

Diversidad y estructura de una comunidad microbiana

Silvia G. Acinas

Departamento de Biología Marina y Oceanografía.
Instituto de Ciencias del Mar, CMIMA-CSIC, Barcelona
sacinas@cmima.csic.es

El resurgir de la Ecología Microbiana

Estamos viviendo un momento privilegiado para la Ecología Microbiana, ya que gracias a la integración de múltiples técnicas procedentes de diversas disciplinas científicas (desde la Biología Molecular hasta la Genómica y la Bioinformática), parece resurgir de entre sus cenizas como el *ave fénix*.

Hoy en día, el estudio de la diversidad microbiana se ha convertido en una de las líneas de investigación más relevantes en el área de la Ecología Microbiana, no sólo por su papel en el conocimiento de la función, estructura y evolución de las poblaciones que componen una comunidad microbiana, sino como una fuente importante de investigación médica y biotecnológica. Durante los últimos cinco años, ha sido factible la inclusión de técnicas de metagenómica a gran escala basadas en el análisis de fragmentos de genomas de microorganismos de ambientes naturales, o el uso de metodologías alternativas de secuenciación como la pirosecuenciación (*pyrosequencing*) (tecnología 454 de Roche) para el estudio de la diversidad en comunidades microbianas naturales. Científicos como O. Bejá, E. F. DeLong o J. C. Venter fueron pioneros en el uso de distintas técnicas metagenómicas para el estudio de comunidades microbianas marinas. Fue Craig Venter en el año 2004 mediante la técnica metagenómica *whole-genome shotgun sequencing* (WGS) basada en la secuenciación al azar de los genomas de una comunidad microbiana, quien hizo posible la secuenciación de más de un millón de genes partir de una muestra del bacterioplancton del Mar de los Sargazos detectando más de 1800 especies o filotipos (secuencias distintas del gen 16S rRNA) (Venter *et al*, 2004). Posteriormente, el proyecto *Global Ocean Sampling* liderado también por Venter ha llevado a cabo la secuenciación masiva de 7,7 millones de genes de genomas bacterianos procedente de 41 muestras de diversos ambientes acuáticos a través de 8000 kilómetros de distancia (Rusch *et al*, 2007). Dicha metodología ha permitido la identificación de más de un millón de nuevos ORFs (*Open Reading Frames*), muchos de ellos con la posibilidad de codificar nuevas enzimas.

Recientemente, también se ha utilizando la tecnología de pirosecuenciación para explorar la estructura de una comunidad microbiana marina en profundidad (Sogin *et al*, 2006) y se han podido secuenciar más de 900000 *tags* (pequeños fragmentos de alrededor de 100 bp de los genes 16S rRNA) de dos comunidades microbianas bentónicas cercanas a una fisura volcánica que se encuentra a más de 1500 m de profundidad (Huber *et al*, 2007).

A pesar del gran avance en distintos aspectos de la Ecología Microbiana, todavía quedan multitud de incógnitas por resolver que harán de esta disciplina una de las más fascinantes de la Microbiología. Aprovechando la oportunidad que se me ofrece aquí, detallaré algunos acontecimientos que durante cinco años de estancia postdoctoral en el laboratorio del Dr. Martin F. Polz en *Massachusetts Institute of Technology* (MIT) de Boston (entre 2001 y 2006), fueron vitales para intentar responder algunas incógnitas sobre la diversidad y estructura de comunidades microbianas en ambientes naturales.

¿Es real toda la diversidad microbiana que observamos?

Para poder responder a esta pregunta, primero es necesario recordar cómo se determina la diversidad microbiana en una muestra ambiental. Allá por el año 2001, cuando todavía Craig Venter no surcaba el mar de los Sargazos en busca de un millón de genes y no disponíamos del “*shotgun technology*” para tal fin, ni tampoco teníamos acceso a la revolucionaria tecnología de pirosecuenciación, *¿cómo se determinaba la diversidad microbiana con técnicas moleculares?* En la actualidad, el acceso a dichas tecnologías es restringido y por lo tanto la mayoría de los laboratorios todavía utilizan la misma estrategia para determinar la diversidad microbiana.

