RECONSTRUCCIÓN DE GENOMAS DE BACTERIAS DEL SUELO MEDIANTE ANÁLISIS DE CAMBIOS DE ABUNDANCIA ENTRE DOS CONDICIONES DE PH PARA EVIDENCIAR MECANISMOS DE ADAPTACIÓN A PH TAXA-ESPECÍFICOS.

Tesis
Entregada A La
Universidad De Chile
En Cumplimiento Parcial De Los Requisitos
Para Optar Al Grado De
Doctor en Ciencias con Mención en Biología
Molecular, Celular Y Neurociencias

Facultad De Ciencias

Por

Jonathan Elías Maldonado Soto

Enero, 2017

Director de Tesis Dr: Mauricio González Canales

FACULTAD DE CIENCIAS

UNIVERSIDAD DE CHILE

INFORME DE APROBACIÓN

TESIS DE DOCTORADO

Se informa a la Escuela de Postgrado de la Facultad de Ciencias que la Tesis de Doctorado presentada por el candidato.

JONATHAN ELÍAS MALDONADO SOTO

Ha sido aprobada por la comisión de Evaluación de la tesis como requisito para optar al grado de Doctor en Ciencias con mención en Biología Molecular Celular y Neurociencias, en el examen de Defensa Privada de Tesis rendido el día 12 de Septiembre del 2016.

Director de Tesis:
Dr. Mauricio González C. .....

Comisión de Evaluación de la Tesis

Dr. Francisco Chávez .....

Dr. Christian González .....

Dr. Nicolás Guiliani .....

Dra. Victoria Guixé .....

Dra. Raquel Quatrini .....

A mi esposa, que en momentos difíciles ha tenido fuerzas para apoyarme y formar juntos una hermosa familia.

Jonathan Elías Maldonado Soto nació el 7 de febrero del año 1982 en Santiago de Chile. Cursó su Educación Media en el Liceo de Aplicación de Santiago, donde obtuvo el premio a la mejor promoción el año 1999. Desde temprana edad presentó interés por la ciencia y la tecnología, lo cual motivó su ingreso al Plan Común de la Escuela de Ingeniería de la Universidad de Chile en el año 2000. Luego de cursar 3 años, debe decidir entre las carreras que ofrece la Universidad y escoge aquella que cumple a cabalidad con sus expectativas de desarrollo, tanto profesional como intelectual. Así, en el año 2003 ingresa a la Carrera de Ingeniería en Biotecnología Molecular dictada en la Facultad de Ciencias de la Universidad de Chile. Los conocimientos adquiridos en dicha carrera le permitieron aportar en diferentes proyectos de investigación, tanto nacionales como internacionales, cuyo fruto se ve reflejado en la participación de varias publicaciones en el área de la biología vegetal, animal y microbiología, un aporte transversal fruto del enfoque bioinformático que lo caracteriza y entusiasma. Con la finalidad de profundizar sus conocimientos, asumir nuevos desafíos en ciencias biológicas y extender sus redes de colaboración, decide el año 2012 tomar el desafío de convertirse en Doctor en Ciencias en el programa más difícil y de mayor excelencia de nuestro país, aquel dictado en la facultad que lo entusiasmó por la investigación. De esta forma entra al Doctorado en ciencias con mención en Biología Molecular, Celular y Neurociencias de la Facultad de Ciencias de la Universidad de Chile. Luego de recorrer 4 años de formación y duro trabajo, su esposa e hijos le acompañan en un cierre de ciclo que abre las puertas a un futuro amplio, posiblemente extranjero y lleno de excitantes oportunidades para seguir haciendo buena ciencia.

AGRADECIMIENTOS

Con respeto y admiración deseo agradecer a mi tutor, Dr. Mauricio González, *el profe*, por aceptar este gran desafío y entregar su apoyo incondicional para el desarrollo de un proyecto ambicioso y con muchas aristas por descubrir. Es un ejemplo de persona cuyo compañerismo y entusiasmo me animaron a sacar el mayor provecho posible de los resultados. Agradezco al Dr. Alejandro Maass y la Dra. Verónica Cambiazo por sus oportunos consejos. Gracias Dinka, Mauricio y Pablo por toda la ayuda entregada, especialmente en esas agotadoras salidas a terreno tomando muestras en el apunante altiplano. Gracias amigos del Laboratorio de Bioinformática y Expresión Génica de INTA por el apoyo. Agradezco de forma especial a la Comisión por perfeccionar este proyecto y por el aporte a mi formación como Doctor en Ciencias.

Agradezco a mi esposa por su gran paciencia y comprensión y a nuestra amada Eli quien nos entrega alegría día a día, consume nuestras energías pero nos llena de vida.

Agradezco a quienes me ayudaron con el financiamiento de esta aventura:

- Centro de Regulación del Genoma (FONDAP 15090007) el cual financió salidas a terreno, muestreo y diversos gastos en determinaciones analíticas.
- Beneficio Gastos Operacionales de CONICYT el cual financió los gastos en

secuenciación de todo lo expuesto en la presente tesis.

- Beca para Estudios de Doctorado Nacional CONICYT la cual financió arancel y manutención.
- Proyecto FONDECYT 1151384, el cual financió mi asistencia a congresos y diversos gastos de operación.

A todos los que aportaron de forma directa o indirecta, ayudando a que este desafío llegara a buen término, les estoy infinitamente agradecido.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTA DE ABREVIATURAS

%N50:	Porcentaje de scaffolds que constituyen el N50 (respecto del número total de scaffolds)
28F:	Partidor del gen del rDNA 16S de E. coli, posición 28, dirección "sentido" (*forward*)
519R:	Partidor del gen rDNA 16S de E. coli, posición 519, dirección "anti-sentido" (reverse).
AD3:	Filo candidado AD3
ArsR:	Proteína con sistema de dos componentes
BlastP:	(software) Herramienta básica de alineamiento local para proteínas (*Basic Local Alignment Search Tool for Proteins*)
BRC1:	Filo candidado BRC1.
CAN:	Comisión de Normalización y Acreditación (Chile)
CCA:	Análisis de correspondencia canónica
CDS:	Secuencia codificante
CE:	Conductividad eléctrica
COG:	Cluster de proteínas ortólogas (*Clusters of Orthologous Groups*)
CONCOCT:	(software) *Clustering cONtigs with COverage and ComposiTion*
contigs:	Secuencias provenientes del ensamble de lecturas (*reads*)
CtaC:	Citocromo C oxidasa
CydA:	Citocromo bd, subunidad A
CydB:	Citocromo bd, subunidad B
CyoB:	Citocromo bo, subunidad B
CyoC:	Citocromo bo, subunidad C
CyoD:	Citocromo bo, subunidad D
CyoE:	Citocromo bo, subunidad E
DNA:	Ácido desoxirribonucleico
F1F0-ATPase:	ATPasa F1F0
FBP:	Filo candiado FBP
FlgS:	Sensor de histidina quinasa
Gad:	Glutamato decarboxilasa
Gb:	Giga pares bases de nucleótidos (109 pb)
GC skew:	Sesgo de GC según fórmula: GC skew = (#G-#C)/(#G+#C)
GC:	Porcentaje de nucleótidos G y C respecto del total de la secuencia.
GN02:	Filo candidado también conocido como BD1-5 y ahora renombrado a Gracilibacteria.

gOTU:	Unidad taxonómica operativa genómica
IDBA-UD:	(software) *Iterative De Bruijn graph de novo Assembler for data with Uneven sequencing Depth*
IOS-ICP:	Espectroscopia óptica de emisión con plasma acoplado de forma inductiva
IUPAC:	*International Union of Pure and Applied Chemistry*
k-mers:	Fragmentos de secuencias con tamaño determinado
m s.n.m.:	metros sobre el nivel del mar
MALT:	(software) *MEGAN alignment tool*
Mb:	Mega pares bases de nucleótidos (106 pb)
MEGAN:	(software) *MEtaGenome ANalyzer*
MLE:	Enzima maloláctica
MO:	Porcentaje de materia orgánica
N50:	Tamaño del scaffold a partir del cual, al sumar el tamaño de todos los scaffolds mayores a él, se obtiene el 50 % de la suma del tamaño de todos los scaffolds.
NCBI:	*National Center for Biotechnology Information* (USA)
NH4:	Amonio
NhaA:	Sensor de pH y efector
NO3:	Nitrato
Nuo:	Oxidoreductasa NAD-quinona
OD1:	Filo candidato conocido ahora como Parcubacteria.
OmpA:	Proteína A de exterior de membrana
OTU:	Unidad taxonómica operativa
pb:	Pares de bases de nucleótidos
PCA:	Análisis de componentes principales
PCR:	Reacción en cadena de la enzima polimerasa.
pvalue:	Valor de probabilidad de error de correlación
QIIME:	(software) *Quantitative Insights Into Microbial Ecology*
rDNA 16S:	Unidad ribosomal 16S
RNA:	Ácido ribonucleico
S1:	Sitio de muestreo 1 con pH 5.88
S2:	Sitio de muestreo 2 con pH 7.48
S3:	Sitio de muestreo 3 con pH 8.51
scaffold:	Conjunto de contigs unidos (ordenados) por información de posición supra ensamble.
SCGs:	Conjunto de genes esenciales de copia única (*single copy genes*).
SlpA:	Proteína acídica de capa superficial
SOM:	Mapas auto-asociativos

TM6:	Filo candidato conocido ahora como Dependentiae.
TM7:	Filo candidato conocido ahora como Saccharibacteria.
TnaA:	Triptófano deaminasa
WPS-2:	Filo candidato WPS-2
WS2:	Filo candidato WS2.
WS3:	Filo candidato conocido ahora como Latescibacteria.

RESUMEN

El desarrollo actual permite enfrentar de forma más sencilla la descripción de vínculos entre las propiedades generales de una comunidad bacteriana compleja y las características de su ambiente. El paso siguiente debiera ser el determinar cuál es la contribución relativa de los miembros a la función global de la comunidad. Una forma de acercarse a este objetivo es lograr una descripción más detallada de la comunidad, por ejemplo, mediante la descripción de miembros de la comunidad, permitiendo distinguir capacidades taxa-específicas que se relacionen de forma directa con propiedades del ambiente. Para abordar este desafío realizamos una caracterización fisicoquímica y taxonómica de la comunidad bacteriana de suelo de 3 sitios cercanos del altiplano del Desierto de Atacama que presentan diferencias significativas en su nivel de pH. Luego utilizamos y perfeccionamos diferentes métodos bioinformáticos a fin de recuperar genomas individuales desde cada metagenoma para finalmente evaluar la existencia de sesgo en sus capacidades metabólicas en relación a mecanismos de homeostasis de pH. Como resultado logramos reconstruir 74 genomas de suelo en calidad borrador: 14 de sitio ácido (pH 5,88), 39 de sitio neutro (pH 7,48) y 21 de sitio alcalino (pH 8,51). Al comparar sus capacidades funcionales es posible distinguir diferencias significativas en categorías asociadas de forma directa o indirecta a mecanismos de homeostasis de pH. Nuestro trabajo revela que el ambiente ejerce una selección sobre los componentes genómicos de estas comunidades bacterianas. Estos nuevos métodos nos ayudarán a entender los procesos que dirigen la adaptación de organismos a ambientes extremos.

ABSTRACT

Current developments allow the easy description of associations between general properties of a complex bacterial community and features of their environment. The next step should be to determine the relative contribution of different community members to the global function of the bacterial community. A way to asses this goal is to obtain a more detailed community description, for example, through the description of community members and the ability to distinguish taxa-specific features related directly with environmental features. To face this challenge, we did a physicochemical and taxonomical characterization of soil bacterial community that belong to 3 near sites of the Atacama Desert plateau that shows significant differences in their pH level. Then we used and improved different bioinformatics methods to recover individual genomes form each metagenome to finally, evaluate the bias existence in their metabolic capacities related to pH homeostasis mechanisms. Be obtained 74 draft soil genomes: 14 of acid soil (pH 5.88), 39 of neutral soil (pH 7.48) and 21 of alkaline soil (pH 8.51). Comparisons of functional capabilities resulted in significant differences in categories directly or indirectly associated to pH homeostasis mechanisms. Our work revealed that environment drives a selection over the genomic component of these bacterial communities. This new methods will help to understood processes that drive organism's adaptations to extreme environments.

INTRODUCCIÓN

1. Metagenómica como herramienta para estudiar la diversidad y funcionalidad microbiana

Estudios basados en el marcador ribosomal 16S han redefinido e incrementado nuestro conocimiento sobre la diversidad microbiana. Cálculos simples de la diversidad microbiana del suelo permiten estimar un rango promedio de 3.000 a 11.000 genomas distintos por gramo de suelo, el cual incluye 10^9 células microbianas individuales. Se considera que los microoganismos contenidos en el suelo pueden conformar un universo de 10^{12} especies bacterianas en el planeta (Torsvik & Ovreas, 2002; Torsvik y col., 2002; Curtis & Sloan, 2004; Cowan y col., 2005; Gans y col., 2005; Hirsch y col., 2010; Hug, Baker, y col., 2016; Locey & Lennon, 2016). En general, menos del 1 % de las células observadas por conteo directo es accesible mediante técnicas de cultivo lo cual limita las opciones de identificar y caracterizar esta gran diversidad de especies (Skinner y col., 1952; Alain & Querellou, 2009; Hirsch y col., 2013). En áreas con alta disponibilidad de nutrientes el número de especies cultivables puede aumentar 10 % (Sørensen, 1997), y porcentajes similares se han obtenido utilizando nuevas estrategias en el diseño de medios de cultivo (Davis y col., 2005; Alain & Querellou, 2009). Sin embargo, la cantidad de bacterias sin aislar es enorme y por lo tanto conocemos solo una pequeña fracción de las características genéticas, fisiológicas y bioquímicas de los

miembros de las comunidades bacterianas del ambiente (Hirsch y col., 2010, 2013). Si bien técnicas como la secuenciación del RNA ribosomal 16S y el análisis de fosfolípidos de membranas celulares (Zelles y col., 1992; Glaser y col., 2004) permiten determinar cambios globales en la composición taxonómica de comunidades bacterianas, ninguno de estos métodos tiene la capacidad o la resolución para describir en detalle la diversidad funcional de sus miembros.

En este escenario, la metagenómica surge como una estrategia experimental que facilita el acceso a la información contenida en los genomas de los microorganismos (Handelsman, 2004; Pettit, 2004; Streit y col., 2004; Streit & Schmitz, 2004). La metagenomica o genómica ambiental ha sido definida como "el análisis genómico de los microorganismos por extracción directa de ADN a partir de ensamble de microorganismos" (Handelsman y col., 1998). En términos generales, la metagenómica consiste en la extracción, clonamiento (según la técnica) y secuenciación del DNA total (metagenoma) de un ecosistema el cual, *a priori*, es considerado como un solo ser vivo (Handelsman y col., 1998). Este tipo de acercamiento permite el estudio de la amplia diversidad de genes individuales codificados en el DNA del conglomerado microbiano y de sus productos, así como el análisis de operones completos (Brady y col., 2001; Courtois y col., 2003). Con esta herramienta es posible responder preguntas ecológicas claves al poder relacionar potenciales funciones metabólicas con el conjunto de individuos que componen la comunidad o sus partes componentes (Torsvik & Ovreas, 2002; Cowan y col., 2005; Lozupone y col., 2012; Steven y col., 2012; Nyyssonen y col., 2013).

Con el advenimiento de la secuenciación masiva a partir del año 2000, se abrió la posibilidad de análisis e interpretaciones genómicas a escala ecológica (Tyson y col., 2004; Venter y col., 2004). En la actualidad y gracias al desarrollo de las tecnologías de secuenciación masiva de segunda y tercera generación basadas en el fraccionamiento total del DNA, ya no es necesario el paso de clonamiento del DNA en un organismo intermediario (Ronaghi, 2001; Ronaghi & Elahi, 2002; Ahmadian y col., 2006; Lauber y col., 2009). Esto permite evitar las restricciones asociadas a la técnica de clonación y a la transformación, y disminuir la manipulación total de la muestra, obteniendo una mejor representación del genoma de la comunidad y de la abundancia relativa de los miembros la comunidad. Lo anterior se ve reforzado por la gran cantidad de DNA que es posible secuenciar hoy en día con menores costos (Kim y col., 2012; Hou y col., 2013; Kilpinen & Barrett, 2013; McCormack y col., 2013). Así, la metagenómica está revolucionando la microbiología con una gran cantidad de información disponible a un bajo costo relativo.

2. Desarrollo de estrategias bioinformáticas para enfrentar el desafío de la interpretación biológica de datos metagenómicos

Uno de los problemas en las investigaciones actuales en metagenómica es la "escala" a la cual se observan y describien los fenómenos. Como fue mencionado, la metagenómica permite describir la estructura de la comunidad (conocer los miembros de la comunidad con sus abundancias relativas) y los procesos funcionales que la definen en su conjunto ('quiénes están ahí' y 'qué están haciendo'). Sin embargo, es mucho más no logra producir una descripción en detallada de las comunidades para entender 'quién

está haciendo qué cosa', información esencial para deducir el ajuste al ambiente y las interacciones entre miembros de una comunidad de microorganismos. Por ejemplo, la reconstrucción de rutas metabólicas de la comunidad bacteriana tiene como supuesto tácito que la comunidad funciona como un meta-organismo, por lo cual, cualquier proteína predicha desde un segmento de DNA secuenciado puede interactuar con una proteína de otro segmento de DNA, y por lo tanto, se puede construir casi cualquier ruta metabólica si los genes que codifican estas proteínas están representados en el metagenoma. Sin embargo, sabemos que este supuesto no representa necesariamente la realidad si consideramos que cada microorganismo dentro la comunidad tiene un espacio finito y discreto que lo define, y que cada microorganismo establece interacciones de distinta naturaleza con los demás miembros de su comunidad, tanto en el espacio como en el tiempo. En particular, la existencia de una membrana plasmática restringe tanto la interacción entre proteínas como el flujo de los metabolitos que se producen debido a su actividad. Por ello, el concepto de interacción extra organismo es correcto sólo en ciertos ámbitos celulares como el "*quorum sensing*", otros fenómenos asociados a multicelularidad y en fenómenos de reciclaje de elementos debido al recambio de individuos de la comunidad.

En atención a lo expuesto, la presente tesis realizamos una descripción metabólica acabada de la potencialidad metabólica de una serie de comunidades de microorganismos, asumiendo un modelo de meta-organismo representado por la sumatoria de los procesos metabólicos (taxa-específicos) de los diversos miembros que componen la comunidad. Ciertamente, cada especie bacteriana tiene una serie de atributos funcionales generales y también otros taxa-específicos codificados en su

genoma. Como el proceso habitual de reconstrucción de un metagenoma no distingue el origen del DNA, estos atributos pasan a formar parte del acervo genético del conjunto, y nuestra interpretación nos lleva a construir super-rutas metabólicas juntando atributos que en su contexto real están separados.

En ausencia de esta consideración, los estudios que realizan la comparación funcional de metagenomas obtenidos de distintos ambientes tienden a subestimar las diferencias funcionales pues, al juntar todos los atributos de los miembros de una comunidad, cualquier ruta parece posible, y se obtiene como resultado que los atributos son iguales entre comunidades incluso de ambientes contrastantes (T. H. M. P. Consortium, 2012; Fierer, Lauber, y col., 2012; Lozupone y col., 2012). La solución a este problema es poder separar, *in silico*, el DNA secuenciado de forma taxa-específica, para luego ensamblar las secuencias de DNA en genomas que representen cada taxa por separado (idealmente a nivel de género o especie). Esto representa un enorme desafío bioinformático pues significa tratar de reconstruir 'taxones individuales' a partir de secuencias de DNA obtenidas de una mezcla de miembros pertenecientes a diferentes especies de la comunidad.

Al inicio de esta tesis Doctoral (abril 2014), muy pocos grupos de investigación habían reportado estrategias experimentales y/o bioinformáticas que permitieran la recuperación de genomas a partir de secuencias metagenomicas complejas. Por ejemplo, el grupo liderado por la Dra. Banfield en la Universidad de Berkeley, California, separó 49 genomas desde una base de datos de 20 Gb de secuencias de DNA a partir de una comunidad aislada del drenaje acido minero (Wrighton y col., 2012). Utilizando

algoritmos de ensamble iterativo, seguido de estrategias de clasificación de secuencias basadas en mapas de auto-organización, logro recuperar 87 grupos de secuencias o "bins", de los cuales 49 pertenecían a genomas de linajes de bacterias no cultivadas y no caracterizadas previamente (*novel*). Esta información permitió identificar estilos de vida fermentativos que no habían sido descritos previamente para bacterias. Un enfoque alternativo es la secuenciación de genomas obtenidos desde células únicas aisladas mediante técnicas de citometría o de microfluidos, y cuyo DNA es posible amplificar y secuenciar gracias a al método de amplificación por desplazamiento múltiple (MDA) (Lasken, 2012). Este tipo de técnicas ha permitido asignar a ciertos miembros de la comunidad nuevas funciones relacionadas con la obtención de energía en proteobacterias pertenecientes a zonas oscuras del océano (Swan y col., 2011) y describir el estilo de vida sacarolítico de bacterias potencialmente degradadoras de celulosa (Dodsworth y col., 2013).

3. Métodos de clasificación de DNA para la construcción de sub-grupos de secuencias con propiedades comunes

Un aspecto decisivo en la recuperación de genomas a partir de secuencias metagenómicas es el grado de complejidad de la muestra analizada. En comunidades bacterianas de baja complejidad (< 100 especies) es posible obtener genomas casi completos de las bacterias dominantes (Tyson y col., 2004; Garcia Martin y col., 2006). Sin embargo, la probabilidad disminuye en comunidades con un mayor número especies, problema que se acentúa cuando la abundancia relativa de una especie cae por debajo

del 1 % (Tringe y col., 2005; Kunin y col., 2008). Un avance en este ámbito han sido las mejoras de las técnicas de secuenciación masiva las cuales han aumentado la cantidad total de secuencia producida o "throughput" por tanda de secuenciación, han aumentado el largo de la lectura y ha aumentado la calidad de las secuencias obtenidas. A pesar de lo anterior, sigue siendo un problema de difícil solución el ensamble de genomas individuales desde poblaciones de baja abundancia en comunidades de alta complejidad (> 100 especies). Por tal motivo, es primordial desarrollar avances en los métodos de clasificación de las secuencias de DNA que representen cada taxa por separado en una muestra metagenómica.

En esta tesis de Doctorado se utilizó la técnica de clasificación reportada por Albertsen y colaboradores (2013), la cual permite clasificar secuencias de bacterias que cambian su proporción de forma efectiva en dos muestras independientes pero relacionadas. Por ejemplo, puede ser una misma comunidad cuyo DNA es extraído con diferentes métodos o distintas comunidades de un mismo ecosistema sujeto a variaciones bióticas o abióticas (ej. diferencia de pH). Esta estrategia de clasificación abre la posibilidad de reconstruir genomas individuales y de de analizar la relación entre sus atributos funcionales y las características del ambiente en que se encuentran. En este contexto, la propuesta de investigación contempló la selección de muestras de comunidades bacterianas que por una parte mantuviesen una comunidad similar en su composición y por otra, que presentaran cambios significativos en la abundancia relativa de sus miembros.

4. Dependencia de la homeostasis de pH en bacterias

Los estudios metagenómicos en suelos que presentan variaciones de pH muestran una fuerte dependencia entre las variaciones en el pH del suelo y la distribución de taxones bacterianos en dichas comunidades microbianas (Villagrán y col., 1981; Bryant y col., 2008; Rousk y col., 2010; Fierer y col., 2011; Fischer y col., 2011; Wang y col., 2011; Fierer, Lauber, y col., 2012; Wang y col., 2012). En general, las bacterias son extremadamente dependientes de la homeostasis del pH debido a que la mayoría de las proteínas tienen rangos determinados de pH en las cuales funcionar (Slonczewski y col., 2009a). Además, la concentración de protones alrededor de la membrana plasmática se relaciona de forma directa con la capacidad de producir energía de las bacterias (Krulwich y col., 2011b). En consideración a estas características, en la presente tesis se propuso realizar un análisis de comunidades microbianas provenientes de suelo con distinto pH. Por ello se desarrollaron y aplicaron herramientas bioinformáticas que permitieran reconstruir genomas de bacterias mediante análisis de cambios de abundancia entre dos condiciones de pH. A este propósito se sumó a nuestro interés por identificar mecanismos de adaptación de especies que conforman un sistema microbiano sometido naturalmente a un gradiente de pH en el Desierto de Atacama.

5. Hipótesis de Trabajo

La hipótesis de trabajo que se desarrolló en esta tesis fue:

En comunidades bacterianas de suelo sometidas naturalmente a condiciones contrastantes de pH, el cambio de abundancia relativa de taxas es una variable que permite discriminar genomas de taxas particulares dentro del metagenoma, revelando a su vez mecanismos taxa-específicos de homeostasis del pH.

6. Objetivo general

Determinar si la diferencia de abundancia entre bacterias presentes en suelo con diferente pH es una herramienta que permite diferenciar de forma eficiente y específica genomas individuales dentro del metagenoma y facilitar la identificación de los elementos génicos vinculados a la de homeostasis de pH propios de cada comunidad.

7. Objetivos específicos

Objetivo 1: Localización de sitios contrastantes en propiedades fisicoquímicas en el Desierto de Atacama y caracterización de variables geoquímicas y ambientales.

El objetivo fue determinar las diferencias entre variables abióticas como temperatura, humedad, pH, salinidad, textura del suelo, contenido de materia orgánica, micro y macro nutrientes de sitios con micro ambientes fisicoquímicos distinguibles y seleccionar los sitios con menor número de variables contrastantes.

Objetivo 2: Caracterización taxonómica y funcional de las comunidades de bacterias de sitios contrastantes para la variable pH de suelo

La caracterización funcional y taxonómica se realizó mediante secuenciación, ensamble y anotación del DNA total obtenido a partir de la comunidad de microorganismos del suelo de cada sitio de muestreo, buscando reconstruir genomas individuales (taxa-específicos) a partir del conjunto de los metagenomas. Esto se realizó identificando y agrupando los cambios de abundancia de las moléculas de DNA ensambladas ("scaffold") en lo que he denominado como "Unidad Taxonómica Operativa genómica" (gOTU). Además, se realizó la caracterización taxonómica de los ambientes seleccionados mediante amplificación y secuenciación del gen codificante para la unidad ribosomal 16S.

Objetivo 3: Identificar y comparar elementos de homeostasis de pH en los genomas ensamblados

El objetivo fue identificar los elementos de homeostasis de pH presentes en cada genoma ensamblado para observar si existe relación entre la abundancia de las taxas extraídas en suelo con diferente pH, y su potencial de homeostasis de pH según los elementos que sean encontrados en cada gOTU.

MATERIALES Y MÉTODOS

1. Descripción de sitio y características generales de sistemas alto andinos

La Puna es un ecosistema altoandino localizado en una planicie desértica sobre los 3500 m.s.n.m. que comprende parte del noreste de Chile, noroeste de Argentina, sureste de Perú y oeste de Bolivia. Se caracteriza por la presencia de cuencas endorreicas, es decir cuencas en las que el agua no tiene salida fluvial hacia el océano y que normalmente forman salares por evaporación. En este ambiente se encuentra el volcán Lascar, uno de los volcanes más activos del norte de Chile andino, cuya última erupción fue el 30 de octubre del 2015, produciendo una importante pluma de ceniza. A sus pies, en la cara sur del volcán se encuentra la laguna Lejía que es un lago oligohalino de poca profundidad con pH alcalino, ubicado sobre los 4000 m s.n.m. Se encuentra rodeado de varios volcanes en una cuenca aislada de gran interés científico debido a su singular localización. Esta zona ha sido designada por el gobierno de Chile como sitio prioritario para la conservación de biodiversidad (Munoz-Pedreros y col., 2013).

2. Sitios de muestreo

Los sitios de muestreo están localizados en la cara sur del volcán Lascar, cerca de la laguna Lejía (23°30'S and 67°42'W) (ver Figura 1.1). El muestreo fue realizado el mismo día, durante el verano austral de la temporada 2014 (mes de Abril). Después de una búsqueda *in situ* evaluando el pH del suelo, seleccionamos tres sitios que presentaban pH contrastante: sitio con pH 5.88 en coordenadas 23°30'07.4''S, 67°43'25.6''W denominado S1, sitio con pH 7,48 en coordenadas 23°30'12.0''S, 67°42'24.0''W denominado S2 y sitio con pH 8,51 en coordenadas 23°30'12.2''S, 67°42'22.9''W denominado S3.

3. Obtención de muestras

Las muestras fueron recolectadas en los tres sitios seleccionados (ver Figura 1.1). En cada sitio se seleccionaron tres sub-sitios localizados a menos de un metro de distancia entre sí y en cada uno se recogieron 500 gr de suelo a 10 cm de profundidad. De cada muestra se destinaron 100 gr de suelo para extracción de DNA y el resto para análisis de composición. Las muestras de suelo destinadas para extracción de DNA fueron recolectadas en condiciones de esterilidad, puestas en tubos plásticos e inmediatamente almacenadas en hielo seco hasta su llegada al laboratorio donde se almacenaron a -80°C.

4. Extracción de DNA

El DNA de suelo fue extraído usando un método en base a CTAB (Bromuro de hexadeciltrimetilamonio) siguiendo el protocolo descrito por Zhou y colaboradores (1996), con las modificaciones indicadas por Prestel y colaboradores (2008). La integridad del DNA fue evaluada mediante electroforesis y luego en una estación Agilent 2200 TapeSation.

5. Amplificación y secuenciación del gen codificante para la subunidad de DNA ribosomal 16S (rDNA 16S)

El gen codificante para rDNA 16S fue amplificado entre las regiones variables 1 y 3 desde el DNA extraído de suelo utilizando los partidores ("primers") 28F (5′-GA GTT TGA TCM TGG CTC AG-3′) and 519R (5′-GWA TTA CCG CGG CKG CTG-3′) (Turner y col., 1999) con código de barras ("barcode") en el partidor "sentido" o "forward". Se utilizó el kit HotStarTaq Plus Master Mix (Qiagen, USA) en las siguientes condiciones: 94°C por 3 minutos, seguido de 28 ciclos de 94°C por 30 segundos, 53°C por 40 segundos y 72°C por 1 minuto, seguido de un paso de elongación final de a 72°C por 5 minutos. Luego de la amplificación los productos de PCR fueron examinados en un gel de agarosa al 2 % para determinar la amplificación exitosa e intensidad de las bandas. Las muestras de los diferentes sitios con sus réplicas (cada una con su respectivo código de barras) fueron mezcladas en proporciones iguales en base a su peso molecular y a la concentración de DNA. Las muestras fueron purificadas usando el sistema

Agencourt AMPure XP (Beckman, USA). Luego las muestras mezcladas y purificadas se usaron para preparar las librerías de DNA siguiendo el protocolo TruSeq DNA sample preparation (Illumina, USA). La secuenciación fue realizada en el laboratorio Molecular Research DNA (MR DNA, www.mrdnalab.com, Shallowater, TX, USA) en la plataforma Illumina MiSeq con una configuración de sobrelapamiento de 2x300 pb siguiendo los protocolos del fabricante para obtener un mínimo de 30.000 secuencias por muestra.

6. Procesamiento de secuencias rDNA 16S

Las secuencias de los amplicones del rDNA 16S fueron procesados y analizados siguiendo protocolos descritos previamente (Dowd y col., 2008; Handl y col., 2011). Brevemente, las secuencias o "lecturas" fueron fusionadas (pares sobrelapados), agrupadas por muestra y luego se eliminaron los segmentos de secuencias "barcodes". Secuencias menores de 150 pb o con asignación ambigua de bases no fueron consideradas para análisis posteriores. Las secuencias aceptadas como válidas fueron agrupadas usando el algoritmo USearch (v6.1.544) con un 4 % de divergencia para remover quimeras y grupos consistentes de solo una secuencia (i.e. singletons) (Edgar, 2010; Edgar y col., 2011). Finalmente las secuencias fueron filtradas por calidad mínima de 30 (q30) con Mothur v.1.22.2 (Schloss y col., 2009). Las secuencias de rDNA 16S obtenidas en la presente tesis fueron depositadas en la base de datos Sequence Read Archive (SRA) del National Center for Biotechnology Information (NCBI) con el código SRP070518.

7. Análisis de secuencias rDNA 16S y asignación taxonómica

Las secuencias fueron analizadas utilizando el programa computacional Quantitative Insights Into Microbial Ecology (QIIME v1.8.0, Caporaso et al., 2010). Brevemente, se utilizó el módulo *pick_closed_reference_otus.py* de QIIME para extraer todas las secuencias que tuvieran coincidencias en la base de datos Greengenes (McDonald D, et al. 2012; versión gg_otus_13_08) con un 97 % de similitud (3 % de divergencia) y la taxonomía fue asignada directamente desde la coincidencia más cercana. El proceso de asignación de Unidades Taxonómicas Operativas (OTUs) se realizó usando el programa USearch v6.1.544 (Edgar, 2010; Edgar y col., 2011) con los parámetros predeterminados por QIIME (-s 0.97 –z True --max_accepts 1 --max_rejects 8 --word_length 8 –minlen 64 --usearch61_sort_method abundance). El proceso de asignación de OTUs mediante "referencia-cercana" generó 2.296 OTUs que comprenden 103.031 secuencias desde las 9 muestras (entre 6.447 y 14.752 secuencias por muestra). Los OTUs sin asignación o con asignación a mitocondria o cloroplastos fueron removidos. Los OTUs con solo 1 secuencia también fueron eliminados.

8. Estimación de la diversidad

En el primer paso del análisis de diversidad alfa (Fierer, Leff, y col., 2012), cada muestra fue sub-muestreada de forma aleatoria y sin reemplazo usando el módulo *alpha_rarefaction.py* de QIIME. El sub-muestreo se realizó a diferentes profundidades en incrementos de 899 secuencias por iteración y 10 iteraciones por cada incremento

hasta alcanzar 6.400 secuencias por muestra. Esta cantidad corresponde al número de secuencias de la muestra más pequeña (muestra S1.2). En cada subgrupo se evaluaron diferentes índices de diversidad alfa (Shannon, Chao1 y Faith's Phylogenetic Diversity o "PD") obteniendo sus correspondientes curvas de rarefacción.

9. Análisis genómico de la comunidad de microorganismos de suelo (metagenoma)

En cada sitio las muestras de DNA extraídas desde los triplicados fueron mezcladas para obtener una muestra representativa de DNA comunitario por sitio. Dicha muestra fue secuenciada mediante tecnología Illumina HiSeq 2500 de extremos apareados (2x150 pb por lectura) en el laboratorio Molecular Research DNA (MR DNA, www.mrdnalab.com, Shallowater, TX, USA) siguiendo los protocolos del fabricante para obtener un mínimo de 15 Gb de datos por muestra.

