats: Modern and Ancient Microorganisms in Stratified Systems*. Springer, pp. 563–582.

van Gemerden, H., 1993. Microbial mats: A joint venture. *Marine Geology*, 113(1-2), p.3–25.

Gilbert, J.A. et al., 2009. Potential for phosphonoacetate utilization by marine bacteria in temperate coastal waters. *Environmental microbiology*, 11(1), p.111–125.

Hacker, J. & Carniel, E., 2001. Ecological fitness, genomic islands and bacterial pathogenicity. *EMBO reports*, 2(5), p.376–381.

Harper J & Hawksworth DL. 1996. Preface. In *Biodiversity: Measurement and Estimation*, Hawksworth DL, Ed. First edition. Springer, London, UK.

Hernández-Montes et al. 2008. The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. *Genome Biology* 9:R95

Hill, M., 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2), p.427–432.

Hubbell, S.P., 2001. *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*, Princeton University Press.

Hunter, S. et al., 2012. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research*, 40(Database issue), p.D306–312.

Hurlbert, S.H., 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, 52(4), p.577–586.

Huson, D.H. et al., 2007. MEGAN analysis of metagenomic data. *Genome research*, 17(3), p.377–86.

Hutchinson GE. 1967. A treatise in limnology. Vol. 2. *Introduction to lake biology and the limnoplankton*. Wiley Intersci. Publ., New York.

INE–SEMARNAP. 2000. Programa de manejo del area de proteccion de flora y fauna Cuatrociénegas. Instituto Nacional de Ecología. 166 pp.

Jensen, L.J. et al., 2009. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(Database issue), p.D412–416.

Johannesson, K.H., Cortés, A. & Kilroy, K.C., 2004. Reconnaissance isotopic and hydrochemical study of Cuatro Ciénegas groundwater, Coahuila, México. *Journal of South American Earth Sciences*, 17(2), p.171–180.

Jones, C.M. & Hallin, S., 2010. Ecological and evolutionary factors underlying global and local assembly of denitrifier communities. *ISME J*, 4(5), p.633–641.

Jorgensen, B.B., Cohen, Y. & Revsbech, N.P., 1986. Transition from Anoxygenic to Oxygenic Photosynthesis in a Microcoleus chthonoplastes Cyanobacterial Mat. *Applied and environmental microbiology*, 51(2), p.408–17.

Jorgensen, B.B. & Des Marais, D.J., 1990. The diffusive boundary layer of sediments: Oxygen microgradients over a microbial mat. *Limnology and oceanography*, p.1343–1355.

Jost, L., 2006. Entropy and diversity. *Oikos*, 113(2), p.363–375.

Kanehisa, M. et al., 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database issue), p.D480–484.

Kembel, S.W. et al., 2011. The Phylogenetic Diversity of Metagenomes. *PLoS ONE*, 6(8), p.e23214.

Keylock, C.J., 2005. Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy. *Oikos*, 109(1), p.203–207.

Kimbrel, J.A. et al., 2010. An improved, high-quality draft genome sequence of the Germination-Arrest Factor-producing Pseudomonas fluorescens WH6. *BMC Genomics*, 11, p.522.

Kunin, V. et al., 2008. Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Molecular systems biology*, 4, p.198.

Lauro, F. et al., 2009. The genomic basis of trophic strategy in marine bacteria. *PNAS,* 106(37):15527-33

Leibold, M.A. & McPeek, M.A., 2006. Coexistence of the niche and neutral perspectives in community ecology. *Ecology*, 87(6), p.1399–1410.

Ley, R.E. et al., 2006. Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. *Appl. Environ. Microbiol*, 72, p.3685–3695.

Li, W., 2009. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC bioinformatics*, 10, p.359.

Li, W. & Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or

nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13), p.1658–9.

Lloyd, M. & Ghelardi, R.J., 1964. A Table for Calculating theEquitability'Component of Species Diversity. *The Journal of Animal Ecology*, p.217–225.

MacArthur, R., 1955. Fluctuations of animal populations and a measure of community stability. *ecology*, 36(3), p.533–536.

MacArthur, R.H., 1957. On the relative abundance of bird species. *Proceedings of the National Academy of Sciences of the United States of America*, 43(3), p.293.

MacArthur, R.H., 1965. Patterns of species diversity. *Biological reviews*, 40(4), p.510–533.