Dicha estrategia se basa en la amplificación por PCR de determinados marcadores moleculares como el gen 16S rRNA, a partir del DNA extraído directamente de los microorganismos presentes en una comunidad microbiana. Posteriormente, se genera una genoteca y se secuencian un número

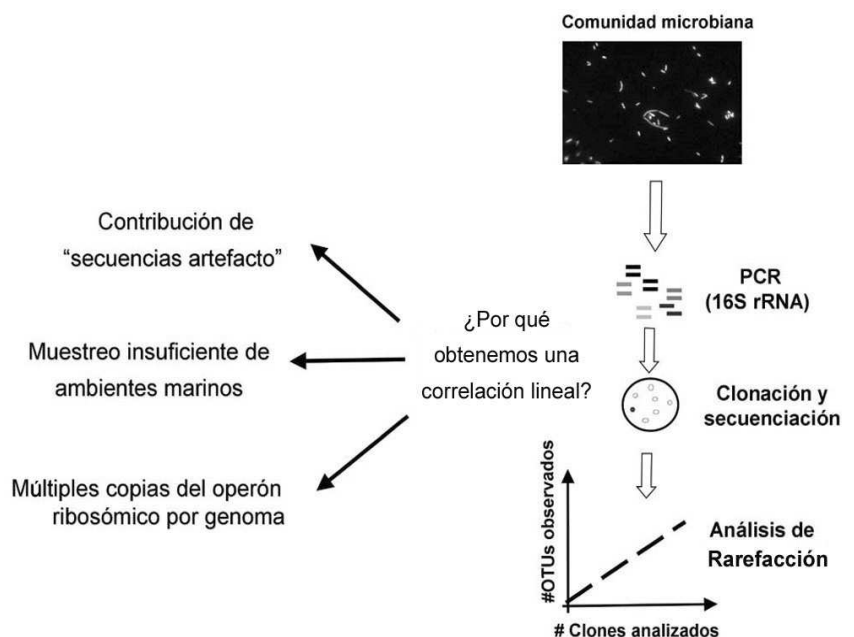
determinado de clones para su análisis. Obviamente, cuantos más clones se secuencian más probabilidades se tiene de recuperar secuencias representativas de los microorganismos existentes (Figura 1). Una de las maneras de analizar el “esfuerzo de muestreo” y por tanto saber si el número de secuencias obtenidas representan una fracción significativa o no de una muestra natural, es por medio de análisis estadísticos como las curvas de rarefacción (el significado original de *rarefaction* en inglés es la reducción de la densidad de una sustancia y ese significado tiene también en español). Los análisis de rarefacción permiten: (i) comparar la diversidad entre diferentes muestras que difieren en tamaño, entendiendo por tamaño el número de clones secuenciados y (ii) extrapolar la diversidad calculando el número hipotético de especies presente en una muestra natural a través de estimadores de riqueza como *Chao-1* (no paramétrico). En estos análisis, el eje de abscisas (X) viene representado por el número de clones analizados, y el eje de ordenadas (Y) por el número de secuencias distintas del gen analizado, en nuestro caso del gen ribosómico 16S rRNA [también se puede expresar como número de filotipos, OTUs (*operational taxonomic units*) o ribotipos pues no existe un consenso de nomenclatura al respecto] proporcionando así una idea de la diversidad existente.

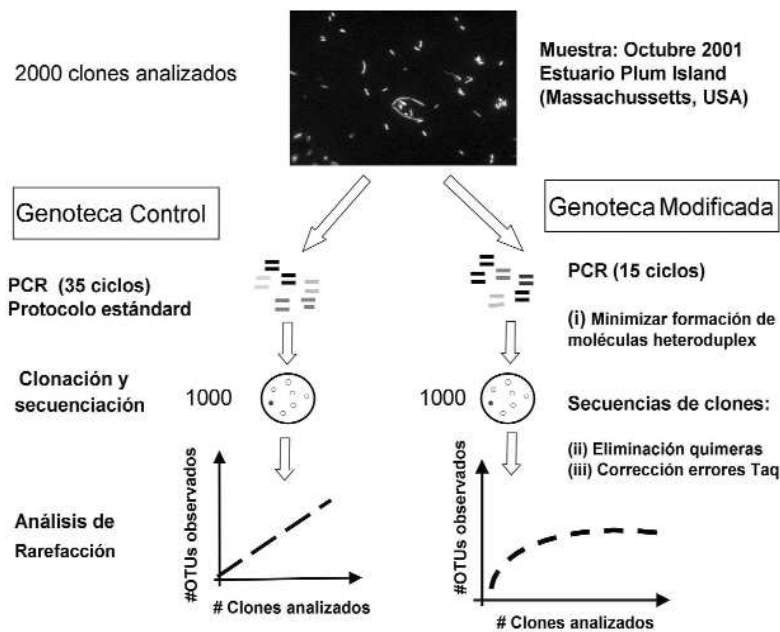
Desafortunadamente, un patrón común que se detecta en muchos de los estudios de diversidad microbiana que utilizan esta estrategia es que el número de clones que se requiere es casi “infinito” pues se observa una correlación lineal entre el número de clones analizados y el número de OTUs



