CENTRO DE INVESTIGACION Y DE ESTUDIOS AVANZADOS DEL INSTITUTO POLITECNICO NACIONAL

UNIDAD IRAPUATO

Análisis de la distribución ecológica del género *Bacillus*, a través de reconstrucción logenómica y genómica comparativa

Tesis que presenta:

M. en C. Ismael Luis Hernádez González

Para Obtener el grado de:

Doctor en Ciencias

En la Especialidad de:

Biotecnología de Plantas

Director de Tesis: Dra. Gabriela Olmedo Álvarez

Irapuato, Guanajuato **Marzo** 2019

El presente trabajo fue realizado por el M. en C. Ismael Luis Hernández González, en el Laboratorio de Bacteriología Molecular, del Departamento de Ingeniería Genética del Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV), Unidad Irapuato, bajo la dirección de la Dra. Gabriela Olmedo Álvarez, investigador titular del Departamento de Ingeniería Genética, y en el Laboratorio ConSequences de la Universidad Wilfried Laurier de Cánada, a cargo del Dr. Gabriel Moreno-Hagelsieb Profesor titular de tiempo completo, y con la asesoría de los dres.: Dr. Luis Eugenio González de la Vara, investigador titular del Departamento de Biotecnología y Bioquímica, el Dr. Luis José Delaye Arredondo investigador titular del Departamento de Ingeniería Genética y el Dr. Cei Abreu Goodger investigador titular de la Unidad de Genómica Avanzada.

El M. en C. Ismael Luis Hernández González (No. CVU378712) contó con el apoyo económico de la beca de estudios de Doctorado del Consejo Nacional de Ciencia y Tecnología (CONACyT).

A mi familia con todo mi cariño
como un reconocimiento por todo su apoyo.
Para Ustedes como una pequeña muestra
de agradecimiento.

Agradecimientos

Al **Consejo Nacional de Ciencia y Tecnología (CONACyT)** por su apoyo mediante la beca de estudios de Doctorado.

A **The Shared Hierarchical Academic Research Computing Network (SHARCNET)** de Canada por las facilidades otorgadas para el uso de su red de cómputo.

A **Canadian Institutes of Health Research (CIHR)**, por su apoyo a través de la subvención al Dr. Gabriel Moreno-Hagelsieb y el **ConSequences Lab** de Wilfried Laurier University.

Resumen

El género *Bacillus* agrupa a bacterias grampositivas, aerobias o anaerobias facultativas, con forma de bastón, que tienen la capacidad de formar endoesporas. Los miembros de este género son considerados ubicuos, dado que ellos han sido aislados de un amplio rango de ambientes tanto terrestres como acuáticos. Sin embargo, no está claro si la presencia de endoesporas halladas en tal variedad de ambientes se debe sólo a la dispersión de las esporas por el aire y/o el agua o a que sus capacidades metabólicas le permiten sobrevivir en estos ambientes. Algunos autores han sugerido la existencia de un grupo de *Bacillus* acuáticos dentro del género. A pesar de la gran cantidad de secuencias de genomas de *Bacillus* disponibles, aisladas de diferentes ambientes, pocos trabajos han analizado la posible relación entre la historia evolutiva de los *Bacillus* spp. y el ambiente donde ellos han sido aislados. En este trabajo, analizamos la relación entre la historia evolutiva de los *Bacillus* spp. y el ambiente del cual ellos han sido aislados, enfocando nuestra atención principalmente sobre los *Bacillus* spp. acuáticos. Además, analizamos si existía una relación entre el ambiente y el contenido funcional de genes. La reconstrucción filogenética usando el genoma núcleo sugirió que los *Bacillus* spp. acuáticos no forman un grupo monofilético. En contraste, un análisis de agrupamiento jerárquico basado en el contenido de genes, agrupó juntos a los *Bacillus* spp. acuáticos que aparecen separados en el árbol filogenético. Estos resultados sugieren que los *Bacillus* spp. aislados de ambientes similares comparten más funciones entre ellos que lo que se podría esperar de sus orígenes polifiléticos. El análisis del contenido de genes también sugiere la presencia de funciones específicos de cada ambiente.

Abstract

The *Bacillus* genus comprises Gram-positive bacteria, aerobic or facultative anaerobic, rod-shaped, that have the ability to form endospores. Members of this genus are considered ubiquitous, since they have been isolated from a wide range of environments, such as terrestrial and aquatic. However, is not clear if their presence in such variety of environments, is due to their ability to colonize new environments, or due to spore resistance and spore dispersal by air and water. Some authors have suggested the existence of an aquatic *Bacillus* group within the genus. Despite the numerous *Bacillus* genome sequences available, isolated from different environments, few works have analyzed the relationship between the evolutionary history of *Bacillus* spp. and the environments where they are found. In this work, we analyzed the relationship between the evolutionary history of *Bacillus* spp. and the environment from which they have been isolated, focusing our attention mainly on the aquatic *Bacillus*. Furthermore, we analyzed if there was a relationship between the environment and the functional gene content. The phylogenetic reconstruction, using the core genome, suggested that the aquatic *Bacillus* do not form a monophyletic group. In contrast, a clustering analysis based on gene content groups together the aquatic *Bacillus* that appear separated in the phylogenetic tree. These results suggest that *Bacillus* isolated from similar environments share more functions between them than would be expected from their polyphyletic origins. The functional gene content analysis also suggests the presence of functions specific for each environment.

Índice de Contenidos

Lista de Figuras

Lista de Tablas

Abreviaturas

- GSS: Genomic Similarity Score (Índice de Similitud Genómica)
- ML: Maximum Likelihood (Máxima verosimilitud)
- COG: Clusters of Orthologous Groups (Grupo de genes ortólogos)
- RBH: Reciprocal Best Hits (Mejor coincidencia recíproca)
- MLST: Multi Locus Sequence Typing (Tipificación por secuencia multi *locus*)
- Jd: Jaccard distance (Distancia de Jaccard)
- PCA: Principal Components Analysis (Análisis de componentes principales)

Introducción

La era Genómica

Durante las últimas décadas, el avance en las tecnologías de secuenciación y de cómputo ha hecho posible obtener en menor tiempo y a menor costo la secuencia de una gran cantidad de genomas bacterianos. La comparación de la secuencia del genoma de organismos distantemente relacionados ha arrojado luz acerca de rasgos importantes que determinan la organización del cromosoma bacterial así como la identificación de los mecanismos que determinan su incomparable diversidad y adaptabilidad [2]. La comparación de los genomas ha mostrado también que las bacterias pueden adaptarse a nuevos ambientes adquiriendo genes de otros organismos [3].

El uso de las secuencias de genomas completos en los análisis evolutivos ha dado origen a un nuevo campo de investigación llamado filgenómica, el cual emplea los principios filogenéticos para dar sentido a los datos genómicos [4]. En la construcción de las filogenias, el uso de los genomas ha mejorado los estudios evolutivos previamente realizados con uno o unos pocos genes, dando paso al empleo de una mayor cantidad de información en el análisis filogenético. Algunos autores han mostrado que el contenido de genes compartido entre organismos está determinado cuantitativamente por la filogenia. Así, especies cercanas tienden a tener genomas similares [5, 6]. Sin embargo, los genes adquiridos por medio de la transferencia horizontal son una fuente de ruido cuando se construyen las filogenias [7, 8].

La transferencia horizontal de genes es el proceso mediante el cual un organismo incorpora material genético de otro organismo perteneciente a una especie diferente. El proceso de transferencia horizontal contrasta con el proceso normal de la herencia vertical en el cual un organismo recibe material genético directamente de los padres. La transferencia horizontal de genes tiene un gran impacto en la evolución de los genomas procariotas, ya que ésta permite que genes evolucionados en contextos diferentes sean combinados en un solo genoma, expandiendo así las vías en las que las bacterias pueden explorar el espacio del contenido de genes.

Además, a través de la transferencia horizontal poblaciones divergentes pueden compartir adaptaciones cuyo valor transciende sus diferencias en capacidades

fisiológicas, estructuras celulares y nichos ecológicos [8]. Las características genotípicas o fenotípicas que son compartidas por un conjunto de organismos pero que no son heredadas de un ancestro común son referidas como homoplasias [9] y son indicativas de evolución convergente. Así, la evolución convergente podría producir especies bacteriales que son funcionalmente similares a pesar de sus diferentes historias evolutivas.

Pangenoma, genoma núcleo y genoma accesorio

Hoy en día es una práctica común caracterizar el contenido total de genes de una especie bacterial mediante el llamado "*pangenoma*", el cual incluye un genoma núcleo compuesto por los genes presentes en todas las cepas y un genoma accesorio compuesto de genes ausentes en una o más cepas de una especie y de genes que son únicos a cada cepa [10].

El genoma núcleo comprende todos los genes necesarios para la reproducción y principalmente es heredado verticalmente [11] los cuales reflejan más estrictamente la filogenia. Por otra parte, el genoma accesorio comprende los genes necesarios para la adaptación a diferentes ambientes. El proceso de creación del genoma periférico ha sido extensamente estudiado en patógenos. En éstos, es común hallar genes de virulencia y patogenicidad, así como genes de resistencia a antibióticos con frecuencia producto del proceso de transferencia horizontal de genes [7, 12].

La enorme variabilidad del repertorio de genes representada en el genoma accesorio proporciona el potencial de adaptación para tener acceso a diferentes nichos ecológicos. Por lo tanto, es de esperarse que el contenido de genes refleje las preferencias ambientales. De esta manera, el contenido de genes es moldeado por la filogenia y el proceso de adaptación al ambiente. Sin embargo, aún no es claro cuál es la relación entre el contenido de genes y el ambiente. Por ejemplo, si existen organismos cercanamente relacionados con diferente contenido de genes en respuesta a adaptaciones ambientales; o si, por el contrario, existen casos de convergencia evolutiva donde organismos filogenéticamente distantes pueden tener genomas similares porque ellos viven en ambientes similares [13].

Antecedentes

El género *Bacillus*

El género *Bacillus* actualmente es uno de los más grandes dentro de la familia *Bacillaceae* con al menos 226 especies propuestas hasta septiembre de 2014 [14]. Este género comprende a las bacterias grampositivas, aerobias estrictas o facultativas, con forma de bastón, y cuya característica principal es su capacidad para formar endoesporas. Las endoesporas, son estructuras altamente refráctiles que pueden tener forma oval, cilíndrica o redonda y que se forman dentro de la célula bacterial. Es mediante este mecanismo, que los *Bacillus* pueden sobrevivir de forma latente al estrés nutricional y otros estreses ambientales [15, 16]. Los miembros de este género, se consideran organismos ubicuos dado que éstos han sido aislados de una serie de ambientes muy diversos, que abarcan tanto ambientes terrestres como acuáticos. Estos organismos han sido aislados de pozas hidrotermales, lagos salinos, suelos alcalinos, columnas de agua, marismas, sedimento del fondo del mar, etc. Sin embargo, no está claro si la presencia de endoesporas halladas en tal diversidad de ambientes se debe sólo a la dispersión de las esporas por el aire y/o el agua o a que sus capacidades metabólicas le permiten sobrevivir en estos ambientes [17, 18].