Magrane, M. & Consortium, U., 2011. UniProt Knowledgebase: a hub of integrated protein data. *Database*, 2011(0), p.bar009–bar009.

Man, D. et al., 2003. Diversification and spectral tuning in marine proteorhodopsins. *The EMBO Journal*, 22(8), p.1725–1731.

Margalef, R., 1958. Information theory in ecology. *General Systems*, 3, p.36–71.

Martinez, A., Tyson, G.W. & DeLong, E.F., 2010. Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environmental microbiology*, 12(1), p.222–238.

Maurer, B.A., 1999. *Untangling ecological complexity: the macroscopic perspective*, University of Chicago Press.

McGill, B.J. et al., 2006. Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution*, 21(4), p.178–185.

McIntosh, R.P., 1967. An index of diversity and the relation of certain concepts to diversity. *Ecology*, p.392–404.

Meyer, F. et al., 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9, p.386.

Minckley, W.L. 1969, Environments of the Bolson of Cuatro Ciengeas, Coahuila, Mexico, with Special Reference to the Aquatic Biota. Texas Western Press, The University of Texas at El Paso, 65 pp.

Minckley, W.L. & Cole, G.A., 1968. Preliminary limnologic information on waters of the Cuatro Cienegas basin, Coahuila, Mexico. *The Southwestern Naturalist*, 13(4), p.421–431.

Mora, C. et al., 2011. How Many Species Are There on Earth and in the Ocean? *PLoS Biol*, 9(8), p.e1001127.

Mulet, M., Lalucat, J. & García-Valdés, E., 2010. DNA sequence-based analysis of the Pseudomonas species. *Environmental microbiology*, 12(6), p.1513–30.

Munday, P.L., 2004. Competitive coexistence of coral-dwelling fishes: the lottery hypothesis revisited. *Ecology*,

85(3), p.623–628.

Naeem, S. et al., 1999. Biodiversity and Ecosystem Functioning: Maintaining Natural Life Support Processes. *Issues in Ecology*, (4).

Naeem, S. & Li, S., 1997. Biodiversity enhances ecosystem reliability. *Nature*, 390(6659), p.507–509.

Nakagawa, S. & Cuthill, I.C., 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82(4), p.591–605.

Nee, S. & Stone, G., 2003. The end of the beginning for neutral theory. *Trends in Ecology & Evolution*, 18(9), p.433–434.

Newcombe, R.G. & others, 1998. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in medicine*, 17(8), p.873–890.

Norse, E.A. & McManus, R.E., 1980. *Ecology and living resources: Biological Diversity*, Washington,D.C.

Ochman, H. et al., 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784), p.299–304.

Overbeek, R. et al., 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17), p.5691–5702.

Pace, N.R., 1997. A Molecular View of Microbial Diversity and the Biosphere. *Science*, 276(5313), p.734–740.

Paerl, H.W. et al., 2003. Hypersaline Cyanobacterial Mats as Indicators of Elevated Tropical Hurricane Activity and Associated Climate Change. *AMBIO: A Journal of the Human Environment*, 32(2), p.87–90.

Paerl, H.W. & Yannarell, A.C., 2010. Environmental Dynamics, Community Structure and Function in a Hypersaline Microbial Mat. In J. Seckbach & A. Oren, eds. *Microbial Mats: Modern and Ancient Microorganisms in Stratified Systems*. Springer, pp. 423–444.

Paine, R.T., 1966. Food web complexity and species diversity. *American Naturalist*, p.65–75.

Parks, D.H. & Beiko, R.G., 2010. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, 26(6), p.715–721.

Peimbert M, Alcaraz LD, Bonilla-Rosso G, Olmedo G García-Oliva F, Eguiarte LE y Souza V. Comparative metagenomics of two microbial mats at Cuatro Ciénegas Basin I: Ancient Lessons on How to Cope with an Environment under Severe Nutrient Stress. *Astrobiology* 12(7):648-658

Philippot, L. et al., 2001. Characterization and transcriptional analysis of Pseudomonas fluorescens denitrifying clusters containing the nar, nir, nor and nos genes. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1517(3), p.436–440.

Philippot, L. et al., 2009. Spatial patterns of bacterial taxa in nature reflect ecological traits of deep branches of

the 16S rRNA bacterial tree. *Environmental microbiology*, 11(12), p.3096–3104.