Silvia G. Acinas se licenció en Biología por la Universidad de Murcia (UM) en 1994 y es Doctora por la Universidad Miguel-Hernández (UMH) de Alicante desde 1999. Fue Premio Extraordinario de doctorado por su tesis: “*Diversidad procariótica en ambientes marinos e hipersalinos*” en el Departamento de Microbiología bajo la dirección del Dr. Francisco Rodríguez Valera. Durante 7 años, desde 1999 hasta el 2006, realizó varias estancias postdoctorales en distintos laboratorios de EEUU y Europa. Gracias a una beca postdoctoral del MEC se incorporó al Departamento del Dr. Martin F. Polz en *Massachusetts Institute of Technology (MIT)* en Boston (EEUU) donde investigó durante 5 años distintos aspectos de la estructura, función y evolución de poblaciones bacterianas en comunidades microbianas marinas y desarrollando nuevas técnicas metagenómicas. Después de su experiencia postdoctoral en EEUU, en el 2006 fue contratada en el Departamento de Microbiología del *NIOO-KNAW*, Holanda, dirigido por el Dr. Lucas Stal donde trabajó sobre la diversidad de picocianobacterias y *Pseudanabaena* como modelo para estudiar la microdiversidad genética en un grupo concreto de microorganismos. Posteriormente, tras conseguir un contrato como investigadora mediante el *Programa Ramón y Cajal* por el área de Biología Molecular, Celular y Genética, se incorporó en el 2007 en el Departamento de Biología Marina y Oceanografía en el Instituto de Ciencias del Mar, en el CMIMA-CSIC de Barcelona. Sus principales líneas de investigación se enfocan hacia el uso de técnicas moleculares y genómicas para el estudio de: (i) diversidad, estructura y evolución de comunidades microbianas en ambientes naturales, y (ii) relevancia y mecanismos que generan microdiversidad genética en poblaciones bacterianas.

Figura 1. Esquema que representa las distintas etapas necesarias para estimar la diversidad de una comunidad microbiana a través de genotecas ambientales y los problemas asociados a las mismas que influyen en la estimación de la diversidad microbiana.





observados (Figura 1 y 2). ¿Por qué? ¿Es realmente la diversidad microbiana tan elevada que no podemos estimarla con las técnicas actuales? Existen múltiples factores que afectan a la estimación de la diversidad microbiana y que hay que tener en cuenta como: (i) la contribución de secuencias generadas por artefactos de la PCR (ii) el muestreo insuficiente de comunidades microbianas de ambientes naturales y (iii) las múltiples copias del gen ribosómico presentes en los genomas bacterianos. Indagar estas cuestiones era crucial para poder realizar estimaciones realistas sobre la diversidad microbiana y para analizar en detalle la estructura de una comunidad microbiana.

Tuve la suerte durante mi estancia en EEUU en el laboratorio de Dr. Polz de estar en el sitio adecuado en el momento idóneo, con elevados recursos, sí, pero más importante, con muchas ganas de meternos en un buen "berenjenal" que requería muchas ganas, esfuerzo, tiempo y, por supuesto, dinero. La estrategia fue diseñar dos genotecas del 16S rRNA de elevado tamaño (1000 clones cada una) de una misma muestra ambiental (Figura 2) para (i) estimar cuántos genomas bacterianos (especies) coexisten en una misma comunidad microbiana (diversidad) y (ii) analizar en detalle cuál era la organización a nivel filogenético de dichos genomas (estructura). Nuestro proyecto fue pionero en su momento (hablamos de fechas anteriores a 2004) por varios motivos:

(i) Supuso el mayor esfuerzo de muestreo realizado hasta ese momento con la secuenciación de unos 2000 clones de un mismo gen y de una misma muestra ambiental.