Las especies de *Bacillus* se clasifican de acuerdo a su ecofisiología en *Bacillus* psicrófilos, termófilos, acidófilos, alcalófilos y halófilos. De acuerdo a su importancia, los *Bacillus* se han clasificado en especies de importancia médica, industrial y ambiental [15]. La importancia de este género bacterial también se ha visto reflejada en el gran número de secuencias de genomas depositadas en la base de datos *GenBank*. Hasta Noviembre de 2017, cerca de 2000 secuencias de genomas completos y preliminares de poco más de 180 especies pertenecientes a este género han sido depositados en la base de datos. Este número presenta un gran sesgo hacia especies de importancia médica, así como de *Bacillus* aislados principalmente de ambientes terrestres, lo que ha limitado el estudio de este género en el contexto de su ecología. Sin embargo, durante los últimos años el número de secuencias de genomas de *Bacillus* aislados de ambientes acuáticos públicamente disponibles ha aumentado.

La comparación a nivel de genomas entre especies del género *Bacillus* se han llevado a cabo principalmente entre genomas particulares del mismo grupo [19, 20, 21, 22, 23, 24, 25, 26], y sólo pocos estudios han intentado incluir la gran diversidad dentro del género en la comparación de genomas y la reconstrucción filogenética [27].

El estudio de los miembros de este género ha sido dirigido principalmente al estudio de los aislados terrestres y patogénicos y pocos trabajos se han enfocado en los aislados de ambientes acuáticos. Sólo recientemente, los *Bacillus* aislados de medios acuáticos han comenzado a ser estudiados de manera sistemática. Sin embargo, la mayoría de estos trabajos se han enfocado principalmente al estudio de la diversidad [28, 29, 30, 31], así como a la búsqueda de nuevos compuestos con actividad biológica [28, 32]. Dado que los *Bacillus* han sido aislados tanto en sistemas terrestres como acuáticos cabe preguntarse: ¿qué características permiten a estas bacterias desarrollarse en tal variedad de ambientes?

Diversos estudios sugieren que los organismos de una misma especie aislados de ambientes distintos pueden ser ecológicamente diferentes, lo que se ve reflejado en su contenido genómico como la presencia de genes ambiente-específicos [33, 34, 35, 36]. Uno podría suponer que los organismos aislados de un mismo ambiente podrían compartir un conjunto de genes que definiera su adaptación al medio. Con la enorme cantidad de información disponible y la diversidad de ambientes de los cuales los *Bacillus* spp. han sido aislados, la comparación a nivel de genoma podría arrojar información sobre la historia evolutiva de los miembros de este género, así como de la presencia de genes ambiente-específicos los cuales pudieran estar involucrados en la adaptación al ambiente.

Bacillus acuáticos y terrestres, ¿una clasificación natural?

El análisis fenotípico y molecular de muestras filogenéticamente diversas de sedimento marino ha sugerido la existencia de un grupo de *Bacillus* estrictamente marinos [30]. Además, mediante el análisis filogenético empleando 821 proteínas conservadas en 20 genomas completos, Alcaraz *et al.* (2010) sugirió la existencia de un grupo de *Bacillus* acuáticos dentro del género [27]. Este grupo incluyó los *Bacillus* aislados de medios acuáticos reportados hasta ese momento: *Bacillus* sp. NRRL B-14911 aislado de columna de agua a 10 m de profundidad en el Golfo de México [29], *Bacillus coahuilensis* m4-4 [37] y *Bacillus* sp. m3-13 [27] ambos aislados de la poza Churince en Cuatro Ciénegas, México.

Actualmente, un número mayor de genomas de especies de *Bacillus* aisladas de ambientes acuáticos han sido secuenciados y depositados en las bases de datos. La

posible existencia de un grupo de *Bacillus* acuáticos, nos ha conducido a pensar que éstos pudieran tener un mismo origen evolutivo. Sin embargo, los recientes reportes de más *Bacillus* aislados de medios marinos y cuya filiación taxonómica determinada mediante el gen 16S rRNA es diferente a la de los miembros del grupo de acuáticos descritos por Alcaraz *et al.* [27] sugieren a su vez que los *Bacillus* acuáticos podrían tener su origen en linajes distintos. Esta nueva información nos permitiría, por una parte: 1) verificar la existencia del grupo de *Bacillus* acuáticos descrito previamente, y 2) estudiar los posibles orígenes evolutivos de los *Bacillus* aislados de ambientes acuáticos de diferentes sitios.

Algunos estudios recientes han establecido, mediante el análisis a nivel de genomas en bacterias, una relación entre las diferencias de hábitat y la presencia de genes ambiente-específicos. Por ejemplo, Luo *et al.*, (2011) mediante la comparación del genoma de una colección de cepas de *Escherichia coli* comensales o patogénicas de mamíferos y aves, y de 9 aislados ambientales tomadas de agua y sedimento, lograron identificar un conjunto de genes ambiente-específicos. En las cepas ambientales se observan todos los genes de la vía para la utilización de dioles como substratos de energía y un gen para la producción de lisozima para el rompimiento de las paredes celulares y ausentes en las cepas comensales o patogénicas, mientras que las cepas comensales y patogénicas incluyen genes involucrados en el transporte y metabolismo de nutrientes como la N-acetilglucosamina, gluconato, y azúcares de 5-C y 6-C como la fucosa, abundantes en el intestino [34]. Por otra parte, la comparación de genomas de *Pseudomonas putida* asociadas con ambientes contaminados y no contaminados mostró que las poblaciones de ambientes contaminados portan genes de resistencia a metales no hallados en las poblaciones de los sitios sin contaminar [36]. En los ambientes marinos *Prochlorococcus marinus* ha divergido en sus adaptaciones en la adquisición de fósforo: las poblaciones presentes en ambientes de bajo fósforo han adquirido un conjunto de genes involucrados en la ingestión, regulación y utilización de fosfatos orgánicos [33]. Estos estudios sugieren que dentro del genoma de las bacterias pueden existir genes que reflejen adaptaciones al medio ambiente en el que viven. Si esta hipótesis es correcta, esperaríamos encontrar, mediante la comparación a nivel de genoma, grupos de genes ambiente-específicos relacionados al hábitat en el que estas bacterias se desarrollan en los diferentes genomas. Los *Bacillus* provenientes de distintos ambientes nos dan la oportunidad de estudiar las relaciones evolutivas de este grupo, así como la de estudiar la existencia de genes ambiente-específicos que pudieran explicar su amplia distribución ecológica. Adicionalmente, este tipo de análisis nos permitiría estudiar los mecanismos que han moldeado la evolución del contenido genético de los miembros del género *Bacillus*.

Hipótesis

La adaptación de los organismos que se desarrollan en ambientes similares, podría verse reflejado en su contenido de genes como un conjunto de genes con funciones ambiente-específicas. Los *Bacillus* provenientes de linajes evolutivos distintos que se desarrollan en ambientes similares podrían compartir un conjunto de funciones específicas relacionadas al ambiente que ellos ocupan.

Objetivos

Objetivo General

Verificar la existencia del grupo de *Bacillus* acuáticos y su origen evolutivo a través de la reconstrucción de la historia evolutiva a nivel de genomas. Así como analizar la posible influencia que el ambiente ha ejercido en el contenido de genes de los *Bacillus*, mediante la comparación del contenido de genes teniendo en cuenta el ambiente natural del cual los *Bacillus* han sido aislados.

Objetivos particulares

1. Seleccionar y construir una base de datos de secuencias de genomas completos y preliminares representativa de las especies del género *Bacillus*.
2. Llevar a cabo la reconstrucción filogenética y filogenómica del género.
3. Clasificar los genomas de acuerdo el ambiente natural del que fueron aislados.
4. Comparar el contenido de genes entre los *Bacillus* que componen la muestra.
5. Analizar la existencia de genes posiblemente relacionados al ambiente de donde fueron aislados.

Materiales y Métodos

Escrutinio y selección de los genomas

Las secuencias genómicas fueron obtenidas de la base de datos RefSeq de NCBI [38]. La selección de genomas completos se llevó a cabo utilizando una herramienta bioinformática en línea (http://microbiome.wlu.ca/research/redundancy/redundancy.cgi). Esta herramienta toma el conjunto de genomas completos depositados en la base de datos *GenBank*, y los clasifica en grupos de genomas empleando diferentes medidas de distancia filogenómica *GSS* (GSS - Genomic Similarity Score) a diferentes umbrales de similitud [39]. Empleando esta herramienta, los genomas completos del género *Bacillus* depositados en la base de datos RefSeq hasta febrero de 2014, fueron agrupados utilizando la medida y el umbral $GSSa = 0{,}95$, el cual corresponde al umbral empleado para definir a la especie en bacterias [39]. Los grupos obtenidos por esta vía se denominaron grupos de genomas completos no-redundantes. Los grupos de genomas no-redundantes, se usaron como base para seleccionar hasta donde fue posible un par de representantes de cada grupo. Los representantes de cada grupo se seleccionaron de acuerdo a su tamaño, en algunos casos los organismos más representativos de cada grupo fueron seleccionados. De esta manera, se construyó una base de datos de genomas completos no-redundantes. Además, un conjunto de secuencias de genomas preliminares (*in draft*) fueron seleccionadas manualmente. Los genomas *in draft* seleccionados incluyen especies no reportadas en la base de datos de genomas completos no-redundantes y especies reportadas como aisladas de ambientes acuáticos. Los organismos utilizados como grupos externos en nuestros análisis, corresponden a los mismos empleados por Alcaraz *et al.* [27].