10. Ensamble y sub-ensambles de metagenomas

Las 3 librerías de secuencias de DNA obtenidas desde las 3 muestras de suelo fueron filtradas por calidad y secuencias adaptadoras mediante el conjunto de programas FASTX-Toolkit (Hannon Lab, http://hannonlab.cshl.edu/fastx_toolkit/). En particular, se utilizó el módulo *fastx_clipper* para eliminar secuencias adaptadoras exigiendo un sobrelapamiento mínimo de 18 nucleótidos y descartando secuencias de tamaño menor a 51 nucleótidos. De forma similar, la calidad de secuencias fue revisada con el módulo *fastq_quality_filter* exigiendo un conteo de calidad (score) mínimo de 20 en al menos el

95 % de la secuencia. Cada muestra fue ensamblada por separado como muestras metagenómicas utilizando el programa computacional IDBA-UD versión 1.1.1 con pre-corrección activada, uso de *k-mers* entre 64 y 124 y salto de 20 *k-mers* (6 pasos de procesamiento) (Peng y col., 2010, 2011, 2012). Para obtener ensambles híbridos (dos o más muestras en conjunto) se utilizó el protocolo descrito por Albertsen *y col.* (2013) con modificaciones tomadas del protocolo descrito por Alneberg *y col.* (2014) y adaptaciones según el tipo de muestras de este proyecto. Brevemente, se realizaron ensambles utilizando las secuencias generadas en las muestras S1 y S2 (metagenoma S1+S2) y las muestras S2 y S3 (metagenoma S3+S3) lo cual incrementa la probabilidad de ensamblar "scaffolds" de buen tamaño y que se encuentren secuencias "sobrelapables" compartidas entre muestras. Posteriormente, las aproximadamente ~$1.4x10^8$ secuencias generadas (lecturas) en cada una de las 3 muestras, mediante tecnología Illumina HiSeq 2500, fueron ubicadas en los "scaffolds" usando el programa computacional Bowtie versión 2.2.2 (Langmead & Salzberg, 2012). El contenido GC de los "scaffolds" y la frecuencia de *k-mers* se calculó mediante los módulos del programa computacional CONCOCT (Alneberg y col., 2014). Todas las variables son utilizadas por CONCOCT para disminuir el número de dimensiones principales mediante un Análisis de Componentes Principales (PCA). En el siguiente paso, las secuencias son agrupadas utilizando un modelo de ajuste de mezcla Gaussiana con una aproximación Bayesiana obteniendo alrededor de 150 grupos por metagenoma. Los grupos de "scaffolds" fueron evaluados en base al número de genes esenciales de copia única (SCGs) presentes en cada grupo (Ciccarelli y col., 2006) utilizando como criterio que cada grupo debe tener como mínimo 18 de los 36 genes esenciales de copia única con un

máximo de 2 copias presentes por gen. Los grupos seleccionados, denominados OTUs genómicos (gOTUS), fueron re-ensamblados utilizando el programa computacional Velvet versión 1.2.10 (Zerbino & Birney, 2008) con *k-mer* de 121, cobertura esperada puesta en modo automático y tamaño del inserto de 900 pb.

11. Anotación funcional y determinación de capacidades funcionales de los metagenomas y genomas individuales

Para llevar a cabo la anotación funcional tanto de los metagenomas como de los genomas individuales se realizó una detección automática de los marcos de lectura abiertos con el programa computacional Prodigal versión 2.6.2 (Hyatt y col., 2012). Luego se realizó la búsqueda por homología utilizando herramienta MALT versión 0.012 de la suite MEGAN versión 5.10.6 (Huson y col., 2011) con la base de datos COG (Clusters of Orthologous Groups) del 2012-2014 (Tatusov y col., 2003; Alneberg y col., 2014) y el algoritmo BlastP con lo cual se obtiene la clasificación taxonómica y funcional de cada proteína. El enriquecimiento de las categorías SEED en cualquiera de los pares de grupos comparados (por ejemplo S1 vs S2) fue calculado con un Test Exacto de Fisher con la asistencia del módulo FISHERTEST del paquete Real-Statistics (http://www.real-statistics.com) para Microsoft Excel. Finalmente, para determinar la presencia de secuencias codificantes para proteínas descritas en mecanismos de homeostasis de pH se realizó BLASTP contra cada genoma reconstruido exigiendo un E-value máximo de $1x10^{-10}$. La lista de proteínas asociadas a pH se realizó manualmente incorporando información experimental reportada en literatura.

12. Análisis de propiedades fisicoquímicas del suelo

En terreno se midió la temperatura y la humedad relativa usando el sensor portátil iButton DS1923 Hygrochron el cual fue localizado a 10 centímetros de profundidad. Las muestras de suelo de cada sector fueron secadas, tamizadas (2 mm) y se les midió pH (G. W. Thomas, 1996) y conductividad eléctrica (Rhoades, 1996) en agua desionizada en proporción 1:1 w/w. El contenido de materia orgánica y carbono total fue determinado por el método de pérdida de peso con oxidación química en dicromato de sodio (Sadzawka, 2006). El N, NH4 y NO3 fueron determinados por el método de destilación de Kjeldahl (Bremner & Mulvaney, 1982; Sadzawka, 2006). La composición elemental fue determinada mediante dispersión de energía de rayos X (EDXRF) luego que las muestras fueran molidas y tamizadas en malla de 75 um (Vanhoof y col., 2004). Las mediciones se realizaron en un equipo Shimadzu EDX-720, disponible en el Laboratorio de Espectrometría de Fluorescencia de Rayos X, Departamento de Geología de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile gracias a la colaboración del Dr. Diego Morata. El equipo fue operado a 15 kV (región de energía Na-Sc) y 50 kV (región de energía Ti-U) con una corriente de 1 mA bajo presión de 30 Pa y un tiempo de adquisición de 100 s por muestra. La fracción biodisponible de metales del suelo fue determinada usando Espectroscopia Óptica de Emisión con Plasma Acoplado de forma Inductiva (IOS-ICP) en el Laboratorio de Servicios Suelo y Foliar de la Facultad de Agronomía e Ingeniería Forestal de la Pontificia Universidad Católica de Chile, de acuerdo a Ettler et al., (2007), Queralt et al., (2005) y métodos establecidos por la Comisión de Normalización y

Acreditación (CAN) de la Sociedad Chilena de Ciencia del Suelo (Zagal & Sadzawka, 2007).

13. Análisis multivariados

Para los análisis multivariados se utilizaron los 2.296 OTUs identificados en las 9 muestras y 18 variables ambientales evaluadas en las mismas 9 muestras. Todos los cálculos se realizaron usando el paquete estadístico R versión 3.1.3 (RCoreTeam, 2015). Para los análisis de varianza se usó análisis de componentes principales (PCA) mediante la función "prcomp" del paquete "stats" versión 3.1.3. Se realizó agrupamiento de OTUs mediante mapas auto-asociativos (SOM) utilizando el paquete "kohonen" versión 2.3.5 para R. Se utilizó el paquete "vegan" versión 3.2.4 para realizar un Análisis de Correspondencia Canónica (CCA) utilizando como base de ordenamiento de la muestras (sitios) una matriz de datos compuesta por la abundancia a nivel de filo de los OTUs identificados, la abundancia de los OTUs seleccionados por análisis SOM y la abundancia de los gOTUs. Se superpuso a dicho ordenamiento la matriz de datos compuesta por las variables ambientales mediante ajuste canónico.

RESULTADOS

1. Caracterización de variables fisicoquímicas del sitio de muestreo

El sitio de muestreo comprende una planicie ubicada a 4000 m s.n.m. en la ladera sur del Volcán Lascar, a unos 200 m al Oeste de la Laguna Lejía en el altiplano del Desierto de Atacama, Chile. El primer objetivo de mi trabajo de tesis tuvo como finalidad inspeccionar las características fisicoquímicas y ambientales generales del lugar para elegir sitios con características fisicoquímicas contrastantes para el estudio. Para ello se analizaron 40 sub-zonas de la planicie buscando diferencias de pH dado que es una variable fácil de medir en terreno y que influye fuertemente en la composición de las comunidades bacterianas. Además, se midió la temperatura y la humedad relativa del suelo. A partir de esta información se seleccionó un transecto de 2 Km de longitud que exhibe un gradiente de pH entre 5.88 (sitio 1, S1) y 8.51 (sitio 3, S3) (Figura 1.1). Se recolectaron 3 muestras de suelo por cada sitio a 10 cm de profundidad, las cuales fueron procesadas como se indica en la sección Materiales y Métodos. La Tabla 1.1 contiene los valores medidos de las variables determinadas en los sitios S1 y S3 y un sitio con pH intermedio denominado S2. En atención a la manera en que pueden afectar el funcionamiento de los miembros de las comunidades de microorganismos, las variables fueron agrupadas en las categorías fisicoquímicas y nutricionales (esta última dividida en micro y macronutrientes). Las variables seleccionadas representan aquellas reportadas en la literatura como las más influyentes en la estructuración de comunidades bacterianas de suelo (Fierer & Jackson, 2006; Andrew y col., 2012; Bachar y col., 2012;

Fierer, Lauber, y col., 2012; Crits-Christoph y col., 2013; Colombo y col., 2014). Aun cuando el análisis en terreno indica que los sitios del transecto no presentan diferencias significativas en variables como altitud, temperatura (Figura 1.2, panel A), humedad relativa (Figura 1.2, panel B) y composición granulométrica de suelo, 13 de las 17 variables del suelo medidas poseen diferencia significativa entre S1 y S3 (Tabla 1.1). En particular, el valor de pH y el contenido de Mg, Ca, Mn y Na en el suelo presentan diferencias significativas en cualquiera de los pares de sitios comparados. Como no existe una norma general respecto a los requerimientos nutricionales de bacterias de suelo desértico, hemos utilizado una norma de requerimientos nutricionales de plantas a fin de obtener una idea de escala entre la disponibilidad y el requerimiento de nutrientes. Esto considerando que la zona posee prácticamente nula vegetación la cual, en otros sistemas, es un factor importante en la estructura de las comunidades bacterianas. De esta forma se observa que los suelos presentan déficit nutricional en P, K, Mg, y Ca y un exceso de B.

La Figura 1.3 grafica el círculo de correlación que se genera del análisis de la relación entre las variables mediante análisis de componentes principales. Este análisis considera el comportamiento de las variables en los 3 sitios y nos permite encontrar cuáles poseen una tendencia similar. Se observa que el pH y Zn, Cu, Fe, Mn y P poseen un comportamiento similar pero inverso, lo cual está dentro de lo esperado pues el pH modula la solubilidad de estos elementos. Estas variables en general poseen un comportamiento de gradiente entre los 3 sitios (ver Tabla 1.1). Otro grupo de variables son Na, Ca, CE, B, S y C las cuales, según se puede observar en la Tabla 1.1, presentan una mayor señal en S3 respecto de los otros 2 sitios.

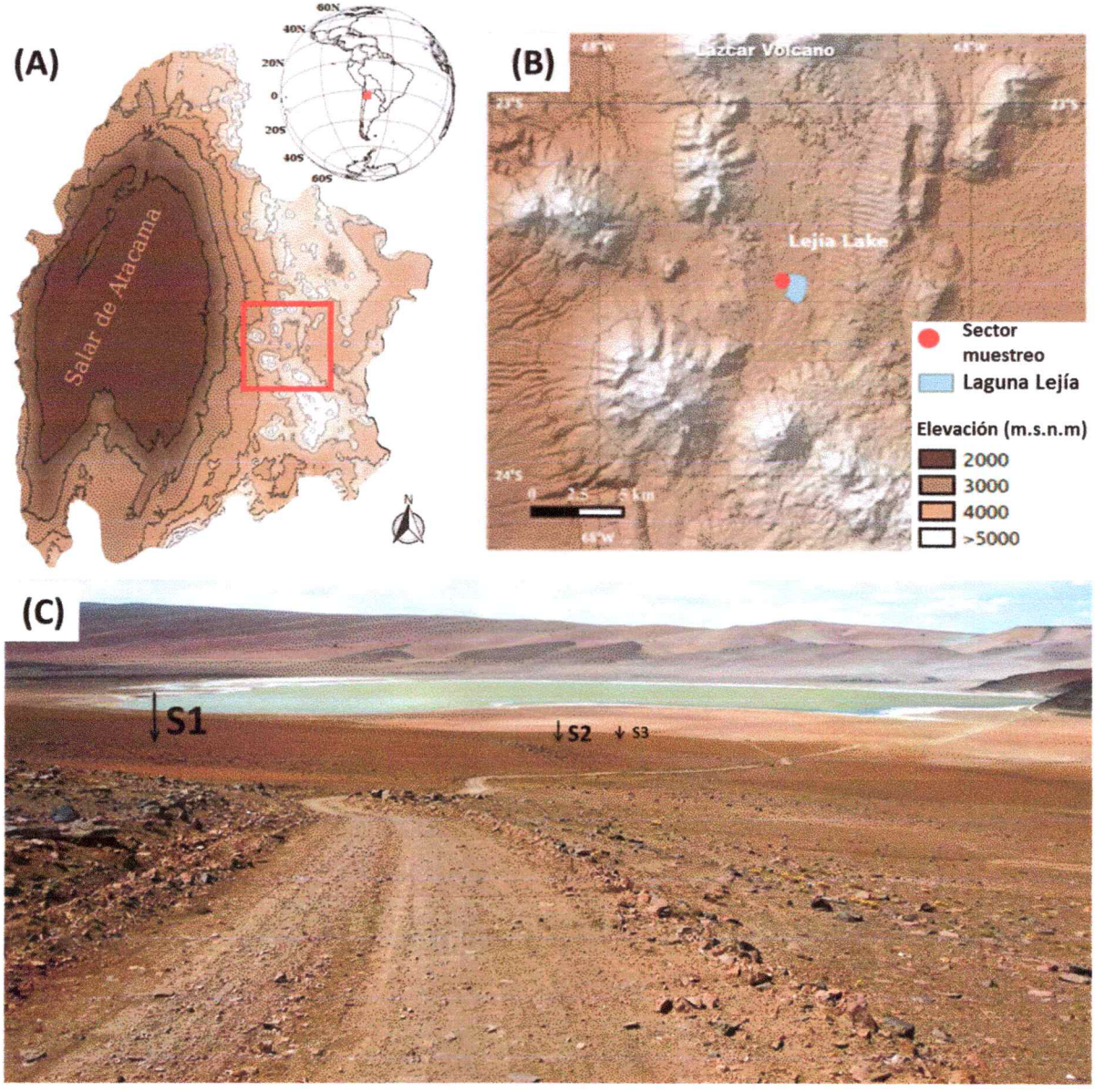


Figura 1.1. Sitio de muestreo. Se encuentra ubicado en el Desierto de Atacama (23°30'S y 67°42'O) en un gradiente de pH de 2 km (5.88 en sitio S1, 7.48 en sitio S2 y 8.51 en sitio S3). Recolectamos 3 réplicas por sitio a una profundidad de 10 cm. **A)** Posición del sector de muestreo en recuadro rojo respecto del Salar de Atacama. **B)** Ampliación del recuadro rojo del panel A para mostrar en punto rojo la posición del sitio de muestreo respecto de la Laguna Lejía. **C)** Panorámica del sector de muestreo demarcando los 3 sitios seleccionados. De fondo se observa la Laguna Lejía.

Tabla 1.1. Caracterización geoquímica de 3 muestras de suelos seleccionados por su variación de pH. Se presentan las características fisicoquímicas y contenido de micro y macro nutrientes disponibles para los 3 sitios. Cada medición representa un promedio del triplicado por sitio con su respectiva desviación estándar. Las letras "a", "b" y "c" representan las relaciones entre los 3 sitios considerando cada variable lo cual fue determinado mediante una prueba T de dos colas con varianza homoscedástica ($p<0.05$). Los nombres de los elementos siguen los códigos estándar IUPAC. Se resalta en negro las mediciones que están en déficit y en rojo las mediciones que están por sobre la norma para cultivos vegetales según (Ahlert, 1998; Jones, 2012) lo cual nos entrega una idea de la riqueza de nutrientes en el suelo a falta de una norma general para bacterias.

Sitio:	S1	S2	S3
Fisicoquímicas			
pH	5.88 ± 0.21 a	**7.48 ± 0.13** b	8.51 ± 0.04 c
Conductividad Eléctrica (mS/cm)	0.04 ± 0.00 a	0.05 ± 0.01 a	0.13 ± 0.00 b
Materia Orgánica (%)	0.24 ± 0.03 a	0.17 ± 0.01 b	0.25 ± 0.01 a
Macronutrientes			
C (mg/Kg)	2433.33 ± 249.44 a	2500.00 ± 141.42 a	3866.67 ± 262.47 b
N (mg/Kg)	24.00 ± 0.00 a	25.50 ± 1.50 a	23.00 ± 4.00 a
NH4 (mg/Kg)	5.00 ± 0.00 a	13.00 ± 0.00 b	8.50 ± 1.50 a
NO3 (mg/Kg)	19.00 ± 0.00 a	12.50 ± 1.50 b	14.50 ± 2.50 b
P (mg/Kg)	**10.87 ± 2.23** a	**5.63 ± 0.52** b	**4.07 ± 0.74** b
K (mg/Kg)	**65.33 ± 13.89** a	**136.00 ± 14.35** b	168.00 ± 13.74 b
Mg (mg/Kg)	**62.40 ± 24.82** a	307.93 ± 51.47 b	**132.08 ± 9.79** c
Ca (mg/Kg)	**406.14 ± 134.23** a	1507.01 ± 113.55 b	7054.08 ± 199.06 c
Micronutrientes			
Cu (mg/Kg)	3.61 ± 1.00 a	1.80 ± 0.92 b	0.57 ± 0.24 b
Fe (mg/Kg)	10.37 ± 2.12 a	5.51 ± 1.84 b	**3.01 ± 0.69** b
Zn (mg/Kg)	**0.65 ± 0.32** a	**0.34 ± 0.20** a	**0.10 ± 0.05** a
Mn (mg/Kg)	7.28 ± 1.86 a	2.04 ± 0.43 b	**0.85 ± 0.08** c
S (mg/Kg)	4.10 ± 1.61 a	2.83 ± 0.24 a	**10.22 ± 1.74** b
Na (mg/Kg)	9.19 ± 1.88 a	16.85 ± 1.08 b	42.13 ± 2.87 c
B (mg/Kg)	**1.80 ± 0.05** a	**1.86 ± 0.05** a	**3.91 ± 0.21** b

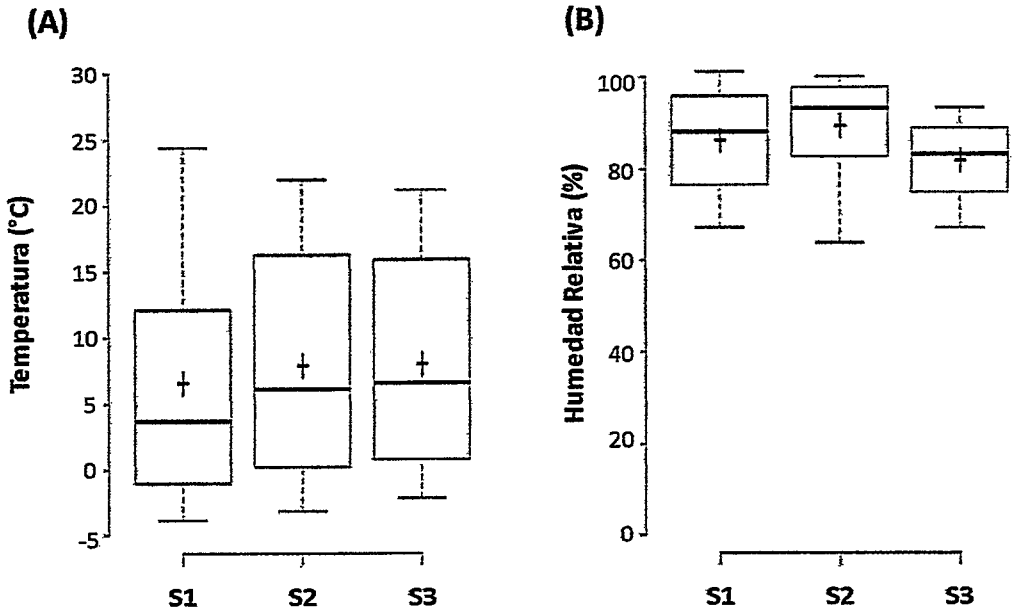


Figura 1.2. Diagrama de cajas con el perfil de temperatura y humedad relativa de la primera capa de suelo. La temperatura y humedad relativa fueron medidas en terreno utilizando el sensor portátil iButton DS1923 Hygrochron posicionado entre 5-10 cm de profundidad durante 2 días al momento del muestreo. **A)** Perfil de temperatura. **B)** Perfil de humedad relativa. El símbolo + representa en ambos gráficos el promedio de la distribución.

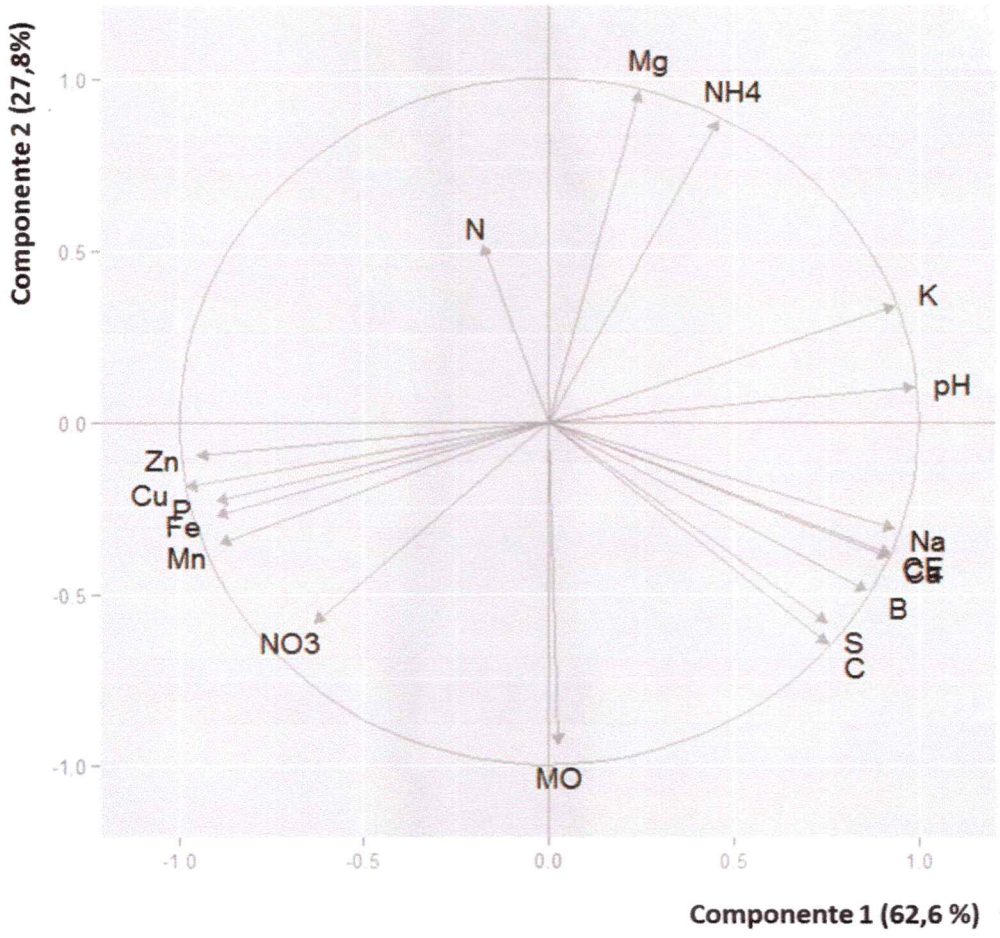


Figura 1.3. Círculo de correlación de variables. Análisis de correlación de todas las variables mediante componentes principales o PCA entre los 3 sitios. Se presenta la relación de acuerdo al Componente 1 y el Componente 2 los cuales representan 62.6 % y 27.8 % de la varianza total, respectivamente.

2. Caracterización taxonómica y funcional de las comunidades de bacterias presentes en el suelo del altiplano del Desierto de Atacama: análisis del efecto asociado a variaciones en el valor de pH del suelo

2.1 Caracterización taxonómica

Como se describe en la sección Materiales y Métodos, a partir del DNA extraído de cada uno de los 3 sitios se amplificó una zona del gen codificante para el rDNA 16S comprendida entre las zonas hipervariables V1 y V3 con los partidores 28F y 519R. Estos amplicones fueron secuenciados mediante tecnología MiSeq en formato 2x300 pb con un mínimo de profundidad de 30.000 lecturas. Algunos descriptores estadísticos generales de los amplicones se incluyen en la Tabla 2.1.1. El número de lecturas obtenidas tras aplicar el filtro de calidad se encuentra entre 30.309 y 44.070, de las cuales el número de secuencias que tienen homología con la base de datos Greengenes va desde 6.447 a 14.742 y el número de OTUs identificados entre 761 y 1199.

En este punto es importante considerar que el número de OTUs identificados puede estar ligado al número de lecturas con homología Tabla 2.1.1, por lo tanto es necesario normalizar los datos por muestra antes de proseguir con el análisis comparativo. Para ello se realiza un procedimiento denominado "rarefacción" en el que el total de lecturas por sitio se analiza mediante sub-muestreos al azar y de tamaño creciente. Luego se procede a realizar el recuento de OTUs (Figura 2.1.2, panel A) o se calcula el índice de diversidad alfa de Shannon (Figuras 2.1.2, panel B), valores que se grafican en función del número de secuencias en los sub-muestreos. Los índices de

diversidad alfa nos dan una idea de la riqueza de especies en una determinada comunidad. En particular, el índice de diversidad alfa de Shannon considera la probabilidad de encontrar un determinado individuo en un ecosistema considerando su abundancia relativa en la comunidad. En éste índice la riqueza de especies está ponderada por el aporte (abundancia) de cada especie en la comunidad. Su magnitud se oscila entre 5 y 9 en comunidades bacterianas provenientes de suelo.

El perfil de rarefacción para el número de OTUs permitió determinar un punto de corte equivalente en los tres sitios analizados (flechas en Figuras 2.1.1, panel A y 2.1.1, panel B) que considera el mayor número de lecturas disponible en la muestra S1 (6400 lecturas). A este número de lecturas la cantidad de OTUs aumenta levemente y el valor de índice de Shannon permanece prácticamente constante (Figuras 2.1.1 panel, C y 2.1.1, panel D), posibilitando entonces los análisis comparativos entre muestras.

La Tabla 2.1.2 contiene los valores de número de OTUs, diversidad alfa (número de especies) y diversidad filogenética (número de especies ponderado por relación filogenética) que describen en su conjunto la comunidad bacteriana presente en los suelos de los 3 sitios estudiados. S3 en comparación a S2 y S1 muestra un 12 % más de OTUs junto a un índice de Shannon y diversidad filogenética levemente superior. Por otra parte, cuando se construye un cladograma con distancia de Pearson a partir del perfil de su abundancia de los OTUs en los tres sitios analizados, se observa que S2 y S3 se agrupan, dejando a S1 en una relación más distante, (Figura 2.1.2). La distancia de Pearson considera la varianza de los datos para reescalar las variables (abundancia de cada OTUs) y comparar los sitios según el perfil de abundancia de OTUs.

Como se observa en el panel A de la Figura 2.1.3, los 3 sitios comparten 554 OTUs lo cual representa el 35.6 % de las muestras mientras que los OTUs únicos son 177, 47 y 124 en S1, S2 y S3 respectivamente. Al revisar las abundancias a nivel de filo por muestra (Figura 2.1.3, panel B) se observa que los taxones predominantes en todas las muestras analizadas son Acidobacteria, Actinobacteria, y Proteobacteria, que las réplicas biológicas poseen un buen grado de similitud y que los sitios (S1, S2 y S3) muestran diferencias evidentes. Las diferencias se manifiestan especialmente en los filos de menor abundancia relativa (inferior al 10 %).

Tabla 2.1.1. Estadística de secuenciación de amplicones rDNA 16S para las muestras de suelo de cada sitio seleccionado. Cada muestra se secuenció por triplicado exigiendo un número mínimo de 30.000 lecturas totales. La técnica utilizada es Illumina MiSeq amplificando 300 pb por cada extremo del DNA ("sentido" y "antisentido") en una configuración de sobrelapamiento con lo cual se obtiene una lectura final de 500 pb aproximadamente. Se presenta el total de lecturas útiles luego de aplicar el filtro de calidad. Las lecturas "mapeadas" corresponden al número de lecturas que exhibieron homología con alguna de las secuencias contenidas en la base de datos Greengenes, mientras que el número de OTUs corresponde al número total de secuencias identificadas en dicha base de datos mediante este proceso de mapeo por homología (sin rarefacción).

Muestra	Lecturas filtradas	Lecturas mapeadas	Número de OTUs
S1.1	30.309	6.517	779
S1.2	29.837	6.447	766
S1.3	31.048	6.650	761
S2.1	42.854	13.345	1.179
S2.2	44.985	13.987	1.199
S2.3	43.731	13.596	1.217
S3.1	40.534	13.448	1.088
S3.2	43.203	14.289	1.074
S3.3	44.070	14.752	1.118

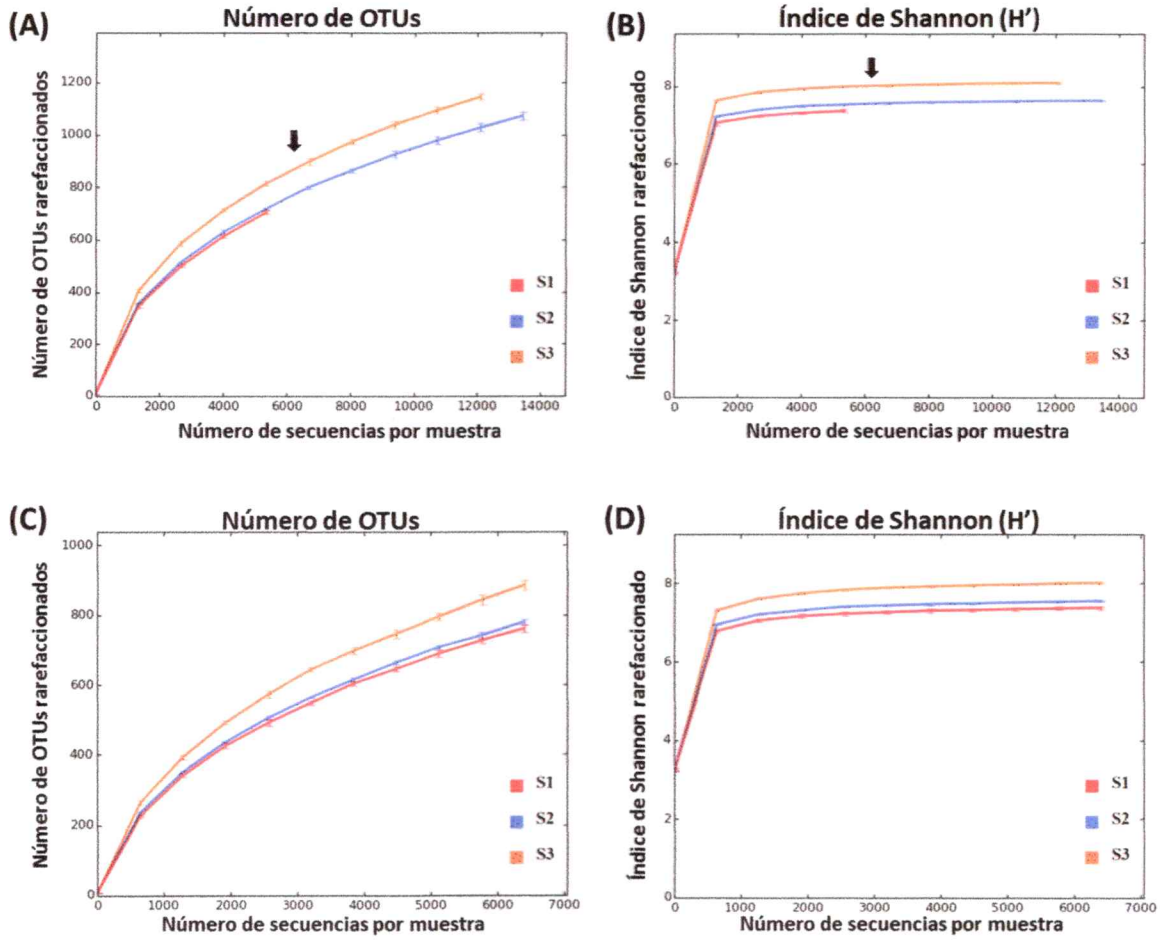


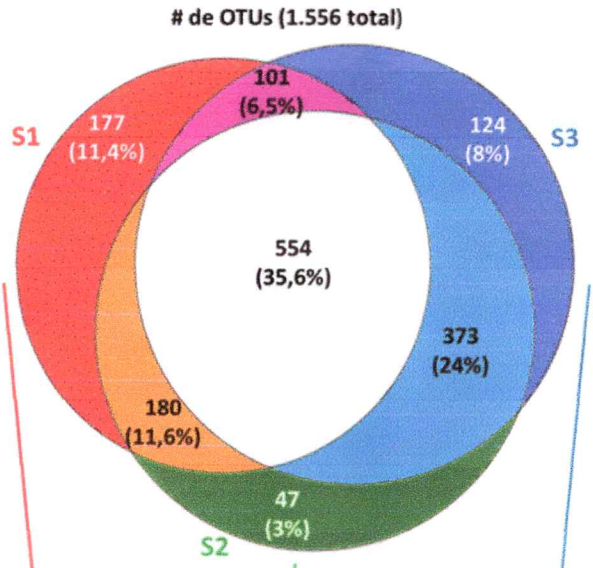
Figura 2.1.1. Promedio de rarefacción de lecturas para cada muestra. El número de OTUs identificados **(A)** y el Índice de Shannon **(B)** fueron calculados a partir del protocolo de rarefacción utilizando el total de lecturas mapeadas (ver detalles en sección Materiales y Métodos). Número de OTUs identificados **(C)** e Índice de Shannon **(D)** con secuencias rarefaccionadas hasta 6.400 secuencias en todas las muestras. En todos los casos, los valores corresponden al promedio y la desviación estándar calculada a partir del triplicado de cada sitio.

Tabla 2.1.2. Índices de diversidad alfa para los OTUs identificados en los 3 sitios. Cada muestra fue rarefaccionada (normalizada) a 6.400 lecturas por 10 iteraciones. Se presenta el resultado promediando los triplicados por sitios con su respectiva desviación estándar. Las letras "a", "b" y "c" representan las relaciones entre los 3 sitios considerando cada índice de diversidad lo cual fue determinado mediante una prueba T de dos colas con varianza homoscedástica ($p<0.05$).

Muestra	Número de OTUs (Riqueza)	Índice de Shannon (H')	Diversidad Filogenética (PD)
S1	761 ± 12 a	7,39 ± 0,05 a	21,77 ± 0,80 a
S2	780 ± 10 a	7,56 ± 0,02 b	23,25 ± 0,12 a
S3	884 ± 17 b	8,01 ± 0,04 c	27,80 ± 0,67 b

Figura 2.1.2. Cladograma de las muestras de cada sitio. El cladograma se realizó utilizando la distancia de Pearson sobre el listado de OTUs identificados y su correspondiente abundancia relativa mediante rarefacción a 6.400 secuencias.