Figura 2. Diseño de dos genotecas del 16S rRNA (control y modificada) a partir de una misma comunidad microbiana marina utilizando distintos protocolos de amplificación de la PCR. En la genoteca "control" se utilizó un protocolo estándar de amplificación con 35 ciclos, mientras que en la genoteca "modificada" se llevó a cabo con 18 ciclos y otros protocolos específicos para minimizar la formación y contribución de secuencias artefactos generados durante el transcurso de la PCR. El objetivo era poder comparar ambas genotecas para (i) estimar la contribución dichos artefactos en la estimación de la diversidad microbiana y (ii) analizar en detalle la estructura de una comunidad microbiana. De cada genoteca (control y modificada) se secuenciaron y analizaron alrededor de 1000 clones.

(ii) Utilizamos diferentes métodos para minimizar y cuantificar la contribución de los artefactos en las estimaciones de la diversidad microbiana.

(iii) Determinamos la proporción de diversidad que era debida a múltiples copias del gen ribosómico.

Este trabajo permitió por primera vez explorar en detalle la diversidad y estructura de una comunidad microbiana y esa aportación al campo fue reconocida con la publicación de un artículo en la revista *Nature* en julio de 2004 (Acinas et al, 2004b).

Efecto de las secuencias artefacto de la PCR en la estimación de la diversidad microbiana.

Hoy en día existen más de 451545 secuencias del gen 16S rRNA depositadas en la base de datos del *Ribosomal Database Project II* (<http://rdp.cme.msu.edu/>) actualizada a fecha de 8 de noviembre de 2007. A pesar del constante incremento en el número de secuencias depositadas, todavía es una incógnita cuantas de ellas representan secuencias erróneas debido a los diferentes artefactos generados durante la PCR. Existen tres tipos de artefactos que se pueden generar durante el transcurso de la PCR:

(a) la formación de quimeras: representan productos de PCR incompletos que pueden hibridar como oligonucleótidos (*primers*) en secuencias heterólogas,

(b) la formación de moléculas heterodúplex: se suelen producir en los últimos ciclos de la PCR cuando puede llegar a existir una limitación en la concentración de oligonucleótidos en

la reacción de amplificación. En estos casos, se puede producir una hibridación cruzada entre secuencias heterólogas formando moléculas heterodúplex,

(c) errores por la polimerasa Taq: introducción de una base errónea, oscilando la tasa de error entre 10^{-4} - 10^{-5} dependiendo de la enzima que se utilice.

En función de estas premisas cabría preguntarse: *¿estamos sobreestimando la diversidad microbiana debido a estos artefactos de la PCR? o ¿Hasta qué nivel estos artefactos perturban la estructura de una comunidad microbiana?*

Aunque existen múltiples estudios de laboratorio que han analizado la contribución de los artefactos de la PCR sobre una población mixta de DNA (4 ó 5 diferentes tipos de DNA), todavía no existía ninguno que intentara trasladar esta problemática a una comunidad microbiana de una muestra natural con la existencia de cientos o miles de tipos diferentes de DNA. Para resolver estas preguntas, diseñamos dos genotecas del 16S rRNA de una misma muestra ambiental utilizando dos estrategias diferentes (Figura 2): la primera genoteca que llamamos “control” se diseñó utilizando un protocolo estándar de amplificación (35 ciclos) utilizado en la mayoría de los laboratorios, mientras la segunda, la genoteca “modificada” se llevó a cabo utilizando protocolos especiales para minimizar la contribución de dichos artefactos de la PCR. Así, por ejemplo, se utilizaron sólo 15 ciclos de amplificación para minimizar la formación de quimeras y moléculas heterodúplex. Seguidamente, se realizó un protocolo extra llamado *PCR reconditioning*, el cual consistía en 3 ó 4 ciclos de amplificación añadiendo oligonucleótidos y polimerasa Taq polimerasa fresca. Con este protocolo se ha comprobado que la formación de moléculas heterodúplex en un modelo de laboratorio se reduce hasta un nivel casi indetectable (1%) (Thompson *et al*, 2002). Finalmente detectamos y corregimos posibles errores producidos por la polimerasa Taq. En los artículos que publicamos al respecto (Acinas *et al*, 2004b; Acinas *et al*, 2005) se pueden ver los protocolos más detallados de esta estrategia e incluyen, además, recomendaciones para generar genotecas minimizando la contribución de los artefactos durante la PCR (Acinas *et al*, 2005).