Pianka, E.R., 1966. Latitudinal gradients in species diversity: a review of concepts. *American Naturalist*, p.33–46.

Pielou, E.C., 1966. Shannon's formula as a measure of specific diversity: its use and misuse. *The American Naturalist*, 100(914), p.463–465.

Pielou, E.C., 1967. The use of information theory in the study of the diversity of biological populations. In *Proc. Fifth Berkeley Symposium on Math. Stat. and Prob*. p. 163–77.

Preston, F., 1948. The commonness, and rarity, of species. *Ecology*, 29(3), p.254–283.

Raes, J. et al., 2007. Prediction of effective genome size in metagenomic samples. *Genome biology*, 8(1), p.R10.

Rényi, A., 1961. On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*. pp. 547–561.

Revsbech, N.P. et al., 1983. Microelectrode Studies of the Photosynthesis and O_2, H_2S, and pH Profiles of a Microbial Mat. *Limnology and Oceanography*, p.1062–1074.

Ricklefs, R.E., 2006. The Unified Neutral Theory of Biodiversity: Do Numbers Add Up? *Ecology*, 87(6), p.1424–1431.

Ricotta, C., 2003. On parametric evenness measures. *Journal of Theoretical Biology*, 222(2), p.189–197.

Root, R.B., 1967. The Niche Exploitation Pattern of the Blue-Gray Gnatcatcher. *Ecological Monographs*, 37(4), p.317–350.

Rusch, D.B. et al., 2007. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol*, 5(3), p.e77.

Sale, P.F., 1977. Maintenance of high diversity in coral reef fish communities. *American Naturalist*, p.337–359.

Shannon, CE. 1948. A mathematical theory of communication. *AT&T TECH Journal* 27:379-423/623-653

Sharpton, T.J. et al., 2011. PhylOTU: A High-Throughput Procedure Quantifies Microbial Community Diversity and Resolves Novel Taxa from Metagenomic Data. *PLoS Comput Biol*, 7(1), p.e1001061.

Shen, K. et al., 2006. Extensive genomic plasticity in Pseudomonas aeruginosa revealed by identification and distribution studies of novel genes among clinical isolates. *Infection and immunity*, 74(9), p.5272–5283.

Schleuter, D., Daufresne, M., Massol, F. & Argillier, C. 2010. A user's guide to functional diversity indices. *Ecological Monographs* 80(3):469-484

Silby, M.W. et al., 2009. Genomic and genetic analyses of diversity and plant interactions of Pseudomonas fluorescens. *Genome biology*, 10(5), p.R51.

Simberloff, D. & Dayan, T., 1991. The Guild Concept and the Structure of Ecological Communities. *Annual Review of Ecology and Systematics*, 22(1), p.115–143.

Simpson, E.H., 1949. Measurement of Diversity. *Nature*, 163, p.688–688.

Souza, V. et al., 2006. An endangered oasis of aquatic microbial biodiversity in the Chihuahuan desert. *Proceedings of the National Academy of Sciences of the United States of America*, 103(17), p.6565–70.

Storey, J.D., Taylor, J.E. & Siegmund, D., 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1), p.187–205.

Tatusov, R.L. et al., 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4, p.41.

Tilman, D., Lehman, C.L. & Thomson, K.T., 1997. Plant Diversity and Ecosystem Productivity: Theoretical Considerations. *Proceedings of the National Academy of Sciences*, 94(5), p.1857–1861.

Tóthmérész, B., 1995. Comparison of different methods for diversity ordering. *Journal of Vegetation Science*, 6(2), p.283–290.

United Nations Environmental Programme, 1992. Convention on Biological Diversity, June 1992.

Venter, J.C. et al., 2004. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667), p.66 –74.

Visscher, P.T., Gritzer, R.F. & Leadbetter, E.R., 1999. Low-Molecular-Weight Sulfonates, a Major Substrate for Sulfate Reducers in Marine Microbial Mats. *Applied and Environmental Microbiology*, 65(8), p.3272–3278.

Visscher, P.T., Prins, R.A. & Gemerden, H., 1992. Rates of sulfate reduction and thiosulfate consumption in a marine microbial mat. *FEMS Microbiology Ecology*, 9(4), p.283–294.

Webb, C.O. et al., 2002. Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, (33), p.475–505.