(A)


de OTUs (1.556 total)
S1
177
(11,4%)
101
(6,5%)
124
(8%)
S3
554
(35,6%)
373
(24%)
180
(11,6%)
47
(3%)
S2

(B)

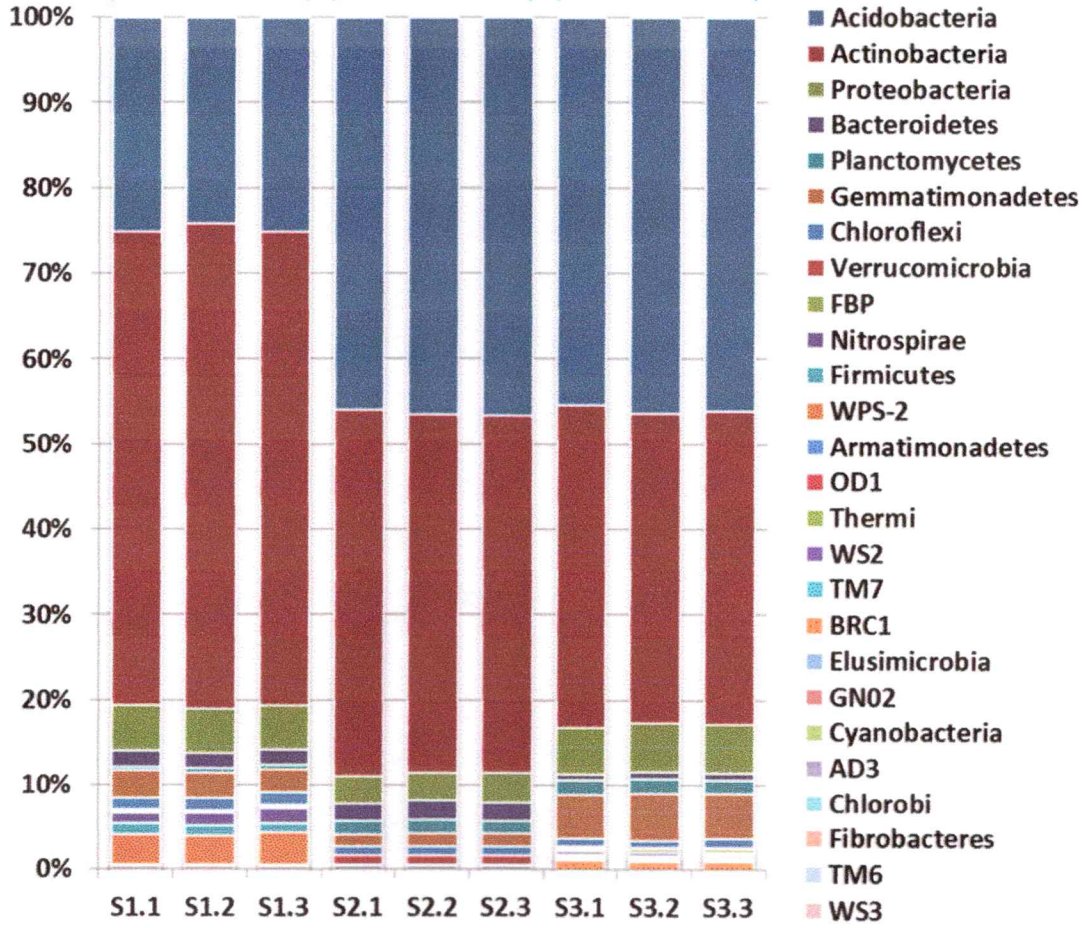

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
S1.1
S1.2
S1.3
S2.1
S2.2
S2.3
S3.1
S3.2
S3.3
Acidobacteria
Actinobacteria
Proteobacteria
Bacteroidetes
Planctomycetes
Gemmatimonadetes
Chloroflexi
Verrucomicrobia
FBP
Nitrospirae
Firmicutes
WPS-2
Armatimonadetes
OD1
Thermi
WS2
TM7
BRC1
Elusimicrobia
GN02
Cyanobacteria
AD3
Chlorobi
Fibrobacteres
TM6
WS3

Figura 2.1.3. Caracterización los OTUs identificados en las tres muestras de suelo. **A)** Diagrama de venn con tamaño proporcional mostrando las intersecciones entre el número de unidades taxonómicas operativas (OTUs) encontradas en muestra S1 (1.012), muestra S2 (1.154) y muestra S3 (1.152). **B)** Abundancia relativa a nivel de filo para cada sitio. Cada barra representa porcentaje de abundancia según amplificación del gen de la unidad ribosomal 16S con su taxonomía a nivel de filo y su abundancia relativa. La leyenda ha sido ordenada en considerando los niveles de abundancia relativa de mayor a menor del sitio S2 (pH=7.48).

Por otra parte, se realizó un análisis de agrupamiento de OTUs mediante mapas auto-asociativos (SOM) con el fin de encontrar aquellos OTUs que presentan un perfil de abundancia similar entre los 3 sitios analizados. La distribución de valores de tres de los grupos de OTUs que presentan mayor separación entre los 3 sitios analizados se muestran en una gráfica de radar (o diagrama de araña) en la Figura 2.1.4. Dichos grupos representan un porcentaje importante de la abundancia total de los sitios a los cuales representan, llegando por ejemplo al 69 % para el grupo representante de la muestra S1. El perfil de abundancia relativa de los grupos g1, g2 y g3 a nivel de filo se muestra en el panel A de la Figura 2.1.5. Al comparar el patrón de abundancia de filos de la Figura 2.1.5. con la graficada en la Figura 2.1.3 se observada que OTUs de 12 de los 26 filos presentes en la muestra total no segregan con aquellos que se asocian a sitios específicos del transecto, incluyendo los filos Armatimonadetes, Chlorobi, Cyanobacteria, Elusimicrobia, Fibrobacteres, GN02, OD1, Thermi, TM6, TM7, WS2 y WS3. Al eliminar los 2 filo más abundantes (Acidobacteria y Actinobacteria) se hace evidente la magnitud de las diferencias en la contribución de cada filo a los grupos g1, g2 y g3, incluyendo la visualización de algunos filos que se presentan en forma exclusiva en uno u otro de los sitios analizados (Figura 2.1.5. panel B).

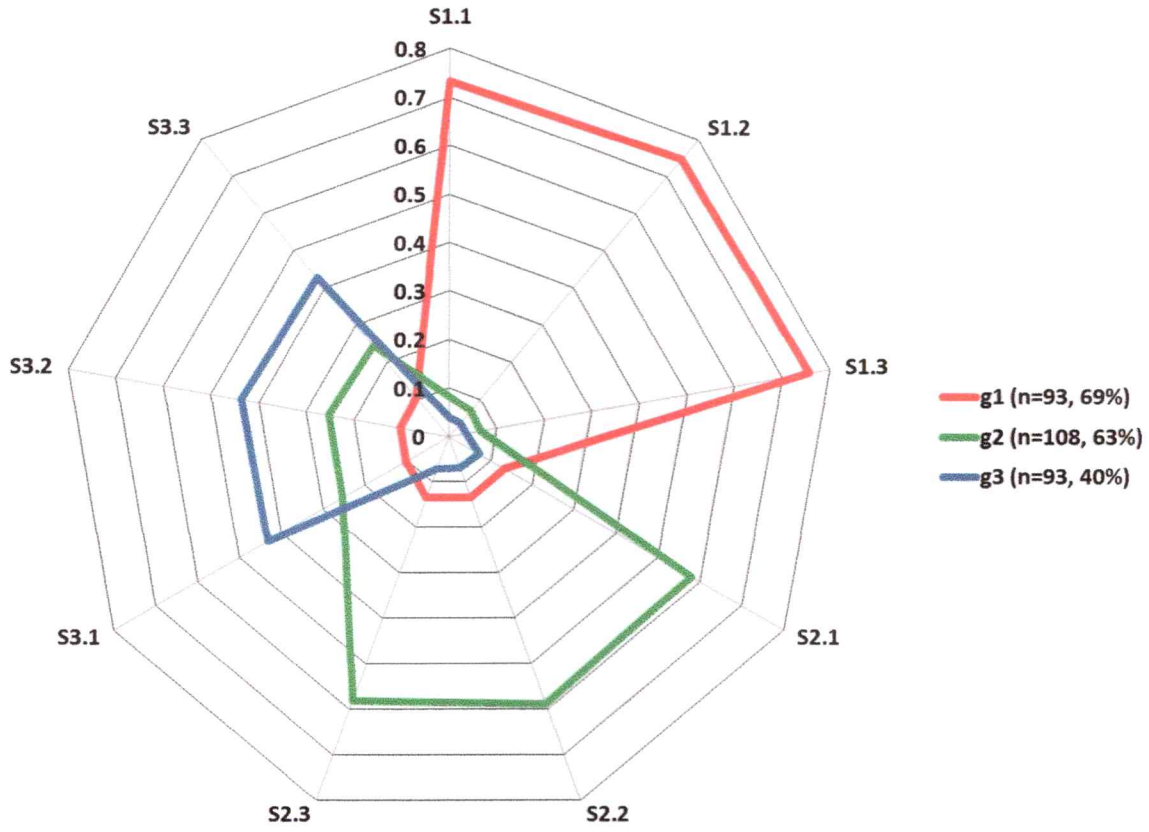


Figura 2.1.4. Distribución de la abundancia relativa de 3 grupos de OTUs seleccionados por presentar un comportamiento diferencial entre los tres sitios del transecto. Los grupos fueron identificados mediante Mapas Auto-Asociativos considerando los 3 sitios analizados con sus respectivas replicas. Los ejes radiales representan desde el centro hacia la periferia el porcentaje creciente de abundancia relativa de los OTUs en cada grupo. El grupo que se distribuye en el área delimitada por la línea de color rojo contiene 93 OTUs y representa el 69 % de la abundancia total de S1. El grupo de color verde contiene 108 OTUs y representa el 63 % de la abundancia total de S2. El grupo de color azul contiene 93 OTUs y representa el 40 % de la abundancia total de S3.

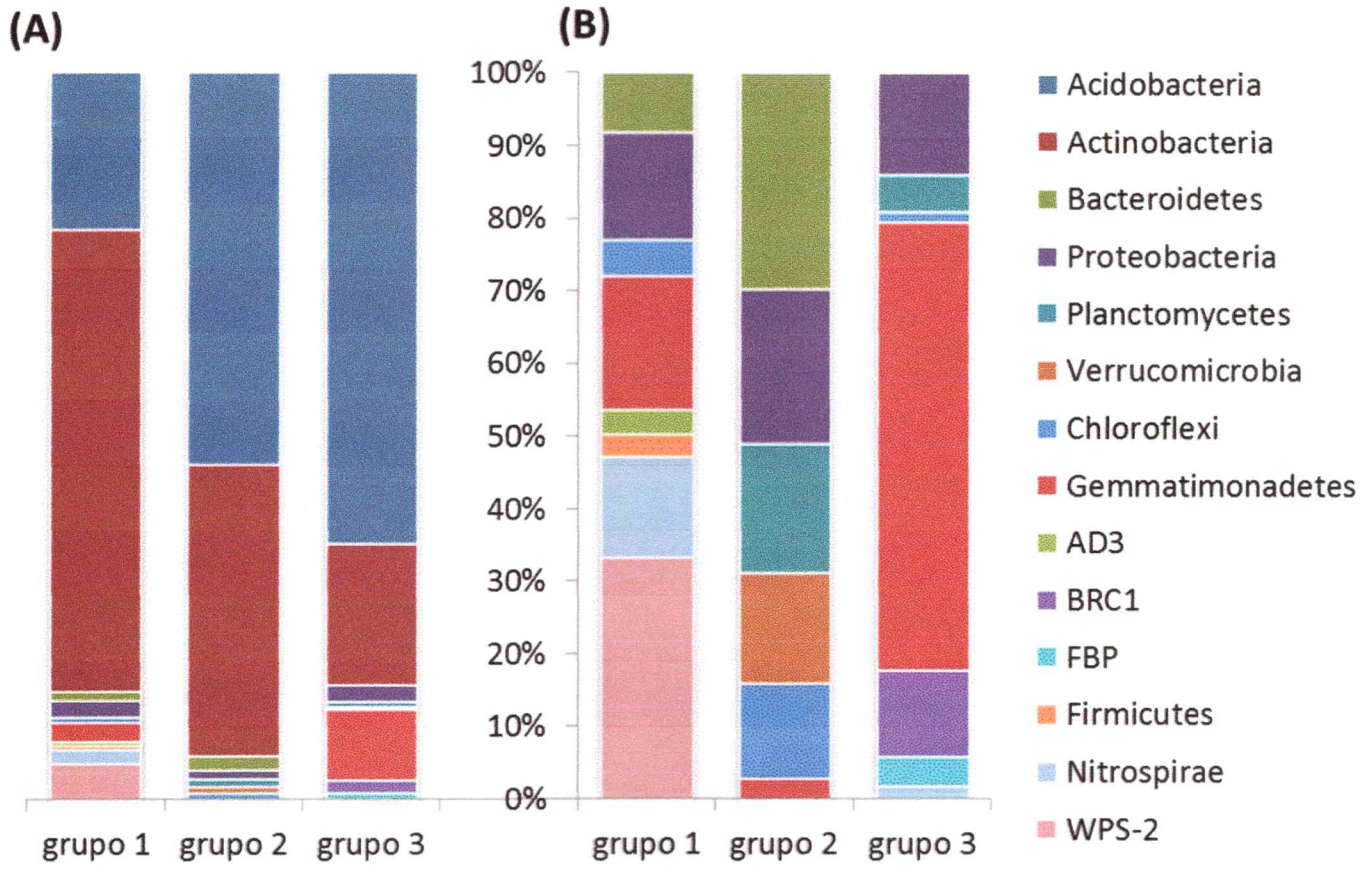


Figura 2.1.5. Caracterización taxonómica a nivel de filo de los 3 grupos de OTUs seleccionados. Cada barra representa el perfil de abundancia relativa a nivel de filo ordenada de mayor a menor utilizando como referencia el grupo g2 (pH=7.48). A) abundancias relativas considerando los 14 filos presentes. B) Abundancias relativas luego de remover los filos más abundantes (Acidobacteria y Actinobacteria).

2.2 Caracterización funcional a partir del análisis metagenómico de la comunidad bacteriana

La caracterización funcional de las bacterias pertenecientes a los 3 sitios seleccionados se realizó a partir de la secuenciación total del DNA extraído desde las muestras de suelo. Se aplicaron dos tipos de análisis: el primero consistió en caracterizar cada muestra como meta-organismo (metodología clásica) y el segundo consistió en discriminar genomas individuales dentro de cada muestra para su posterior caracterización funcional (objetivo general de esta tesis).

En la Tabla 2.2.1 se presentan el resumen de los datos crudos obtenidos luego de la secuenciación masiva de DNA extraído desde las tres muestras de suelo y el resultado del proceso de limpieza por calidad y secuencias adaptadoras.

Tabla 2.2.1. Estadísticas de secuenciación metagenómica para las 3 muestras de suelos seleccionados. Cada muestra se secuenció dos veces para lograr un tamaño total mínimo de 20 Gb (detalles en la sección Materiales y Métodos). Sufijo "_F" representa secuenciación por extremo "sentido" (forward) y sufijo "_R" representa secuenciación por extremo "anti-sentido" (reverse).

	Datos crudos			Datos procesados (calidad/adaptadores)		
Nombre de la Muestra	**Número de secuencias**	**Tamaño total (Gb)**	**Tamaño promedio (pb)**	**Número de secuencias**	**Tamaño total (Gb)**	**Tamaño promedio (pb)**
S1.1_F	33.087.050	4,9	148,8			
S1.1_R	33.087.050	4,9	148,8			
Total S1.1	66.174.100	9,8	148,8	62.816.508	8,7	138,2
S1.2_F	34.211.335	5,1	148,8			
S1.2_R	34.211.335	5,1	148,8			
Total S1.2	68.422.670	10,2	148,8	64.977.269	9,0	138,3
Total S1	**134.596.770**	**20,0**	**148,8**	**127.793.777**	**17,7**	**138,3**
S2.1_F	36.360.999	5,4	148,8			
S2.1_R	36.360.999	5,4	148,8			
Total S2.1	72.721.998	10,8	148.8	69.330.751	9,6	138,1
S2.2_F	37.634.627	5,6	148,8			
S2.2_R	37.634.627	5,6	148,8			
Total S2.2	75.269.254	11,2	148.8	71.776.738	9,9	138,2
Total S2	**147.991.252**	**22,0**	**148,8**	**141.107.489**	**19,5**	**138,2**
S3.1_F	39.313.488	5,8	148,8			
S3.1_R	39.313.488	5,8	148,8			
Total S3.1	78.626.976	11,7	148,8	75.268.496	10,3	137,0
S3.2_F	40.759.140	6,1	148,8			
S3.2_R	40.759.140	6,1	148,8			
Total S3.2	81.518.280	12,1	148,8	78.022.217	10,7	137,0
Total S3	**160.145.256**	**23,8**	**148,8**	**153.290.713**	**21,0**	**137,0**

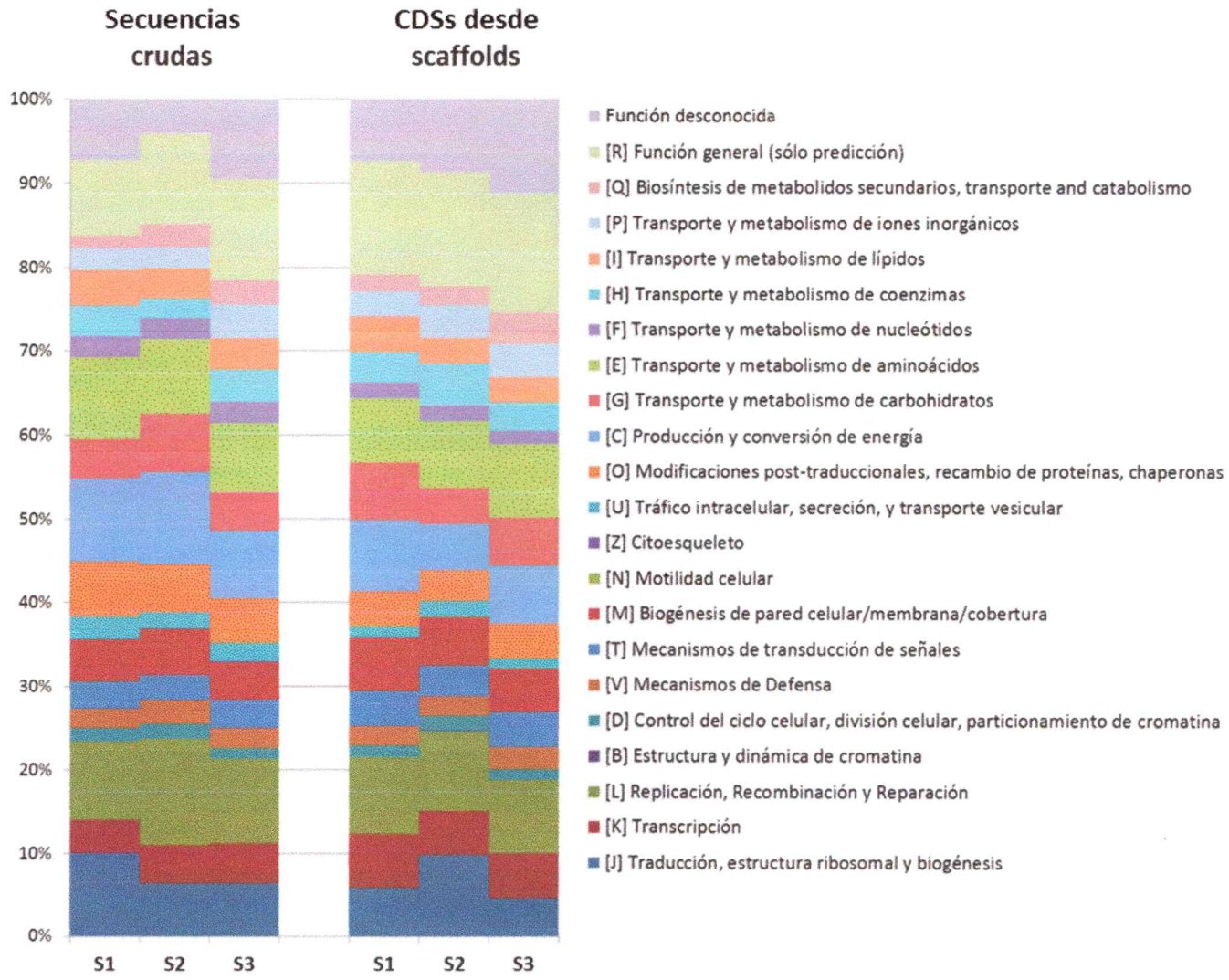
En primer lugar se procedió al ensamble de todas las secuencias obtenidas por sitio. Con este propósito se probaron diversos programas como Velvet, WGS-Assembler, IDBA-UD y CLC, de los cuales, el que entregó mejores resultados fue IDBA-UD (Peng y col., 2011). El mínimo de secuencias utilizadas por el ensamblador para construir los "scaffolds" fue de 32.2 % respecto del total disponible para la muestra a S3 lo cual es superior al estándar del 10 % descrito en publicaciones de metagenomas (Tabla 2.2.2). La estrategia de ensamble incluyo el uso de secuencias de pares de muestras (S2 + S1 y S2 + S3). Este ensamble se realizó con el propósito de maximizar la posibilidad de crear "scaffolds" de mayor tamaño y potenciar el ensamble de secuencias de DNA que están presentes en ambas muestras según se describe unos párrafos más adelante.

Tabla 2.2.2. Estadística de ensambles metagenómicos para las tres muestras de suelos seleccionados. Todas las muestras fueron ensambladas por separado. Además se realizaron ensambles de pares de muestras "S2 + S1" y "S2 + S3", en los cuales utilizaron las secuencias de ambas muestras.

	S1	S2	S3	S2 + S1	S2 + S3
Tamaño total (Mb)	823	830	678	1685	1542
Número de scaffolds	893.733	892.653	714.882	1.832.894	1.656.500
Tamaño promedio (pb)	920	929	948	919	931
Tamaño máximo (pb)	306.936	459.495	1.145.089	567.355	1.145.083
N50 (pb) y %N50	1.283 (13%)	1.334 (12%)	1.384 (11%)	1.290 (13%)	1.343 (11%)
% de lecturas utilizadas	38,3%	33,2%	32,2%	36,6%	34,5%
% GC	67,3%	66,6%	64,7%	66,9%	65,7%

Los 3 metagenomas fueron anotados funcionalmente contra la base de datos COG y el resultado fue comparado mediante el programa computacional MEGAN. En la Figura 2.2.1 se grafica la proporción relativa de categorías COG en el metagenoma de cada sitio.

Figura 2.2.1. Caracterización funcional (categorías COG) de los metagenomas. Se representa la abundancia relativa de los genes pertenecientes a las principales categorías funcionales presentes en los metagenomas tanto desde un análisis directo usando las secuencias pre-ensamble, y un análisis compuesto que considera los CDSs predichos a partir de los "scaffolds" generados.



En segundo lugar se procedió a la reconstrucción de genomas individuales. Para discriminar genomas individuales se procedió según el protocolo descrito por Albertsen *et* al (2013). Considerando que dicho método fue elaborado para discriminar genomas en muestras de baja complejidad (OTUs < 40, crecidas en un bioreactor), en este trabajo de tesis se realizó un esfuerzo importante en evaluar e implementar diversas modificaciones a este protocolo orientados a permitir el análisis de muestra con mayor complejidad (OTUs > 800). La modificación que tuvo mejor resultado fue tomar como base un ensamble híbrido de muestras de dos condiciones a comparar (S2 + S1 o S2 + S3). Este manejo de los datos fue extraordinariamente útil en incrementar el tamaño de los "scaffolds" y aumentar el número de secuencias del ensamble que están presentes en ambas muestras. En cambio, al hacer un ensamble hibrido con todas las muestras o con las muestras S1 y S3 se produjo una gran cantidad de quimeras de difícil resolución. Esto nos muestra que juntar poblaciones es una buena estrategia sólo hasta cierto límite el cual está determinado por el nivel de similitud de las mismas.

A partir de los ensambles de pares de muestras (S2 + S1 o S2 + S3) se obtuvieron 2 nuevos metagenomas con las características descritas en las últimas 2 columnas de la Tabla 2.2.2. Posteriormente las secuencias de los 3 sitios fueron mapeadas por separado para determinar la cobertura promedio de cada "scaffold" en cada muestra. También se determinó el contenido de GC y la frecuencia de tetrámeros de cada "scaffold". Las 3 variables fueron incluidas en un análisis de componentes principales (PCA) para identificar aquellos "scaffolds" que segregan juntos siguiendo los procedimientos descritos anteriormente (Albertsen y col., 2013; Alneberg y col., 2014). En el panel A de la Figura 2.2.2 se observa como ejemplo el agrupamiento de "scaffolds" de la muestra

"S2 + S3" graficado en función de los 2 primeros componentes principales donde cada "scaffold" está representado por un punto. Aquellos puntos de igual color componen un grupo o "cluster" que co-segrega en el espacio del PCA (espacio delimitado por elipses segmentadas). Debido a que estos grupos de "scaffolds" son utilizados para reconstruir estructuras genómicas que no representan necesariamente un único genoma (pueden ser representantes de un grupo de genomas o especies muy similares), se tomará la precaución de considerar dichos grupos como una unidad genómica operativa (conceptual o virtual) toda vez que su utilidad es similar al concepto de OTU. Por ello hemos acuñado la denominación gOTUs. Se obtuvieron 150 gOTUs que en su conjunto reclutaron 35 % de las bases secuenciadas y un 8 % del total de "scaffolds" ensamblados. Luego se evaluó la "completitud" de los gOTUs mediante la búsqueda y conteo de 36 genes descritos como: 1) presentes en todos los genomas de bacterias secuenciados y 2) que aparecen presentes en solo una copia (Alneberg y col., 2014). Basados en una inspección preliminar, se decidió trabajar con aquellos gOTUs que contenían al menos la mitad de los 36 genes, los cuales a su vez, no debían tener más de dos copias por "cluster" de "scaffolds". Los gOTUs que complían con este criterio fueron resaltados en el recuadro rojo del panel B de la Figura 2.2.2. Finalmente, considerando los dos ensambles híbridos, se obtuvieron 74 gOTUs los cuales fueron descompuestos en sus secuencias originales y re-ensamblados como conjunto (por gOTUs) mediante el programa computacional Velvet según lo recomendado por Alneberg y col. (2014).

Al hacer un recuento del número de secuencias por gOTUs se observaron diferencias en los niveles de abundancia relativa (cobertura) en los distintos suelos

examinados. Este dato permite clasificar los 74 gOTUs en 3 grandes grupos en función de su abundancia relativa (cobertura) de reads de cada sitio: grupo 1 más abundante en S1 (n=14), grupo 2 más abundante en muestra S2 (n=39) y grupo 3 más abundante en S3 (n=21) (Figura 2.2.3). En la Figura 2.2.4 y a modo de ejemplo, se representa en forma gráfica un ejemplo de 4 gOTUs con mayor abundancia en el S1 y 4 gOTUs del S3.

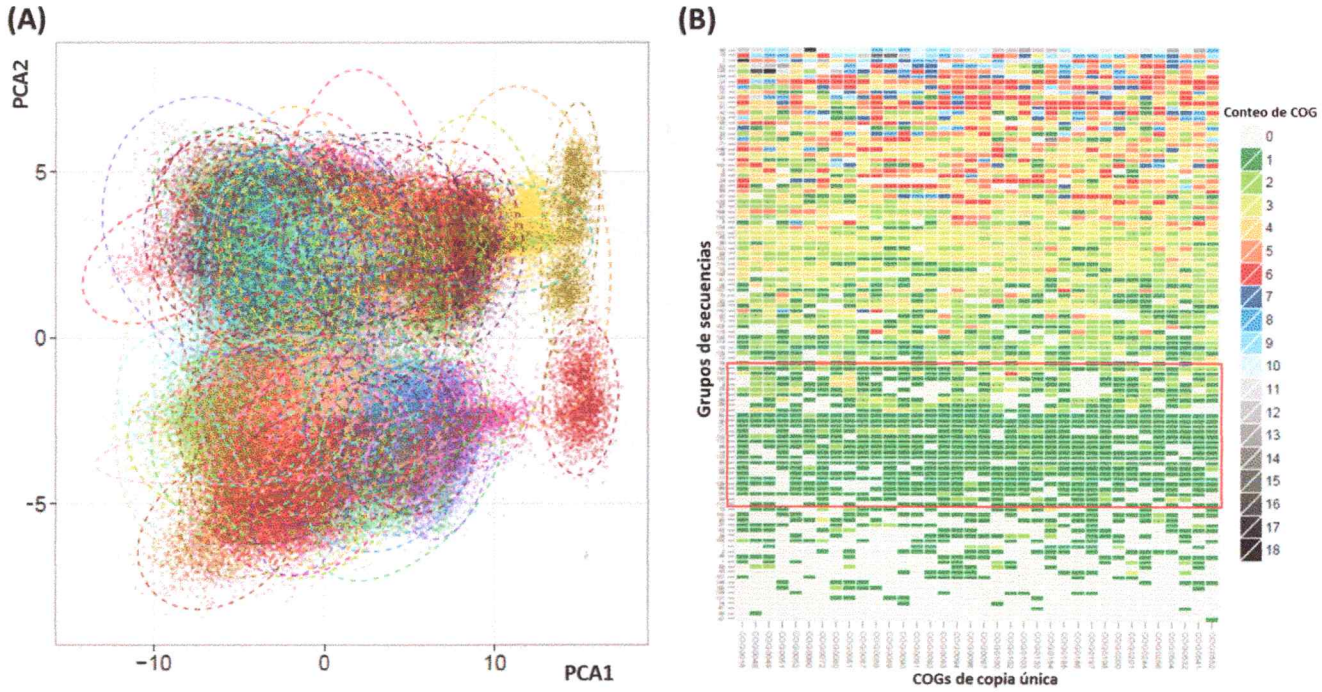


Figura 2.2.2. Criterios utilizados en la selección de los gOTUs de la muestra "S2 + S3". A) Agrupamiento de "scaffolds" mediante PCA. Cada "scaffold" está representado por un punto y su color representa el agrupamiento (o "clustering") realizado el cual también se destaca con óvalos del mismo color. **B)** Evaluación del número de copias de 36 genes considerados de copia única (columnas) [ver descripción en el texto] en los 150 gOTUs identificados (filas). El código de colores indica el número de copias identificado. Se destaca con un rectángulo rojo los "clusters" seleccionados por tener al menos 18 genes de copia única con 1 o 2 copias.

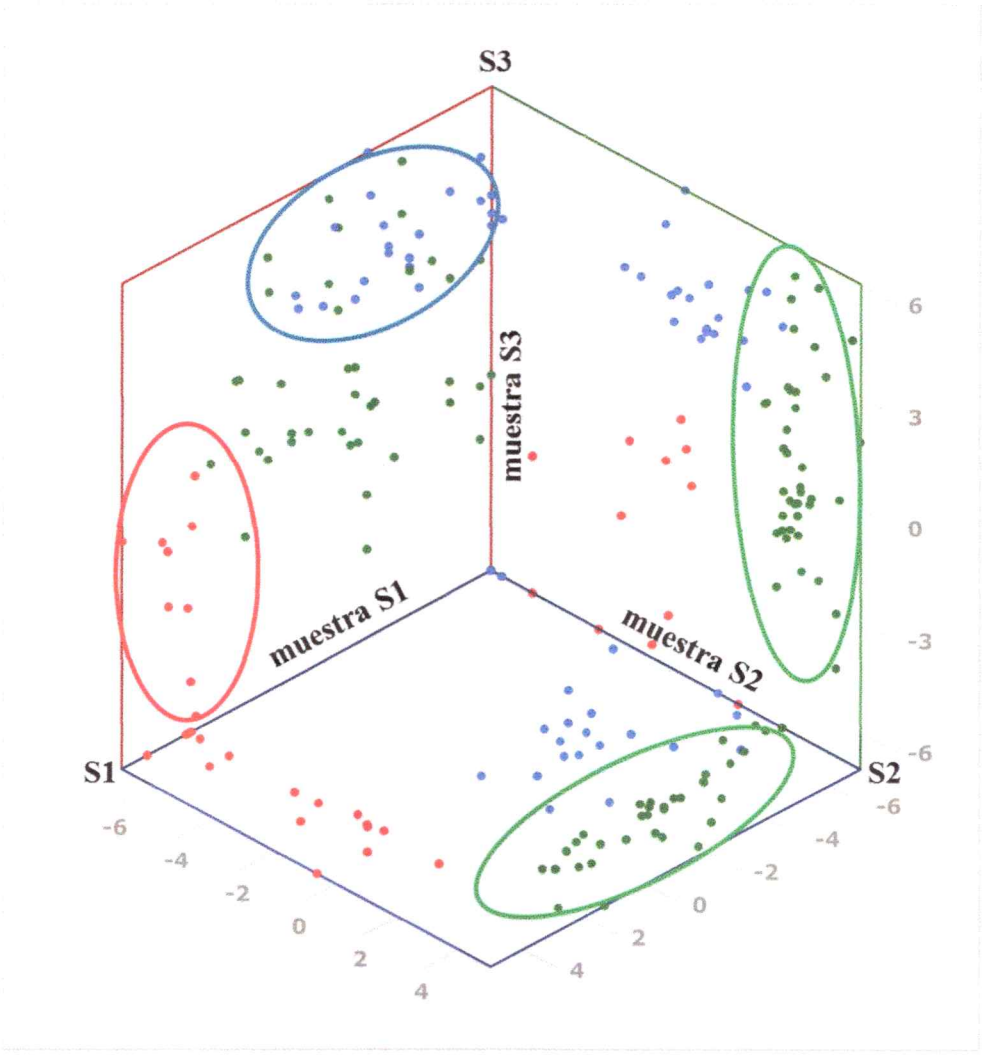


Figura 2.2.3. Representación de la cobertura promedio de cada gOTU identificado. El gráfico representa la cobertura promedio de los 74 gOTUs identificados según las 3 comparaciones posibles: S1/S3, S1/S2 y S3/S3. Cada eje representa la cobertura en la muestra S1, S2 o S3 según corresponda. Los gOTUs que segregan en la proximidad de los ejes de S1, S2 o S3 se demarcan dentro de óvalos rojo, verde y azul respectivamente.