De la comparación de ambas genotecas se pudo concluir que utilizando el protocolo estándar con 35 ciclos de amplificación estábamos sobreestimando la diversidad presente. Esto fue evidente al comparar las dos genotecas ya que el número de:

(i) Secuencias quiméricas detectadas con diferentes herramientas bioinformáticas tenía

una incidencia de un 13% en la genoteca control, mientras que en la genoteca modificada esa cifra disminuyó al 3%, es decir, 4 veces menos (Acinas *et al*, 2005),

(ii) Secuencias distintas del gen 16S rRNA observadas (OTUs, filotipos o ribotipos) disminuyó significativamente, pasando de un 76% en el control a un 48% en la genoteca modificada. Igualmente el número de especies estimadas por Chao-1 en la genoteca control fue el doble pasando de 3381 a 1633 en la genoteca modificada (Figura 3A).

Estos resultados demostraron que es imprescindible tomar precauciones a la hora de generar genotecas si el objetivo es realizar estudios de diversidad y analizar la estructura de una comunidad microbiana, ya que de otra manera estaremos generando un número significativo de secuencias erróneas debido a los artefactos de la PCR.

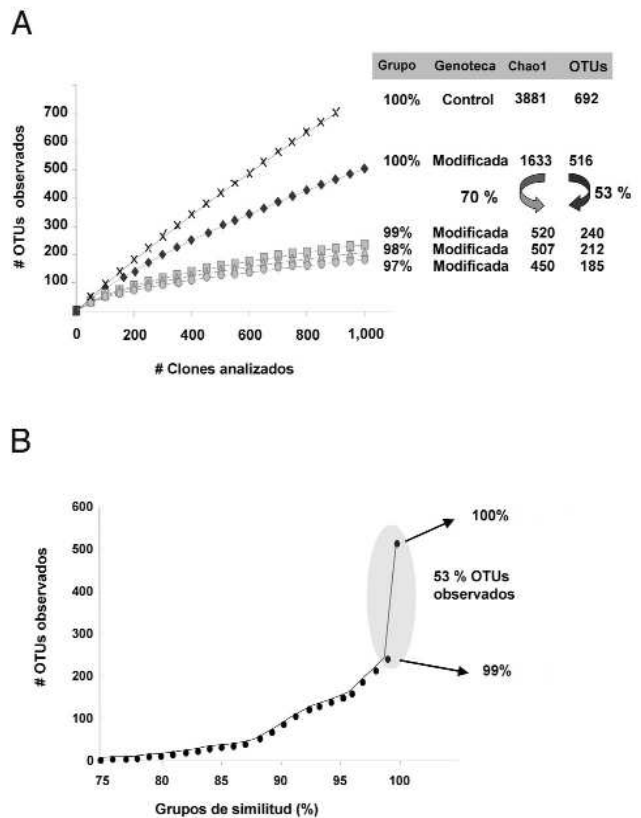


Figura 3. A. Análisis de rarefacción comparando la genoteca control y modificada a distintos niveles de similitud, en donde se observa el número de especies estimadas por Chao-1 y el número de OTUs observados en ambas genotecas. **B.** Gráfica que ilustra el número de OTUs observado en la genoteca modificada a distintos grupos de similitud (desde 100% al 75%). Más de la mitad de lo OTUs se concentran en grupos de 99% de similitud (grupos de microdiversidad) revelando que la fracción predominante de diversidad es microdiversidad en esta comunidad microbiana.

Muestreo insuficiente de comunidades microbianas

Hasta la publicación de Venter en abril de 2004, y nuestro artículo de julio de 2004, el estudio de la diversidad microbiana se basaba en el análisis de un número muy limitado de clones a partir de genotecas ambientales. En esta fase es en donde se observa los diferentes niveles de financiación de cada laboratorio pues no es lo mismo secuenciar 100 clones que 1000, y, por supuesto, las respuestas que obtienes tampoco son las mismas. Curiosamente, a pesar del desconocido potencial de secuenciación de Venter en el año 2004 con la secuenciación de 1,66 millones de carreras (secuenciación parcial de genes de aproximadamente 800 bp de longitud), sólo pudo identificar 1400 genes del 16S rRNA, similar en número a las 1067 secuencias de nuestra genoteca modificada. Sólo en el reciente estudio de Rusch (Rusch *et al*, 2007) con la secuenciación de 7,7 millones de carreras, se ha incrementado el número de genes ribosómicos a 4125. Sin embargo, utilizando la pirosecuenciación se pueden detectar hasta 900000 *tags* del gen 16S rRNA representativas de una comunidad bacteriana (Huber *et al*, 2007).