White, A.K. & Metcalf, W.W., 2004. Two C—P Lyase Operons in Pseudomonas Stutzeri and Their Roles in the Oxidation of Phosphonates, Phosphite, and Hypophosphite. *Journal of Bacteriology*, 186(14), p.4730–4739.

Whittaker, R.H., 1965. Dominance and diversity in land plant communities. *Science*, 147(3655), p.250.

Wilson, J.B., 1999. Guilds, functional types and ecological groups. *Oikos*, 86(3), p.507–522.

Wohl, D.L., Arora, S. & Gladstone, J.R., 2004. Functional Redundancy Supports Biodiversity and Ecosystem Function ina a Closed and Constant Environment. *Ecology*, 85(6), p.1534–1540.

Wolaver, B.D. et al., 2008. Delineation of Regional Arid Karstic Aquifers: An Integrative Data Approach. *Ground Water*, 46(3), p.396–413.

Wootton, J.T., 2005. Field parameterization and experimental test of the neutral theory of biodiversity. *Nature*, 433(7023), p.309–312.

Wu, M. & Eisen, J. a, 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome biology*, 9(10), p.R151.

Yang, H. et al., 2005. Niche heterogeneity determines bacterial community structure in the termite gut (Reticulitermes santonensis). *Environmental microbiology*, 7(7), p.916–932.

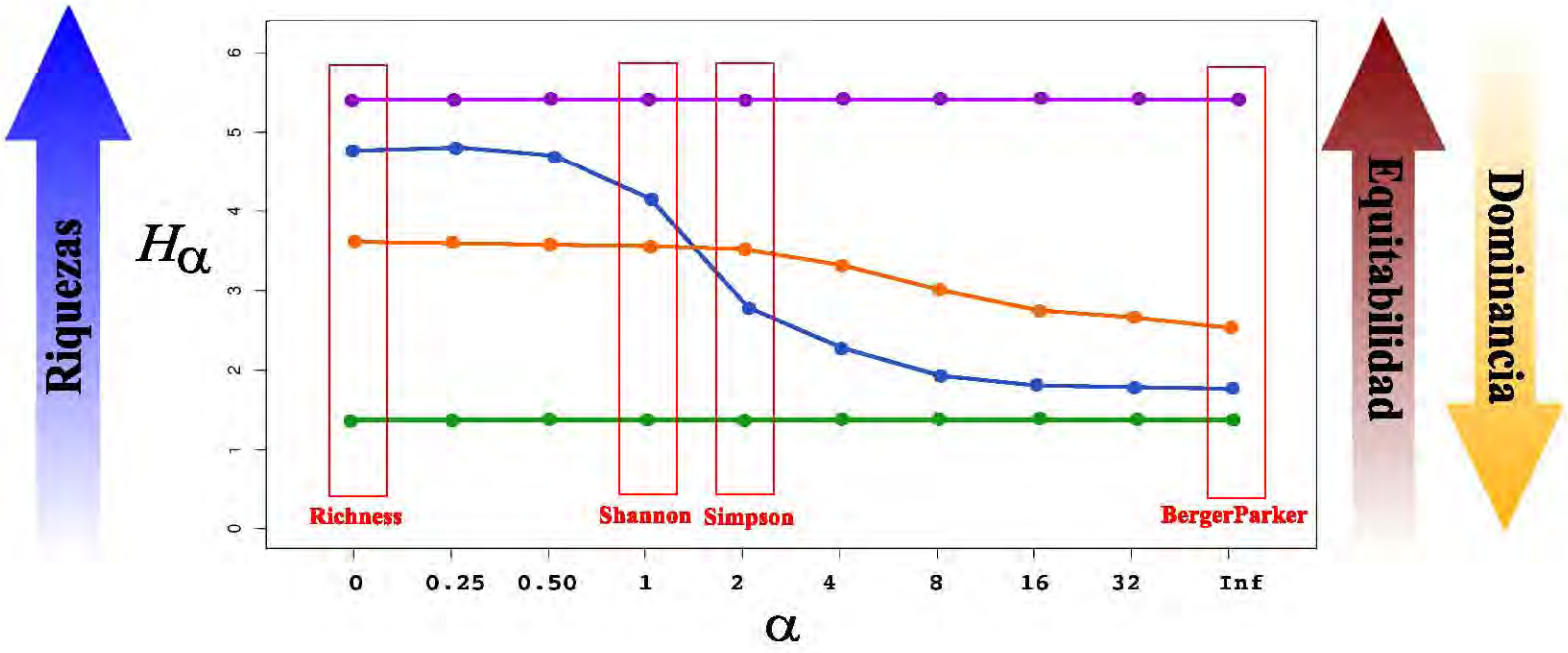
Yannarell, A.C., Steppe, T.F. & Paerl, H.W., 2007. Disturbance and recovery of microbial community structure and function following Hurricane Frances. *Environmental microbiology*, 9(3), p.576–583.