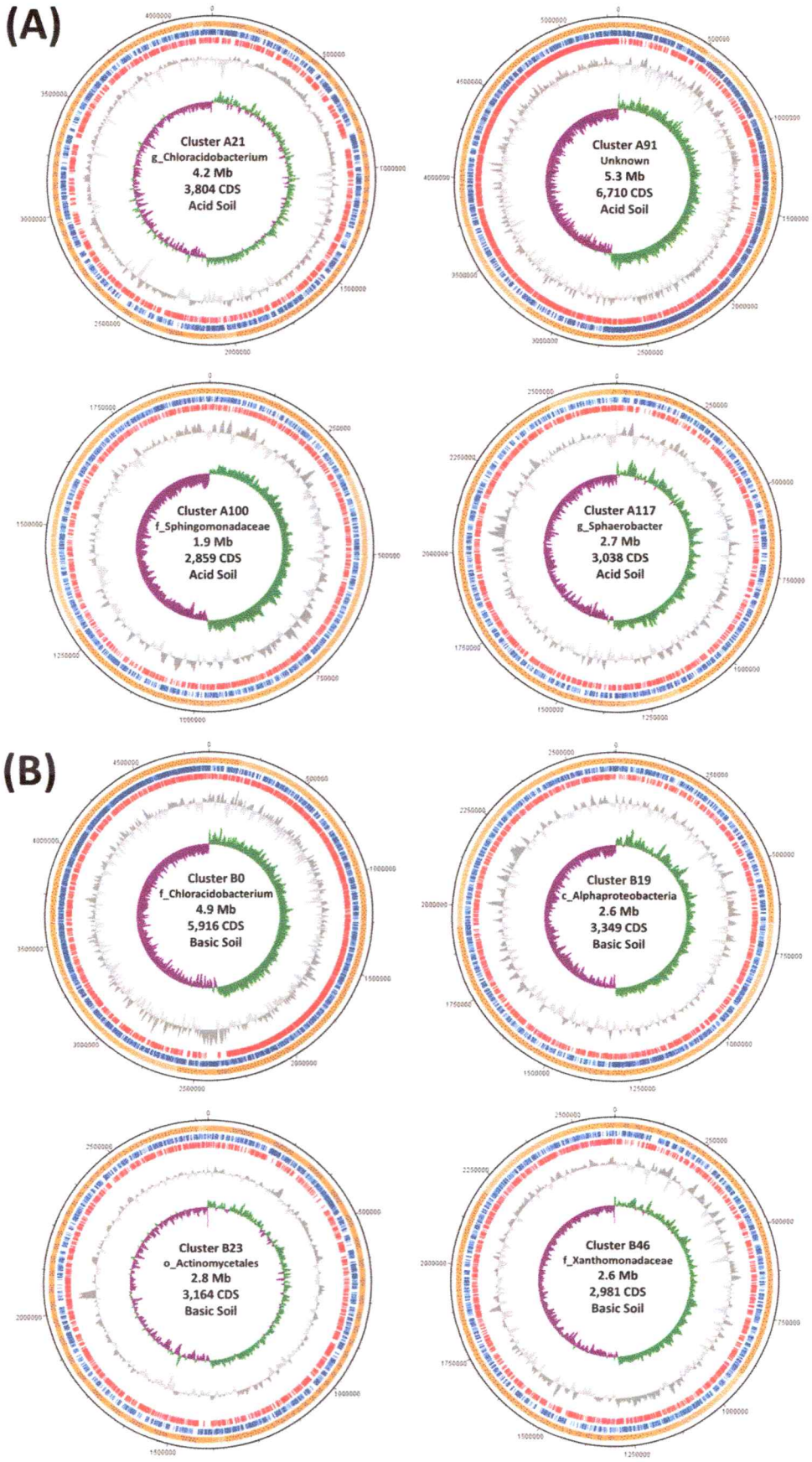

(A)
Cluster A21
g_Chloracidobacterium
4.2 Mb
3,804 CDS
Acid Soil
Cluster A91
Unknown
5.3 Mb
6,710 CDS
Acid Soil
Cluster A100
f_Sphingomonadaceae
1.9 Mb
2,859 CDS
Acid Soil
Cluster A117
g_Sphaerobacter
2.7 Mb
3,038 CDS
Acid Soil
(B)
Cluster B0
f_Chloracidobacterium
4.9 Mb
5,916 CDS
Basic Soil
Cluster B19
c_Alphaproteobacteria
2.6 Mb
3,349 CDS
Basic Soil
Cluster B23
o_Actinomycetales
2.8 Mb
3,164 CDS
Basic Soil
Cluster B46
f_Xanthomonadaceae
2.6 Mb
2,981 CDS
Basic Soil

Figura 2.2.4. Ejemplo de genomas reconstruidos. Cada grafica corresponde a un gOTU. El círculo naranjo representa los "scaffolds", los cuales están ordenados por sesgo de GC (GC skew) y luego por largo. En círculos azul y rojo se representa los CDSs en sentido forward y reverse respectivamente. En círculo gris se representa el contenido de GC respecto al promedio y el círculo interno representa el sesgo de GC respecto del promedio. **A)** 4 genomas de muestra S1. **B)** 4 genomas de muestra S3.

El genoma de los 74 gOTUs fue anotado funcionalmente contra la base de datos COG y el resultado fue comparado mediante el programa computacional MEGAN. Algunos estadísticos de la primera fase de anotación se detallan en la Tabla 2.2.3 la cual contiene un resumen de los datos generales obtenidos para los 74 gOTUs incluyendo su clasificación taxonómica. La Figura 2.2.5 organiza los gOTUs en un cladograma considerando el número de gOTUs identificado en cada taxón separados en base a su distribución preferente entre los sitios analizados. Los datos indican que los gOTUs pertenecen a 7 filos donde el filo Actinobacteria es el más abundante en los 3 grupos de gOTUs seguido de Acidobacteria y Proteobacteria. El género candidato *Chloracidobacterium* (bacteria fotosintéticas del filo Acidobacteria) es el más representado con 14 miembros de la muestra S2, 6 de la S3 y ninguno en la S1.

En la Figura 2.2.6 se grafica la relación de vecindad entre gOTUs y diferentes ensambles de metagenomas. La vecindad entre los gOTUs y metagenomas se realizó utilizando las categorías funcionales SEED (ver detalles en sección Materiales y Métodos). Los círculos encierran aquellos casos de agrupamiento entre gOTUs que tienen una distribución preferente al mismo tipo de pH. Por ejemplo, A99, A91 y A58

(en el extremo inferior del gráfico), representan tres gOTUs abundantes en S1. Sin embargo, se observa que el mapeo de secuencias pre-ensamble (muestras con nombres terminados en "reads") y el mapeo de CDSs desde el ensamble (muestras con nombres terminados en "idba") de los tres sitios analizados presentan un alto grado de similitud funcional (Figura 2.2.6, rectángulos de línea segmentada). Este resultado indica que la descripción funcional de los metagenomas está fuertemente ligada a la técnica utilizada y no representa necesariamente lo que ocurre con la población a nivel de genomas particulares. Siguiendo la línea de análisis asociado a la Figura 2.2.6, se realizó un análisis más detallado de los genes que codifican para proteínas vinculadas a la respuesta a estrés de los genomas individuales reconstruidos (Figura 2.2.7). Este grupo de proteínas fue seleccionado considerando su potencial importancia para la adaptación de los organismos que se encuentran en ambientes extremos. Se designó como grupo 1 al resultado de la asignación funcional usando sólo secuencias (procedimiento clásico). El grupo 2 representa la asignación funcional usando los genes predichos desde el ensamble de secuencias (metagenomas). El grupo 3 representa la asignación funcional usando los genes de todos los gOTUs mezclados como un solo metagenoma. El grupo 4 representa la asignación funcional usando los genes predichos en cada genoma reconstruido. En la Figura 2.2.7 (panel A) se observa que la abundancia relativa de los genes de respuesta a estrés es similar entre los tres sitios analizados (barra roja, verde y azul) e independiente de la estrategia de ensamble de los metagenomas (columnas 1, 2 y 3). Sin embargo, la abundancia relativa de los genes de respuesta a estrés muestra una mayor variación entre los gOTUs (columna 4). La categoría de respuesta a estrés se compone de varias sub-categorías relacionadas, entre ellas encontramos estrés oxidativo

y estrés térmico que contienen el mayor número de genes en la mayoría de los gOTUs reconstruidos con un comportamiento similar al descrito para estrés oxidativo (Figura 2.2.7). Las otras subcategorías, al contener una abundancia relativa menor, presentan una mayor diferencia entre los genomas ensamblados (columnas 1, 2 y 3) y entre los gOTUs (columna 4). Por ejemplo, la categoría estrés ácido en la columna 1 no está representada, y en la columna 2 sólo está representada en la muestra básica. En cambio, al observar los gOTUs encontramos que la categoría está representada en el 20 % de los genomas reconstruidos y de forma transversal a la muestra de origen. Por lo anterior, la descripción de los metagenomas de las columnas 1, 2 e incluso 3 son necesarias pero no suficientes al momento de realizar descripciones y comparaciones entre comunidades en detalle. Las diferencias entre cada categoría serán presentadas con mayor detalle en el capítulo 3 de la presente tesis.

Tabla 2.2.3. Características genómicas generales de los gOTUs ensamblados. La inicial del nombre indica su origen: A = pH ácido (S1), B = pH básico (S3), AN o BN = pH neutro (S2). La tabla está ordenada según el nombre de los genomas.

Nombre	Tamaño (Mb)	# de Scaffolds	largo N50 (pb)	# de CDSs	# de genes con COG	ID taxonomía	Taxonomía
A9	24.02	9,644	2,981	30,166	432	2	Bacteria;
A21	4.20	289	24,183	3,804	141	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
A25	6.65	1,301	5,862	6,923	259	940557	Fibrobacteres/Acidobacteria group;Acidobacteria;Acidobacteriia;Acidobacteriales;Acidobacteriaceae;Granulicella;
A38	4.00	109	68,743	3,590	612	2	Bacteria;
A58	3.61	608	7,343	4,053	103	2056	Chloroflexi <phylum>;Thermomicrobia;Sphaerobacteridae;Sphaerobacterales;Sphaerobacterineae;Sphaerobacteraceae;Spha
A59	1.48	1,032	1,379	1,952	77	173479	Gemmatimonadetes;Gemmatimonadetes <class>;Gemmatimonadales;Gemmatimonadaceae;Gemmatimonas;
A72	8.89	733	18,602	10,262	399	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
A77	8.48	5,801	1,424	12,601	221	191494	Actinobacteria <phylum>;Actinobacteria;Rubrobacteridae;Solirubrobacterales;Conexibacteraceae;Conexibacter;
A91	5.31	1,951	3,138	6,710	484	2	Bacteria;
A99	4.68	607	9,440	5,298	536	2	Bacteria;
B0	4.86	2,417	2,124	5,916	121	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
B6	11.24	1,009	23,252	11,818	321	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
B16	5.06	566	16,549	5,224	84	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
B19	2.63	1,008	2,885	3,349	75	28211	Proteobacteria;Alphaproteobacteria;
B22	1.29	713	1,881	1,854	675	42255	Actinobacteria <phylum>;Actinobacteria;Rubrobacteridae;Rubrobacterales;Rubrobacterineae;Rubrobacteraceae;Rubrobacter
B23	2.78	197	24,843	3,164	156	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
B34	5.38	132	65,992	4,685	114	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
B46	2.61	742	4,488	2,981	674	32033	Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;
B66	2.31	461	5,251	2,781	292	2070	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;Pseudonocardineae;Pseudonocardiaceae;
B80	5.67	259	46,028	5,108	59	332162	Fibrobacteres/Acidobacteria group;Acidobacteria;Solibacteres;Solibacterales;Solibacteraceae;Candidatus Solibacter;
B83	2.06	422	5,884	2,410	144	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
B92	4.94	2,907	1,713	6,558	63	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
B102	2.70	910	4,686	3,678	112	95818	unclassified Bacteria;Candidatus Saccharibacteria;
B105	2.33	1,405	1,683	3,013	53	239934	Chlamydiae/Verrucomicrobia group;Verrucomicrobia;Verrucomicrobiae;Verrucomicrobiales;Verrucomicrobiaceae;Akkermar
B109	4.60	327	23,155	4,795	655	2	Bacteria;
B111	8.13	5,741	1,383	12,041	36	1224	Proteobacteria;
B117	3.08	1,208	2,846	3,774	608	2	Bacteria;

B126	5.08	1,804	3,038	6,469	155	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
B128	11.93	519	79,216	10,215	99	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
B143	4.92	3,504	1,353	7,216	142	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
B145	6.80	723	11,956	7,523	141	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
A100	1.94	1,144	1,746	2,859	272	41297	Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;
A117	2.68	576	5,423	3,038	116	2056	Chloroflexi <phylum>;Thermomicrobia;Sphaerobacteridae;Sphaerobacterales;Sphaerobacterineae;Sphaerobacteraceae;Sphae
A135	6.37	1,930	3,584	7,399	104	173479	Gemmatimonadetes;Gemmatimonadetes <class>;Gemmatimonadales;Gemmatimonadaceae;Gemmatimonas;
A139	3.78	938	5,309	4,843	517	42255	Actinobacteria <phylum>;Actinobacteria;Rubrobacteridae;Rubrobacterales;Rubrobacterineae;Rubrobacteraceae;Rubrobacter;
AN14	18.43	10,615	1,564	26,855	149	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
AN17	1.61	759	2,345	2,063	244	1236	Proteobacteria;Gammaproteobacteria;
AN19	6.18	2,162	3,636	7,672	116	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
AN22	1.96	356	5,987	2,239	688	2	Bacteria;
AN23	8.32	4,262	2,050	10,979	58	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
AN34	1.46	894	1,631	2,081	281	2070	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;Pseudonocardineae;Pseudonocardiaceae;
AN35	1.78	919	2,021	2,267	142	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
AN60	3.54	602	8,649	3,636	138	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
AN66	3.45	2,449	1,369	4,667	69	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
AN84	7.49	467	24,244	7,701	99	41297	Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;
AN85	2.70	351	9,523	3,033	205	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
AN92	6.26	4,132	1,478	8,869	337	2	Bacteria;
AN98	4.21	480	11,072	4,137	58	332162	Fibrobacteres/Acidobacteria group;Acidobacteria;Solibacteres;Solibacterales;Solibacteraceae;Candidatus Solibacter;
AN102	0.99	568	1,853	1,545	165	191494	Actinobacteria <phylum>;Actinobacteria;Rubrobacteridae;Solirubrobacterales;Conexibacteraceae;Conexibacter;
AN112	4.94	337	21,744	4,442	122	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
AN114	7.35	249	76,388	6,916	133	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
AN118	4.15	286	21,692	3,699	135	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
AN129	4.16	786	5,428	4,992	210	2070	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;Pseudonocardineae;Pseudonocardiaceae;
AN132	2.43	583	5,261	2,943	160	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
AN141	15.04	10,485	1,415	23,538	94	191494	Actinobacteria <phylum>;Actinobacteria;Rubrobacteridae;Solirubrobacterales;Conexibacteraceae;Conexibacter;

BN5	10.92	1,050	19,798	11,152	73	41297	Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;
BN21	9.93	6,581	1,507	14,680	107	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
BN37	2.66	741	3,962	2,716	211	126	Planctomycetes;Planctomycetia;Planctomycetales;Planctomycetaceae;
BN51	5.85	530	20,116	5,413	114	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
BN58	2.63	180	34,331	2,855	284	41297	Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;
BN61	3.42	120	41,954	3,530	608	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
BN70	4.17	275	24,143	3,712	139	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
BN71	4.36	910	5,641	5,098	119	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
BN73	5.29	3,309	1,572	7,515	234	191494	Actinobacteria <phylum>;Actinobacteria;Rubrobacteridae;Solirubrobacterales;Conexibacteraceae;Conexibacter;
BN82	3.78	560	8,320	4,377	206	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
BN84	8.09	5,617	1,406	11,898	41	1224	Proteobacteria;
BN110	7.29	258	73,495	6,906	133	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
BN115	3.97	2,282	1,806	5,091	144	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
BN120	4.62	2,827	1,622	6,803	613	42255	Actinobacteria <phylum>;Actinobacteria;Rubrobacteridae;Rubrobacterales;Rubrobacterineae;Rubrobacteraceae;Rubrobacter;
BN123	10.63	758	37,852	11,483	121	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;
BN127	4.47	2,688	1,682	5,865	65	458032	Fibrobacteres/Acidobacteria group;Acidobacteria;unclassified Acidobacteria;Candidatus Chloracidobacterium;
BN130	5.28	389	20,707	5,335	69	332162	Fibrobacteres/Acidobacteria group;Acidobacteria;Solibacteres;Solibacterales;Solibacteraceae;Candidatus Solibacter;
BN134	2.01	351	6,466	2,314	700	2	Bacteria;
BN144	7.83	4,866	1,642	12,170	269	2037	Actinobacteria <phylum>;Actinobacteria;Actinobacteridae;Actinomycetales;

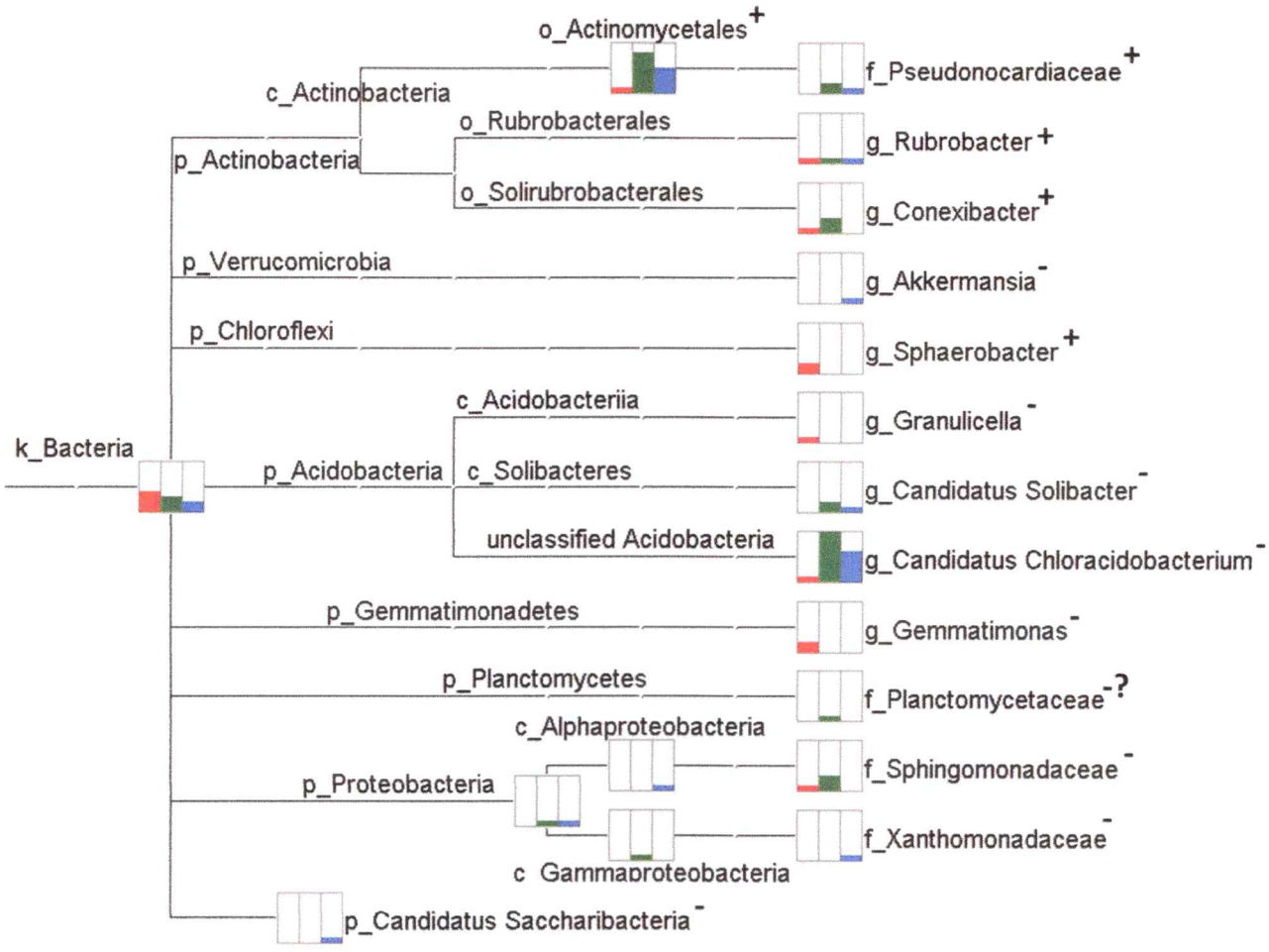


Figura 2.2.5. Relación filogenética de los gOTUs y número de gOTUs identificado en cada taxón. La identificación de taxonomía se realizó mediante el mapeo de CDSs de cada genoma contra la base de datos COG de NCBI. En el cladograma los gOTUs se contabilizan considerando su abundancia preferente en los sitios S1, S2 y S3, representados por barrar rojas, verdes y azules respectivamente. Cada barra representa como máximo 10 elementos. La única barra que superaba ese valor es el género Candidato *Chloracidobacterium* en la muestra S2 con 14 gOTUs, quedando reducida a 10 para mejorar la visualización general. Los taxas incluyen como prefijo la letra inicial del nivel taxonómico al cual corresponde y con símbolo "+" o "–" se ha indicado su carácter gram positivo o negativo respectivamente.

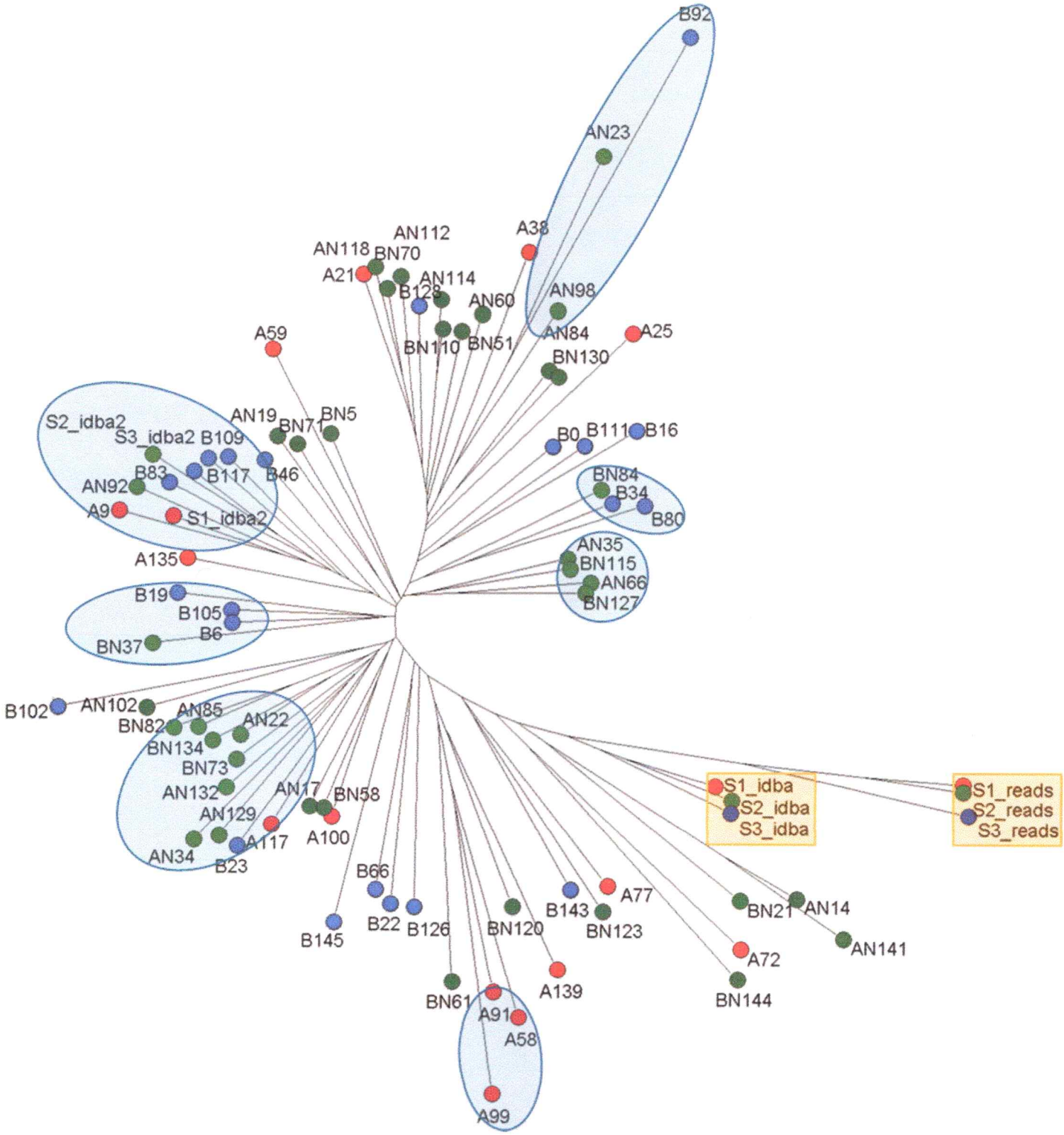

B92
AN23
AN112
A38
AN118
BN70
A21
AN114
B128
AN60
AN98
A59
BN110
BN51
AN84
A25
BN130
S2_idba2
AN19
BN71
BN5
S3_idba2
B109
B0
B111
B16
B83
B117
B46
AN92
BN84
B34
A9
S1_idba2
B80
A135
AN35
BN115
AN66
BN127
B19
B105
B6
BN37
B102
AN102
AN85
AN22
BN82
BN134
BN73
AN132
AN17
BN58
AN129
A100
A117
AN34
B23
S1_idba
S2_idba
S3_idba
S1_reads
S2_reads
S3_reads
B66
A77
B143
BN21
AN14
B22
B126
BN120
BN123
B145
AN141
A72
A139
BN144
BN61
A91
A58
A99

Figura 2.2.6: Árbol de vecindad (*Neighbor-tree*) utilizando categorías SEED y distancia de Bray-Curtis de los gOTUs y los diferentes metagenomas ensamblados. Este análisis considera la abundancia que cada gOTU presenta en todas las categorías funcionales SEED hasta el nivel 2 y las compara mediante distancia Bray-Curtis para determinar vecindad. Círculos rojos, verdes y azules representan muestras provenientes de suelo S1, S2 y S3 respectivamente. Muestras terminadas en "reads" representan el análisis a partir de las lecturas originales. Muestras con nombres terminados en "idba" representan el análisis de los ensambles de metagenomas Muestras con nombres terminados en "idba2" representan metagenomas construidos desde el conjunto de los gOTUs. A modo representativo se han encerrado en círculos o cuadrados los elementos que pertenecen a un mismo clado.

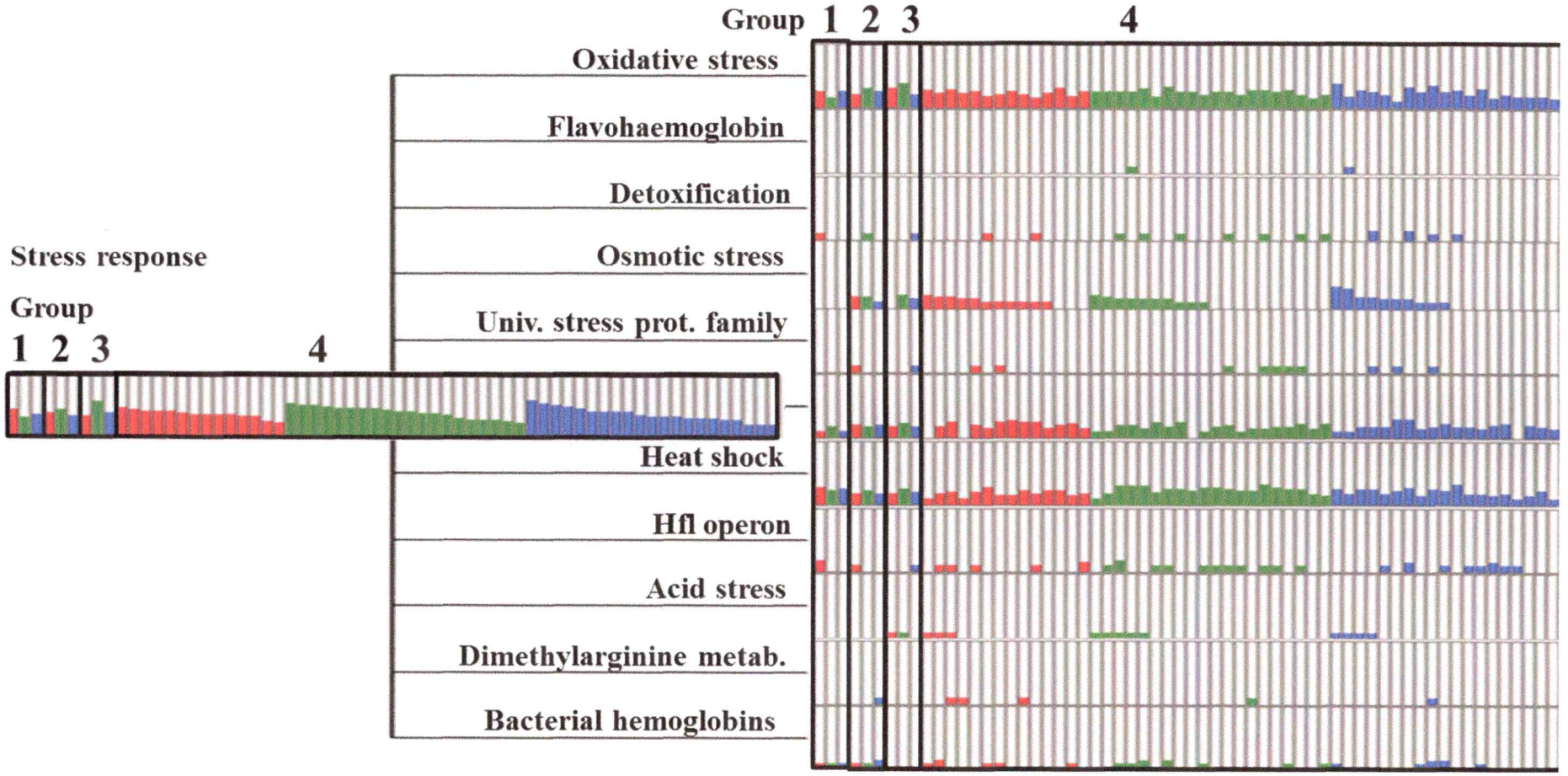

Group 1 2 3 4
Oxidative stress
Flavohaemoglobin
Detoxification
Osmotic stress
Univ. stress prot. family
Heat shock
Hfl operon
Acid stress
Dimethylarginine metab.
Bacterial hemoglobins
Stress response
Group
1 2 3 4

Figura 2.2.7: Comparación funcional para la categoría SEED Respuesta Estrés y sus categorías internas. Cada barra representa la abundancia relativa de los genes de cada categoría. El grupo 1 representa la asignación funcional usando sólo secuencias (procedimiento clásico). El grupo 2 representa la asignación funcional usando los genes predichos desde el ensamble de secuencias (metagenomas). El grupo 3 representa la asignación funcional usando los genes de todos los gOTUs mezclados como un solo metagenoma. El grupo 4 representa la asignación funcional usando los genes predichos en cada genoma reconstruido. Barras rojas, verdes y azules representan la asociación de cada columna (muestras) a sitios S1, S2 y S3 respectivamente.

2.3 Relación entre las variables ambientales y los cambios de abundancia relativa de OTUs identificados y gOTUs reconstruidos

Con el propósito de generar una visión integrada de los datos de secuencias genómicas y de variables ambientales obtenidos en los 3 sitios muestreados (S1, S2 y S3), se realizó un Análisis de Correspondencia Canónica (CCA), una técnica de análisis de regresión multivariado (Palmer, 1993; Legendre & Legendre, 1998) que permite analizar la relación entre dos grupos de variables. El primer grupo de variables incluye la abundancia relativa de: los filos identificados, los OTUs seleccionados por SOM y los gOTUs. El segundo grupo de variables consiste en las variables ambientales (fisicoquímicos y nutrientes) determinadas para las mismas muestras. En ambas matrices las variables son columnas mientras que los sitios son filas. Básicamente la técnica de CCA limita la ordenación de la matriz de factores ambientales por una regresión lineal múltiple de las abundancias de filos. Los resultados de este análisis se presentan en un gráfico separado en 2 paneles para simplificar su interpretación (Figura 2.3.1). La gráfica incluye los 3 grupos de OTUs seleccionados mediante SOM (ver Figura 2.1.4) (círculos abiertos) y los genomas reconstruídos (ver Figura 2.2.3) (círculos cerrados). Se observa en el panel A de la Figura 2.3.1 que los sitios se ordenan en 3 vértices, cada uno con sus respectivas réplicas. Cada filo posee una ordenación (posición) particular dependiendo de su tendencia a estar en una u otra muestra donde destacan los siguientes filos: AD3 y WPS3 en el sitio S1, OD1 en el sitio S2 y Fibrobacteres, TM6 y BRC1 en sitio S3. Se observa que los OTUs elegidos mediante SOM efectivamente poseen una preferencia hacia distintos sitios y los genomas reconstruidos poseen una preferencia mucho más marcada, en especial para S1 (puntos rojos) y S3 (puntos azules). Respecto

de las variables ambientales, se observa en el panel B de la Figura 2.3.1, que diferentes grupos de variables presenta ordenación diferencial entre los sitios, indicando que otros parámetros ambientales segregan junto con el pH entre sitios muestreados.