Análisis filogenético basado en el gen 16S rRNA

Debido a que los miembros del género *Bacillus* poseen múltiples copias del gen 16S rRNA, las secuencias del gen 16S rRNA de cada genoma fueron agrupadas usando un umbral de 97 % (el umbral utilizado para definir la unidad taxonómica

operativa en bacterias) empleando el programa *cd-hit-est* [40]. Las secuencias 16S rRNA elegidas por cd-hit-est como representante del grupo con el mayor número de secuencias en cada genoma, fueron seleccionadas para el análisis. Las secuencias fueron alineadas empleando modelos de covarianza mediante el software Infernal 1.1 [41]. Con el objeto de eliminar las posiciones con mayor incertidumbre dentro del alineamiento, el alineamiento se enmascaró utilizando el programa Zorro [42]. Las posiciones dentro del alineamiento con valores menores al recomendado por los autores ($\leq$ 0,40) fueron eliminadas del alineamiento mediante un *script* escrito en Perl. El modelo de sustitución se calculó empleando el programa jModelTest [43]. La reconstrucción filogenética por máxima verosimilitud se llevó a cabo empleando el software PhyML [44] utilizando el modelo GTR+I+G+F con los parámetros $I = 0{,}643$, $G = 0{,}485$ y $F = 0{,}24508, 0{,}22268, 0{,}30847, 0{,}22378$.

Reconstrucción filogenómica utilizando marcadores filogenómicos.

El uso de marcadores filogenómicos ha sido propuesto como una forma de resolver las inconsistencias observadas en las reconstrucciones filogenéticas basadas en un sólo gen. Con el objeto de alcanzar una mejor resolución en nuestro análisis, la búsqueda de marcadores filogenómicos se llevó a cabo utilizando el paquete AMPHORA2 [45]. El corazón de AMPHORA2 es su base de datos, la cual consiste de alineamientos curados manualmente y de perfiles basados en modelos ocultos de Markov (HMMs) de 31 proteínas universalmente conservadas. Las secuencias de proteína de 11 marcadores conservados en todos los genomas que componen nuestra muestra, fueron alineadas y podadas mediante el script MarkerAlignTrim.pl del paquete AMPHORA2. Posteriormente los alineamientos podados fueron concatenados para construir una súper-matriz con 2088 residuos de aminoácidos. La súper-matriz se utilizó para determinar el modelo de sustitución mediante el programa ProtTest 3.4 [46]. La reconstrucción por máxima verosimilitud se realizó a través del programa PhyML [44], usando el modelo de sustitución LG+I+G ($alpha = 0{,}86$, $p - inv = 0{,}36$)

Reconstrucción filogenómica empleando el genoma núcleo.

Para determinar los genes que componen el genoma núcleo de los organismos en nuestra base de datos, la ortología de los genes entre los diferentes genomas fue establecida empleando el método del mejor "hit" recíproco (RBH) como se

describe en [47]. Las secuencias ortólogas fueron filtradas para obtener dos grupos: el *Genoma núcleo* donde las secuencias se conservan en el 100 % de los genomas, y el *Genoma núcleo 70* en el cual los genes ortólogos se encuentran presentes en al menos el 70 % de los genomas. Las secuencias en ambos grupos fueron tratadas siguiendo la misma metodología: la secuencia de proteínas de los genes ortólogos identificados, se alinearon utilizando el programa MUSCLE [48]. Cada uno de los alineamientos fue enmascarado mediante el programa Zorro [42] para eliminar las posiciones más inciertas. Las posiciones con valores menores al recomendado por los autores (0,40) fueron removidas del alineamiento mediante un script de Perl. Los alineamientos podados fueron entonces concatenados para construir una súper matriz. El modelo de sustitución se calculó mediante el programa ProtTest 3.4 [46].

La reconstrucción por máxima verosimilitud del genoma núcleo se realizó usando el programa PhyML empleando el modelo de sustitución LG+I+G ($alha$ = 0,241 , $p - inv$ = 0,826). Debido a las limitaciones computacionales, la reconstrucción filogenómica basada en el Genoma núcleo 70, se llevó a cabo utilizando el programa RAxML (argumentos: -m PROTGAMMALGF -# 100) implementado en la plataforma *CIPRES science Gateway (Cyberinfrastructure for Phylogenetic Research)* [49].

Reconstrucción por el método de *Neighbor-joining* empleando los índices de similitud genómica (GSS).

Se ha mostrado que los índices de similitud genómica (GSS) y otras medidas filogenómicas coinciden con distancias filogenéticas computacionalmente más demandantes [27, 50]. Los índices de similitud genómica (GSS) entre todos los organismos en nuestra base de datos fueron calculados como se describe en [51]. Empleando una matriz de distancia basada en el índice de similitud genómica (GSS), el dendrograma fue calculado mediante el programa "*neighbor*" del paquete Phylip [52]. Los valores de soporte estadístico de las ramas en este árbol se calcularon mediante máxima verosimilitud usando el software *WeightLESS* [53].

Análisis de congruencia.

La congruencia entre la topología de los árboles se analizó empleando la diferencia simétrica de Robinson-Fould [54] usando el programa Treedist incluído en el paquete Phylip [52]. El grado de resolución de cada uno de los árboles se determinó contando el número de nodos con valor de *bootstrap* $\geq$ 80; en el caso del árbol obtenido mediante la distancia *GSS*, los nodos con valores significativos

corresponden a $P - val \leq 0{,}05$. Los árboles fueron analizados mediante un script escrito en Python empleando el paquete *ete2*.

Determinación del Contenido Funcional

El contenido funcional de un organismo se define como el conjunto de genes cuyos productos tienen una función biológica asignada. Normalmente, la asignación de la función se realiza, tomando como base lo que se conoce experimentalmente en organismos modelo, mediante la búsquedas por homología de secuencia y del establecimiento de las relaciones de ortología entre los genes de distintas especies u organismos. El análisis del contenido funcional se llevó a cabo empleando las coincidencias a dos tipos de familias de proteínas con funciones conservadas tales como la clasificación de Grupos Ortólogos (COG) [55] y de Figfams [56].

La clasificación por COG para los genes en nuestra base de datos se llevó a cabo usando el programa RPSBLAST [57] y las matrices posición específica (PSSMs) COG V1.0 [57]. Las asignaciones de COGs fueron evaluadas para descartar aquellas con alineamientos que cubren menos del 70 % de PSSMs. Nosotros permitimos un translape máximo entre COGs alineados dentro de una proteína $\leq$10 %. La clasificación del contenido funcional a través de Figfams se llevó a cabo empleando las anotaciones de los genomas hechas con RAST (Rapid Annotations using Subsystems Technology) [58] obtenidas de la base de datos PATRIC [59]. Los genomas no presentes en la base de datos PATRIC fueron anotados localmente usando la interfaz *myRAST-Toolkit* [58].

Análisis de agrupamiento utilizando el contenido funcional

Para comparar los contenidos funcionales, los índices de Jaccard fueron calculados como: $J(A, B) = (A \cap B)/(A \cup B)$ donde: A y B son el conjunto de categorías funcionales COG/Figfams presentes en el genoma A y B, respectivamente. La distancia de Jaccard por lo tanto se define como: $Jd(A, B) = 1 - J(A, B)$.

Cada una de las matrices conteniendo las distancias de Jaccard (Jd) para COG y Figfams fueron utilizadas para llevar a cabo un agrupamiento jerárquico (*Clustering*) evaluando diferentes métodos de agrupamiento. La calidad de la estructura de los agrupamientos obtenidos con los diferentes métodos de agrupamiento, se evaluó empleando el coeficiente de aglomeración. Este coeficiente se define como uno menos la relación promedio de disimilitud de un elemento al grupo con el que se fusiona en primer lugar, a la disimilitud de su fusión en el paso final. Si el coeficiente es cercano a cero, el algoritmo no halló una estructura de grupo natural

(los datos consisten de un solo grupo), si este es cercano a 1, entonces existen grupos bien definidos. Los grupos obtenidos fueron además evaluados por sus valores *Silhouette* [60]. Estos valores reflejan la distancia de cada elemento de un grupo a los otros elementos del mismo grupo, comparadas a las distancias en contra de los miembros de los otros grupos. Los valores van de -1 a 1, con 1 indicando que el elemento claramente pertenece al grupo donde ha sido colocado.

Análisis de componentes principales

El análisis de componentes principales (Principal Component Analysis o PCA por sus siglas en inglés) se llevó a cabo empleando la matriz de frecuencias de las categorías COG y Figfams. El número de componentes se seleccionó graficando la varianza de cada uno de los componentes principales (*Scree plot*) y/o reteniendo los componentes para los cuales la varianza es mayor a la unidad (criterio de Kaiser).

Clasificación de los ambientes naturales de aislamiento

La información sobre el ambiente del cual los organismos presentes en nuestra base de datos fueron aislados se obtuvo de distintas fuentes. Principalmente, de la información asociada a cada genoma en la base de datos NCBI [61], de la base de datos HAMAP [62], de la base de datos PATRIC [59] y, cuando la información estaba ausente en las fuentes anteriores, de la literatura. Los ambientes hallados, fueron clasificados siguiendo los mismos criterios de Parter, *et al.* [63]. Así, los ambientes fueron clasificados como *Acuático* para *Bacillus* que viven en agua dulce o marina, *Terrestres* aquellos que viven en el suelo, *Facultativos* para aquellos que han sido aislados asociados a un hospedero y *Especializados* para aquellos organismos aislados de ambientes como ventilas termales y una cepa aislada del aire. La categoría *No Clasificada* fue agregada para ubicar aquellas cepas cuyo ambiente de aislamiento no pudo ser determinado.

Asociación ambiente natural - agrupamiento por contenido funcional

Para establecer si existe una asociación entre el ambiente natural del cual los *Bacillus* fueron aislados y los agrupamientos obtenidos utilizando la distancia de Jaccard basada en COG/Figfams, los dendrogramas de COG fueron cortados de $k = 3$ a $k = 22$ grupos, mientras que el dendrograma Figfam se cortó de $k = 3$ a

$k = 16$ grupos. Basados en la distribución hipergeométrica, un valor-p (*p-val*) fue calculado. Los *p-val* obtenidos, fueron corregidos usando el método de la "Tasa de falsos descubrimiento" (False Discovery Rate - FDR), mediante la función *P-adjust* implementada en R [64].

Funciones grupo específicas

Para investigar si los agrupamientos observados poseen funciones grupo específicas, los 3 principales grupos en el dendrograma obtenido con la Jd y la matriz de frecuencias de COG se compararon usando el paquete *ShotgunFunctionalizeR* [65]. El paquete *ShotgunFunctionalizeR* permite comparar dos grupos con múltiples muestras. La comparación de las frecuencias de COG, se realizó de forma pareada entre grupos, y los COG con *p-val* corregidos < 0.05 fueron seleccionados. Adicionalmente, los COG seleccionados fueron revisados para verificar su presencia en al menos el 80 % de los miembros del grupo. Entonces se verificó la sobrerepresentación o subrepresentación del COG respecto a los otros dos grupos.