Yooseph, S. et al., 2007. The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS Biol*, 5(3), p.e16.

Apéndice I.
Métricas de Diversidad Utilizadas y su Interpretación

Métrica	Ecuación	Términos	Significado
Chao	$E_{(S)}=S_o+\frac{f_1^2}{2f_2}$	S_o = Número de especies observadas f_1 = Número de especies observadas una vez f_2 = Número de especies observadas dos veces	Límite inferior del número total de especies esperadas en una muestra
Rarefacción	$E_{(S)}=\sum_{i=1}^{S}(1-q_i)$ $q_i=\frac{\binom{N-x_i}{n}}{\binom{N}{n}}$	N = Número total de individuos n = Número de individuos en la muestra x_i = individuos de la especie i q_i = Probabilidad de que la especie xi no se observe en un tamaño n	Número total de especies esperadas en función de los individuos muestreados
Shannon	$H=-\sum_{i=1}^{S} p_i \ln(p_i)$	p_i = Número de individuos en la especie i S = Número total de especies	Entropía o información de una muestra a partir del número de especies y su abundancia
Simpson	$D=\sum_{i=1}^{S} p_i^2$	p_i = Número de individuos en la especie i S = Número total de especies	Probabilidad de que dos individuos muestreados aleatoriamente sean de la misma especie
Perfil de Rényi	$H_\alpha=\frac{\ln(\sum_{i=1}^{s} p_i^\alpha)}{1-\alpha}$	p_i = Número de individuos en la especie i S = Número total de especies α = Parámetro de escala	Familia de cuantificadores de diversidad que ponderan la riqueza y abundancia en función de un parámetro de escala

Apéndice II.
Guía para la Interpretación de los Perfiles de Rényi



Los Perfiles de Entropía de Rényi fueron desarrollados en el campo de la teoría de la información como una generalización del índice de información de Shannon siguiendo un parámetro de escala (alfa). Este parámetro de escala es una medida variable que otorga una ponderación decreciente al número de categorías (especies) y una ponderación creciente a las abundancias relativas (individuos dentro de éstas especies). Algunos momentos particulares de esta función convergen con ciertos índices clásicos de diversidad: una medida de la riqueza específica es obtenida cuando alfa=0; el índice de dominancia de Simpson se obtiene cuando alfa=2; y la dominancia máxima del índice de Berger-Parker se obtiene cuando alfa tiende a infinito. Un caso especial ocurre cuando alfa=1 porque no es posible resolver con una solución simple. El límite de la función de Rényi cuando alfa tiende a 1 es lo que conocemos como el índice de Shannon, y como tal es un momento particular de la ecuación en donde las categorías y las abundancias relativas tienen exactamente el mismo peso en la ponderación.

En la representación gráfica, mientras más arriba se encuentre el perfil, más diversa es la comunidad a la que pertenece. Asimismo, los valores altos a la izquierda del perfil (debajo de alfa=1) revelan una mayor riqueza específica, mientras que los valores altos a la derecha del perfil revela una mayor equitabilidad y menor dominancia. Una comunidad es entonces más diversa que otra si todos y cada uno de los puntos del perfil se encuentran por encima del perfil de la otra (perfil púrpura en la figura), y será menos diversa si todos sus puntos ocurren por debajo de los de la otra (perfil verde). Cuando los perfiles se intersectan, no pueden ser ordenados en cuanto a su diversidad total, pues uno de ellos tendrá una mayor riqueza específica (perfil azul) mientras que el otro tendrá menos especies pero con mayor equitabilidad (perfil naranja) que su intersecto, es decir que tendrá menos riqueza pero sus individuos estarán más homogéneamente distribuidos entre todas las especies.