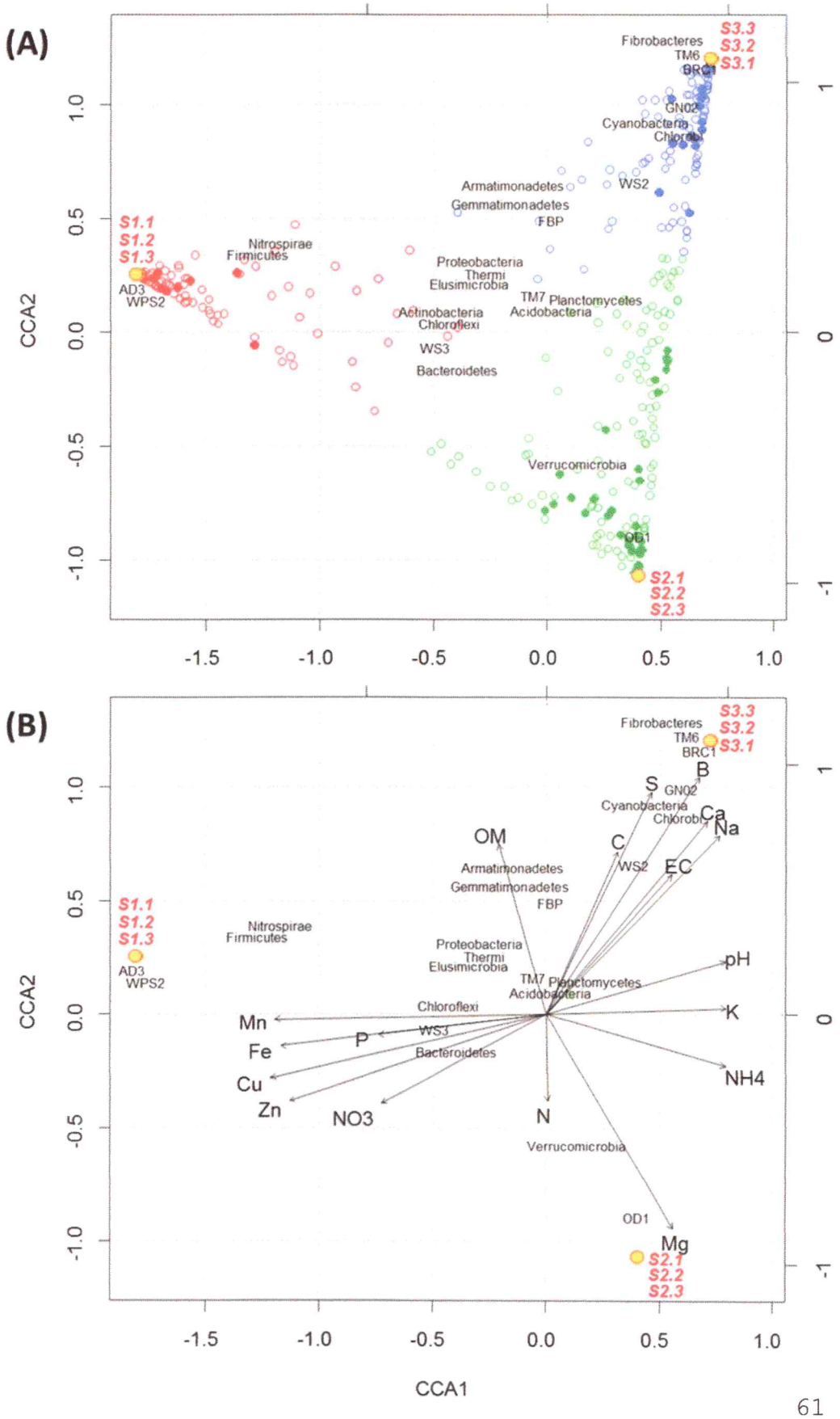

(A)
Fibrobacteres
TM6
BRC1
S3.3
S3.2
S3.1
GN02
Cyanobacteria
Chloroflexi
Armatimonadetes
Gemmatimonadetes
FBP
WS2
S1.1
S1.2
S1.3
Nitrospirae
Firmicutes
AD3
WPS2
Proteobacteria
Thermi
Elusimicrobia
TM7
Planctomycetes
Acidobacteria
Actinobacteria
Chloroflexi
WS3
Bacteroidetes
Verrucomicrobia
OD1
S2.1
S2.2
S2.3
CCA2
(B)
B
S
C
Ca
Na
OM
EC
pH
K
NH4
Mn
Fe
Cu
Zn
P
NO3
N
Mg
CCA1

Figura 2.3.1. Biplot del Análisis de Correspondencia Canónica entre filos identificados y factores ambientales determinados en las 3 muestras de suelo con sus respectivos triplicados. El análisis se realizó utilizando la abundancia relativa de filos en cada muestra por triplicado agregando además la abundancia de los OTUs seleccionados por SOM y la abundancia relativa de los genomas reconstruidos. Como segunda matriz se utilizaron los datos de factores ambientales determinados para las mismas muestras. **A)** representa el ordenamiento de filos, OTUs y genomas reconstruídos. **B)** representa el ordenamiento de filos y las correlaciones del ordenamiento de los sitios con los factores ambientales. Círculos amarillos representan los sitios de muestreo con sus triplicados. Círculos abiertos representan los OTUS seleccionados por SOM coloreados según preferencia de sitio con rojo, verde y azul para sitios S1 (n=93), S2 (n=108) y S3 (n=93) respectivamente. Círculos cerrados representan los genomas reconstruidos coloreados según preferencia de sitio con rojo, verde y azul para sitios S1 (n=14), S2 (n=39) y S3 (n=21) respectivamente. Las flechas de cada factor ambiental representan la dirección e intensidad de la correlación respecto de los 3 sitios.

3. Análisis comparativo de los elementos de homeostasis de pH en los genomas ensamblados

3.1 Análisis de enriquecimiento en categorías SEED

Para revisar las posibles implicancias funcionales asociadas a las diferencias ambientales descritas y avanzar en el entendimiento de los mecanismos de homeostasis de pH involucrados se realizó un análisis de enriquecimiento de funciones en base a la categorización SEED obtenida en el capítulo 2.2. Se empleó para ello un Test Exacto de Fisher con punto de corte de valor p (probabilidad de error) menor a 0.001 (indicado como "p value" en las tablas 3.1.1 a 3.1.4). En la Tabla 3.1.1 se presenta el listado de las 14 categorías SEED enriquecidas en los genomas más abundantes en sitio S1 respecto de S2 y en la Tabla 3.1.2 se presenta las 15 categorías enriquecidas en el sentido contrario, es decir, más presentes en S2 respecto de S1. Descontando las 4 categorías comunes que se intersectan en Nivel 1 de la categorización SEED (Sub-sistemas de agrupamiento, Transporte de Membrana, Metabolismo de Nitrógeno, Respiración), se observa que en Nivel 1, existen 6 categorías enriquecidas únicas de S1:

- Aminoácidos y Derivados
- Carbohidratos
- Cofactores, Vitaminas, Grupos Prostéticos, Pigmentos
- Metabolismo de Compuestos Aromátics
- Nucleósidos y nucleótidos
- Metabolismo de Azufre

Y existen 7 categorías únicas de S2:

- Metabolismo de DNA
- Adquisición de hierro y metabolismo
- Metabolismo de Proteínas
- Regulación y Señalización Celular
- Metabolismo de RNA
- Metabolismo Secundario
- Virulencia, Enfermedad y Defensa

En la Tabla 3.1.3 se presenta la única categoría enriquecida en S3 respecto de S2: "Resistencia a antibióticos y compuestos tóxicos", mientras que en la Tabla 3.1.4 se presenta también la única categoría enriquecida en S2 respecto de S3: "Ácidos orgánicos".

Tabla 3.1.1. Listado de categorías SEED enriquecidas en genomas más abundantes en S1 respecto de S2. Se realizó el análisis sobre las categorías del nivel 2 de la categorización SEED y en la tabla se presenta tanto el nivel 2 como el nivel 1. El punto de corte del Test Exacto de Fisher fue un pvalue máximo de 0.001 para la comparación S1 vs S2. Se presenta también el pvalue obtenido en las otras comparaciones. En negrita se encuentran resaltados los pvalue menores al punto de corte.

Categoría SEED		pvalue		
Nivel 2	Nivel 1	**S1 vs S2**	S3 vs S2	S1 vs S3
Aminoácidos de cadena ramificada	Aminoácidos y derivados	**6.7E-05**	3.1E-03	**3.1E-09**
Di y oligosacáridos	Carbohidratos	**1.3E-14**	7.2E-01	**1.3E-10**
Glicosido-hidrolasas	Carbohidratos	**1.7E-04**	1.4E-01	6.2E-02
Monosacáridos	Carbohidratos	**1.4E-09**	9.3E-01	**1.7E-07**
Transporte putativo de grupo hemin	Subsistemas basados en agrupamiento	**1.3E-12**	3.5E-01	**9.3E-07**
Tetrapiroles	ıctores, Vitaminas, Grupos Prostéticos, Pigmentos	**2.0E-04**	2.4E-01	3.0E-02
Transportadores ABC	Transporte de Membrana	**7.4E-15**	4.1E-02	**1.6E-20**
Degradación de Gentisare	Metabolismo de compuestos aromáticos	**9.3E-05**	1.0E+00	**9.8E-04**
Metabolismo de intermediarios aromáticos centrales	Metabolismo de compuestos aromáticos	**-2.2E-15**	1.0E+00	**1.4E-13**
Nitritoreductasa disimilatoria	Metabolismo de Nitrógeno	**7.1E-04**	5.9E-01	3.6E-02
Metabolismo de Hidantoin	Nucleósidos y Nucleótidos	**4.9E-16**	1.0E+00	**8.6E-12**
Factores de maduración de dehidrogenasas de monóxido de carbono	Respiración	**4.3E-04**	9.6E-02	**1.2E-05**
Reacciones aceptoras de electrónes	Respiración	**3.3E-05**	6.0E-02	5.3E-02
Formato hidrogenasa	Respiración	**1.3E-04**	8.2E-01	**7.4E-04**
Asimilación de azufre orgánico	Metabolismo de Azufre	**7.8E-15**	7.6E-01	**8.7E-17**

Tabla 3.1.2. Listado de categorías SEED enriquecidas en genomas más abundantes en S2 respecto de S1. Se realizó el análisis sobre las categorías del nivel 2 de la categorización SEED y en la tabla se presenta tanto el nivel 2 como el nivel 1. El punto de corte del Test Exacto de Fisher fue un pvalue máximo de 0.001 para la comparación S1 vs S2. Se presenta también el pvalue obtenido en las otras comparaciones. En negrita se encuentran resaltados los pvalue menores al punto de corte.

Categoría SEED		pvalue		
Nivel 2	Nivel 1	**S1 vs S2**	S3 vs S2	S1 vs S3
Cluster de genes conservado asociado con Met-tRNA_formiltransferasa	Subsistemas basados en agrupamiento	**1.1E-04**	8.4E-01	**4.5E-04**
CRISPRs	Metabolismo de DNA	**1.7E-08**	1.3E-02	2.9E-03
Sistema de modificación/Restricción	Metabolismo de DNA	**5.3E-15**	1.1E-02	**2.3E-06**
Adquisición de hierro en Vibrio	Adquisición de hierro y metabolismo	**3.7E-07**	4.7E-01	**5.7E-07**
Sideróforos	Adquisición de hierro y metabolismo	**2.8E-04**	4.8E-01	**1.6E-04**
Sistema de secreción de proteínas, Tipo II	Transporte de Membrana	**6.4E-06**	1.5E-01	7.6E-03
Sistemas de transporte Ton y Tol	Transporte de Membrana	**8.9E-20**	1.7E-03	**7.0E-26**
Fijación de Nitrógeno	Metabolismo de Nitrógeno	**9.3E-04**	8.8E-02	1.5E-01
Biosíntesis de proteínas	Metabolismo de Proteínas	**3.3E-06**	1.8E-01	4.3E-03
Señalización cAMP en bacteria	Regulación y señalizacion Celular	**5.0E-14**	6.1E-01	**1.1E-09**
Hidrogenasa inducida por monóxido de carbono	Respiración	**1.2E-04**	1.0E+00	**5.1E-04**
Genes asociados a intrones del Grupo II	Metabolismo de RNA	**9.8E-05**	4.2E-02	1.0E-01
Compuestos biológicos activos en metazoos para defensa y diferenciación	Metabolismo secundario	**4.0E-06**	1.7E-01	5.2E-03
Invación y resistencia intracelular	Virulencia, Enfermedad y Defensa	**9.1E-04**	8.3E-01	1.5E-03

Tabla 3.1.3. Listado de categorías SEED enriquecidas en genomas más abundantes en S3 respecto de S2. Se realizó el análisis sobre las categorías del nivel 2 de la categorización SEED y en la tabla se presenta tanto el nivel 2 como el nivel 1. El punto de corte del Test Exacto de Fisher fue un pvalue máximo de 0.001 para la comparación S3 vs S2. Se presenta también el pvalue obtenido en las otras comparaciones. En negrita se encuentran resaltados los pvalue menores al punto de corte.

Categoría SEED		pvalue		
Nivel 2	Nivel 1	S1 vs S2	S1 vs S3	**S2 vs S3**
Resistencia a antibióticos y compuestos tóxicos	Virulencia, Enfermedad y Defensa	1.8E-01	**1.7E-05**	**2.0E-04**

Tabla 3.1.4. Listado de categorías SEED enriquecidas en genomas más abundantes en S2 respecto de S3. Se realizó el análisis sobre las categorías del nivel 2 de la categorización SEED y en la tabla se presenta tanto el nivel 2 como el nivel 1. El punto de corte del Test Exacto de Fisher fue un pvalue máximo de 0.001 para la comparación S3 vs S2. Se presenta también el pvalue obtenido en las otras comparaciones. En negrita se encuentran resaltados los pvalue menores al punto de corte.

Categoría SEED		pvalue		
Nivel 2	Nivel 1	S1 vs S2	S1 vs S3	**S2 vs S3**
Acidos orgánicos	Carbohidratos	8.8E-03	**3.4E-07**	**7.9E-04**

3.2 Análisis de presencia de mecanismos particulares de homeostasis de pH en los genomas reconstruidos

Utilizando como base lo descrito principalmente por Krulwich y cols. (2011b) y Casey y cols. (2010), se seleccionaron 26 proteínas que cuentan con evidencia experimental, en diferentes bacterias, sobre su relación funcional con la homeostasis a pH (Tabla 3.2.1). Utilizando como templado la secuencia primaria de estas proteínas, se examinó su presencia en los genomas reconstruidos mediante análisis de homología BLASTP y luego se realizó un Test Exacto de Fisher con punto de corte de valor p (probabilidad de error) menor a 0.01. Los resultados se presentan en la Tabla 3.2.2, en la cual se observa que en los genomas predominantes de la muestra S1 (suelo ácido) hay un enriquecimiento de la proteína CyoB respecto de S2 y S3 y de la proteína FlgS respecto de S2. Por otro lado, los genomas predominantes en la muestra S3 tienen un enriquecimiento de la proteína SlpA respecto de los genomas de S1 y S2.

Tabla 3.2.1. Listado de proteínas relacionadas con la homeostasis de pH en diferentes microorganismos modelo.

Proteína	Nombre	Especie	Función	Referencias
ArsR	two-component systems TCS ArsRS	*H. pylori*	Sensa el pH del medio	(Wen y col., 2007)
ArsRS_Hp1	two-component systems TCS ArsRS 1	*H. pylori*	"	"
ArsRS_Hp2	two-component systems TCS ArsRS 2	*H. pylori*	"	"
ArsRS_Hp3	two-component systems TCS ArsRS 3	*H. pylori*	"	"
CtaC	cytochrome c oxidase subunit 2	*B. pseudofirmus*	Captura de protones / retención en superficie por tener bajo pI	(Mesbah y col., 2009) (Kennedy y col., 2001) (Knight y col., 2004)
CydA	cytochrome bd subunit A	*E. coli*	Bomba independiente de protón (minimiza pérdida de protones en PMF)	(Slonczewski y col., 2009b) (Maurer y col., 2005) (Stancik, 2002)
CydB	cytochrome bd subunit B	*E. coli*	Bomba independiente de protón (minimiza pérdida de protones en PMF)	"
CyoA	cytochrome bo subunit A	*E. coli*	Bomba de protones para PMF	(Slonczewski y col., 2009b)
CyoB	cytochrome bo subunit B	*E. coli*	"	"
CyoC	cytochrome bo subunit C	*E. coli*	"	"
CyoD	cytochrome bo subunit D	*E. coli*	"	"
CyoE	cytochrome bo subunit E	*E. coli*	"	"
F1F0-ATPase_Eh1	F1F0-ATPase	*E. hirae*	inactiva en pH básico	(Kobayashi y col., 1986) (Ikegami, 1999) (Kakinuma, 1998)
F1F0-ATPase_Eh2	F1F0-ATPase	*E. hirae*	"	"

F1F0-ATPase_Sm	F1F0-ATPase	*S. mutans*	Dirección hidrolítica (protones son expulsados de la célula)	"
FlgS	orphan histidine kinase sensor	*H. pylori*	Probablemente responde al pH citoplasmático	(Wen, 2003)
GadB	glutamate decarboxylase-β	*E. coli*	Consumo de protones para producir GABA	(Slonczewski y col., 2009b) (Foster, 2004) (Gut, 2006)
GadC	glutamate/γ-aminobutyrate antiporter	*E. coli*	Eflujo de GABA effux en intercambio por glutamado	"
MLE	malolactic enzyme	*S. mutans*	Consumo de protones en fermentación malolactica (no respiratoria)	(Sheng & Marquis, 2007)
NhaA	a dual pH-sensing effector	*E. coli*	Sensor de pH y transportador	(Padan, 2008)
Nuo	NADH–ubiquinone oxidoreductase	*E. coli*	Sensor de pH y transportador	(Slonczewski y col., 2009b)
NuoA	NADH-quinone oxidoreductase subunit A	*E. coli*	"	"
NuoB	NADH-quinone oxidoreductase subunit B	*E. coli*	"	"
OmpA	Outer membrane protein A	*A. ferrooxidans*	Prepelente de protones (alto pI)	(Tomb, 1997) (Chi, 2007)
SlpA	teichuronic acids + acidic S-layer protein	*B. pseudofirmus*	Incrementa la concentración de protones cerca de la superficie	(Gilmour, 2000) (Aono y col., 1999)
TnaA	tryptophan deaminase	*E. coli*	contribuye a la $\Delta\psi$, incrementando la fuga de protones	(Blankenhorn y col., 1999) (Yohannes y col., 2004)

Tabla 3.2.2. Listado de proteínas buscadas en los 74 gOTUs. Las columnas S1, S2 y S3 contienen el número de genes totales que tuvieron homología en los gOTUS de cada sitio muestreado. La comparación entre pares de sitios se realizó mediante Test Exacto de Fisher. En negrita las proteínas que tuvieron pvalue menor a 0.01.

Código de la Proteína	% de genes identificados			pvalue		
	S1	S2	S3	S1 vs S2	S1 vs S3	S2 vs S3
ArsR	8%	9%	8%	7.2E-01	9.3E-01	7.8E-01
ArsRS_Hpylori1	8%	9%	8%	7.2E-01	9.3E-01	7.8E-01
ArsRS_Hpylori2	0%	0%	0%	1.0E+00	1.0E+00	1.0E+00
ArsRS_Hpylori3	22%	20%	19%	4.6E-01	1.1E-01	2.7E-01
CtaC	4%	4%	4%	6.7E-01	6.3E-01	9.2E-01
CydA	0%	1%	0%	9.9E-02	3.6E-01	1.0E-02
CydB	0%	0%	0%	1.0E-01	1.0E+00	1.0E-01
CyoA	3%	2%	3%	4.3E-01	8.9E-01	2.5E-01
CyoB	**9%**	**6%**	**6%**	**2.7E-03**	**2.5E-03**	6.1E-01
CyoC	5%	4%	4%	2.6E-01	5.8E-01	6.1E-01
CyoD	0%	0%	0%	3.3E-01	1.0E+00	3.4E-01
CyoE	2%	3%	3%	2.2E-01	2.7E-01	1.0E+00
F1F0-ATPase_Ehirae1	6%	5%	6%	6.5E-01	5.6E-01	2.6E-01
F1F0-ATPase_Ehirae2	5%	5%	7%	8.6E-01	3.3E-01	1.7E-01
F1F0-ATPase_Smutans	7%	6%	8%	7.5E-01	3.3E-01	1.4E-01
FlgS	**6%**	**9%**	**7%**	**4.4E-03**	4.6E-01	4.8E-02
GadB	0%	0%	0%	5.5E-01	1.0E+00	1.0E+00
GadC	1%	0%	0%	1.8E-01	1.4E-01	7.6E-01
MLE	1%	1%	1%	4.5E-01	1.0E+00	5.8E-01
NhaA	1%	1%	1%	5.4E-02	1.4E-01	8.6E-01
Nuo	3%	3%	2%	8.0E-01	3.4E-01	1.5E-01
NuoA	4%	5%	4%	2.8E-01	1.0E+00	2.9E-01
NuoB	4%	4%	4%	1.0E+00	7.3E-01	6.9E-01
OmpA	2%	1%	2%	5.2E-01	1.0E+00	6.3E-01
SplA	0%	0%	**3%**	1.0E+00	**3.7E-09**	**3.7E-12**
TnaA	0%	0%	0%	1.0E+00	1.0E+00	7.2E-01
Genes identificados	912	1869	957			

DISCUSIÓN

1. Caracterización de variables fisicoquímicas como primer paso para la caracterización de comunidades microbianas

Con la idea de encontrar un sitio con características ambientales y/o geoquímicas contrastantes y extremas para la vida nos dedicamos a pesquisar un sector de grandes contrastes geológicos cercano al Salar de Atacama a 2.500 m s.n.m., en pleno Desierto de Atacama, el cual ha sido estudiado desde diferentes perspectivas (Betancourt y col., 2000; Kuch y col., 2002; Nester y col., 2007; Díaz y col., 2016) y donde tenemos una ventana de oportunidad científica gracias a los proyectos colaborativos del Centro de Regulación del Genoma establecidos en el lugar. Ello nos permitió contar con una primera descripción de la geoquímica y de la flora de la zona, específicamente de una zona de fácil acceso camino al pueblo de Talabre que asciende desde los 2.200 hasta los 4.500 m s.n.m. de altura bordeando las faldas del Volcán Lazcar y las orillas de la Laguna Lejía en un trayecto de 50 Km (Díaz y col., 2016). Los datos previos de la zona (artículo en preparación) nos indican que además del gradiente de altura, temperatura y humedad existe un marcado gradiente de pH que va desde 9 en la zona baja a 5 en la zona alta (Mandakovic *et al*, 2016. *en preparación*). Es esta propiedad la que nos llamó la atención como punto de inicio para el estudio comparativo de comunidades bacterianas de suelo desértico pues influye fuertemente en la composición de

comunidades bacterianas (Fierer & Jackson, 2006; Baker-Austin & Dopson, 2007; Lauber y col., 2009; Rousk y col., 2010; Fierer y col., 2011) . Con el objetivo de reducir el número de variables involucradas en las diferencias que pudiéramos encontrar, específicamente temperatura y humedad debido a la gran diferencia de altura entre los sitios contrastantes, decidimos buscar gradientes de pH en zonas de menor amplitud dentro del mismo transecto analizando el pH del suelo de sub-zonas determinadas al azar en diferentes puntos del transecto altitudinal. De esta forma logramos encontrar una segunda zona con gradiente de pH situada en la parte superior del transecto donde los sitios con pH extremo están a la misma altura (± 30 m) a una distancia lineal de 2 Km.

En esta zona se eligieron 3 sitios, 2 contrastantes y uno intermedio, de acuerdo al pH determinado *in situ* el cual fue posteriormente corroborado en el laboratorio. Una inspección visual en terreno no permite detectar diferencias obvias entre los 3 sitios seleccionados o con otros sitios de este paisaje altiplánico. La temperatura del suelo durante las horas de prospección y muestreo no presentó diferencias significativas entre sitios. La estructura física de suelo recolectado está dominada por arena (aproximadamente 90 %, compuesta de partículas entre 0.02 y 0.2 mm), además contiene un 3 % de limo (partículas entre 0.002 y 0.02 mm) y 7 % de arcilla (partículas menores a 0,002 mm). Esto nos posiciona en la zona con mayor porcentaje de arena del transecto, similar a lo encontrado en zonas del desierto absoluto (a 2.200 m s.n.m.). Los suelos arenosos por su alta capacidad de drenaje retienen con dificultad la humedad y los nutrientes (Ahlert, 1998). Al revisar la disponibilidad de nutrientes en el suelo de los 3 sitios en la Tabla 1.1 se aprecia que en general los macro y micronutrientes se encuentran en déficit considerando como norma el requerimiento de cultivos vegetales

(Ahlert, 1998; Jones, 2012). Se utiliza como parámetro de comparación el contenido de nutrientes en el suelo necesario para el cultivo de plantas debido a que para microorganismos de ambientes naturales no existe un consenso generalizado ni detalles específicos respecto de los requerimientos mínimos o limitantes de nutrientes en suelo. En los tres sitios, el suelo presenta una concentración de P inferior a 20 mg/Kg, una concentración de Zn inferior a 1 mg/Kg y una concentración de B superior a 1.55 mg/Kg indicando que toda la zona presenta condiciones poco favorables para la vida macroscópica. A lo anterior se suman las deficiencias específicas por sitio y por elemento. Por ejemplo, K, Mg y Ca se encuentran especialmente disminuidos en el sitio S1 teniendo menos de la mitad del mínimo requerido, el cual es 145 mg/Kg, 195 mg/Kg y 2.405 mg/Kg respectivamente (Ahlert, 1998; Jones, 2012) . Por su parte el sitio S3 suma al P y al Zn el déficit de Mg, Fe y Mn los cuales debieran estar por sobre los 195 mg/Kg, 4.5 mg/Kg y 1 mg /Kg respectivamente. Si bien estos límites han sido establecidos para plantas, es razonable considerar que en general los mecanismos moleculares, incluyendo transportadores de membrana, asociados a la incorporación de estos átomos en plantas son compartidos por bacterias (y levaduras) y por lo tanto estos valores nos proporcionan una idea comparativa del estatus de nutrientes del suelo al cual están enfrentadas las comunidades bacterianas de esta zona. Por otro lado, la temperatura y humedad de los sitios a 10 cm bajo suelo, fue similar entre sitios presentando una diferencia en torno al promedio, inferior a 2°C y 5 % de humedad respectivamente, esto considerando que la diferencia mínima con otros lugares de la zona supera los 5°C al descender 100 m y la humedad relativa cae 10 % al descender 500 m. Considerando lo observado *in situ* es de suponer que las condiciones ambientales

de los sitios también son comparables a lo largo del año. Esta zona forma parte de un ecosistema Alto Andino el cual se caracteriza por poseer temperaturas extremadamente bajas, viento y nieve (Villagrán y col., 1981; Cabrol y col., 2009; Díaz y col., 2016), y cuya diversidad de organismos está limitada a quienes puedan resistir tales condiciones (Fischer y col., 2011). La presencia de vegetación es mínima, limitándose a gramíneas perennes de tipo manojo distribuidas en baja densidad. Son suelos que permanecen cubiertos de nieve como mínimo 4 meses en el año pero cuya temperatura puede superar los 25°C en temporada estival disminuyendo a temperaturas bajo cero durante la noche tal como se observa en las mediciones realizadas durante los días de muestreo. Lo anterior, sumado al aislamiento geográfico de la cuenca de la Laguna Legía nos posiciona en un escenario de gran interés ecológico y biotecnológico.

El análisis de PCA de las variables fisicoquímicas y los contenidos de nutrientes en los sitios (Figura 1.3) indica que la disponibilidad de P, Cu, Fe, Zn y Mn es inversamente proporcional al incremento de pH mientras que la disponibilidad de K sigue al pH sugiriendo que, como ha sido descrito en otros suelos (Jones, 2012), la solubilidad de esos elementos depende del nivel de pH. Por otro lado, la conductividad eléctrica y el contenido de Na, Ca, B, S y C no se relacionan de forma directa con el valor de pH del suelo. Estas variables exhiben una mayor correlación con S3, es decir, su variación no se ajusta al gradiente de pH. Es importante destacar que de los 3 sitios en estudio, el sitio S3 posee los niveles más altos de B y S, alcanzando valores reportados como tóxicos para otros organismos como las plantas, cuyos límites son 1.55 mg/Kg y 8 mg/Kg respectivamente (Ahlert, 1998; Jones, 2012). La fuente más probable de estos metales es la emanación de gases del volcán Lascar, que produce una acumulación

mayor en el sitio S3 el cual se encuentra en una cota inferior respecto de la cuenca incluyendo la pendiente asociada a los sitios S1 y S2 (con una diferencia de 30 m de altura) con lo cual un efecto de lavado por lluvia podría movilizar elementos siguiendo la pendiente de las laderas de la cuenca hacia sitios del tipo S3.

Las diferencias significativas en los contenidos de micro y macronutrientes entre los sitios en estudio (resultado no esperado al inicio del estudio), indica la importancia de considerar otros parámetros del suelo al momento de interpretar los resultados del análisis de la estructura y función de las comunidades de microorganismos de estos suelos.

En conclusión, los resultados sobre las características fisicoquímicas de los sitios seleccionados nos sitúan en un ambiente útil para abordar el estudio de comunidades de microorganismos en un ambiente no intervenido por actividad humana. Los suelos de esta cuenca altiplánica representan además un modelo de perturbación ambiental en forma de un gradiente de pH, variable que ha sido identificada como el factor que mejor explica la diversidad de bacterias en ambientes terrestres (Fierer y col., 2011; Fierer, Leff, y col., 2012).

Nuestra caracterización se suma a la recientemente reportada por Díaz y colaboradores (Díaz y col., 2016), quienes evaluaron desde una perspectiva paleobotánica la distribución de especies vegetales en el transecto Talabres-Lejia.

2. Caracterización taxonómica y funcional de las comunidades de bacterias de sitios contrastantes para la variable pH de suelo

2.1 Caracterización taxonómica

Diversos estudios han examinado la estructura taxonómica de los microbiomas del suelo en el Desierto de Atacama. Estos trabajos se han orientado principalmente a identificar parámetros fisicoquímicos que simulan ambientes en otros planetas, la relación entre diversidad y humedad del ambiente, incluyendo el efecto de un gradiente de humedad en la orientación norte-sur (Andrew y col., 2012; Azua-Bustos y col., 2012; Moreno y col., 2012; Narasingarao y col., 2012; Neilson y col., 2012; Viles, 2012; Crits-Christoph y col., 2013; Davila y col., 2013; Paulino-Lima y col., 2013; Piubeli y col., 2015). Sin embargo los lugares de muestreo se sitúan en regiones del desierto denominadas "desierto absoluto", las cuales son predominantes y prácticamente no presentan precipitaciones durante todo el año. Por otra parte, la estepa alto andina, si bien presenta un mayor nivel de precipitaciones se caracteriza por poseer temperaturas extremadamente bajas, viento y nieve, con un balance de agua crítico (tasa de precipitación/evaporación) debido a la exposición a altas temperaturas en verano y a las cada vez menos abundantes precipitaciones de invierno (Vuille & Bradley, 2000; Demergasso y col., 2010) y están expuestas a una radiación UV de 1.65 veces más que a nivel del mar (Cabrol y col., 2009). A pesar de que el altiplano es un ecosistema extremo, que se extiende a lo largo de la cordillera de los Andes muy poco se sabe de su microbiología. El estudio más reciente en el sector se abocó al análisis de las

comunidades de microorganismos del agua y sedimentos de la Laguna Lejía (Demergasso y col., 2010) seguido de otro con enfoque en el zooplancton (Munoz-Pedreros y col., 2013). Desde un punto de vista metodológico, los estudios de las comunidades de microorganismos se han realizado mediante técnicas de Electroforesis en Gel con Gradiente Denaturante (DGGE) o Polimorfismos en la Longitud de los Fragmentos de Restricción (T-RFLP) cuyas resoluciones presentan limitaciones cuando se intenta describir muestras complejas (muchas especies en baja abundancia). Debido a la baja resolución de las técnicas utilizadas hasta el momento, decidimos enfocar nuestros esfuerzos en una identificación profunda a nivel tanto de rDNA 16S como de metagenómica utilizando las técnicas de secuenciación masiva siguiendo los protocolos actuales de identificación microbiana. En primera instancia se probaron diversos "kits" comerciales y protocolos de extracción de DNA de suelo llegando a la conclusión que ninguno de los "kits" comerciales nos entregaba un DNA íntegro y en cantidad adecuada, a diferencia del método en base a CTAB. Lo interesante de este punto es que resultó no ser un problema de nuestra muestra pues al revisar la información proporcionada por los fabricantes encontramos que la calidad asegurada bajo sus propios controles (información publicada en sus sitios web) considera la obtención de DNA de integridad variable pero "amplificable". Lo anterior no es problema en estudios cualitativos destinados principalmente a identificación pero es delicado en estudios cuantitativos destinados a comparación y a análisis metagenómicos. Por ello, y considerando los resultados comparativos entre diferentes protocolos de extracción de DNA (no mostrados), utilizamos un protocolo en base a CTAB el cual fue optimizado para muestras de desierto (Prestel y col., 2008). Se escogió para la amplificación del gen

codificante para la unidad ribosomal 16S una zona que comprende los primeros 500 pares de bases del gen debido a que contiene las regiones híper-variables V1 a V3 y es considerada la zona con mayor representación en las bases de datos lo cual nos permitió un nivel adecuado de identificación (McDonald y col., 2012; Quast y col., 2013).

Al revisar publicaciones recientes sobre comunidades microbianas de muestras complejas no existe consenso sobre el número mínimo de lecturas que se debe utilizar para la descripción total de los OTUs presentes, dado que este valor depende de la complejidad de la muestra (número de genomas diferentes y riqueza filogenética). En términos prácticos, en nuestras muestras de suelo no fue posible llegar a un número constante de OTUs, incluso utilizando esfuerzos de secuenciación que alcanzaron las 100.000 lecturas por muestra, lo cual pudimos comprobar empíricamente al realizar el análisis mezclando las lecturas de los triplicados (datos no mostrados). Este fenómeno puede ser explicado si consideramos la gran diversidad microbiana presente en este tipo de matriz biológica, en la cual 1 g de suelo puede contener entre 3.000 a 10.000 microorganismos diferentes (Torsvik & Ovreas, 2002; Torsvik y col., 2002; Curtis & Sloan, 2004; Cowan y col., 2005; Gans y col., 2005; Hirsch y col., 2010). Mediante análisis exploratorios con muestras de la zona logramos determinar que para asegurar un mínimo de 5.000 lecturas mapeadas a la base de datos Greengenes con un 99 % de identidad se necesitan 30.000 lecturas totales. Durante la secuenciación se obtuvo un número de lecturas superior a 40.000 lecturas para la mayoría de las muestras a excepción de la muestra S1 donde los triplicados tuvieron lecturas en torno a las 30.000 lecturas. Si bien cumple con límite propuesto, la diferencia respecto de las otras 6 muestras puede tener origen en la naturaleza fisicoquímica del suelo desde el cual se

extrajo la muestra (pH 5.88). En términos comparativos el rendimiento de extracción de DNA en esta muestra siempre fue un 10 % menor a los otros suelos probablemente debido a la influencia residual del bajo pH sobre el método de extracción, algo que continuó influyendo en la amplificación de rDNA 16S en la que se obtuvo una menor cantidad de amplicón total respecto de las otras muestras a pesar de mezclar cantidades equimolares de DNA. A pesar de lo anterior, al aplicar los filtros de calidad y largo nos quedamos con una población en torno al valor esperado según se observa en la Tabla 2.1.1, con un número de lecturas mapeadas que van desde las 6.447 hasta las 14.752. Debido a esta diferencia de lecturas mapeadas entre las muestras no es posible realizar una identificación de OTUs y comparación de muestras directa sin antes normalizar el número de lecturas. Esto se realiza extrayendo en forma reiterada (al menos 10 veces) una selección al azar de lecturas (secuencias) desde las muestras en un ciclo que parte en 10 lecturas y termina en 6.400 lecturas. Este proceso es conocido como rarefacción. El resultado, utilizando lecturas normalizadas, nos indica que el sitio S3 (pH 8.51) es quien posee un mayor número de OTUs respecto de los otros dos sitios y también posee mayores índices de diversidad Shannon y diversidad filogenética PD, lo cual contrasta con lo reportado por Fierer (2006) y Lauber (2009) quienes, en una prospección de sitios a nivel planetario observaron que en general la riqueza (número de OTUs) y la diversidad eran mayores en pH neutro que en pH extremo. Sin embargo, al examinar los datos publicados (Fierer & Jackson, 2006) con los declarados en esta tesis considerando el mismo rango de pH encontrado en nuestros sitios el cual no es extremo, encontramos que la correlación entre pH y riqueza así como la relación entre pH y diversidad son igualmente no significativas, sugiriendo que las comunidades de microorganismos de

suelo ajustan su riqueza y diversidad en ambientes que alcanzan mayores rangos de variación de pH. Sin embargo también es posible que variaciones, locales más que planetarias, en factores nutricionales u otros factores fisicoquímicos modulen los parámetros de riqueza y diversidad en nuestra comunidad. En este contexto, los datos sugieren que la mayor riqueza observada en el sitio S3 está relacionada con las particulares condiciones ambientales de este sitio. En particular, llama la atención que las especies dominantes son más variadas según lo indica el índice de Shannon y, en general, también lo son los demás miembros de la comunidad según lo indica la diversidad filogenética PD.