Múltiples copias del gen ribosómico en genomas bacterianos.

Otra de las preguntas clave para estimar la diversidad de una comunidad microbiana utilizando el gen ribosómico era: *¿Qué proporción de la diversidad que observamos es debido a la existencia de múltiples copias del gen 16S rRNA en genomas bacterianos?* Nuestro estudio también abarcó esta cuestión mediante el análisis del número de genes parálogos del 16S rRNA presentes en los genomas bacterianos disponibles en las bases de datos. Entonces, sólo existían 97 genomas bacterianos y se detectó un total de 242 copias del gen 16S rRNA, lo cual nos proporcionaba un valor de una sobreestimación de 2,5 veces (242/97) (Acinas *et al*, 2004a). Una comparación alternativa fue realizada a partir de los datos del metagenoma del Mar de los Sargazos en la que se comparó el número de genes ribosómicos (1400) con el número de genes *recA* de copia única en los genomas bacterianos (600), obteniendo una estimación casi idéntica a la nuestra (2,3). Otros genes alternativos que se utilizan como marcadores moleculares son aquellos de los que se sabe que tienen sólo una copia en los genomas bacterianos y que se encuentran conservados evolutivamente. Un ejemplo son los genes funcionales *recA* (recombinasa A), *rpoB* (subunidad B de la RNA polimerasa) o el gen *gyrB* (subunidad B de la DNA girasa)

aunque su uso aún es limitado debido a que el número de secuencias depositadas es todavía muy inferior al de los genes ribosómicos y por tanto su valor filogenético restringido.

¿Cuántos genomas bacterianos (“especies”) coexisten en una misma comunidad microbiana?

Como he comentado anteriormente, el esfuerzo de Venter en el año 2004 y el nuestro con respecto al número de genes identificados del 16S rRNA fue bastante similar (1400 vs. 1067). A pesar de utilizar distintas estrategias (metagenoma vs. genoteca), nuestros resultados en cuestión a la diversidad estimada de una comunidad microbiana son comparables. Así, en el metagenoma del mar de los Sargazos, se detectaron 1164 filotipos y se estimó la existencia de unas 1800 especies en esa comunidad microbiana. Nosotros detectamos 516 filotipos y estimamos unas 1633 especies bacterianas que coexistían en una misma comunidad microbiana. Sin embargo, utilizando la pirosecuenciación para comparar la diversidad entre dos comunidades microbianas se han podido detectar hasta 18500 filotipos con la predicción de la existencia de unas 37000 especies entre ambas comunidades (Huber *et al*, 2007).

Gracias al uso de las técnicas metagenómicas ha sido posible indagar sobre el contenido genómico, fisiología y relevancia ecológica de microorganismos no cultivados de diversos ambientes naturales (Bejá *et al*, 2001; DeLong *et al*, 2006). La aplicación de esta tecnología ofrece una visión más global y realista de la diversidad genética de una comunidad microbiana ya que no se limita al estudio de genes individuales o a un grupo determinados de ellos. También por ese motivo, dichas técnicas no resultan las más idóneas para analizar en profundidad la estructura de una comunidad microbiana. Por el contrario, la tecnología de pirosecuenciación, con capacidad para secuenciar entre 2 ó 3 órdenes de magnitud superior a la tradicional clonación y secuenciación, en la que se pueden secuenciar cientos de miles de pequeños fragmentos de genes, se prevé como una aproximación más adecuada para tal fin (Sogin *et al*, 2006, Huber *et al*, 2007). Sin embargo, todavía tiene que ser optimizada ya que el hecho de que se limiten a 100 bp del gen 16S rRNA (200 bp como máximo) implica que sólo se pueda asignar un valor taxonómico de “Phylum-Class-Order” de menor resolución al que se obtiene con los 800 bp de promedio en la secuenciación de un clon del gen 16S rRNA de manera tradicional. Actualmente sigue siendo una aproximación poco accesible para la mayoría de laboratorios modestos por su

elevado coste, pero existen convocatorias como la del *International Census of Marine Microbes* (ICoMM), que han financiado proyectos en los que se incluya el uso de la pirosecuenciación con muestras ambientales.

“Microdiversidad”: fracción predominante de una comunidad microbiana marina.