Resultados y Discusión

Para investigar la potencial relación entre los orígenes evolutivos de los miembros del género *Bacillus* y el ambiente que ellos ocupan, llevamos a cabo la reconstrucción filogenética y filogenómica de este género bacterial y la comparamos con la información recopilada sobre sus ambientes naturales de aislamiento. Además, llevamos a cabo un análisis de agrupamiento jerárquico basado en el contenido funcional para investigar el potencial efecto del ambiente en el contenido de genes de estas bacterias. La figura 1 muestra la metodología empleada en este estudio.

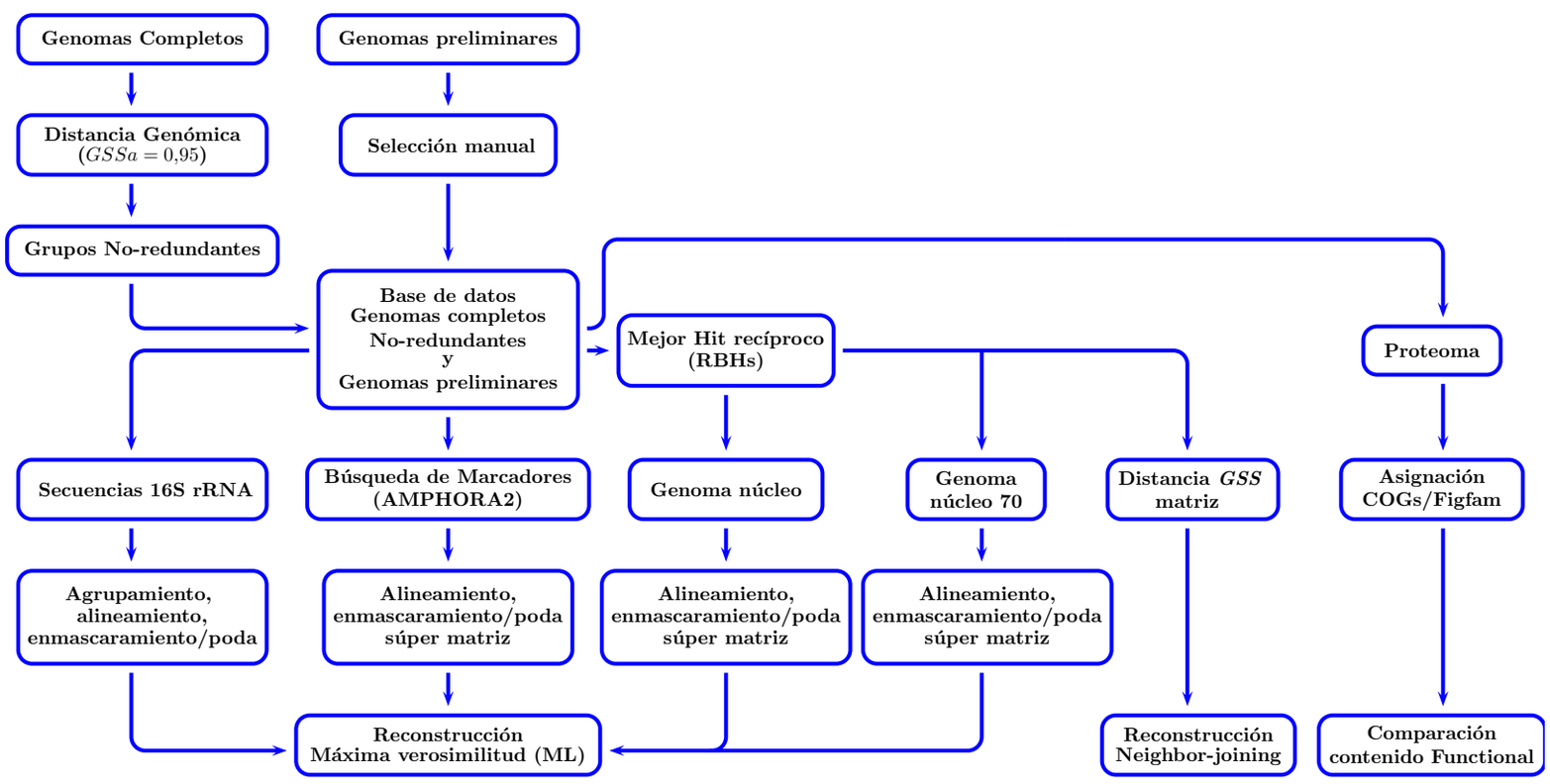

Figura 1: Esquema de la metodología seguida en este trabajo.

Selección de genomas

Durante los últimos años, el número de genomas completos y preliminares depositados en la base de datos *GenBank* se ha visto aumentado significativamente. Tan solo hasta diciembre de 2017 aproximadamente 2082 secuencias de genoma completos y preliminares pertenecientes a especies del género *Bacillus* habían sido depositadas en la base de datos de genomas RefSeq de NCBI [38]. Sin embargo, este número de genomas presenta enormes sesgos hacia algunas especies dentro del género, con algunas especies que tienen un gran número de cepas representadas, mientras que otras especies tienen solo un representante. Con el fin de obtener un conjunto de genomas que reflejara la diversidad del género *Bacillus* presente en las bases de datos, la selección se llevó a cabo siguiendo los criterios descritos en los materiales y métodos. Los genomas completos fueron seleccionados empleando el índice de similitud genómica ($GSSa = 0{,}95$). El Apéndice A muestra los 33 grupos obtenidos utilizando el *GSS*, así como los genomas seleccionados de cada grupo. De esta manera, se seleccionaron 50 genomas completos. Por otra parte, 29 genomas preliminares se seleccionaron manualmente (Materiales y Métodos). Además, 4 genomas fueron seleccionados como grupos externos, los cuales corresponden a los mismos empleados por Alcaraz *et al.* [27]. En total, 83 genomas fueron seleccionados para llevar a cabo nuestros análisis. La tabla 1, muestra los 83 genomas empleados en este estudio.

Tabla 1: Especies empleadas en este estudio

Organismo	**Bioproject**	**Tamaño (Mb)**	**GC (%)**	**Estatus**
B. alcalophilus ATCC 27647	PRJNA171835	4.22	37.20	In draft
B. amyloliquefaciens DSM 7	PRJEA41719	3.98	46.10	Completo
B. amyloliquefaciens TA208	PRJNA64581	3.94	45.80	Completo
B. amyloliquefaciens plantarum FZB42	PRJNA13403	3.92	46.50	Completo
B. amyloliquefaciens plantarum UCMB5036	PRJEB1155	3.91	46.60	Completo
B. anthracis Ámes Ancestor'	PRJNA10784	5.50	35.26	Completo
B. anthracis Ames	PRJNA309	5.23	35.40	Completo
B. anthracis Sterne	PRJNA10878	5.23	35.40	Completo
B. aquimaris TF12	PRJNA71387	4.03	37.3	In draft
B. atrophaeus 1942	PRJNA46075	4.17	43.20	Completo
B. azotoformans LMG 9581	PRJNA80827	4.22	37.20	In draft
B. bataviensis LMG 21833	PRJNA77725	5.37	39.60	In draft

continua en la página siguiente

Tabla 1 – *Continuación de la página previa*

Organismo	**Bioproject**	**Tamaño (Mb)**	**GC (%)**	**Estatus**
B. cellulosilyticus DSM 2522	PRJNA38423	4.68	36.50	Completo
B. cereus AH187	PRJNA17715	5.60	35.52	Completo
B. cereus ATCC 10987	PRJNA74	5.43	35.52	Completo
B. cereus ATCC 14579	PRJNA384	5.43	35.31	Completo
B. cereus B4264	PRJNA17731	5.42	35.30	Completo
B. cereus E33L	PRJNA12468	5.84	35.17	Completo
B. cereus F837/76	PRJNA15716	5.29	35.40	Completo
B. cereus FRI-35	PRJNA171769	5.38	35.45	Completo
B. cereus G9842	PRJNA17733	5.74	35.05	Completo
B. cereus Q1	PRJNA16220	5.51	35.50	Completo
B. cereus anthracis CI	PRJNA36309	5.49	35.27	Completo
B. clausii KSM-K16	PRJNA13291	4.30	44.80	Completo
B. coagulans 2-6	PRJNA61501	3.07	47.30	Completo
B. coagulans 36D1	PRJNA15679	3.55	46.50	Completo
B. coahuilensis m2-6	PRJNA84423	3.21	34.9	In draft
B. coahuilensis m4-4	PRJNA19551	3.38	38.00	In draft
B. coahuilensis p1.1.43c	PRJNA84425	3.41	33.6	In draft
B. cytotoxicus NVH 391-98	PRJNA13624	4.09	35.89	Completo
B. halodurans C-125	PRJNA235	4.20	43.70	Completo
B. horikoshii DSM 8719	PRJNA72395	4.67	41.3	In draft
B. infantis NRRL B-14911	PRJNA212797	4.88	46.00	Completo
B. isronensis B3W22	PRJNA173035	4.02	38.80	In draft
B. licheniformis 9945A	PRJNA49115	4.38	45.90	Completo
B. licheniformis DSM 13 = ATCC 14580	PRJNA13082	4.22	46.20	Completo
B. macauensis ZFHKF-1	PRJNA167832	3.74	40.60	In draft
B. megaterium DSM 319	PRJNA42425	5.10	38.10	Completo
B. megaterium QM B1551	PRJNA30165	5.52	37.97	Completo
B. megaterium WSH-002	PRJNA71447	5.08	38.15	Completo
B. methanolicus MGA3	PRJNA49595	3.40	38.50	In draft
B. mojavensis RO-H-1	PRJNA68567	3.95	43.7	In draft
B. mycoides DSM 2048	PRJNA29701	5.56	35.20	In draft
B. oceanisediminis 2691	PRJNA167766	5.76	40.80	In draft
B. pseudofirmus OF4	PRJNA28811	4.25	39.89	Completo
B. pumilus SAFR-032	PRJNA20391	3.70	41.30	Completo
B. selenitireducens MLS10	PRJNA13376	3.59	48.70	Completo
B. stratosphericus LAMA 585	PRJNA176166	3.71	41.20	In draft
B. subtilis PY79	PRJNA225627	4.03	43.80	Completo
B. subtilis gtP20b	PRJNA53249	4.21	44.00	In draft
B. subtilis spizizenii TU-B-10	PRJNA68561	4.21	43.80	Completo
B. subtilis spizizenii W23	PRJNA38713	4.03	43.90	Completo
B. subtilis subtilis 168	PRJNA76	4.22	43.50	Completo
B. thuringiensis Al Hakam	PRJNA18255	5.31	35.41	Completo
B. thuringiensis BMB171	PRJNA43631	5.64	35.19	Completo