Es importante mencionar que los antecedentes sobre la formación geológica de los suelos de la cuenca del volcán Lazcar, indican que este terreno formó parte de una misma laguna más extensa y profunda la cual, luego de un incremento progresivo de su drenaje y evaporación, ha dado paso a un amplio sector alcalino y de salinidad intermedia. Lo anterior nos permite suponer que el perfil base del asentamiento bacteriano en la zona corresponde a microorganismos con capacidades de resistencia a pH alto (mayor a pH 8) y salinidad, mientras que la aparición de nuevas zonas de colonización mediadas por el material efluente de las emisiones volcánicas depositadas en el terreno dieron paso a nuevas etapas de colonización y adaptación. Esto explicaría la alta riqueza y diversidad observada en suelo S3 respecto de los otros sitios. Las relaciones de abundancia entre filos de los sitios será retomado más adelante en un Análisis de Correlación Canónica entre abundancia de especies y variables ambientales. Al utilizar los OTUs como elemento de comparación entre los sitios (Figura 2.1.3, panel A) se observa que todos los sitios comparten un 35,6 % de los OTUs totales (554

elementos) lo cual comprueba nuestro supuesto sobre la redundancia de genomas bacterianos entre los 3 sitios debido a su proximidad a pesar de las significativas diferencias ambientales. Según lo observado en trabajos de escala global, el nivel de intersección a nivel de OTUs cae en función de la distancia entre las muestras (Fierer & Jackson, 2006; Lauber y col., 2009). Los sitios S2 y S3 comparten un 24 % de los OTUs totales por lo cual quedan en el mismo clado según la Figura 2.1.2. Estos sitios están a unos 100 metros de distancia por lo cual era esperable que compartieran mayor número de genomas respecto del sitio S1 que se encuentra a unos 2.000 metros de distancia. Probablemente por la misma razón el sitio S1 posee el mayor número de OTUs únicos con un 11.4 % del total.

Finalmente y en el contexto de la caracterización taxonómica, nuestros resultados nos permitieron identificar 26 filos de los cuales los más abundantes son Acidobacteria, Actinobacteria y Proteobacteria comprometiendo más del 80 % de las muestras, lo cual concuerda con lo descrito en otros sitios del desierto (Gómez-Silva y col., 2008; Neilson y col., 2012; Crits-Christoph y col., 2013; Paulino-Lima y col., 2013; Piubeli y col., 2015). Interesantemente, de los 23 filos que presentaron los menores valores de abundancia relativa, la mayoría aún son filos candidatos (raros) (Solden y col., 2016).

La abundancia relativa de los OTUs identificados muestra el patrón típico encontrado en la mayoría de las muestras de suelo siguiendo una distribución del tipo power-low, característica de ensambles bacterianos en los cuales unos pocos OTUs dan cuenta de la mayor parte de la biomasa de la comunidad (Bailey y col., 2013).

Con el fin de seleccionar los OTUs característicos de cada sitio (aunque no necesariamente exclusivos) se realizó un agrupamiento mediante mapas auto-asociativos

conocido por sus siglas en inglés como SOM o red de Kohonen (Kohonen, 2001), el cual es un tipo de red neural artificial, la cual es entrenada con los datos. En nuestro caso con los valores de abundancia relativa de los OTUs, de modo de producir una representación discreta de los datos, en este caso, grupos de OTUs que comparten variaciones equivalentes en sus valores de abundancia entre los sitios analizados. SOM utiliza un algoritmo competitivo (opuesto al de aprendizaje prueba/error) que preserva las propiedades topológicas de los datos originales mediante una función de vecindad. Esto lo hace particularmente útil en datos asociados a entidades interactuantes como las comunidades biológicas (Kohonen, 2001). Al observar los grupos formados encontramos 3 cuyos OTUs presentan abundancias diferenciables entre los 3 sitios según se observa en la Figura 2.1.4 y representan porcentajes importantes de la abundancia total de la comunidad: 69 % en la muestra S1, 63 % en la muestra S2 y 40 % en la muestra S3. La caída en la representación del grupo S3 puede estar ligada nuevamente al menor número de OTUs únicos, según ya se ha observado, lo cual significa que gran parte de su población puede habitar en los otros sitios examinados. La caracterización taxonómica de estos sub-grupos nos presenta un escenario más acotado pues están representados sólo 14 filos de muestra original. El grupo 1 es el único que posee los filos AD3, Firmicutes y WPS-2 mientras que el grupo 3 es el único que posee OTUs de los filos BRC1 y FBP. Estos filos son considerados "materia oscura", pues por ahora se conoce muy poco sobre su biología (Rinke y col., 2013; Solden y col., 2016). Por lo tanto, nuestros datos sobre su pertenencia a grupos de OTUs conocidos pueden ser de ayuda para futuros proyectos de investigación en cuanto a sus propiedades funcionales.

En conclusión, la información taxonómica permitió caracterizar la estructura de las comunidades de microorganismos de los 3 sitios junto con establecer sub-grupos que comparten patrón de abundancia específico para cada sitio, información que será contrastada con: 1) los datos de variables ambientales y de nutrientes, y 2) la información funcional que se obtuvo desde los genomas reconstruidos.

2.2 Ensamble y caracterización funcional de metagenomas

A la fecha, existen pocos reportes que describen procedimientos para reconstruir genomas desde metagenomas complejos, siendo los trabajos realizados en muestras de agua y de sedimento marino (Delmont y col., 2015; Sunagawa y col., 2015; Hug, Thomas, y col., 2016) lo más parecido al nivel de complejidad que exhiben los suelos analizados en esta tesis de Doctorado. Considerando lo anterior, el grupo de Per Nielsen estableció que es necesaria una base de información (volumen de secuenciación de DNA) mínima de 10 Gb de datos con lecturas superiores a 100 pb para obtener genomas con abundancias inferiores al 0.1 % desde muestras de baja complejidad (< 40 OTUs, crecidas en bioreactor) (Albertsen y col., 2013). Por ello y considerando que el número de OTUs en las muestras de suelo oscila entre los 750 y 1200 nos pusimos como meta obtener al menos 15 Gb de datos con lecturas de 150 pb. Los resultados indican que, luego de descartar secuencias sin información y perder un 7 % en el tamaño de las lecturas por problemas de calidad, nuestro análisis se ejecutó sobre un conjunto de secuencias mayor a 17.7 Gb, con lecturas de largo promedio de 138,3 pb lo cual nos permitió eludir restricciones de ensamble de genomas por problemas de baja cobertura.

Para ensamblar los metagenomas se realizaron pruebas con tres programas ("ensambladores") dedicados a este propósito con el fin de comparar y detectar eventuales inconsistencias en los resultados: Velvet (Zerbino & Birney, 2008), CLC Genomics Workbench (Qiagen, USA) y IDBA-UD (Peng y col., 2012) (datos no mostrados). Si bien Velvet es uno de los programas más utilizados en ensamble de microorganismos, su algoritmo no está optimizado para enfrentar ensamble de secuencias con una gran variación en la profundidad de lecturas por molécula ensamblada. Esto tiene sentido pues la idea de Velvet es ensamblar genomas únicos provenientes de muestras homogéneas. Algo similar ocurre con CLC, el cual si bien tiene un mejor desempeño que Velvet con nuestras muestras, su algoritmo no está preparado para lidiar con cambios bruscos de profundidad de lecturas por lo cual excluye mucha información (>90 %). En cambio, IDBA-UD es un programa computacional cuyo algoritmo está pensado para muestras metagenómicas por lo cual logra mejores resultados, ocupando un mayor porcentaje de los datos ingresados y obteniendo mayor número de secuencias ensambladas con largos que superan las 10.000 pares de bases. Finalmente, todas las muestras metagenómicas fueron ensambladas con este programa logrando un uso total de datos en torno al 32 % mientras que los otros ensambladores no utilizaban más del 10 % de datos.

La información total obtenida en nuestros ensambles fue en promedio 770 Mb lo cual pareciera mantener una relación al número inicial de lecturas generadas en la etapa de secuenciación pues, al mezclar el conjunto de secuencias de dos muestras, se obtiene un poco más del doble de datos y de "scaffolds". Esto se ha observado con anterioridad (Albertsen y col., 2013; Carr y col., 2013; Alneberg y col., 2014; Nielsen y col., 2014;

Delmont y col., 2015), siguiendo un efecto incremental similar al descrito con el número de lecturas provenientes de amplificación rDNA 16S donde a pesar de los esfuerzos en secuenciación no se llega a *plateau*. Este efecto tiene su explicación en que existe un gran número de genomas contenidos en el suelo, de los cuales sólo podemos observar cierto porcentaje que además está sobre-representado (genomas más abundantes). Al aumentar los esfuerzos en secuenciación seguimos viendo los genomas abundantes y van apareciendo siempre nuevos genomas con menor representación de abundancia.

Existen al menos un par de formas de realizar una caracterización funcional a partir de datos metagenómicos. Por una parte se pueden utilizar las lecturas sin ensamblar lo cual asegura una caracterización total de lo que existe en las muestras sin sesgo de ensamble. El problema es que al usar secuencias cortas (< 150 pb) se produce una pérdida de especificidad, y por lo tanto, una menor potencia de análisis con mayor probabilidad de error. La otra opción es realizar una caracterización a partir de los genes predichos (CDSs) desde los "scaffolds" ensamblados. Esta última opción permite una caracterización más específica pues los elementos a comparar constituyen genes de largo completo y la información de vecindad de los genes en "scaffolds" genera una capa de información importante para evaluar presencia de operones. Al realizar la caracterización de nuestros metagenomas utilizando ambos tipos de aproximaciones y analizar los resultados en cuanto a la clasificación COG (Figura 2.2.1) se obtiene que tanto para secuencias crudas como para CDSs existen 8 categorías que están presentes en el grupo de las 10 más abundantes de cada muestra. Es decir, estas 8 categorías funcionales representan la mayor abundancia en todas las muestras, en ambos tipos de

análisis:

- [J] Traducción, estructura ribosonal y biogénesis
- [L] Replicación, recombinación y reparación
- [M] Biogénesis de pared celular/membrana/cobertura
- [C] Producción y conversión de energía
- [G] Transporte y metabolismo de carbohidratos
- [E] Transporte y metabolismo de aminoácidos
- [R] Función generales (sólo predicción)
- Función desconocida

De estas 8 categorías, 3 se intersectan al revisar las 5 categorías más abundantes en todas las muestras, en ambos tipos de análisis:

- [L] Replicación, recombinación y reparación
- [E] Transporte y metabolismo de aminoácidos
- [R] Función general (sólo predicción)

En los perfiles se observa que en general la magnitud de las diferencias funcionales, entre las tres muestras, es de menor envergadura a las diferencias observadas en la composición taxonómica, lo cual es similar a lo descrito anteriormente para tracto intestinal, mares y suelos (T. H. M. P. Consortium, 2012; Fierer, Leff, y col., 2012; Sunagawa y col., 2015). Con el fin de obtener mayor información sobre los componentes de los metagenomas y revisar si estos resultados siguen siendo válidos ahondamos en la descripción de las comunidades mediante el ensamble del mayor número de genomas de calidad que se puedan rescatar desde dichas muestras.

2.3 Recuperación de genomas individuales desde metagenomas

La recuperación de genomas individuales desde una muestra compleja se ha realizado principalmente en base a métodos dependientes de la composición del DNA, como por ejemplo el porcentaje de GC (Mande y col., 2012). Un indicador más sofisticado y específico consiste en la construcción de mapas emergentes de auto-organización basados en la frecuencia de tetranucleótidos (Tetranucleotide frequency–based emergent self-organizing maps: ESOMs). El concepto detrás de este indicador es que los genomas de distintas especies poseen una frecuencia particular de organización de sus nucleótidos, la cual puede ser usada como una huella dactilar para agrupar segmentos de DNA. Este indicador fue aplicado para obtener los genomas de las 49 poblaciones menos abundantes de un acuífero modificado por presencia de acetato (Wrighton y col., 2012). Sin embargo, muchos de los genomas obtenidos no se lograron separar de forma clara, lo cual se observa por la presencia de múltiples copias de genes de copia única dentro de los subgrupos de secuencias separados por ESOMs, lo cual puede derivar en una incorrecta interpretación de su evolución, fisiología y ecología.

Es importante mencionar que el método de clasificación de Albertsen y colaboradores (2013) es independiente de su composición y puede ser usado como primer paso de clasificación, seguido de la clasificación mediante otros indicadores como ESOMs. Ellos, uilizando un metagenoma proveniente de una comunidad de bacterias de un bioreactor, cuyo DNA fue extraído de dos formas diferentes, obtuvieron 31 grupos de poblaciones lo cual incluía especies de baja abundancia (<1 % de la abundancia relativa). Diecinueve de estos grupos fueron refinados en genomas

completos o casi completos. Al probar este nuevo método con el set de datos de secuencia metagenómica derivadas de una muestra de agua ambiental (Wrighton y col., 2012), lograron una mejor separación de las especies o sub-grupos de secuencias, lo cual permite mejorar la calidad y acabado de ensambles individuales. Nuestro procedimiento de reconstrucción de genomas aprovechó la oportunidad de contar con comunidades de microorganismos que varían la magnitud de la abundancia relativa de sus miembros según lo observado por rDNA 16S. El principio detrás del método es que los segmentos de DNA que pertenecen a una misma especie experimentarán el mismo cambio de abundancia promedio entre distintas condiciones, pues el cambio de biomasa de la especie es directamente proporcional a su cantidad de DNA, lo cual permitirá agrupar los segmentos de DNA o "scaffolds" en función de este cambio.

Si bien, el protocolo aplicado para discriminar genomas individuales dentro de los metagenomas ya ensamblados consideran lo descrito por Albertsen *et* al (2013), incluye diversas modificaciones para mejorar la capacidad de discriminación en genomas a partir de una muestra compleja (OTUs >800). La modificación que tuvo mejor resultado fue tomar como base un ensamble híbrido de muestras de dos condiciones a comparar (S2 + S1 o S2 + S3). Lo anterior con el propósito de maximizar la posibilidad de crear "scaffolds" de mayor tamaño y potenciar el ensamble de secuencias de DNA que están presentes en ambas muestras.

A partir de los ensambles híbridos de las muestras se obtuvieron 2 nuevos metagenomas con las características descritas en las últimas 2 columnas de la Tabla 2.2.2. Posteriormente las secuencias de todas las muestra fueron mapeadas por separado para determinar la cobertura promedio de cada "scaffold" en cada muestra. También se

determinó el contenido de GC y la frecuencia de tetrámeros de cada "scaffold". Las 3 variables fueron convertidas en componentes principales (PCA) para realizar el agrupamiento de "scaffolds" (Albertsen y col., 2013; Alneberg y col., 2014). En el panel A de la Figura 2.2.2 se observa como ejemplo el agrupamiento de "scaffolds" de la muestra "S2 + S3" graficado en función de los 2 primeros componentes principales donde cada "scaffold" está representado por un punto de distinto color según el grupo al cual fue asignado. Se obtuvieron 150 "clusters" o gOTUs que incluyen 35 % de las bases totales y 8 % del total de "scaffolds", lo cual incrementa significativamente la eficiencia en el uso de secuencias en comparación a lo que hemos logrado rastrear a la fecha en la literatura. Se evaluó el grado de acabado de los gOTUs mediante la búsqueda y conteo de 36 genes que en general se presentan como copia única en otras especies (Alneberg y col., 2014). Utilizando los criterios mencionados en la sección de resultados (Figura 2.2.2, panel B), se logró recuperar 74 gOTUs (o "clusters" de "scaffolds") los cuales fueron desarmados a sus secuencias originales y re-ensamblados como conjunto (por "cluster") mediante el programa computacional Velvet para optimizar la calidad de ensamble de cada gOTU. El número de CDSs recuperados se situó en un rango acotado dentro de lo esperado, con un promedio de 4.500 genes. Doce gOTUs superaron los 10.000 genes de los cuales 6 genomas que superan los 10Mb. Dichos casos le dan fuerza al concepto gOTUs pues los genomas reconstruidos no constituyen necesariamente un único genoma, pero si constituyen una unidad taxonómica (genomas que representan cepas de una especie o especies cercanas del mismo género) que por el momento no es posible resolver en sus miembros (cepas/especies) originales. Para evitar problema de

sesgo por tamaño en los análisis comparativos se realizó, en todos los casos, una normalización por tamaño de muestra.

El indicador de abundancia relativa de cada gOTU, calculado en las 3 muestras, permitió posicionar los gOTUs en un espacio tridimensional como se observa en la Figura 2.2.3. Este análisis permitió clasificar los 74 gOTUs en 3 grupos según su distribución diferencial con alguno de los 3 sitios (S1, S2 o S3). Esto nos permite agrupar 14 genomas con mayor abundancia en S1, 39 con mayor abundancia en muestra S2 y 21 con mayor abundancia en muestra S3. El mayor número de genomas reconstruidos en S2 apoya la idea que el número de "scaffolds" se relaciona directamente con el número inicial de lecturas generadas en la etapa de secuenciación pues, al mezclar el conjunto de secuencias de dos muestras, se obtiene un poco más del doble de datos y de "scaffolds": uno proveniente del ensamble S1 con S2 y otro proveniente del ensamble S3 con S2. Como la intersección de ambos ensambles es la muestra S2, al reconstruir genomas desde cada ensambles hay una doble ganancia para la muestra S2 obteniendo el doble de gOTUs que con las otras muestras. La posibilidad de agrupar los genomas reconstruidos según su distribución diferencial entre sitios, y lo que ya sabemos en cuanto a la importante intersección de OTUs entre los 3 sitios, nos lleva a pensar que existe un sesgo en el método de clasificación de "scaffolds" el cual priorizaría la separación de aquellos que poseen una alta diferencia de cobertura entre las muestra evaluadas. Esta hipótesis debe ser contrastada con el comportamiento de la población (OTUs) lo cual se discute más adelante en el Análisis de Correspondencia Canónica.

La estructura de la mayoría de los genomas reconstruidos sigue lo esperado en

cuanto a una uniformidad en el contenido de GC y la paridad esperada en el sesgo de GC (GC skew: G-C/G+C) la cual cambia de signo en torno al 50 % del largo total del genoma en todos los gOTUs construidos. Considerando que en general esta posición en las bacterias coincide con los sitios de origen y termino de la replicación [revisado por Grigoriev (1998) y por Bentley y Parkhill (2004)] nuestro resultado sustenta la estrategia de análisis utilizada en la discriminación de "scaffolds", y refuerza la propuesta de utilizar los parámetros dependientes de secuencia elegidos (porcentaje de GC y frecuencia de tetrámeros) y los parámetros independientes de secuencia (abundancia relativa entre sitios).

El proceso de anotación taxonómica nos permitió identificar el 88 % de los gOTUs, logrando llegar a nivel de género en 37 casos, siendo el género candidato *Chloracidobacterium* el más común con 21 gOTUs, la mayoría [14 gOTUs] perteneciente a S2 (Tabla 2.2.2 y Figura 2.2.5). Éste es un género de reciente clasificación dentro del filo Acidobacteria donde se propone el nombre Blastocatellia como clase candidata para agruparlo. Posee una especie secuenciada denominada *Chloracidobacterium thermophilum* encontrada en fuentes termales de Yellowstone y Nuevo México la cual es propuesta como un nuevo grupo de bacteria fotosintética aeróbica dependiente de clorosomas (Garcia Costas y col., 2012; Hallenbeck y col., 2016). La similitud entre los genomas encontrados en este trabajo y el genoma secuenciado está en el rango del 50 % y 60 % a nivel nucleotídico por lo cual es probable que los genomas reconstruidos pertenezcan a un nuevo tipo de bacteria cuya especie filogenéticamente más próxima sería *Chloracidobacterium thermophilum*, bacteria con la cual no comparte nicho ambiental.

2.4 Análisis integrado de datos mediante superposición de variables

Debido a la gran cantidad de datos obtenidos para los 3 sitios, tanto abióticos como bióticos, es necesaria una reducción de dimensiones para integrar la información en un modelo descriptivo del ecosistema. Si bien existen diferentes métodos para llevar a cabo esta tarea, el más específico es el Análisis de Correspondencia Canónica. Empleando este método es posible priorizar el ordenamiento de los sitios en función de las variables bióticas (abundancias de filos) y, en función del resultado, permite evaluar la matriz de variables ambientales para determinar cómo se relacionan con el ordenamiento previo (Legendre & Legendre, 1998). En nuestro caso la matriz inicial contiene los filos identificados en cada sitio con su abundancia relativa considerando triplicados. A esta matriz le hemos sumado los OTUs seleccionados mediante SOM también con sus abundancias relativas en triplicado y los gOTUs reconstruidos con la abundancia relativa calculada en base al número de lecturas promedio por "scaffold" para cada genoma con abundancia diferencial entre los sitios S1, S2 y S3. Al graficar los datos en el sistema de coordenadas biplot, que incluye todas las variables y sus relaciones, se observa el comportamiento de las variables biológicas (Figura 2.3.1 paneles A y B). Respecto del ordenamiento de sitios podemos ver que éstos forman un triángulo, en el cual cada vértice corresponde a los tres sitios analizados con sus respectivas réplicas las cuales se ubican muy cercanas entre sí. Esto nos indica que la variabilidad biológica en cada sitio es muy baja respecto de la variabilidad con los otros sitios, sugiriendo que la distribución diferencial de los gOTUS generados se explica por el ajuste funcional de estos genomas a las particularidades fisicoquímicas de cada uno de

los sitios analizados. Como se observó en la Figura 2.1.3, existe un núcleo de filos que está presente en todas las muestras los cuales en este el CCA se mantienen en torno al centro del triángulo; pero a la vez, existe ciertos filos que pertenecen de forma exclusiva a algún sitio o tienen sesgo importante en alguno de los vértices.

De particular interés resulta la presencia exclusiva en algunos sitios de filos considerados "materia oscura", los cuales representan bacterias poco conocidas hasta el momento y cuyo potencial promete enriquecer las posibilidades metabólicas conocidas hasta ahora (Rinke y col., 2013; Hug, Thomas, y col., 2016). Al revisar el ordenamiento de los OTUs seleccionados por SOM encontramos que poseen una tendencia de ordenamiento a hacia los sitios correspondientes (S1, S2 y S3) lo cual tiene sentido considerando que los perfiles fueron seleccionado por esa propiedad. Esta tendencia es mucho más fuerte en los genomas reconstruidos, los cuales se ordenan en su mayoría muy cerca de los vértices del triángulo. De esta forma se observa que el método de reconstrucción efectivamente tiene un sesgo por el cual agrupa "scaffolds" que poseen una alta diferencia en los niveles de abundancia entre las muestras siendo que, según el análisis de OTUs, sabemos que existe un gran número de genomas con abundancias intermedias. Este sesgo resulta beneficioso en nuestro caso pues nos permite identificar gOTUs únicos para cada sitio y esta condición puede estar relacionada con una capacidad metabólica de especiación a nicho y/o con déficit de capacidades para sobrevivir en condiciones diferentes a las de su nicho.

Al revisar las variables ambientales que correlacionan con el ordenamiento de sitios y con las condiciones diferenciales que pueden definir la condición de vida de

dichos gOTUs, es importante considerar la magnitud de los vectores fisicoquímicos y/o nutricionales (panel B, Figura 2.3.1). Considerando que los vectores entregan una idea de la fuerza de correlación, la orientación en cambio indica el tipo de relación (directa o indirecta). Finalmente, el ángulo que forma el vector respecto de la línea imaginaria que va del centro a los vértices del triángulo indica si esa variable es importante para el ordenamiento de los sitios. En atención a estos criterios nuestro datos indican el sitio S1 posee una correlación positiva y marcada con la disponibilidad de los micronutrientes Mn, Fe, Cu y Zn y posee una relación negativa y más débil con pH y K. Este es un resultado esperado considerando el pH del suelo determina en gran medida la solubilidad y, por lo tanto, la disponibilidad de estos micronutrientes (Andrews y col., 2003; Jones, 2012; Colombo y col., 2014). Estas variables permiten separar la muestra S1 de S2 y S3 y, por lo tanto, podrían definir las comunidades bacterianas presentes en S1. Por otro lado, S3 posee una correlación positiva fuerte con B, S, Ca, Na y un poco más débil con C y EC. Tal como fue discutido anteriormente, este sitio es el que posee mayor rastro de las emisiones volcánicas del Lazcar, además de ser el sitio más cercano al borde de la alcalina Laguna Lejía. La alta concentración de especies de origen volcánico medidas nos hace pensar que también pueda tener la más alta concentración de metales pesados y, por lo tanto, su comunidad bacteriana y en particular los genomas exclusivos deben poseer mecanismos que les permitan enfrentar niveles tóxicos de metales y sales.

3. Representación diferencial de los elementos de homeostasis de pH en los genomas ensamblados

3.1 Análisis comparado de los elementos de homeostasis de pH en los genomas ensamblados de S1, S2 y S3 utilizando categorías SEED.

El análisis de las variables fisicoquímicas y nutricionales indica que los tres sitios del estudio presentan características ambientales diferenciales. En paralelo, el análisis de las comunidades bacterianas sugiere que estas diferencias repercuten marcadamente, tanto en la estructura de las comunidades, como en el repertorio de sus funciones moleculares. Con el propósito de evaluar esta posibilidad, se utilizaron los genes predichos desde los genomas reconstruidos como un indicador de capacidades funcionales que pueden ser comparadas en grupo. De esta forma se genera una matriz de abundancia, en la cual las filas contienen las diferentes categorías funcionales, las columnas corresponden a los 74 gOTUs y cada celda contiene el número de genes que codifica para proteínas con una función clasificada las categorías SEED, normalizados por el total genes que ingresaron al conteo. Los datos indican que los metagenomas se agrupan formando parte de la misma rama en el árbol de vecindad (Figura 2.2.6), situación que se repite independientemente del protocolo de ensamble utilizado. Este resultado coincide con lo observado en metagenomas de microbiomas de humanos (H. M. P. Consortium, 2012), pues los atributos funcionales son similares entre las personas a pesar de la variación en la estructura de la comunidad. Interesantemente, un estudio a escala global describiendo metagenomas de ecosistemas marinos coincide con este

resultado (Sunagawa y col., 2015). En este último estudio, las notables diferencias fisicoquímicas entre los ecosistemas marinos se corresponden con diferencias en la composición taxonómica de los microbiomas, sin embargo la mayoría de los genes constituyen un núcleo funcional compartido. Por lo tanto, nuestro resultado suma un nuevo ambiente en el cual la estructura taxonómica de las comunidades se diferencia entre sitios con características fisicoquímicas diferenciales (incluyendo la composición de nutrientes), pero no se diferencian en cuanto a sus atributos funcionales. Por otra parte, la comparación de los atributos funcionales entre algunos de los gOTUs generados si permite detectar diferencias entre los diferentes tipos de suelo analizado (Figura 2.2.6). En este mismo contexto, una comparación en detalle de la categoría funcional "respuesta a estrés" y sus categorías internas, entre estos genomas reconstruidos y los metagenomas ensamblados (Figura 2.2.7), revela nuevamente una baja variabilidad entre metagenomas de los tres sitios estudiados y un mayor grado de diferenciación entre los gOTUs reconstruidos. Ambos resultados confirman nuestra hipótesis de trabajo en cuanto a que el discriminar genomas de taxas particulares dentro del metagenoma, permite identificar propiedades funcionales taxa-específicos. Dicha matriz nos permitió observar las relaciones entre los genomas en la Figura 2.2.6 y el detalle de la categoría respuesta a estrés de la Figura 2.2.7 según lo discutido anteriormente. Como uno de los objetivos de la tesis es la comparación a nivel de sitio, la matriz fue simplificada agrupando los genomas pertenecientes al mismo sitio para lo cual se sumó el conteo relativo de cada categoría SEED generando una matriz con 3 columnas: S1, S2 y S3. Dicha matriz nos permitió evaluar la existencia de categorías enriquecidas en cada muestra mediante Test Exacto de Fisher. El análisis se realizó en el nivel 2 de la

categorización SEED lo cual nos permite tener un nivel intermedio de descripción con una potencia representativa a nivel de p menor a 0.001, es decir, el número de elementos a comparar es suficiente para establecer una correlación significativa.

Siguiendo este razonamiento, los gOTUs fueron agrupados por sitio a fin de comparar el número de elementos que suman en cada categoría SEED e identificar la existencia de categorías enriquecidas en cada muestra mediante Test Exacto de Fisher considerando como sitio de referencia a la muestra S2 (suelos de pH neutro), la cual representa la condición más amigable para el establecimiento bacteriano según lo descrito en la introducción. En base a lo anterior, la terminología de comparación se basa en el enriquecimiento (sobre-representación) o empobrecimiento (sub-representación) de características funcionales respecto de S2. De esta forma vemos que S1 posee un enriquecimiento en 15 categorías y un empobrecimiento de 14 categorías SEED de nivel 2, mientras en que S3 sólo una categoría está enriquecida y una disminuida usando el mismo nivel de categorización. Esta notoria diferencia se correlaciona con lo observado al comparar las semejanzas de los sitios a nivel de taxonomía (Figura 2.1.2), pues la composición de la comunidad de S3 se asemeja con S2 indicando una mayor similitud taxonómica, algo que se evidencia al observar que el nivel de intersección de OTUs de ambos sitios llega al 24 % lo cual es el doble de la intersección entre S1 y S2 (Figura 2.1.2, panel A). En efecto, de los 21 gOTUs asignados a S3 por poseer allí una mayor abundancia, 20 están presentes también en S2 y 18 están presentes también en S1. Al realizar esta misma comparación pero centrados en S1 se observa que de los 14 gOTUs asignados a S1 por poseer allí una mayor abundancia, 13 también están presentes en S2 y 9 están presentes también en S3. Por lo

tanto, la similitud funcional entre los genomas de sitios S2 y S3 tiene un fuerte asidero en la similitud de composición de ambos sitios. A pesar de lo anterior resalta en los resultados que la categoría funcional enriquecida en S3 respecto de S2 sea "resistencia a antibióticos y compuestos tóxicos". Al avanzar un nivel dentro de la categoría encontramos que la sub-categoría responsable de esta diferencia es "resistencia a cobalto-zinc-cadmio" la cual en general involucra respuesta a metales pesados. Sabemos que la disponibilidad de Zn no presenta una diferencia significativa entre los tres sitios pero exhibe una tendencia mayor en S1. Dado lo anterior, debemos suponer que esta categoría debe estar activada en respuesta a Cadmio, Cobalto y otros metales pesados. Como se discutió anteriormente, el sitio S3 presenta la mayor disponibilidad de azufre y boro la cual se asocia a una mayor influencia de gases volcánicos, por lo cual es de esperar que la zona también posea una alta disponibilidad de metales pesados a los cuales las bacterias del sector se encuentran habituadas.

Respecto al sitio S1 destacan varias categorías funcionales sobre-representadas, las cuales se discuten a continuación. En primer lugar se ubica la categoría "asimilación orgánica de azufre". Esa categoría está relacionada con las formas que utilizan los organismos para almacenar y transportar azufre en las células. Al observar las sub-categorías encontramos que la razón de este enriquecimiento se debe a que 5 gOTUs de S1 poseen proteínas asociadas al transporte de glutatión. Éste cofactor es un iso-tripéptido producto de la condensación de cisteína en extremo C-terminal del glutamato y una adición posterior de glicina que es utilizado por muchos organismos como reservorio de azufre. Además, por sus propiedades redox y nucleofílicas, determina el tono reductor del citoplasma y participa en reacciones de reducción, constituyendo una

línea de defensa ante especies reactivas de oxígeno (ROS), xenobioticos y metales pesados (Mendoza-Cózatl y col., 2005). Como se ha discutido anteriormente, existen reportes indicando que la zona tiene una importante presencia de metales pesados por lo cual la presencia de esta molécula ya representaría una ventaja, pero en particular el sitio S1, por su bajo pH, representa una zona donde las bacterias se enfrentan a concentraciones elevadas de ROS producto de los mecanismos de adaptación que utilizan modificaciones de la cadena respiratoria para mantener la fuerza protón motriz (Krulwich y col., 2011b). Lo anterior se observa en S1 en las 3 categorías sobre-representadas en gOTUs asociadas a "Respiración" las cuales se contraponen con las sobre-representadas en S2. De esta forma, la presencia de glutatión en algunos gOTUs de S1 podría tener relación directa con el pH del medio y la presencia de metales pesados. Existe otra categoría sobre-representada en S1 que está involucrada con la disponibilidad de azufre: la categoría "nitrito reductasa disimilatoria". Ésta, en sus sub-categorías, se relaciona con la presencia de la proteína uroporphyrinogen III, un precursor de grupo hemo que permite la síntesis del grupo prostético "siroheme (sirohaem)" (Hansen y col., 1997). Éste grupo es usado por algunas enzimas para realizar la reducción de azufre y nitrógeno (Murphy y col., 1974), siendo esencial en *S. cereviseae* para la asimilación de azufre al convertir el sulfito en sulfido, el cual puede ser incorporado en compuestos orgánicos como la homocisteína (D. Thomas & Surdin-Kerjan, 1997). Considerando todo lo anterior creemos que la presencia de glutatión es central en los gOTUs identificados en S1. La categoría "trasportadores putativos de *hemin*", sobre-representada en S1 respecto de S2 y S3, involucra todo lo relacionado con el transporte, permeabilidad y unión a lípidos del grupo prostético *hemin* el cual es el

grupo *hemo* presente en peroxidasas, catalasas y citocromos.

Otra categoría sobre-representada en S1 respecto de S2 y S3 es "transportadores ABC" cuya categoría interna apunta a los "transportadores ABC de aminoácidos de cadena ramificada" que corresponden a leucina, isoleucina y valina. Estos transportadores están asociados a la obtención de estos aminoácidos desde otros organismos, es decir una actividad principalmente simbionte y/o patógena (Prell y col., 2009; Hsu-Ming y col., 2012; Belitsky, 2015). Taxonómicamente ninguno de los 74 genomas reconstruidos puede ser vinculado de forma directa a estos estilos de vida por lo cual no es clara la diferenciación encontrada en esta categoría respecto de los 3 sitios. Sorprendentemente la categoría de "síntesis y degradación de aminoácidos ramificados" también está sobre-representada lo cual nos entrega una nueva arista para revisar en trabajos futuros.