Uno de los resultados más relevantes de nuestro estudio fue poder analizar en detalle la estructura de una comunidad microbiana mediante el análisis de los 1067 genes del 16S rRNA de nuestra genoteca modificada. Para ello, realizamos curvas de rarefacción a distintos grupos de similitud (Figura 3A y 3B), de manera que en el eje de abscisas se representaba el número de clones analizados y en el eje de ordenadas se mostraba el número de OTUs (o filotipos) observados a distintos niveles de similitud desde el 100 % hasta un 75 % que fue el porcentaje en la cual se agruparon todas las secuencias en un único OTU (Figura 3B). Lo más sorprendente fue detectar que alrededor de un 53 % de los OTUs se agruparon en el nivel del 99 % de similitud. Traducido a otras palabras significaba que la mayoría de las secuencias recuperadas eran muy parecidas entre sí, como bacterias “primas hermanas” con una divergencia nucleotídica menor del 1% y que por lo tanto la fracción predominante de diversidad en nuestra muestra ambiental era “microdiversidad”. Estos mismos resultados se observaban cuando se comparaba el número de especies estimadas por Chao-1 entre los grupos de 100 y 99 % de similitud, pasando de 1633 a 520 especies estimadas. Eso significaba que el 70 % de las especies estimadas se concentraban en esos grupos a 99 % de similitud (grupos de microdiversidad o “micro-clusters”). Observamos que la predominancia de microdiversidad (“99 % micro-clusters”) no era un fenómeno casual, sino que era un patrón reproducible y así se ha observado en otros estudios de otras comunidades microbianas marinas (Pommier *et al*, 2007) e incluso procedentes de comunidades microbianas de marismas (Klepac-Ceraj *et al*, 2004). Estos grupos de microdiversidad no se pueden explicar ni por errores de la PCR, ni por las múltiples copias del gen ribosómico y constituyen un patrón observado independientemente de la estrategia usada, ya que también se ha detectado en el metagenoma del Mar de los Sargazos. Una de las hipótesis posibles es que estos grupos de microdiversidad pueden representar importantes unidades de diferenciación como “especies” o “ecotipos de especies”, dentro de una comunidad microbiana. La relevancia de la microdiversidad se ha puesto de manifiesto también en

los recientes estudios de Rusch (Rusch *et al*, 2007) y Huber (Huber *et al*, 2007) utilizando ambos dos estrategias muy diferentes a las genotecas (metagenomas y pirosecuenciación, respectivamente). Uno de mis proyectos actuales en el ICM es determinar la relevancia ecológica y variación genómica de estos grupos de microdiversidad en determinadas poblaciones bacterianas e indagar sobre los mecanismos por los cuales se que generan. ¡La emoción esta servida!

Bibliografía

Acinas SG, Marcelino LA, Klepac-Ceraj V and Polz MF. 2004a. Divergence and Redundancy of 16S rRNA Sequences in Genomes with Multiple *rrn* Operons. *J Bact* 186:2629-2635.

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL and Polz MF. 2004b. Fine Scale Phylogenetic Architecture of a Complex Bacterial Community. *Nature* 430:551-554.

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V and Polz MF. 2005. PCR-induced sequence error and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol* 71: 8966- 8969.

Béjà O, Spudich EN, Spudich JL, Leclerc M, DeLong EF. 2001. Proteorhodopsin phototrophy in the ocean. *Nature* 411: 786-9.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM. 2006. Community genomics among stratified microbial assemblages in the ocean’s interior. *Science* 311: 496-503.

Huber JA, Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML. 2007. Microbial population structures in the deep marine biosphere. *Science* 318: 97-100.

Klepac-Ceraj V, Bahr M, Crump BC, Teske AP, Hobbie JE, Polz MF. 2004. High overall diversity and dominance of microdiverse relationships in salt marsh sulphate-reducing bacteria. *Environ Microbiol* 6: 686-98.

Pommier T, Canbäck B, Riemann L, Boström KH, Simu K, Lundberg P, Tunlid A, Hagström A. 2007. Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol* 16:867-80.

Rusch DB, Halpern AL, ... and Venter JC. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5(3):e83.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proc Natl Acad Sci U S A* (32):12115-20.

Thompson JR, Marcelino LA, Polz MF. 2002. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by ‘reconditioning PCR’. *Nucleic Acids Res* 30:2083-8.

Venter JC, Remington K, ... and Smith HO. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:58-60.