continua en la página siguiente

Tabla 1 – *Continuación de la página previa*

Organismo	**Bioproject**	**Tamaño (Mb)**	**GC (%)**	**Estatus**
B. thuringiensis HD-771	PRJNA171845	6.44	35.04	Completo
B. thuringiensis MC28	PRJNA167562	6.69	34.92	Completo
B. thuringiensis chinensis CT-43	PRJNA43737	6.15	35.12	Completo
B. thuringiensis finitimus YBT-020	PRJNA60447	5.68	35.38	Completo
B. thuringiensis konkukian 97-27	PRJNA10877	5.31	35.36	Completo
B. thuringiensis kurstaki HD73	PRJNA185468	5.91	35.19	Completo
B. thuringiensis thuringiensis IS5056	PRJNA187142	6.77	34.91	Completo
B. toyonensis BCT-7112	PRJNA225857	5.03	35.55	Completo
B. weihenstephanensis KBAB4	PRJNA13623	5.87	35.52	Completo
Bacillus sp. 10403023	PRJEA70827	4.60	36.8	In draft
Bacillus sp. 1NLA3E	PRJNA53255	4.82	38.00	Completo
Bacillus sp. 2_A_57_CT2	PRJNA40003	5.88	40.90	In draft
Bacillus sp. 5B6	PRJNA79215	3.90	46.60	In draft
Bacillus sp. 7_6_55CFAA_CT2	PRJNA40005	5.75	34.90	In draft
Bacillus sp. B14905	PRJNA18949	4.50	37.60	In draft
Bacillus sp. BT1B_CT2	PRJNA40001	4.37	45.70	In draft
Bacillus sp. HYC-10	PRJNA162763	3.61	41.30	In draft
Bacillus sp. JS	PRJNA79217	4.12	43.90	Completo
Bacillus sp. L1(2012)	PRJNA182346	3.86	40.80	In draft
Bacillus sp. M 2-6	PRJNA161543	3.80	41.10	In draft
Bacillus sp. NRRL B-14911	PRJNA13545	5.12	45.70	In draft
Bacillus sp. SG-1	PRJNA19283	3.95	42.10	In draft
Bacillus sp. m3-13	PRJNA38237	4.14	40.70	In draft
Bacillus sp. p15.4	PRJNA384653	4.86	43.7	In draft
Geobacillus sp. Y412MC52	PRJNA30797	3.67	52.31	Completo
Listeria innocua Clip11262	PRJNA86	3.09	37.35	In draft
Listeria monocytogenes 4b CLIP 80459	PRJEA32207	2.91	38.10	Completo
Oceanobacillus iheyensis HTE831	PRJNA284	3.63	35.70	Completo

Reconstrucción de la historia evolutiva del género *Bacillus*

Para investigar la potencial relación entre el orígen evolutivo de los *Bacillus* y el ambiente que ellos ocupan, la reconstrucción filogenética y filogenómica del género se llevó a cabo mediante máxima verosimilitud (ML) empleando 4 grupos de datos diferentes: 1) las secuencias del gen 16S rRNA, 2) un conjunto de marcadores filogenómicos a nivel de proteínas, 3) el conjunto de proteínas conservadas en el 100 % de los genomas que forman el llamado *Genoma núcleo*, y 4) el *Genoma 70*, que correspon al conjunto de proteínas conservadas en al menos el 70 % de los genomas. Adicionalmente, se llevó a cabo la reconstrucción filogenética por distancia basada en el contenido de genes empleando el índice de similitud genómica (GSS) como medida de distancia. Los árboles obtenidos de cada reconstrucción fueron comparados entre sí para analizar la conservación de las topologías.

Reconstrucción filogenética basada en el gen 16S rRNA

La secuencia del gen 16S rRNA ha sido extensamente utilizada para la identificación y clasificación taxonómica de las bacterias, por lo que ha sido considerado el estándar de oro. En general, en este árbol los valores de la prueba de remuestreo o *bootstrap* son bajos (*bootstrap* < 80) y sólo algunos clados presentan valores significativos (*bootstrap* ≥ 80) (Tabla 2). El árbol obtenido de la reconstrucción filogenética por máxima verosimilitud (ML) se muestra en la figura 2. Como se puede observar también, el árbol obtenido recupera los grupos de *B. cereus* y *B. subtilis* los cuales han sido ampliamente descritos en la literatura. En el caso de los *Bacillus* aislados de ambientes acuáticos, estos resultados sugieren un orígen polifilético puesto que se encuentran distribuidos en diferentes clados del árbol.

Tabla 2: Número de nodos con valores de *bootstrap* ≥ 80

	16S rRNA	Marcadores AMPHORA	Genoma Núcleo	Genoma 70	Distancia *GSS* *
Num. Nodos ≥ 80	22	44	69	74	46

*Los valores en el árbol GSS están basados en máxima verosimilitud ($p \leq 0{,}05$) no en valores *bootstrap*.

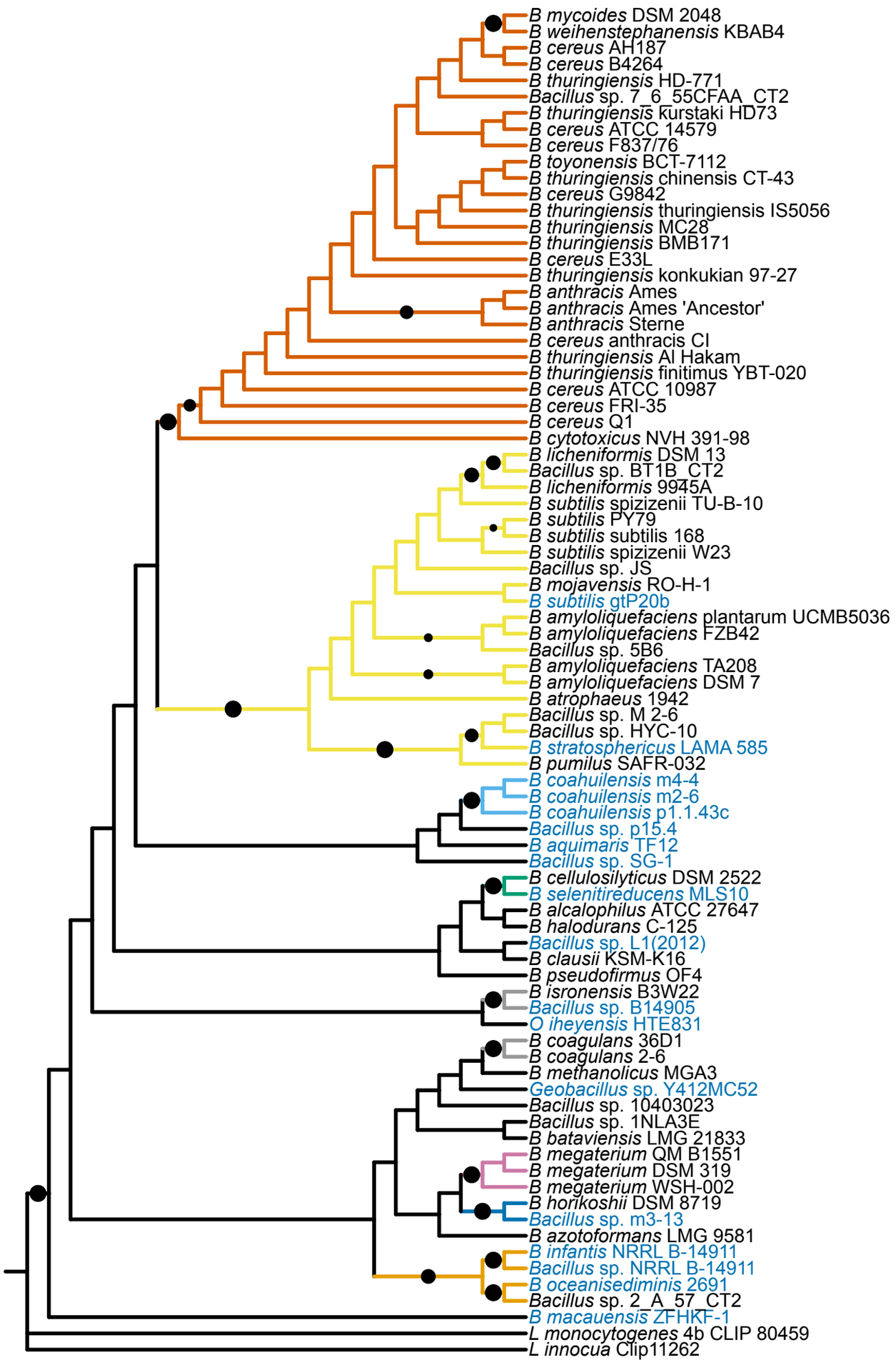

Figura 2: Reconstrucción filogenética por ML basada en el gen 16S rRNA. El nombre de las especies aisladas de ambientes acuáticos se indica en color azul. Los valores de la prueba de *bootstrap* se indican como puntos sobre los nodos, sólo se muestran los valores ≥ 80.

No obstante, es posible observar la existencia de un clado formado exclusivamente de *Bacillus* aislados de ambientes acuáticos. Sin embargo, este grupo no presenta apoyo estadístico significativo (≥ 80).

Reconstrucción filogenómica mediante el uso de marcadores

Con frecuencia la información que proporciona un solo gen es insuficiente para obtener un firme apoyo estadístico para nodos particulares dentro de una filogenia [66]. El uso de marcadores filogenómicos ha sido propuesto como una forma de resolver las inconsistencias observadas en las reconstrucciones filogenéticas basadas en un sólo gen. Por esta razón, llevamos a cabo la reconstrucción filogenómica por máxima verosimilitud utilizando los alineamientos concatenados de 11 marcadores identificados en los 83 genomas empleados en nuestro estudio (Materiales y Métodos). La reconstrucción filogenómica utilizando los marcadores, resultó en un árbol con un mayor número de nodos con valores de *bootstrap* > 80, comparado con el árbol obtenido utilizando el gen 16S rRNA (Tabla 2). Los grupos *B. cereus* y *B. subtilis* observados en el árbol obtenido con el gen 16S rRNA también fueron observados en este árbol, con algunos rearreglos menores. Además, fue posible observar en este árbol cinco grupos adicionales con un apoyo significativo de bootstrap: los grupos *B. megaterium*, *B. clausii*, *B. methanolicus*, *B. coagulans* y *B. isronensis* (figura 3).