Respecto a las categorías sobre-representadas en S2 respecto de S1 destacan los mecanismos de adquisición y metabolismo de hierro con las categorías "adquisición de hierro en *Vibrio*", "sideróforos" y "sistemas de transporte tipo Ton y Tol (específicamente tonB)". Debido a los distintos tipos de nicho en donde vive *Vibrio cholerae* y su absoluto requerimiento de hierro se sabe que posee múltiples sistemas de adquisición de hierro que incluyen el sistema de transporte dependiente de tonB y el uso de distintos tipos de sideróforos (Wyckoff y col., 2007), los cuales están en su mayoría sobre-representados en S2 y también en S3 (datos no mostrados). Según los datos de disponibilidad de nutrientes (Tabla 1.1) sabemos que los sitios S2 y S3 presentan menor disponibilidad de Fe respecto del sitio S1, lo cual es una de las variables influyentes en la estructura de las comunidades bacterianas inversamente asociada al pH del suelo

(Figura 2.3.1). De esta forma, los datos de capacidad funcional de los genomas reconstruidos sugieren la influencia directa de condiciones ambientales sobre la selección de miembros de la comunidad con capacidades específicas.

Algo que pareciera ser específico de S2 son los mecanismos de "resistencia a invasión intracelular", a la cual se suman otras categorías sobre-representadas como el sistema "CRISPR" (Barrangou y col., 2007; Marraffini & Sontheimer, 2008, 2010), los "compuestos biológicamente activos en metazoos para la defensa y diferenciación celular (sulfatos esteroides)" (Marinho y col., 2012) y "sistemas de modificación mediados por restricción molecular (enzimas de restricción de tipo I y tipo III)" (Westra y col., 2012; Chen y col., 2016). En su mayoría se trata de mecanismos que previenen la transferencia horizontal de DNA lo cual podría indicar que en la comunidad bacteriana de S2 existe mayor competencia entre especies y posiblemente también haya mayor presencia viral.

Respecto a la fijación de nitrógeno, S2 presenta gOTUs con presencia de distintos miembros del operón Nif (Merrick & Edwards, 1995), como nifS, nifU y nifA. Según se observa en el panel B de la Figura 2.3.1, la comunidad bacteriana de S2 está influenciada por la disponibilidad de N en un vector que se descompone en NO3 y NH4. Si bien no es posible establecer por ahora el tipo de mecanismo específico, sabemos por taxonomía que los gOTUs de S1 no pertenecen a los grupos clásicos de fijación simbionte y tampoco es un sector donde existan plantas relacionadas con esta función, por lo tanto es probable que se trate de mecanismos de fijación de nitrógeno en vida libre (N_2).

3.2 Análisis comparado de proteínas asociadas homeostasis de pH en los genomas ensamblados de S1, S2 y S3

Mientras la mayoría de los eucariontes exhiben un pH interno de 7,3 y dependen fuertemente de un pH externo de 7,4 (Casey y col., 2010), algunas bacterias neutralofílicas pueden crecer con un pH externo entre 5,5 y 9,0 pero generalmente mantienen su pH citoplasmático cercano al rango de 7,5-7,7 (Padan y col., 2005; Slonczewski y col., 2009a). Estas bacterias pueden ser expuestas a pH más extremo y sobrevivir, pero inhibirán su capacidad reproductiva para concentrarse en mantener activa la maquinaria de homeostasis de pH. Por el contrario, las bacterias acidofílicas son capaces de crecer y reproducirse entre pH 1,0 y 3,0 con un crecimiento moderado entre pH 3,0 y 5,0, y las bacterias alcalofílicas entre pH 10,0 y 13,0 con un crecimiento moderado entre pH 9,0 y 10,0 (Gerday y col., 2007; Gerday & Glansdorff, 2009). Las bacterias extremófilas, como las que se encuentran en las zonas escogidas del Desierto de Atacama, usan varias de las estrategias presentes en bacterias neutrhalofílicas, pero con adaptaciones para responder a desafíos más extremos (Slonczewski y col., 2009a). En general, las estrategias utilizadas son las siguientes:

- Regulación del consumo de protones o generación de protones por enzimas metabólicas
- Expulsión o captura de protones
- Cambios en la permeabilidad de membrana a protones
- Adaptaciones genómicas (uso de codones)

- Mejores mecanismos de reparación de DNA y proteínas

Cada una de estas estrategias involucra elementos funcionales particulares que varían entre especies y que han sido descritos a partir de organismos modelo para cada rango de pH (Baker-Austin & Dopson, 2007; Slonczewski y col., 2009a; Krulwich y col., 2011a).

En base a estos mecanismos se seleccionaron 13 tipos de proteínas asociadas a respuesta a pH en bacterias (Tabla 3.2.1) las cuales se usaron como molde para encontrar ortólogos en cada genoma reconstruido. Al agrupar el resultado de homología de genes según el sitio donde se observó cada genoma se observa que la mayoría de las proteínas poseen representantes en los genomas de los 3 sitios (Tabla 3.2.2). Ello nos muestra que en el ámbito de las capacidades funcionales asociadas a respuesta a pH evaluadas, no existe un sesgo evidente (todo o nada) entre los sitios. Al revisar los resultados en profundidad mediante un Test Exacto de Fisher se observa una diferencia significativa en presencia de la subunidad B de la proteína citocromo bo (CyoB). Esta proteína es una bomba que expulsa protones al exterior y cuya actividad se ve incrementada en *E. coli* ante un desafío ácido (Slonczewski y col., 2009a). Al observar el contexto de esta proteína tenemos que el complejo citocromo bo (subunidades A, B, C, E) está presente en el 82 % de los genomas de sitio S1, mientras que está en el 74 % de los genomas de S2 y en el 75 % de los genomas de S3. Esto nos muestra una tendencia general respecto de la presencia de este complejo en los sitios con una mayor presencia en el sitio ácido (S1), lo cual concuerda con lo esperado según lo descrito en *E. coli* (Padan y col., 2005; Slonczewski y col., 2009a). Otra de las proteínas con

presencia significativa en S1 es FlgS la cual es un sensor de histidina que pertenece a la familia de los sensores de dos componentes y que probablemente responde a variaciones de pH en el periplasma. Se ha establecido que esta proteína es parte esencial del mecanismo de regulación de pH citoplasmático mediado por la expresión de la enzima ureasa en la bacteria gram negativa *H. pylori*. El mecanismo involucra la interacción entre FlgS y ArsRS (otro sensor dependiente de pH) para regular el pH del periplasma. De esta forma la bacteria logra sobrevivir oscilaciones entre pH ácido y neutro propios del ambiente gástrico. La presencia significativa FlgS en los gOTUs de la muestra S1 podría dar indicios de una capacidad de homeostasis de pH mediada por regulación periplasmática pero de una forma diferente a la descrita en *H. pylori*, algo que abre nuevas posibilidades de exploración en el área (Krulwich y col., 2011b). Al revisar la incidencia de FlgS en los genomas reconstruidos se observa que está presente en el 80 % de los gOTUs de S1, el 67 % de S2 y 70 % de S3 indicando nuevamente una tendencia favorable en las comunidades de pH ácido.

Respecto los gOTUs presentes en pH acalino (muestra S3), se observa que éstos tienen significativamente más presencia de la proteína acídica de capa de superficie (Acidic surface-layer protein: SlpA), quien es el material estructural de la capa monomolecular de superficie (capa-S) de algunas bacterias (Krulwich y col., 2011b). Estos polímeros pueden unir cationes, lo cual contribuiría a la captura de H^+ y Na^+, incrementando su disponibilidad cerca de la superficie contribuyendo a la homeostasis de pH y al trabajo bio-energético en condiciones de pH alcalino (Padan y col., 2005). Específicamente son sólo 3 genomas reconstruidos los que poseen dicha proteína: gOTU B23 del orden actinomycetales con 14 ortólogos, gOTU B117 sin asignación

taxonómica con 11 ortólogos y gOTU B6 del orden actinomycetales con 4 ortólogos. Si bien la presencia no es exclusiva de estos genomas pues también se encontró en gOTU AN22 y BN134, ambos de muestra neutra, en ellos sólo existe un ortólogo, mientras que no se encontraron ortólogos en gOTUs de muestra ácida.

Considerando que éste es el estado del arte en cuanto a respuesta a pH, nuestros resultados nos indican que a nivel general los mecanismos descritos en organismos modelos no representan necesariamente una polarización de capacidades funcionales en respuesta a pH en organismos de comunidades naturales las cuales, según hemos encontrado mediante el análisis diferencial de las categorías SEED, presentan diferencias significativas como conjunto en categorías completas. Sin duda es necesario tener en consideración otras capas de información para corroborar lo encontrado hasta ahora en la respuesta específica a pH, como por ejemplo evaluar la expresión de los genes y/o la presencia proteínas lo cual requeriría desplegar nuevas capacidades de muestreo en el terreno.

CONCLUSIONES

El cambio de abundancia relativa de taxas es una herramienta útil para discriminar genomas de taxas particulares dentro de metagenomas pertenecientes a comunidades bacterianas de suelo sometidas a condiciones contrastantes de pH.

Los resultados de la caracterización fisicoquímica de los sitios seleccionados permitieron establecer un ambiente útil para abordar el estudio de comunidades de microorganismos en un sector no intervenido por actividad humana. Además, los suelos de esta cuenca altiplánica representan un modelo de perturbación ambiental en forma de un gradiente de pH, variable que explica gran parte de la diversidad de bacterias en ambientes terrestres.

El análisis taxonómico permitió caracterizar la estructura de las comunidades de microbianas de los 3 sitios estudiados, junto con establecer sub-grupos que comparten un patrón de abundancia sitio-específico. Además, siguiendo un modelo clásico de análisis de comunidades bacterianas, se ensambló el metagenoma de cada muestra con un porcentaje de utilización de lecturas en torno al 32 % lo cual es superior a lo descrito en muestras similares. Al revisar los perfiles de categorías funcionales obtenidos desde la anotación funcional de los metagenomas se observa que la magnitud de las diferencias funcionales entre las 3 muestras, es de menor envergadura a las diferencias observadas en la composición taxonómica, lo cual es similar a lo descrito hasta el momento para metagenomas de tracto intestinal, mar y suelo.

Utilizando criterios dependientes e independientes de la composición de secuencia fue posible recuperar 74 genomas (gOTUs), los cuales presentan una distribución diferencial respecto de los 3 sitios (S1, S2 o S3). De esta forma se obtienen 14 genomas con mayor abundancia en S1, 39 genomas con mayor abundancia en S2 y 21 con mayor abundancia en S3. Al evaluar diferentes parámetros de calidad en los genomas reconstruidos los resultados sustentan la estrategia de análisis utilizada en la discriminación de "scaffolds". A la vez se refuerza la propuesta de utilizar una mezcla de métodos para maximizar la recuperación de genomas desde muestras complejas. El proceso de identificación taxonómica nos permitió realizar una asignación al 88 % de los gOTUs, logrando llegar a nivel de género en 37 casos.

El análisis integrado de datos nos permitió analizar las correlaciones entre la composición de las comunidades y las variables fisicoquímicas del sector, observando que la variabilidad biológica en cada sitio es baja respecto de la variabilidad entre sitios. Ello sugiere que la distribución diferencial de los gOTUs recuperados se explica por una adaptación funcional de estos microorganismos a las particularidades fisicoquímicas de los sitios analizados. De igual forma, a pesar que existe un gran número de OTUs compartidos entre los 3 sitios, se observa que ciertos filos poseen una fuerte tendencia a estar en sitios particulares, especialmente filos representantes de bacterias poco conocidas hasta ahora ("dark matter") y cuyo potencial promete enriquecer las posibilidades metabólicas conocidas.

El sitio S1 presentó una correlación positiva con los micronutrientes Mn, Fe, Cu y Zn pero negativa y más débil con pH y K, lo cual es de esperar considerando que el pH determina en gran medida la solubilidad de esos metales. Por otro lado, S3 posee una

correlación positiva con B, S, Ca, Na, C y Conductividad Eléctrica, siendo este el sitio con mayor aporte de material proveniente de emisiones volcánicas del Lazcar y el más cercano a la Laguna Legía, lo cual nos permite predecir que posee una mayor presencia de metales pesados respecto a los otros sitios.

Al categorizar la anotación funcional de los gOTUs en base a SEED y comparar la distribución de funciones entre los sitios se encontraron correlaciones positivas con las características fisicoquímicas de los suelos, varias de ellas ligadas de forma directa al pH del suelo. Destaca en el sitio S1 la presencia de proteínas asociadas a la síntesis de glutatión, el cual se asocia a la asimilación de azufre, resistencia a metales pesados y resistencia a ROS, el cual puede ser producido producto de mecanismos de homeostasis de pH (ajustes en la cadena respiratoria). En cambio, en S2 destacan las categorías asociadas a mecanismos de adquisición de hierro, cuya disponibilidad es menor comparada con S1 posiblemente debido a la diferencia de pH.

La búsqueda de proteínas relacionadas con los mecanismos de homeostásis de pH descritos en organismos modelos en los gOTUs de cada sito no reveló una polarización de capacidades funcionales en respuesta a pH. De hecho existe una presencia transversal de estas proteínas en los gOTUs de los 3 sitios, a pesar que mediante la comparación de categorías SEED se observan enriquecimientos de categorías asociadas a homeostasis de pH. Sin duda, es necesario tener en consideración otras capas o niveles de información para corroborar lo encontrado hasta ahora en la respuesta específica a pH. A pesar de lo anterior destaca en el sitio S1 la presencia de la proteína CyoB, existiendo una tendencia en este sitio a la presencia completa del complejo citocromo bo el cual es una bomba de protones utilizada por distintos

organismos para mantener la FPM ante condiciones de pH bajo. Por otro lado, la proteína acídica de capa de superficie SlpA se encuentra sobre-representada en algunos gOTUs de S3, sugiriendo que en estos genomas puede existir un mecanismo de atracción superficial de protones ante condiciones de pH alto.

Finalmente, hemos encontrado que existen relaciones entre las categorías funcionales enriquecidas en los gOTUs y las características del ambiente de cada sitio, las cuales tienen una importante asociación a la variación de pH.

Las capacidades generadas durante el desarrollo de la tesis no sólo entregaron nuevas fórmulas para potenciar la reconstrucción desde comunidades complejas, además demostraron la utilidad de estas herramientas para establecer vínculos entre las capacidades funcionales de los miembros de una comunidad bacteriana y su relación con las características del ambiente donde viven.

REFERENCIAS BIBLIOGRÁFICAS

Ahlert, Robert C. (1998). An introduction to soils for environmental professionals Duane L. Winegardner, CRC Press, Inc., Boca Raton, FL, (1996), 270 pages, [ISBN No.: 1-87371-939-5]. *Environmental Progress, 17*(2), A7-A8. doi: 10.1002/ep.670170205

Ahmadian, A., Ehn, M., & Hober, S. (2006). Pyrosequencing: History, biochemistry and future. *Clinica Chimica Acta, 363*(1-2), 83-94. doi: 10.1016/j.cccn.2005.04.038

Alain, Karine, & Querellou, Joël. (2009). Cultivating the uncultured: limits, advances and future challenges. *Extremophiles, 13*(4), 583-594. doi: 10.1007/s00792-009-0261-3

Albertsen, M, Hugenholtz, P, Skarshewski, A, Nielsen, K. L., Tyson, Gene W., & Nielsen, Per H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotech, 31*(6), 533-538. doi: 10.1038/nbt.2579

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nat Methods, 11*(11), 1144-1146. doi: 10.1038/nmeth.3103

Andrew, D. R., Fitak, R. R., Munguia-Vega, A., Racolta, A., Martinson, V. G., & Dontsova, K. (2012). Abiotic factors shape microbial diversity in Sonoran Desert soils. *Applied and Environmental Microbiology, 78*(21), 7527-7537. doi: 10.1128/AEM.01459-12

Andrews, S. C., Robinson, A. K., & Rodriguez-Quinones, F. (2003). Bacterial iron homeostasis. *FEMS Microbiol Rev, 27*(2-3), 215-237.

Aono, R., Ito, M., & Machida, T. (1999). Contribution of the cell wall component teichuronopeptide to pH homeostasis and alkaliphily in the alkaliphile Bacillus lentus C-125. *J. Bacteriol., 181*, 6600-6606.

Azua-Bustos, A., Urrejola, C., & Vicuna, R. (2012). Life at the dry edge: microorganisms of the Atacama Desert. *FEBS Lett, 586*(18), 2939-2945. doi: 10.1016/j.febslet.2012.07.025

Bachar, A., Soares, M. I. M., & Gillor, O. (2012). The

Effect of Resource Islands on Abundance and Diversity of Bacteria in Arid Soils. *Microbial Ecology, 63*(3), 694-700. doi: 10.1007/s00248-011-9957-x

Bailey, V. L., Fansler, S. J., Stegen, J. C., & McCue, L. A. (2013). Linking microbial community structure to beta-glucosidic function in soil aggregates. *ISME J, 7*(10), 2044-2053. doi: 10.1038/ismej.2013.87

Baker-Austin, C., & Dopson, M. (2007). Life in acid: pH homeostasis in acidophiles. *Trends Microbiol, 15*(4), 165-171. doi: 10.1016/j.tim.2007.02.005

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science, 315*(5819), 1709-1712. doi: 10.1126/science.1138140

Belitsky, B. R. (2015). Role of branched-chain amino acid transport in Bacillus subtilis CodY activity. *J Bacteriol, 197*(8), 1330-1338. doi: 10.1128/jb.02563-14

Bentley, S. D., & Parkhill, J. (2004). Comparative genomic structure of prokaryotes. *Annu Rev Genet, 38*, 771-792. doi: 10.1146/annurev.genet.38.072902.094318

Betancourt, J. L., Latorre, C., Rech, J. A., Quade, J., & Rylander, K. A. (2000). A 22,000-Year Record of Monsoonal Precipitation from Northern Chile's Atacama Desert. *Science, 289*(5484), 1542-1546.

Blankenhorn, D., Phillips, J., & Slonczewski, J. L. (1999). Acid- and base-induced proteins during aerobic and anaerobic growth of Escherichia coli revealed by two-dimensional gel electrophoresis. *J. Bacteriol., 181*, 2209-2216.

Brady, Sean F., Chao, Carol J., Handelsman, Jo, & Clardy, Jon. (2001). Cloning and Heterologous Expression of a Natural Product Biosynthetic Gene Cluster from eDNA. *Organic Letters, 3*(13), 1981-1984. doi: 10.1021/ol015949k

Bremner, J.M., & Mulvaney, C.S. (1982). Total nitrogen. In A. L. Page, R. H. Miller & D. R. Kennedy (Eds.), *Methods of soil analysis* (Vol. 2, pp. 595-624). Madison, Wisc.: Agron. Monogr. 9, Am. Soc. Agron.

Bryant, J. A., Lamanna, C., Morlon, H., Kerkhoff, A. J., Enquist, B. J., & Green, J. L. (2008). Microbes on mountainsides: Contrasting elevational patterns of bacterial and plant diversity. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 11505-11511. doi:

10.1073/pnas.0801920105
Cabrol, Nathalie A., Grin, Edmond A., Chong, Guillermo, Minkley, Edwin, Hock, Andrew N., Yu, Youngseob, Bebout, Leslie, Fleming, Erich, Haeder, Donat P., Demergasso, Cecilia, Gibson, John, Escudero, Lorena, Dorador, Cristina, Lim, Darlene, Woosley, Clayton, Morris, Robert L., Tambley, Cristian, Gaete, Victor, Galvez, Matthieu E., Smith, Eric, Peate, Ingrid Ukstins, Salazar, Carlos, Dawidowicz, G., & Majerowicz, J. (2009). The High-Lakes Project. *Journal of Geophysical Research-Biogeosciences, 114*. doi: 10.1029/2008jg000818
Caporaso, J. Gregory, Kuczynski, Justin, Stombaugh, Jesse, Bittinger, Kyle, Bushman, Frederic D., Costello, Elizabeth K., Fierer, Noah, Peña, Antonio Gonzalez, Goodrich, Julia K., Gordon, Jeffrey I., Huttley, Gavin A., Kelley, Scott T., Knights, Dan, Koenig, Jeremy E., Ley, Ruth E., Lozupone, Catherine A., McDonald, Daniel, Muegge, Brian D., Pirrung, Meg, Reeder, Jens, Sevinsky, Joel R., Turnbaugh, Peter J., Walters, William A., Widmann, Jeremy, Yatsunenko, Tanya, Zaneveld, Jesse, & Knight, Rob. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods, 7*(5), 335-336. doi: 10.1038/nmeth.f.303
Carr, Rogan, Shen-Orr, Shai S., & Borenstein, Elhanan. (2013). Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution. *PLoS Comput Biol, 9*(10), e1003292. doi: 10.1371/journal.pcbi.1003292
Casey, Joseph R., Grinstein, Sergio, & Orlowski, John. (2010). Sensors and regulators of intracellular pH. *Nat Rev Mol Cell Biol, 11*(1), 50-61.
Ciccarelli, Francesca D., Doerks, Tobias, von Mering, Christian, Creevey, Christopher J., Snel, Berend, & Bork, Peer. (2006). Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science, 311*(5765), 1283-1287. doi: 10.1126/science.1123061
Colombo, Claudio, Palumbo, Giuseppe, He, Ji-Zheng, Pinton, Roberto, & Cesco, Stefano. (2014). Review on iron availability in soil: interaction of Fe minerals, plants, and microbes. *Journal of Soils and Sediments, 14*(3), 538-548. doi: 10.1007/s11368-013-0814-z
Consortium, Human Microbiome Project. (2012). Structure, function and diversity of the healthy human

microbiome. *Nature, 486*(7402), 207-214. doi: 10.1038/nature11234

Consortium, The Human Microbiome Project. (2012). Structure, function and diversity of the healthy human microbiome. *Nature, 486*(7402), 207-214. doi: 10.1038/nature11234

Courtois, S., Cappellano, C. M., Ball, M., Francou, F. X., Normand, P., Helynck, G., Martinez, A., Kolvek, S. J., Hopke, J., Osburne, M. S., August, P. R., Nalin, R., Guerineau, M., Jeannin, P., Simonet, P., & Pernodet, J. L. (2003). Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl Environ Microbiol, 69*(1), 49-55.

Cowan, D., Meyer, Q., Stafford, W., Muyanga, S., Cameron, R., & Wittwer, P. (2005). Metagenomic gene discovery: past, present and future. *Trends in Biotechnology, 23*(6), 321-329. doi: 10.1016/j.tibtech.2005.04.001

Crits-Christoph, A., Robinson, C. K., Barnum, T., Fricke, W. F., Davila, A. F., Jedynak, B., McKay, C. P., & Diruggiero, J. (2013). Colonization patterns of soil microbial communities in the Atacama Desert. *Microbiome, 1*(1), 28. doi: 10.1186/2049-2618-1-28

Curtis, T. P., & Sloan, W. T. (2004). Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology. *Curr Opin Microbiol,* 7(3), 221-226. doi: 10.1016/j.mib.2004.04.010

Chen, K., Stephanou, A. S., Roberts, G. A., White, J. H., Cooper, L. P., Houston, P. J., Lindsay, J. A., & Dryden, D. T. (2016). The Type I Restriction Enzymes as Barriers to Horizontal Gene Transfer: Determination of the DNA Target Sequences Recognised by Livestock-Associated Methicillin-Resistant Staphylococcus aureus Clonal Complexes 133/ST771 and 398. *Adv Exp Med Biol, 915,* 81-97. doi: 10.1007/978-3-319-32189-9_7

Chi, A. (2007). Periplasmic proteins of the extremophile Acidithiobacillus ferrooxidans: a high throughoutput proteomics analysis. *Mol. Cell. Proteomics, 6,* 2239-2251.

Davila, A. F., Hawes, I., Ascaso, C., & Wierzchos, J. (2013). Salt deliquescence drives photosynthesis in the hyperarid Atacama Desert. *Environ Microbiol Rep, 5*(4), 583-587. doi: 10.1111/1758-2229.12050

Davis, K. E. R., Joseph, S. J., & Janssen, P. H. (2005).

Effects of growth medium, inoculum size, and incubation time on culturability and isolation of soil bacteria. *Applied and Environmental Microbiology, 71*(2), 826-834.

Delmont, Tom O, Eren, A Murat, Maccario, Lorrie, Prestat, Emmanuel, Esen, Özcan, Pelletier, Eric, LePaslier, Denis, SIMONET, Pascal, & Vogel, Timothy. (2015). Reconstructing Rare Soil Microbial Genomes using in situ Enrichments and Metagenomics. *Frontiers in Microbiology, 6*. doi: 10.3389/fmicb.2015.00358

Demergasso, Cecilia, Dorador, Cristina, Meneses, Daniela, Blamey, Jenny, Cabrol, Nathalie, Escudero, Lorena, & Chong, Guillermo. (2010). Prokaryotic diversity pattern in high-altitude ecosystems of the Chilean Altiplano. *Journal of Geophysical Research-Biogeosciences, 115*. doi: 10.1029/2008jg000836

Díaz, Francisca P., Frugone, Matías, Gutiérrez, Rodrigo A., & Latorre, Claudio. (2016). Nitrogen cycling in an extreme hyperarid environment inferred from δ15N analyses of plants, soils and herbivore diet. *Scientific Reports, 6*, 22226. doi: 10.1038/srep22226

Dodsworth, J. A., Blainey, P. C., Murugapiran, S. K., Swingley, W. D., Ross, C. A., Tringe, S. G., Chain, P. S. G., Scholz, M. B., Lo, C., Raymond, J., Quake, S. R., & Hedlund, B. P. (2013). Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat Commun, 4*, 1854. doi: 10.1038/ncomms2884

Dowd, S. E., Callaway, T. R., Wolcott, R. D., Sun, Y., McKeehan, T., Hagevoort, R. G., & Edrington, T. S. (2008). Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP). *BMC Microbiol, 8*, 125. doi: 10.1186/1471-2180-8-125

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics, 26*(19), 2460-2461. doi: 10.1093/bioinformatics/btq461

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics, 27*(16), 2194-2200. doi: 10.1093/bioinformatics/btr381

Ettler, V., Mihaljevic, M., Sebek, O., & Grygar, T. (2007). Assessment of single extractions for the determination of mobile forms of metals in highly polluted soils and sediments--analytical and thermodynamic approaches.

Anal Chim Acta, 602(1), 131-140. doi: 10.1016/j.aca.2007.09.017

Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci U S A, 103*(3), 626-631. doi: 10.1073/pnas.0507535103

Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J, 6*(5), 1007-1017. doi: 10.1038/ismej.2011.159

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., Owens, S., Gilbert, J. A., Wall, D. H., & Caporaso, J. G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci USA, 109*(52), 21390-21395. doi: 10.1073/pnas.1215210110

Fierer, N., McCain, C. M., Meir, P., Zimmermann, M., Rapp, J. M., Silman, M. R., & Knight, R. (2011). Microbes do not follow the elevational diversity patterns of plants and animals. *Ecology, 92*(4), 797-804.

Fischer, A., Blaschke, M., & Bässler, C. (2011). Altitudinal gradients in biodiversity research: the state of the art and future perspectives under climate change aspects. *Forest Ecology, Landscape Research and Nature Conservation, 11*, 35-47.

Foster, J. W. (2004). Escherichia coli acid resistance: tales of an amateur acidophile. *Nature Rev. Microbiol., 2*, 898-907.

Gans, J., Wolinsky, M., & Dunbar, J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science, 309*(5739), 1387-1390. doi: 10.1126/science.1112665

Garcia Costas, A. M., Liu, Z., Tomsho, L. P., Schuster, S. C., Ward, D. M., & Bryant, D. A. (2012). Complete genome of Candidatus Chloracidobacterium thermophilum, a chlorophyll-based photoheterotroph belonging to the phylum Acidobacteria. *Environ Microbiol, 14*(1), 177-190. doi: 10.1111/j.1462-2920.2011.02592.x

Garcia Martin, H., Ivanova, N., Kunin, V., Warnecke, F., Barry, K. W., McHardy, A. C., Yeates, C., He, S., Salamov, A. A., Szeto, E., Dalin, E., Putnam, N. H., Shapiro, H. J., Pangilinan, J. L., Rigoutsos, I., Kyrpides, N. C., Blackall, L. L., McMahon, K. D., &

Hugenholtz, P. (2006). Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol, 24*(10), 1263-1269. doi: 10.1038/nbt1247

Gerday, C., & Glansdorff, N. (2009). *Extremophiles*: Eolss Publishers Co Ltd.

Gerday, C., Glansdorff, N., & Microbiology, American Society for. (2007). *Physiology and Biochemistry of Extremophiles*: ASM Press.

Gilmour, R. (2000). Two-dimensional gel electrophoresis analyses of pH-dependent protein expression in facultatively alkaliphilic Bacillus pseudofirmus OF4 lead to characterization of an S-layer protein with a role in alkaliphily. *J. Bacteriol., 182*, 5969-5981.

Glaser, B., Turrion, M. B., & Alef, K. (2004). Amino sugars and muramic acid - biomarkers for soil microbial community structure analysis. *Soil Biology & Biochemistry, 36*(3), 399-407. doi: 10.1016/j.soilbio.2003.10.013

Gómez-Silva, Benito, Rainey, FredA, Warren-Rhodes, KimberleyA, McKay, ChristopherP, & Navarro-González, Rafael. (2008). Atacama Desert Soil Microbiology. In P. Dion & C. Nautiyal (Eds.), *Microbiology of Extreme Soils* (Vol. 13, pp. 117-132): Springer Berlin Heidelberg.

Grigoriev, A. (1998). Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res, 26*(10), 2286-2290.

Gut, H. (2006). Escherichia coli acid resistance: pH-sensing, activation by chloride and autoinhibition in GadB. *EMBO J., 25*, 2643-2651.

Hallenbeck, P. C., Grogger, M., Mraz, M., & Veverka, D. (2016). Draft Genome Sequence of the Photoheterotrophic Chloracidobacterium thermophilum Strain OC1 Found in a Mat at Ojo Caliente. *Genome Announc, 4*(1). doi: 10.1128/genomeA.01570-15

Handelsman, J. (2004). Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews, 68*(4), 669-685.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry and Biology, 5*(10), R245-R249.

Handl, S., Dowd, S. E., Garcia-Mazcorro, J. F., Steiner, J. M., & Suchodolski, J. S. (2011). Massive parallel 16S

rRNA gene pyrosequencing reveals highly diverse fecal bacterial and fungal communities in healthy dogs and cats. *FEMS Microbiol Ecol, 76*(2), 301-310. doi: 10.1111/j.1574-6941.2011.01058.x

Hansen, J., Muldbjerg, M., Cherest, H., & Surdin-Kerjan, Y. (1997). Siroheme biosynthesis in Saccharomyces cerevisiae requires the products of both the MET1 and MET8 genes. *FEBS Lett, 401*(1), 20-24.

Hirsch, P. R., Mauchline, T. H., & Clark, I. M. (2010). Culture-independent molecular techniques for soil microbial ecology. *Soil Biology and Biochemistry, 42*(6), 878-887.

Hirsch, P. R., Mauchline, T. H., & Clark, I. M. (2013). Culture-Independent Molecular Approaches to Microbial Ecology in Soil and the Rhizosphere. *Molecular Microbial Ecology of the Rhizosphere, Two Volume Set*, 45.

Hou, R., Yang, Z. X., Li, M. H., & Xiao, H. S. (2013). Impact of the next-generation sequencing data depth on various biological result inferences. *Science China-Life Sciences, 56*(2), 104-109. doi: 10.1007/s11427-013-4441-0

Hsu-Ming, W., Naito, K., Kinoshita, Y., Kobayashi, H., Honjoh, K., Tashiro, K., & Miyamoto, T. (2012). Changes in transcription during recovery from heat injury in Salmonella typhimurium and effects of BCAA on recovery. *Amino Acids, 42*(6), 2059-2066. doi: 10.1007/s00726-011-0934-y

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree of life. *Nat Microbiol, 1*, 16048. doi: 10.1038/nmicrobiol.2016.48

Hug, L. A., Thomas, B. C., Sharon, I., Brown, C. T., Sharma, R., Hettich, R. L., Wilkins, M. J., Williams, K. H., Singh, A., & Banfield, J. F. (2016). Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environ Microbiol, 18*(1), 159-173. doi: 10.1111/1462-2920.12930

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N., & Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res,*

21(9), 1552-1560. doi: 10.1101/gr.120618.111

Hyatt, Doug, LoCascio, Philip F., Hauser, Loren J., & Uberbacher, Edward C. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics, 28*(17), 2223-2230. doi: 10.1093/bioinformatics/bts429

Ikegami, M. (1999). Enterococcus hirae vacuolar ATPase is expressed in response to pH as well as sodium. *FEBS Lett., 454*, 67-70.

Jones, J.B. (2012). *Plant Nutrition and Soil Fertility Manual, Second Edition*: Taylor & Francis.

Kakinuma, Y. (1998). Inorganic cation transport and energy transduction in Enterococcus hirae and other streptococci. *Microbiol. Mol. Biol. Rev., 62*, 1021-1045.

Kennedy, S. P., Ng, W. V., Salzberg, S. L., Hood, L., & DasSarma, S. (2001). Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Res., 11*, 1641-1650.

Kilpinen, H., & Barrett, J. C. (2013). How next-generation sequencing is transforming complex disease genetics. *Trends in Genetics, 29*(1), 23-30. doi: 10.1016/j.tig.2012.10.001

Kim, J., Lee, S., Shin, H., Kim, S. C., & Cho, B. K. (2012). Elucidation of bacterial genome complexity using next-generation sequencing. *Biotechnology and Bioprocess Engineering, 17*(5), 887-899. doi: 10.1007/s12257-012-0374-x

Knight, C. G., Kassen, R., Hebestreit, H., & Rainey, P. B. (2004). Global analysis of predicted proteomes: functional adaptation of physical properties. *Proc. Natl Acad. Sci. USA, 101*, 8390-8395.

Kobayashi, H., Suzuki, T., & Unemoto, T. (1986). Streptococcal cytoplasmic pH is regulated by changes in amount and activity of a proton-translocating ATPase. *J. Biol. Chem., 261*, 627-630.

Kohonen, T. (2001). *Self-Organizing Maps*: Springer Berlin Heidelberg.

Krulwich, T. A., Sachs, G., & Padan, E. (2011a). Molecular aspects of bacterial pH sensing and homeostasis. *Nat Rev Microbiol, 9*(5), 330-343. doi: 10.1038/nrmicro2549

Krulwich, T. A., Sachs, G., & Padan, E. (2011b). Molecular aspects of bacterial pH sensing and homeostasis. *Nat Rev Micro, 9*(5), 330-343. doi: 10.1038/nrmicro2549

Kuch, M., Rohland, N., Betancourt, J. L., Latorre, C., Steppan, S., & Poinar, H. N. (2002). Molecular analysis of an 11,700-year-old rodent midden from the Atacama Desert, Chile. *Mol Ecol, 11*(5), 913-924.

Kunin, Victor, Copeland, Alex, Lapidus, Alla, Mavromatis, Konstantinos, & Hugenholtz, Philip. (2008). A Bioinformatician's Guide to Metagenomics. *Microbiology and Molecular Biology Reviews, 72*(4), 557-578. doi: 10.1128/mmbr.00009-08

Langmead, B, & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Meth, 9*(4), 357-359. doi: 10.1038/nmeth.1923

Lasken, Roger S. (2012). Genomic sequencing of uncultured microorganisms from single cells. *Nat Rev Micro, 10*(9), 631-640.