Reconstrucción filogenómica basada en el genoma núcleo

Como con el uso de marcadores, la utlización del así llamado genoma núcleo para la reconstrucción filogenómica tiene el potencial de proporcionar una mayor resolución del árbol. El genoma núcleo se define como el conjunto de genes compartidos entre todas las cepas de una misma especie de bacteria, mientras que el conjunto de genes no conservados constituye el genoma accesorio [10, 67]. Estos conceptos han sido extendidos para abarcar otros niveles taxónomicos [27, 68]. Recientemente, el análisis filogenético basado en el genoma accesorio (el conjunto de todos los genes presentes en al menos un porcentaje definido de las cepas de una misma especie) ha sido propuesto como una forma para mejorar el apoyo filogenético de un árbol medido a través de las probabilidades *bootstrap* [69].

La reconstrucción filogenómica basada en el *Genoma Núcleo* se llevó a cabo utilizando los alineamientos concatenados de 196 proteínas ortólogas (Materiales y Métodos). Como era de esperarse el número de genes en el genoma núcleo resultó en un número menor a los 814 genes reportados por Alcaraz *et al.* [27] dado que el número de genes compartidos tiende a disminuir cuando el número de genomas en estudio aumenta [10]. El árbol obtenido con el Genoma Núcleo presentó un alto número de nodos con valores de bootstrap ≥ 80 en comparación con el árbol

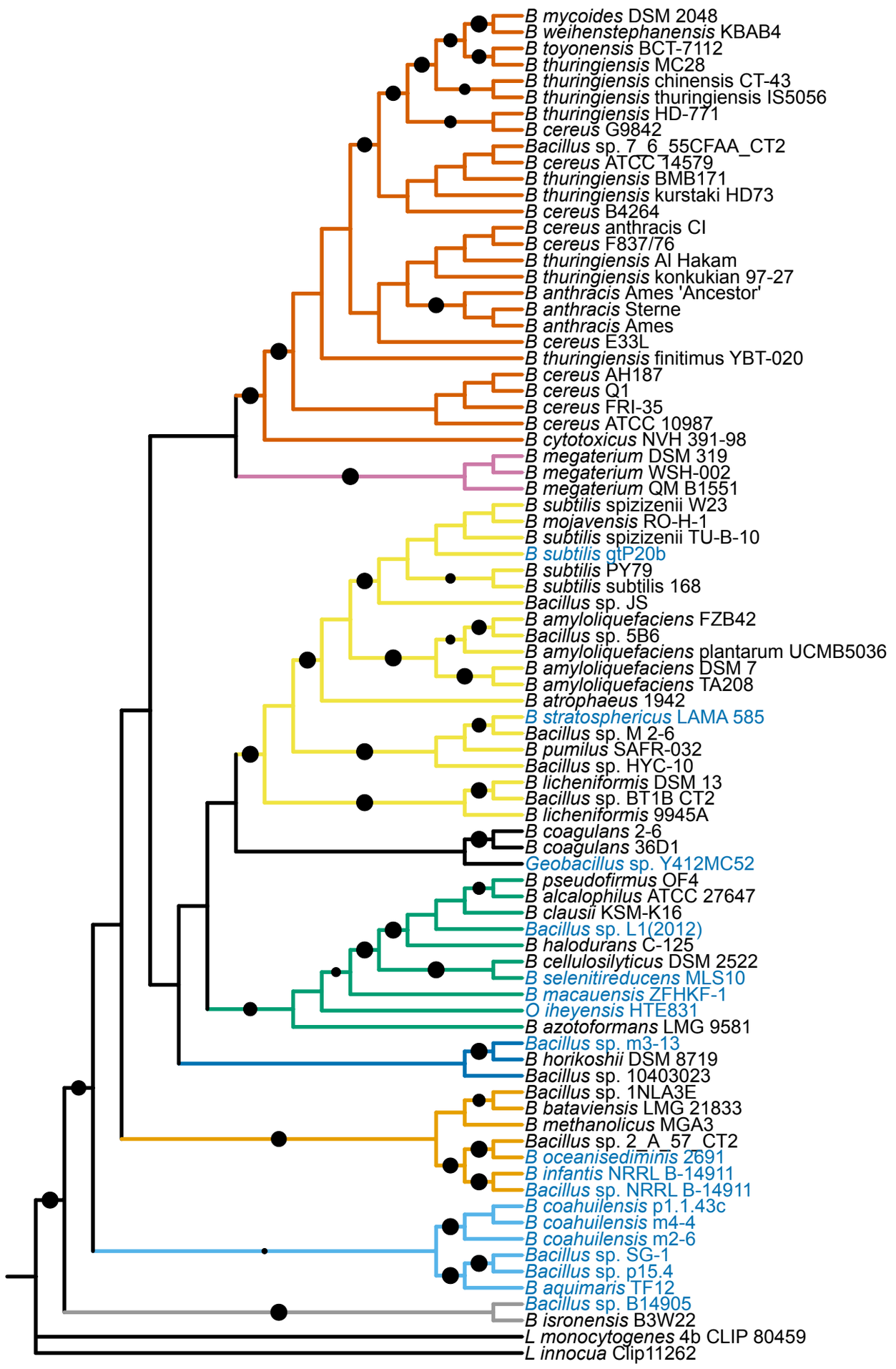

Figura 3: Árbol filogenómico basado en la secuencia de proteína de 11 marcadores. Los 11 marcadores fueron identificados usando el paquete AMPHORA2. El nombre de las especies aisladas de ambientes acuáticos se indican en color azul. Los valores de la prueba de *bootstrap* se indican como puntos sobre los nodos, sólo los valores ≥ 80 son mostrados.

obtenido con el gen 16S rRNA y el árbol obtenido con 11 marcadores filogenómicos (Tabla 2). El árbol del *Genoma 70*, se obtuvo empleando 437 grupos ortólogos conservados en al menos el 70 % de las especies. Este árbol resultó en un número mayor de nodos con valores de bootstrap $\geq$ 80 (Tabla 2) en comparación con los anteriores y con una topología muy similar a la del genoma núcleo (Tabla 3). Los árboles obtenidos usando los genes ortólogos (*Genoma núcleo* y *Genoma 70*) resultaron en nueve grupos bien definidos y bien apoyados por la prueba de remuestreo (figura 4 y 5).

El clado más grande en ambos árboles, agrupó las especies cercanamente relacionadas *B. cereus*, *B. thuringiensis*, *B. anthracis*, *B. weihenstephanensis*, *B. mycoides*, *B. toyonensis*, y *B. cytotoxicus*, las cuales se ha propuesto forman el grupo *Bacillus cereus sensu lato* [70, 71, 72]. Estudios previos empleando electroforesis de enzimas multilocus (MEE) [73], la tipificación de secuencia multilocus (MLST) [74, 75] y la comparación genómica [69, 76] han establecido una cercana relación entre estos organismos.

El segundo grupo más grande incluyó las especies *B. subtilis*, *B. amyloliquefaciens*, *B. licheniformis*, *B. athrophaeus*, *B. mojaviensis*, *B. pumilus*, así como algunos *Bacillus* no clasificados a nivel de especie: *Bacillus* sp. HYC–10, *Bacillus* sp. M 2–6, *Bacillus* sp. BT1B_CT2, *Bacillus* sp. 5B6, y *Bacillus* sp. JS; ademaś en este grupo se encontraron dos aislados marinos fueron incluidos: *B. stratosphericus* y *B. subtilis* gtP20b. Estudios previos, sugieren que *Bacillus pumilus*, *B. stratosphericus*, *Bacillus* sp. HYC–10 (ahora *Bacillus xiamenensis* [77]), y otras especies no incluidas en este estudio forman el clado *B. pumilus* [77, 78, 79]. Taxonómicamente, las especies del grupo *B. pumilus* forman parte del clado *B. subtilis* [80].

Un tercer grupo comprendió las cepas *B. alcalophilus*, *B. pseudofirmus*, *B. halodurans*, *B. clausii*, *B. selenitireducens*, *B. cellulosilyticus*, *B. macauensis*, *B. azotoformans* y el aislado marino *Bacillus* sp. L1(2012). Dos de los organismos empleados como grupos externos fueron localizados en este clado: *Oceanobacillus iheyensis* y *Geobacillus* sp. Y412MC52, sin embargo, su posición dentro del clado no mostró apoyo estadístico significativo. En el caso de *Geobacillus* sp. Y412MC52, el análisis de los genomas núcleo y accesorio sugiere que el genoma núcleo de los *Geobacillus* despliegua una clara homología al de los *Bacillus* [81], esto prodría explicar el porque *Geobacillus* no se localiza como un grupo externo.

Los otros clados correspondieron a los grupos menos representados. Dentro de estos grupos, las cepas de *B. megaterium* forman un grupo monofilético localizado más cercano al clado de *B. subtilis*. Adyacente a este grupo, *B. horkoshii*, *Bacillus* sp. 10403023, y el *bacillus* acuático *Bacillus* sp. m3-13, formaron otro grupo, mientras que *B. methanolicus*, *B. bataviensis*. *Bacillus* sp. 1NLA3E y *Bacillus* sp. 2_A_57_CT2 se agruparon con las cepas marinas *B. oceanisediminidis*, *B.*

infantis, y *Bacillus* sp. NRRL B-14911.

Basal a este grupo, se observó un clado compuesto exclusivamente de *Bacillus* acuáticos aislados tanto de ambientes marinos como no marinos. Este grupo esta formado por *B. aquimaris* TF12, *Bacillus* sp. SG1, y *Bacillus* sp. P15.4, así como de las cepas de la especie *Bacillus coahuilensis*. Los dos últimos clados corresponden al grupo formado por *B. isronensis* y *Bacillus* sp. B14905 una cepa marina. Por último, se tiene el grupo compuesto por las cepas de *B. coagulans*, los *Bacillus* más pequeños reportados a la fecha.

Es importante notar que el grupo de *Bacillus* acuáticos descrito previamente [27], no se observó en nuestro análisis dado que las especies que lo componían (*B. coahuilensis* m4–4, *Bacillus* sp. NRRL B14911 y *Bacillus* sp. m3–13) en nuestras reconstrucciones se hallaron dispersas en tres diferentes clados (figura 4 y 5). La disrupción de este grupo podría ser consecuencia del mayor número de cepas aisladas de diversos ambientes, disponibles para este estudio. La distribución en nuestro árbol sugiere que los *Bacillus* acuáticos son polifiléticos, ya que múltiples clados incluyen al menos un *Bacillus* acuático. No obstante, se puede observar un grupo compuesto exclusivamente de especies aisladas de ambientes acuáticos (figura 4 y 5). Estas cepas aisladas de sitios geográficos distintos, apoyan la hipótesis de un sólo orígen evolutivo para al menos algunos de los *Bacillus* acuáticos.

En general, los árboles obtenidos usando el gen 16S rRNA, los 11 marcadores filogenéticos, el *Genoma Núcleo* y el *Genoma 70*, recrean los principales grupos y difieren en la proporción de clados apoyados estadísticamente por un *bootstrap* ≥ 80. La Tabla 2 muestra el número de nodos con apoyo estadístico ≥ 80 para cada uno de los árboles obtenidos por ML. Es evidente de esta tabla que el número de nodos con valores de *bootstrap* ≥ 80 aumenta en relación al número de genes utilizados en cada reconstrucción. A pesar de las diferencias en los valores de *bootstrap* entre los diferentes árboles, todos los árboles muestran distribuciones similares con respecto al orígen polifilético de los *Bacillus* acuáticos. Además, el clado formado exclusivamente por *Bacillus* acuáticos se observa en todas las reconstrucciones filogéneticas (Figura 2, 3, 4 y 5).

Los resultados obtenidos en nuestras reconstrucciones son consistentes con los resultados obtenidos en otros estudios, los cuales han empleado los alineamientos concatenados de 814 proteínas de 20 genomas completos [27]; con aquellos usando los alineamientos concatenados de 157 genes de copia única de la familia *Bacillaceae* [82] y con los obtenidos empleando 20 proteínas de mantenimiento y ribosomales de 34 especies del género *Bacillus* [80], todos los cuales recrean los principales clados dentro del género: el *clado subtilis*, el *clado cereus* y el *clado clausii*.

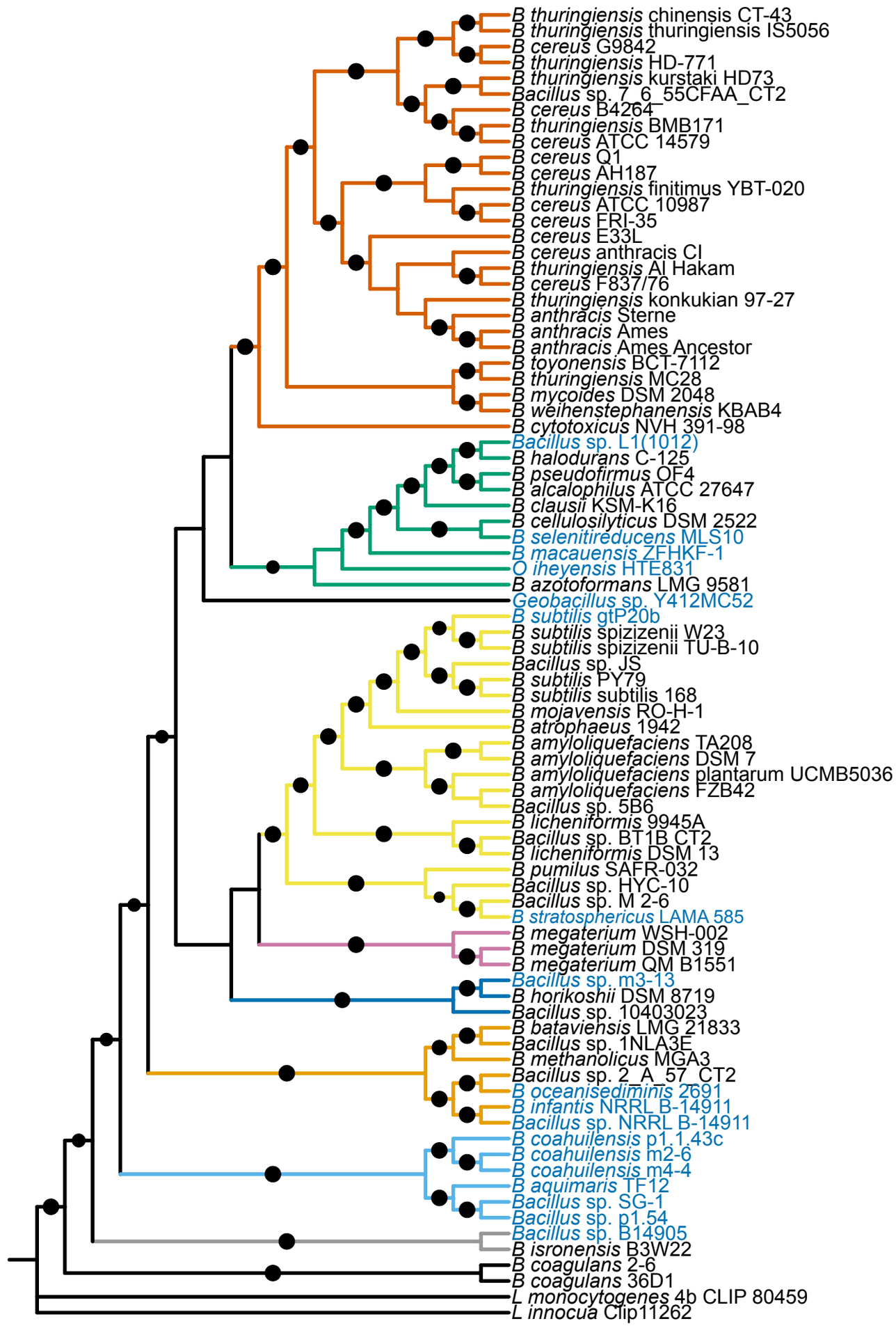

Figura 4: Reconstrucción filogenómica utililizando el genoma núcleo. El árbol se obtuvo empleando el alineamiento concatenado de 196 proteínas que conforman el genoma núcleo. Los nueve grupos claramente definidos se muestran mediante diferentes colores sobre las ramas. Las especies aisladas de ambientes acuáticos se indican en color azul. Los valores de *bootstrap* se muestran como puntos sobre los nodos, sólo los valores ≥ 80 son mostrados. Se observa el gran número de nodos internos con valores de *bootstrap* significativos.

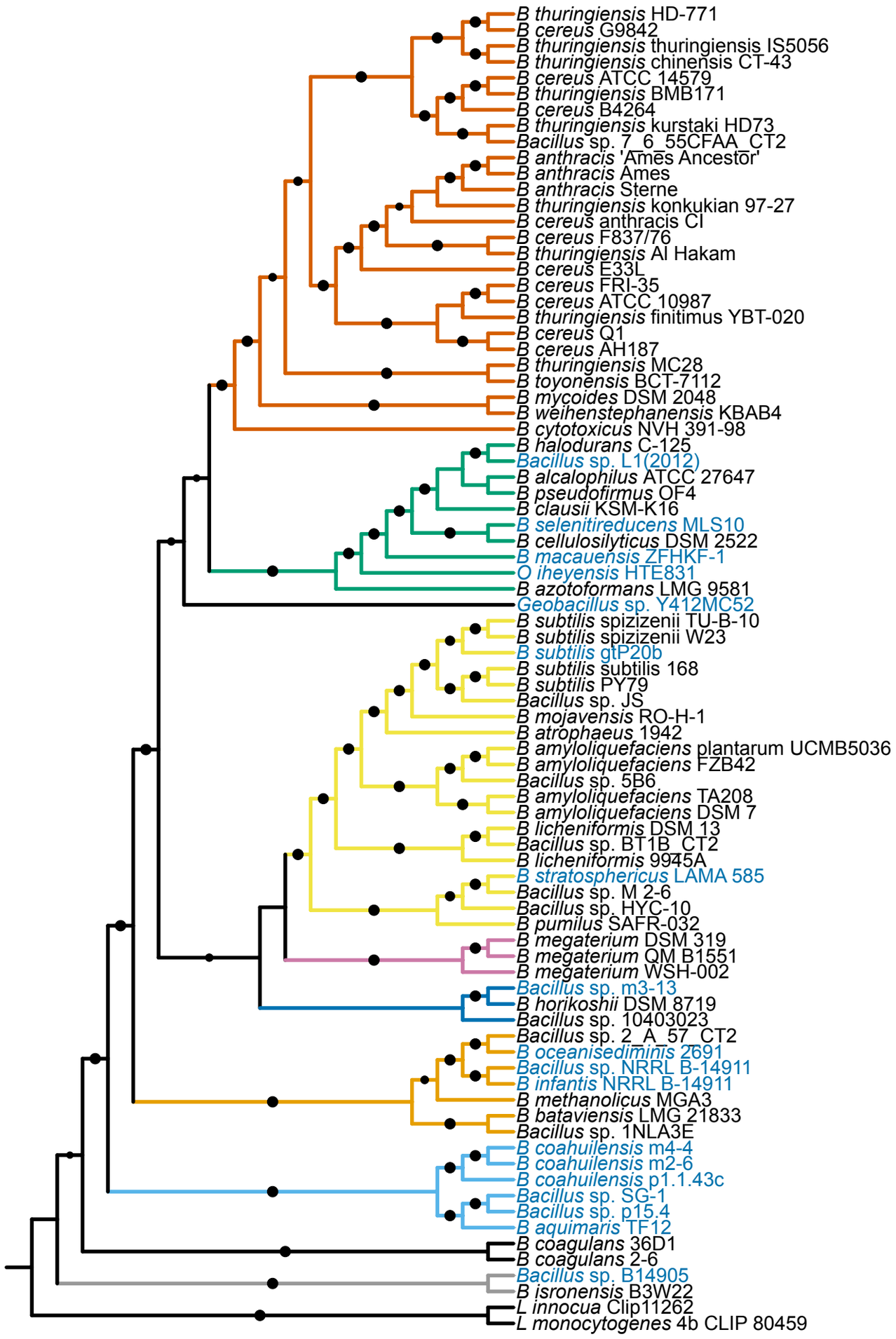

Figura 5: Reconstrucción filogenómica basada en el genoma 70. El árbol se obtuvo empleando el alineamiento concatenado de 437 proteínas conservadas en al menos el 70 % de los genomas. El nombre de las especies aisladas de ambientes acuáticos se indican en color azul. Los colores de las ramas reflejan la posición que el organismo ocupa en el árbol obtenido con el Genoma Núcleo (Figura 4). Los puntos sobre los nodos representan los valores de *bootstrap* ≥ 80. El tamaño de los puntos es proporcional al valor de *bootstrap*.