Lauber, C. L., Hamady, M., Knight, R., & Fierer, N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol, 75*(15), 5111-5120. doi: 10.1128/AEM.00335-09

Legendre, P., & Legendre, L.F.J. (1998). *Numerical Ecology*: Elsevier Science.

Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A, 113*(21), 5970-5975. doi: 10.1073/pnas.1521291113

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., & Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature, 489*(7415), 220-230. doi: 10.1038/nature11550

Mande, Sharmila S., Mohammed, Monzoorul Haque, & Ghosh, Tarini Shankar. (2012). Classification of metagenomic sequences: methods and challenges. *Briefings in Bioinformatics, 13*(6), 669-681. doi: 10.1093/bib/bbs054

Marinho, P. R., Simas, N. K., Kuster, R. M., Duarte, R. S., Fracalanzza, S. E., Ferreira, D. F., Romanos, M. T., Muricy, G., Giambiagi-Demarval, M., & Laport, M. S. (2012). Antibacterial activity and cytotoxicity analysis of halistanol trisulphate from marine sponge Petromica citrina. *J Antimicrob Chemother, 67*(10), 2396-2400. doi: 10.1093/jac/dks229

Marraffini, L. A., & Sontheimer, E. J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science, 322*(5909), 1843-1845. doi: 10.1126/science.1165771

Marraffini, L. A., & Sontheimer, E. J. (2010). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet, 11*(3), 181-190. doi: 10.1038/nrg2749

Maurer, L. M., Yohannes, E., Bondurant, S. S., Radmacher, M., & Slonczewski, J. L. (2005). pH regulates genes for flagellar motility, catabolism, and oxidative stress in Escherichia coli K-12. *J. Bacteriol., 187*, 304-319.

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., & Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution, 66*(2), 526-538. doi: 10.1016/j.ympev.2011.12.007

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R., & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J, 6*(3), 610-618. doi: 10.1038/ismej.2011.139

Mendoza-Cózatl, David, Loza-Tavera, Herminia, Hernández-Navarro, Andrea, & Moreno-Sánchez, Rafael. (2005). Sulfur assimilation and glutathione metabolism under cadmium stress in yeast, protists and plants. *FEMS Microbiology Reviews, 29*(4), 653-671. doi: 10.1016/j.femsre.2004.09.004

Merrick, M. J., & Edwards, R. A. (1995). Nitrogen control in bacteria. *Microbiol Rev, 59*(4), 604-622.

Mesbah, N., Cook, G., & Wiegel, J. (2009). The halophilic alkalithermophile Natranaerobius thermophilus adapts to multiple environmental extremes using a large repertoire of Na+(K+)/H+ antiporters. *Mol. Microbiol., 74*, 270-281.

Moreno, M. L., Piubeli, F., Bonfa, M. R., Garcia, M. T., Durrant, L. R., & Mellado, E. (2012). Analysis and characterization of cultivable extremophilic hydrolytic bacterial community in heavy-metal-contaminated soils from the Atacama Desert and their biotechnological potentials. *J Appl Microbiol, 113*(3), 550-559. doi: 10.1111/j.1365-2672.2012.05366.x

Munoz-Pedreros, Andres, De Los Rios, Patricio, & Moeller, Patricia. (2013). Zooplankton in Laguna Lejía, a high-altitude Andean shallow lake of the Puna in northern Chile. *Crustaceana, 86*(13-14), 1634-1643. doi: 10.1163/15685403-00003265

Murphy, M. J., Siegel, L. M., Tove, S. R., & Kamin, H. (1974). Siroheme: a new prosthetic group participating in six-electron reduction reactions catalyzed by both sulfite and nitrite reductases. *Proc Natl Acad Sci U S A, 71*(3), 612-616.

Narasingarao, Priya, Podell, Sheila, Ugalde, Juan A., Brochier-Armanet, Celine, Emerson, Joanne B., Brocks, Jochen J., Heidelberg, Karla B., Banfield, Jillian F., & Allen, Eric E. (2012). De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J, 6*(1), 81-93.

Neilson, J. W., Quade, J., Ortiz, M., Nelson, W. M., Legatzki, A., Tian, F., LaComb, M., Betancourt, J. L., Wing, R. A., Soderlund, C. A., & Maier, R. M. (2012). Life at the hyperarid margin: novel bacterial diversity in arid soils of the Atacama Desert, Chile. *Extremophiles, 16*(3), 553-566. doi: 10.1007/s00792-012-0454-z

Nester, P. L., Gayo, E., Latorre, C., Jordan, T. E., & Blanco, N. (2007). Perennial stream discharge in the hyperarid Atacama Desert of northern Chile during the latest Pleistocene. *Proc Natl Acad Sci U S A, 104*(50), 19724-19729. doi: 10.1073/pnas.0705373104

Nielsen, H. Bjorn, Almeida, Mathieu, Juncker, Agnieszka Sierakowska, Rasmussen, Simon, Li, Junhua, Sunagawa, Shinichi, Plichta, Damian R., Gautier, Laurent, Pedersen, Anders G., Le Chatelier, Emmanuelle, Pelletier, Eric, Bonde, Ida, Nielsen, Trine, Manichanh, Chaysavanh, Arumugam, Manimozhiyan, Batto, Jean-Michel, Quintanilha dos Santos, Marcelo B., Blom, Nikolaj, Borruel, Natalia, Burgdorf, Kristoffer S., Boumezbeur, Fouad, Casellas, Francesc, Dore, Joel, Dworzynski, Piotr, Guarner, Francisco, Hansen, Torben, Hildebrand, Falk, Kaas, Rolf S., Kennedy, Sean, Kristiansen, Karsten, Kultima, Jens Roat, Leonard, Pierre, Levenez, Florence, Lund, Ole, Moumen, Bouziane, Le Paslier, Denis, Pons, Nicolas, Pedersen, Oluf, Prifti, Edi, Qin, Junjie, Raes, Jeroen, Sorensen, Soren, Tap, Julien, Tims, Sebastian, Ussery, David W., Yamada, Takuji, Meta, H. I. T. Consortium, Renault, Pierre, Sicheritz-Ponten, Thomas, Bork, Peer, Wang, Jun, Brunak, Soren, & Ehrlich, S. Dusko. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using

reference genomes. *Nat Biotech, 32*(8), 822-828. doi: 10.1038/nbt.2939

Nyyssonen, Mari, Hultman, Jenni, Ahonen, Lasse, Kukkonen, Ilmo, Paulin, Lars, Laine, Pia, Itavaara, Merja, & Auvinen, Petri. (2013). Taxonomically and functionally diverse microbial communities in deep crystalline rocks of the Fennoscandian shield. *ISME J, 8*(1), 126-138. doi: 10.1038/ismej.2013.125

Padan, E. (2008). The enlightening encounter between structure and function in the NhaA Na+-H+ antiporter. *Trends Biochem. Sci., 33*, 435-443.

Padan, E., Bibi, E., Ito, M., & Krulwich, T. A. (2005). Alkaline pH homeostasis in bacteria: new insights. *Biochim Biophys Acta, 1717*(2), 67-88. doi: 10.1016/j.bbamem.2005.09.010

Palmer, Michael W. (1993). Putting Things in Even Better Order: The Advantages of Canonical Correspondence Analysis. *Ecology, 74*(8), 2215-2230. doi: 10.2307/1939575

Paulino-Lima, I. G., Azua-Bustos, A., Vicuna, R., Gonzalez-Silva, C., Salas, L., Teixeira, L., Rosado, A., Leitao, A. A., & Lage, C. (2013). Isolation of UVC-tolerant bacteria from the hyperarid Atacama Desert, Chile. *Microb Ecol, 65*(2), 325-335. doi: 10.1007/s00248-012-0121-z

Peng, Yu, Leung, Henry C. M., Yiu, S. M., & Chin, Francis Y. L. (2010). IDBA - A Practical Iterative de Bruijn Graph De Novo Assembler. In B. Berger (Ed.), *Research in Computational Molecular Biology: 14th Annual International Conference, RECOMB 2010, Lisbon, Portugal, April 25-28, 2010. Proceedings* (pp. 426-440). Berlin, Heidelberg: Springer Berlin Heidelberg.

Peng, Yu, Leung, Henry C. M., Yiu, S. M., & Chin, Francis Y. L. (2011). Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics, 27*(13), i94-i101. doi: 10.1093/bioinformatics/btr216

Peng, Yu, Leung, Henry C. M., Yiu, S. M., & Chin, Francis Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics, 28*(11), 1420-1428. doi: 10.1093/bioinformatics/bts174

Pettit, R. K. (2004). Soil DNA libraries for anticancer drug discovery. *Cancer Chemotherapy and Pharmacology, 54*(1), 1-6.

Piubeli, F., de Lourdes Moreno, M., Kishi, L. T., Henrique-

Silva, F., Garcia, M. T., & Mellado, E. (2015). Phylogenetic Profiling and Diversity of Bacterial Communities in the Death Valley, an Extreme Habitat in the Atacama Desert. *Indian J Microbiol, 55*(4), 392-399. doi: 10.1007/s12088-015-0539-3

Prell, J., White, J. P., Bourdes, A., Bunnewell, S., Bongaerts, R. J., & Poole, P. S. (2009). Legumes regulate Rhizobium bacteroid development and persistence by the supply of branched-chain amino acids. *Proc Natl Acad Sci U S A, 106*(30), 12477-12482. doi: 10.1073/pnas.0903653106

Prestel, E., Salamitou, S., & DuBow, M. S. (2008). An examination of the bacteriophages and bacteria of the Namib desert. *J Microbiol, 46*(4), 364-372. doi: 10.1007/s12275-008-0007-4

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res, 41*(Database issue), D590-596. doi: 10.1093/nar/gks1219

Queralt, I., Ovejero, M., Carvalho, M. L., Marques, A. F., & Llabrés, J. M. (2005). Quantitative determination of essential and trace element content of medicinal plants and their infusions by XRF and ICP techniques. *X-Ray Spectrometry, 34*(3), 213-217. doi: 10.1002/xrs.795

RCoreTeam. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Rhoades, J.D. (1996). Salinity: Electrical conductivity and total dissolved solids. In D. L. Sparks (Ed.), *Methods of soil analysis: Chemical methods* (Vol. 3, pp. 417-435). ASA and SSSA: Madison, WI.

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

Ronaghi, M. (2001). Pyrosequencing sheds light on DNA

sequencing. *Genome Research, 11*(1), 3-11. doi: 10.1101/gr.11.1.3

Ronaghi, M., & Elahi, E. (2002). Pyrosequencing for microbial typing. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences, 782*(1-2), 67-72. doi: 10.1016/s1570-0232(02)00693-1

Rousk, J., Baath, E., Brookes, P. C., Lauber, C. L., Lozupone, C., Caporaso, J. G., Knight, R., & Fierer, N. (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J, 4*(10), 1340-1351. doi: 10.1038/ismej.2010.58

Sadzawka, Angélica R.; Carrasco R, Maria Adriana; Gez Z, Renato; Mora G, Maria de la Luz; Flores P, Hugo; Neaman, Alexander. (2006). *Métodos de análisis recomendados para los suelos de Chile*. (2006 ed.). Instituto de Investigaciones Agropecuarias, Serie ACTAS INIA N°34, Santiago de Chile.

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., & Weber, C. F. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology, 75*(23), 7537-7541. doi: 10.1128/aem.01541-09

Sheng, J., & Marquis, R. E. (2007). Malolactic fermentation by Streptococcus mutans. *FEMS Microbiol. Lett., 272*, 196-201.

Skinner, F. A., Jones, P. C., & Mollison, J. E. (1952). A comparison of a direct- and a plate counting technique for the quantitative estimation of soil micro-organisms. *J Gen Microbiol, 6*(3-4), 261-271.

Slonczewski, J. L., Fujisawa, M., Dopson, M., & Krulwich, T. A. (2009a). Cytoplasmic pH measurement and homeostasis in bacteria and archaea. *Adv Microb Physiol, 55*, 1-79, 317. doi: 10.1016/s0065-2911(09)05501-5

Slonczewski, J. L., Fujisawa, M., Dopson, M., & Krulwich, T. A. (2009b). Cytoplasmic pH measurement and homeostasis in bacteria and archaea. *Adv. Microb. Physiol., 55*, 1-79.

Solden, L., Lloyd, K., & Wrighton, K. (2016). The bright side of microbial dark matter: lessons learned from

the uncultivated majority. *Current Opinion in Microbiology, 31*, 217-226. doi: 10.1016/j.mib.2016.04.020

Sørensen, J. (1997). Modern soil microbiology. *The Rhizosphere as a Habitat for Soil Microorganisms*, 21-45.

Stancik, L. M. (2002). pH-dependent expression of periplasmic proteins and amino acid catabolism in Escherichia coli. *J. Bacteriol., 184*, 4246-4258.

Steven, B., Gallegos-Graves, L., Starkenburg, S. R., Chain, P. S., & Kuske, C. R. (2012). Targeted and shotgun metagenomic approaches provide different descriptions of dryland soil microbial communities in a manipulated field study. *Environmental Microbiology Reports, 4*(2), 248-256. doi: 10.1111/j.1758-2229.2012.00328.x

Streit, W. R., Daniel, R., & Jaeger, K. E. (2004). Prospecting for biocatalysts and drugs in the genomes of non-cultured microorganisms. *Current Opinion in Biotechnology, 15*(4), 285-290.

Streit, W. R., & Schmitz, R. A. (2004). Metagenomics - The key to the uncultured microbes. *Current Opinion in Microbiology, 7*(5), 492-498.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G., & Bork, P. (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science, 348*(6237), 1261359. doi: 10.1126/science.1261359

Swan, B. K., Martinez-Garcia, M., Preston, C. M., Sczyrba, A., Woyke, T., Lamy, D., Reinthaler, T., Poulton, N. J., Masland, E. D. P., Gomez, M. L., Sieracki, M. E., DeLong, E. F., Herndl, G. J., & Stepanauskas, R. (2011). Potential for Chemolithoautotrophy Among Ubiquitous Bacteria Lineages in the Dark Ocean.

Science, 333(6047), 1296-1300. doi: 10.1126/science.1203690

Tatusov, Roman L., Fedorova, Natalie D., Jackson, John D., Jacobs, Aviva R., Kiryutin, Boris, Koonin, Eugene V., Krylov, Dmitri M., Mazumder, Raja, Mekhedov, Sergei L., Nikolskaya, Anastasia N., Rao, B. Sridhar, Smirnov, Sergei, Sverdlov, Alexander V., Vasudevan, Sona, Wolf, Yuri I., Yin, Jodie J., & Natale, Darren A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics, 4*, 41-41. doi: 10.1186/1471-2105-4-41

Thomas, D., & Surdin-Kerjan, Y. (1997). Metabolism of sulfur amino acids in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev, 61*(4), 503-532.

Thomas, G.W. (1996). Soil pH and soil acidity. In S. D.L. (Ed.), *Methods of soil analysis: Chemical methods* (Vol. 3, pp. 475-490). Madison WI: Soil Science Society of America, American Society of Agronomy.

Tomb, J. F. (1997). The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature, 388*, 539-547.

Torsvik, V., & Ovreas, L. (2002). Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol, 5*(3), 240-245.

Torsvik, V., Ovreas, L., & Thingstad, T. F. (2002). Prokaryotic diversity--magnitude, dynamics, and controlling factors. *Science, 296*(5570), 1064-1066. doi: 10.1126/science.1071698

Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., Bork, P., Hugenholtz, P., & Rubin, E. M. (2005). Comparative metagenomics of microbial communities. *Science, 308*(5721), 554-557. doi: 10.1126/science.1107851

Turner, S., Pryer, K. M., Miao, V. P., & Palmer, J. D. (1999). Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J Eukaryot Microbiol, 46*(4), 327-338.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature, 428*(6978), 37-43.

Vanhoof, C., Corthouts, V., & Tirez, K. (2004). Energy-dispersive X-ray fluorescence systems as analytical tool for assessment of contaminated soils. *J Environ Monit, 6*(4), 344-350. doi: 10.1039/b312781h

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H., & Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science, 304*(5667), 66-74. doi: 10.1126/science.1093857

Viles, H. A. (2012). Microbial geomorphology: A neglected link between life and landscape. *Geomorphology, 157*, 6-16. doi: 10.1016/j.geomorph.2011.03.021

Villagrán, C., Armesto, J. J., & Kalin Arroyo, M. T. (1981). Vegetation in a high Andean transect between Turi and Cerro León in Northern Chile. *Vegetatio, 48*(1), 3-16. doi: 10.1007/BF00117356

Vuille, Mathias, & Bradley, Raymond S. (2000). Mean annual temperature trends and their vertical structure in the tropical Andes. *Geophysical Research Letters, 27*(23), 3885-3888. doi: 10.1029/2000GL011871

Wang, Jianjun, Soininen, Janne, He, Jizheng, & Shen, Ji. (2012). Phylogenetic clustering increases with elevation for microbes. *Environmental Microbiology Reports, 4*(2), 217-226. doi: 10.1111/j.1758-2229.2011.00324.x

Wang, Jianjun, Soininen, Janne, Zhang, Yong, Wang, Beixin, Yang, Xiangdong, & Shen, Ji. (2011). Contrasting patterns in elevational diversity between microorganisms and macroorganisms. *Journal of Biogeography, 38*(3), 595-603. doi: 10.1111/j.1365-2699.2010.02423.x

Wen, Y. (2003). Acid-adaptive genes of Helicobacter pylori. *Infect. Immun., 71*, 5921-5939.

Wen, Y., Feng, J., Scott, D. R., Marcus, E. A., & Sachs, G. (2007). The HP0165-HP0166 two-component system (ArsRS) regulates acid-induced expression of HP1186 α-carbonic anhydrase in Helicobacter pylori by activating the pH-dependent promoter. *J. Bacteriol., 189*, 2426-2434.

Westra, E. R., Swarts, D. C., Staals, R. H., Jore, M. M., Brouns, S. J., & van der Oost, J. (2012). The CRISPRs, they are a-changin': how prokaryotes generate adaptive

immunity. *Annu Rev Genet, 46*, 311-339. doi: 10.1146/annurev-genet-110711-155447

Wrighton, K. C., Thomas, B. C., Sharon, I., Miller, C. S., Castelle, C. J., VerBerkmoes, N. C., Wilkins, M. J., Hettich, R. L., Lipton, M. S., Williams, K. H., Long, P. E., & Banfield, J. F. (2012). Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. *Science, 337*(6102), 1661-1665. doi: 10.1126/science.1224041

Wyckoff, E. E., Mey, A. R., & Payne, S. M. (2007). Iron acquisition in Vibrio cholerae. *Biometals, 20*(3-4), 405-416. doi: 10.1007/s10534-006-9073-4

Yohannes, E., Barnhart, D. M., & Slonczewski, J. L. (2004). pH-dependent catabolic protein expression during anaerobic growth of Escherichia coli K-12. *J. Bacteriol., 186*, 192-199.

Zagal, E., & Sadzawka, A. R. (2007). *Protocolo de Métodos de Análisis para Suelos y Lodos.* Universidad de Concepción, Facultad de Agronomía, Chillán. Comisión de Normalización y Acreditación, Sociedad Chilena de la Ciencia del Suelo. Servicio Agrícola y Ganadero, Gobierno de Chile. Retrieved from http://www.sag.cl/sites/default/files/METODOS LODOS SUELOS.pdf

Zelles, L., Bai, Q. Y., Beck, T., & Beese, F. (1992). Signature fatty acids in phospholipids and lipopolysaccharides as indicators of microbial biomass and community structure in agricultural soils. *Soil Biology and Biochemistry, 24*(4), 317-323.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res, 18*(5), 821-829. doi: 10.1101/gr.074492.107

Zhou, J., Bruns, M. A., & Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol, 62*(2), 316-322.

ANEXOS

ANEXO 1. Publicación.

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Genome announcement

Complete genome sequence of *Microbacterium* sp. CGR1, bacterium tolerant to wide abiotic conditions isolated from the Atacama Desert

Dinka Mandakovic[a,b,1], Pablo Cabrera[a,b,1], Rodrigo Pulgar[a,b], Jonathan Maldonado[a,b], Pamela Aravena[a,b], Mauricio Latorre[a,b], Verónica Cambiazo[a,b], Mauricio González[a,b,*]

[a] Laboratorio de Bioinformática y Expresión Génica, INTA—Universidad de Chile, El Líbano 5524 Santiago, Chile
[b] Fondap Center for Genome Regulation (CGR), Avenida Blanco Encalada 2085 Santiago, Chile

ARTICLE INFO

Article history:
Received 14 October 2015
Accepted 22 October 2015
Available online 29 October 2015

Keywords:
Microbacterium
Atacama Desert
Abiotic factors tolerance
Riboflavin
Arsenic biosensor

ABSTRACT

Microbacterium sp. CGR1 (RGM2230) is an isolate from the Atacama Desert that displays a wide pH, salinity and temperature tolerance. This strain exhibits riboflavin overproducer features and traits for developing an environmental arsenic biosensor. Here, we report the complete genome sequence of this strain, which represents the first genome of the genus *Microbacterium* sequenced and assembled in a single contig. The genome contains 3,634,864 bp, 3299 protein-coding genes, 45 tRNAs, six copies of 5S-16S-23S rRNA and a high genome average GC-content of 68.04%.

Bacteria isolated from extreme environments show remarkable tolerance to life-threatening conditions, making them important sources of prominent genes useful for biotechnological approaches. In this context, the Atacama Desert is especially attractive, since microorganism communities growing there are exposed to challenging environmental conditions, such as extremely low water availability, nutrient-poor soils, extreme solar radiation, large temperature oscillations, elevated salinity and high levels of arsenic (Crits-Christoph et al., 2013; Smedley and Kinniburgh, 2002). Recently, we recovered 30 isolates with different morphologies and abiotic tolerant features from a pH and salinity gradient transect located at 4480 m.a.s.l in the Atacama Desert, with ranging day/night temperatures from −3.9 °C to 24.4 °C. Among the isolates, a Gram positive, motile and yellow-intense pigmented bacterium, identified as a *Microbacterium* strain by 16S rDNA gene sequencing, had one of the widest pH (5–12) and salinity (0–7%) tolerance ranges, and so it was selected for whole genome sequencing. This isolate was named *Microbacterium* sp. CGR1 (RGM2230).

Genomic DNA from *Microbacterium* sp. CGR1 was purified from exponential growth cultures (1 mL, OD_{600}:0.5) using the DNeasy Blood & Tissue Kit for DNA (Qiagen). Genome sequencing was performed at GCB Genome Sequencing Shared Resource (Duke University) using four single-molecule-real-time (SMRT) cells of the PacBio RSII (Pacific Biosciences, Menlo Park, CA) platform with a 15 kb to 20 kb insert library and XL/C2 chemistry, producing a total of 299,922 reads post-filter with mean read length of 12,794 bp, N50 size of 17,569 bp and a total of 3,837,473,443 bp. *De novo* assembly was conducted using HGAP version 3 (SMRT Analysis version 2.3) with default parameters (Chin et al., 2013), resulting in a complete contig of 3,634,864 bp with coverage of 717*x*, representing the first genome of the genus *Microbacterium* assembled in a single contig.

Microbacterium sp. CGR1 was annotated using the NCBI Prokaryotic Genome Annotation Pipeline released 2013 (Tatusova et al., 2013) and approved on August 11th, 2015. We predicted 51 RNA genes (6 rRNA and 45 tRNA), a total of 3299 protein-coding genes and a high genome average GC-content of 68.04 mol% (see Table 1).

The genome of *Microbacterium* sp. CGR1 revealed the presence of the complete set of genes for riboflavin biosynthesis. Riboflavin (vitamin B2) is a yellow water-soluble vitamin produced by all plants, fungi and many microorganisms, but not by higher animals including humans. Vitamin B2 is known to be the central component of the cofactor's flavin adenine dinucleotide (FAD) and flavin mononucleotide (FMN), which are necessary to all flavoproteins. For this reason, riboflavin is required for a large variety of cellular processes. Traditional chemical synthesis of riboflavin is now being replaced by low cost and less energy wasting commercially competitive biotechnological processes that use ascomycetes *Ashbya*

* Corresponding author at: Laboratorio de Bioinformática y Expresión Génica, INTA—Universidad de Chile, El Líbano 5524 Santiago, Chile.
E-mail address: mgonzale@inta.uchile.cl (M. González).
[1] These authors have contributed equally to this work.

http://dx.doi.org/10.1016/j.jbiotec.2015.10.020
0168-1656/

ANEXO 2. Resumen congreso "XXVIII Reunión Anual de la Sociedad de Biología Celular". Realizado entre los días 25 y 29 de Octubre 2015 en la ciudad de Puerto Varas. Chile.

Discrimination of taxa-specific genomes of soil bacteria through analysis of abundance changes between two natural environmental conditions

J. Maldonado[1,2], D. Mandakovic[2], M. Latorre[2], P. Cabrera[2], V. Cambiazo[2] and M. González[2].

[1]Doctorado en Ciencias, mención Biología Molecular, Celular y Neurociencias. Facultad de Ciencias, Universidad de Chile.

[2]Laboratorio de Bioinformática y Expresión Génica, INTA, Universidad de Chile

jomaldon@gmail.com

Introduction:

Current developments in metagenomics allow us to describe links between community properties and features of their environment. The next step will be to determine the relative contribution of different taxa to the global function of the bacterial community in the environment. The goal of the present work is to use and develop bioinformatics tools to discriminate taxa-specific genomes inside a metagenome to understand the relationship between changes on relative abundance in the bacterial community members that face abiotic variations, and their functional capacities responsive to the changes.

Material and Methods:

We extract DNA form soil of 2 near sites of the Atacama Desert that present differences in environmental variables and nutrient composition. This DNA was sequenced and samples were assembled together. The scaffolds were binning by abundance, GC content and other variables. These subgroups were reassembled to obtain final genomes that represent taxa-specific genomes.

Results:

We have been able to assembly 3 draft genomes of novel bacteria form each site with contrasting abundances. Functional analysis of those genomes reveals significant differences from the corresponding metagenome pattern and from the genomes of the contrasting site.

Discussion:

Current results on big metagenome projects like Human Gut or Global Ocean Microbiome shows that functional pattern don't change despite the origin or composition of samples. Our work reveals that the environment drives a selection over the components of a community, and those components, don't follow necessarily the

community functional pattern. This new methods of taxa-specific discrimination of genomes will help us to understand the processes that explain the adaptation of these organisms to extreme environments.

Investigation founded by FONDAP CGR15090007 and CONICYT Doctoral Fellowship.

ANEXO 3. Resumen congreso internacional "16th International Symposium on Microbial Ecology, ISME 2016". Realizado entre los días 21 y 26 de Agosto 2016 en la ciudad de Montreal. Canadá.

Discrimination of taxa-specific genomes of soil bacteria through analysis of abundance changes between two contrasting environmental conditions

J. Maldonado[1,2], D. Mandakovic[2], M. Latorre[2], P. Cabrera[2], V. Cambiazo[2] and M. González[2].

[1]Doctorado en Ciencias, mención Biología Molecular, Celular y Neurociencias. Facultad de Ciencias, Universidad de Chile.

[2]Laboratorio de Bioinformática y Expresión Génica, INTA, Universidad de Chile

jomaldon@gmail.com

Current developments in metagenomics allow us to describe associations between community properties and features of their environment. The next step should be to determine the relative contribution of different taxa to the global function of the bacterial community. The goal of the present work is to use and develop bioinformatic tools to discriminate taxa-specific genomes inside a metagenome in order to understand the relationship between changes in the relative abundance of bacterial community members that face abiotic variations, and their functional capacities responsive to these changes.

We extracted DNA form soil of 2 near sites of the Atacama Desert that present differences in environmental variables and nutrient composition. This DNA was sequenced and samples were assembled together. The scaffolds were binned by abundance, GC content and other variables. These subgroups were reassembled to obtain final genomes that represent taxa-specific genomes.

We have been able to assembly 55 draft genomes of novel bacteria of contrasting abundances among sample sites. Functional analysis of those genomes revealed significant differences from the corresponding metagenome functional patterns.

Our work reveals that the environment drives a selection over the components of a community, and those components do not necessarily follow the community functional pattern. This new approach of taxa-specific discrimination of genomes will help to understand the processes that cause the adaptation of these organisms to extreme environments.

Investigation founded by FONDAP CGR15090007 and CONICYT Doctoral Fellowship.

ANEXO 4. Publicación en preparación para revista Scientific Reports

Local soil features regulate bacterial community composition at the Lascar volcano secluded basin (Atacama Desert)

Jonathan Maldonado[1,2], Dinka Mandakovic[1,2], Mauricio Latorre[1,2,3], Pablo Cabrera[1,2], Verónica Cambiazo[1,2], Mauricio González[1,2,*].

[1]Laboratorio de Bioinformática y Expresión Génica, INTA-Universidad de Chile, El Líbano 5524, Santiago, Chile.

[2]Fondap Center for Genome Regulation (CGR), Avenida Blanco Encalada 2085, Santiago, Chile.

[3]Mathomics, Center for Mathematical Modeling, Universidad de Chile, Santiago, Chile

*Corresponding author: Mauricio González mgonzale@inta.uchile.cl

Key words: Atacama Desert; bacterial diversity, soil microbiome, 16S rRNA

Despite the high effort recently added to the survey of soil bacterial communities using high throughput sequencing of 16S rRNA gene and culture-independent methods, yet few studies have analyzed the influence of local environmental factors over soil microbial diversity. The central objective of this study was to characterize the soil microbiomes from three geographically related sites, but with significant differences in soil physicochemical parameters and nutrients composition, located at 4,480 m a.s.l. at the summit of Lascar volcano in central region of the Atacama Desert. In a first step, we described the microbial community by deep sequencing of the hypervariable regions V1-V3 of 16S rRNA gene and compared their diversity pattern with different environmental variables. We mapped 15,768 reads to Greengenes database with a 3 % cutoff, classifying 2,077 total OTUs and 799 to 977 OTUs per sample after rarefaction (9,000 reads). Community structure parameters such as richness and phylogenetic diversity were correlated with pH, electrical conductivity, and available soil fractions of P, K, Ca, Fe and Zn. Canonical correspondence and SOM clustering analyses allowed us to identify OTUs differentially distributed among the three sampled soils. Taken together, our results indicate that the differences in local soil features rather than shared environmental variables (temperature, relative humidity, UV radiation) among sampled sites had a higher impact defining a large fraction of the soil microbiome composition,

including several OTUs from candidate phyla or "dark matter". The identification of differential bacteria among sites is relevant for future genomic studies that explain how bacteria fit to contrasting soil abiotic factors.

ANEXO 5. Publicación en preparación para revista Scientific Reports

Reconstructing Rare Soil Microbial Genomes using Metagenomic

Jonathan Maldonado[1,2], Mauricio González[1,2,*].

[1]Laboratorio de Bioinformática y Expresión Génica, INTA-Universidad de Chile, El Líbano 5524, Santiago, Chile.
[2]Fondap Center for Genome Regulation (CGR), Avenida Blanco Encalada 2085, Santiago, Chile.
*Corresponding author: Mauricio González mgonzale@inta.uchile.cl

Key words: soil, metagenomics, environmental genomics

Current developments in metagenomics allow us to describe associations between community properties and features of their environment. The next step should be to determine the relative contribution of different taxa to the global function of the bacterial community. The central objective of the present work is to use and develop bioinformatic tools to discriminate taxa-specific genomes inside a metagenome in order to understand the relationship between changes in the relative abundance of bacterial community members that face abiotic variations, and their functional capacities responsive to these changes.
We extracted DNA form soil of 3 near sites of the Atacama Desert that present differences in environmental variables and nutrient composition. This DNA was sequenced and samples were assembled together. The scaffolds were binned by abundance, GC content and other variables. These subgroups were reassembled to obtain final genomes that represent taxa-specific genomes.
We have been able to assembly 74 draft genomes of novel bacteria of contrasting abundances among sample sites. Functional analysis of those genomes revealed significant differences from the corresponding metagenome functional patterns.
Our work reveals that the environment drives a selection over the components of a community, and those components do not necessarily follow the community functional pattern. This new approach of taxa-specific discrimination of genomes will help to understand the processes that cause the adaptation of these organisms to extreme environments.

ANEXO 6. Publicación en preparación para revista Soil Biology and Biochemistry

Bacterial diversity pattern and cultivable-community from Lejía lake shore, Atacama Desert

Dinka Mandakovic[1,2], Rodrigo Pulgar[1,2,3], Pablo Cabrera[1,2], Jonathan Maldonado[1,2], Mauricio Latorre[1,2,4], Verónica Cambiazo[1,2,3], Mauricio González[1,2,*]

[1]Laboratorio de Bioinformática y Expresión Génica, INTA-Universidad de Chile, El Líbano 5524, Santiago, Chile.
[2]Fondap Center for Genome Regulation (CGR), Avenida Blanco Encalada 2085, Santiago, Chile.
[3]Laboratorio de Genómica Aplicada, INTA-Universidad de Chile, El Líbano 5524, Santiago, Chile.
[4]Mathomics, Center for Mathematical Modeling (CMM), Universidad de Chile, Beauchef 851 (northern building) 6th floor, Santiago, Chile.
*Corresponding author: Mauricio González mgonzale@inta.uchile.cl

The Atacama Desert is a recognized extreme environment for its severe aridity, oscillating temperatures and intense UV radiation, environmental conditions that are exacerbated in the high elevated saline lakes from the Altiplano. Despite the harsh environmental features displayed in these lakes, some groups of bacteria manage to reside in these settings. In order to deeply characterize the bacterial composition of one of these lakes, we identified by Next Generation Sequencing (NGS) using a culture-independent approach the complete bacterial community present at Lejía lake soil (LLS) shore. We identified around one thousand bacteria distributed in twenty-six different phyla, including those in low abundance. Then, to recover a highly represented bacterial community from Lejía lake shore in the laboratory, we implemented culture-based approaches combined with NGS technology. For this, we developed two agar culture media, one based in supplementation of a minimal medium (SM) and another based on an extract of soluble fraction of soil directly obtained from LLS (SEM). The NGS of 16S rRNA gene of the bacterial cultivable-community showed that SEM resulted better than SM to recover bacterial community richness and diversity. Moreover, the number of different bacteria recovered after culturing in SEM accounted for 12.9 % of the total bacteria from LLS, while only 7.8 % was recovered from SM. Thus, the capture of bacteria using SEM was significantly higher than by SM. On the other hand, measurements of the elements in the soil and in culture emphasized the relevance of recapitulating the physicochemical composition of the environment to better recover microbial communities.

In this study, we were able to describe for the first time the microbial community inhabiting the Lejía lake soil shore and its complete cultivable-community by NGS technology. The combination of cultivation techniques coupled with high throughput sequencing approaches allowed the cultivation of important previously uncultured bacteria and makes possible *in vitro* studies of environmental microbial communities and their ecological interactions.