Reconstrucción filogenética empleando el contenido de genes

El índice de similitud genómica (*Genomic Similarity Score – GSS*) se basa principalmente en la conservación del contenido de genes, incluyendo así, la información del genoma núcleo y el genoma accesorio. El *GSS* fue calculado empleando la mejor coincidencia recíproca (RBH) entre cada par de genomas como se describe en [39, 51] (Materiales y Métodos). La reconstrucción filogenética se llevó a cabo utilizando la matriz de distancia *GSS* y el método de reconstrucción *neigbor-joining*. En general, el árbol obtenido con la distancia *GSS* recupera en gran parte los grupos obtenidos empleando el Genoma Núcleo y el Genoma 70 (figura 6). Este resultado es congruente con la observación que el contenido de genes compartido es determinado cuantitativamente por la filogenia [5]. Sin embargo, del árbol GSS se puede observar que es posible seleccionar un clado que contiene 10 de los 18 *Bacillus* clasificados como acuáticos, mezclado con sólo seis *Bacillus* no acuáticos (Figura 6). Los genomas acuáticos son tan dispersos en la filogenía del Genoma Núcleo y el Genoma 70 que estos diez genomas no pueden ser reunidos en un solo clado sin considerar la mayor parte del árbol. La Figura 7 muestra como los *Bacillus* acuáticos se agrupan mejor en el árbol *GSS*. Dado que la medida *GSS* está basada en los *scores* de BLAST de cada RBH para cada par de genomas, el agrupamiento de los genomas acuáticos sugiere un contenido de genes compartido más alto entre los organismos acuáticos de lo que pudiera esperarse de sus orígenes polifiléticos. Tal contenido de genes compartido podría ser explicado por un proceso de convergencia. En otras palabras, las diferencias entre los grupos obtenidos empleando el índice *GSS* y los clados obtenidos en la reconstrucción filogenómica podría tener un componente homoplásico quizás influido por el ambiente.

La convergencia funcional y/o metabólica ha sido observada en endosimbiontes, donde bacterias de distintos orígenes filogenéticos han mostrado convergencia hacia perfiles funcionales similares. Por ejemplo, el endosimbionte *Xiphinematobacter* del nemátodo daga *Xiphinema americanum*, un ectoparásito migratorio de numerosos cultivos, mostró convergencia evolutiva con los endosimbiontes hallados en insectos que se alimentan de savia, posiblemente debido a la similitud en sus modos de alimentación [83]. También, un alto grado de convergencia metabólica ha sido observado entre bacterias endosimbiontes muy distantemente relacionadas de artrópodos que se alimentan de sangre, y de la sanguijuela *Haementeria officinalis* [84]. La convergencia adaptativa de los genes transferidos horizontalmente también ha sido observada en dos patógenos restringidos a humanos [85].

El grado de resolución de cada uno de los árboles se evaluó mediante el conteo del número de nodos con valores de bootstrap $\geq$ 80, con excepción del árbol obtenido empleando la distancia *GSS*, donde los valores significativos corresponden a los valores obtenidos por máxima verosimilitud ($p \leq 0{,}05$) (Tabla 2).

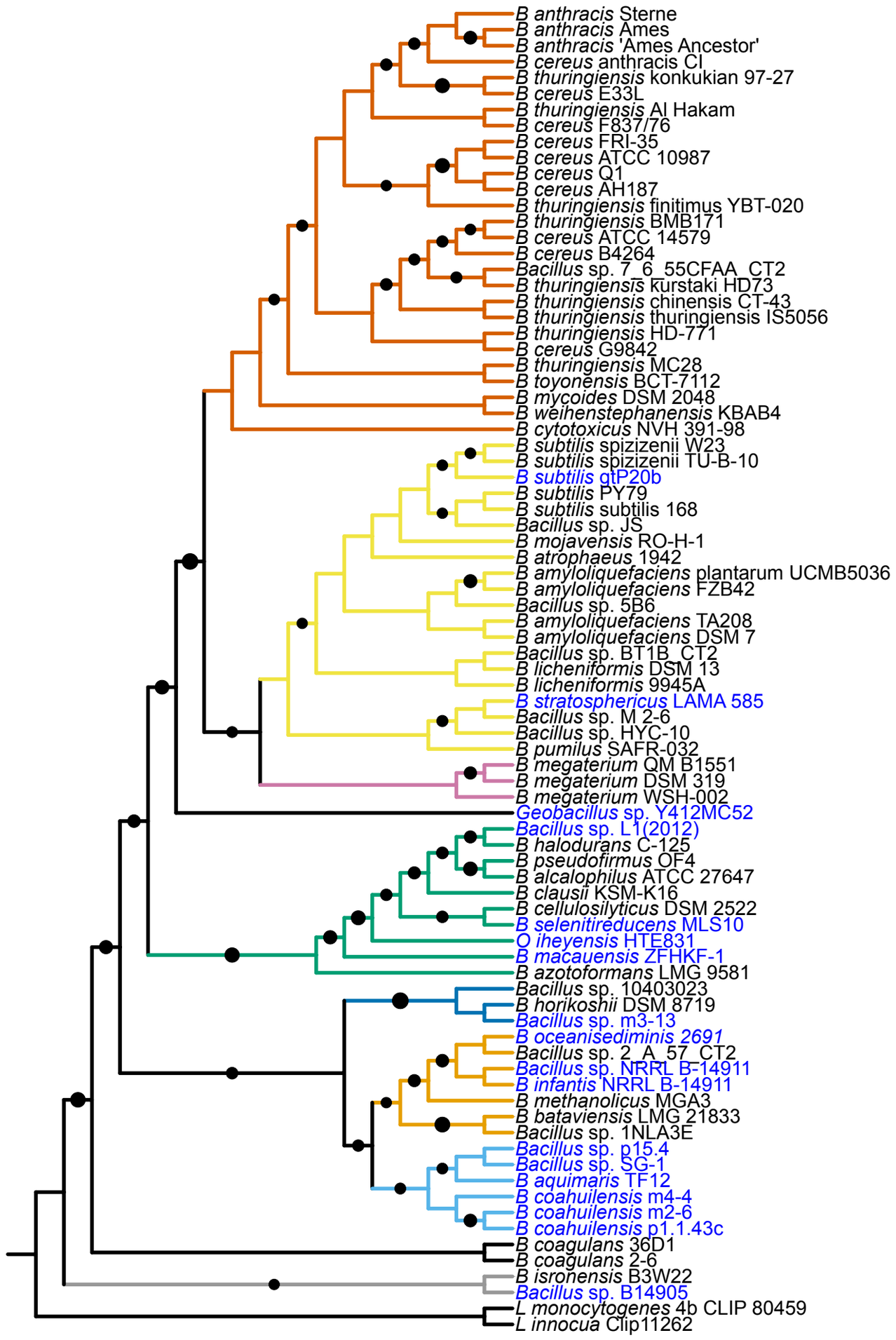

Figura 6: Reconstrucción filogenómica por distancia usando el índice de similitud genómica (*GSS*). Árbol obtenido por *neigbor-joining* utilizando la distancia *GSS*. Los nombres de las especies aisladas de ambientes acuáticos se indican en color azul. Los colores de las ramas reflejan la posición que el organismo ocupa en el árbol obtenido con el genoma núcleo (Figura 4). Los puntos sobre los nodos muestran los valores de apoyo estadístico obtenidos por máxima verosimilitud (Materiales y Métodos). Sólo los $p - val \leq 0{,}05$ son mostrados.

a) GSS Tree **b) Core70 Tree**

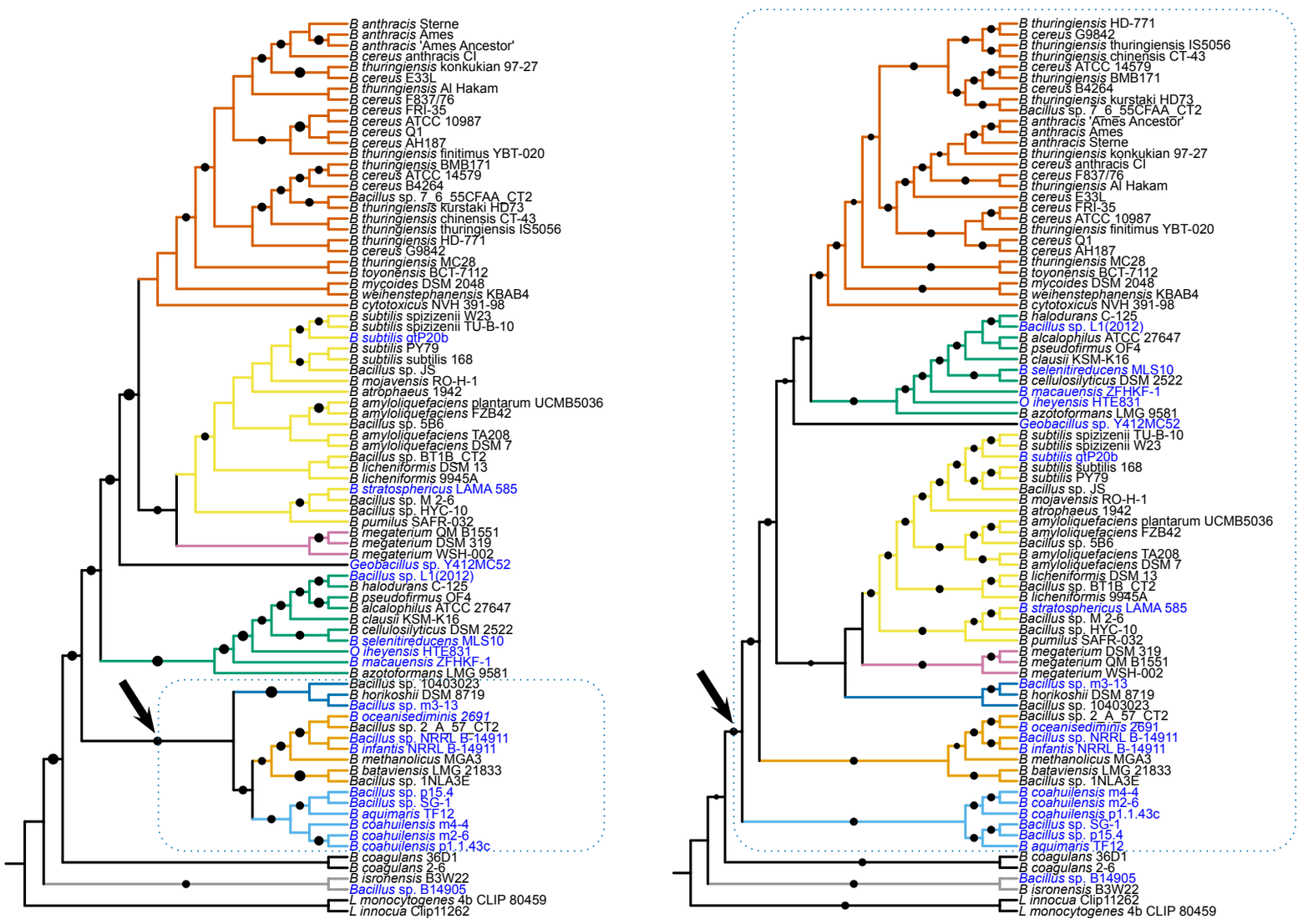

Figura 7: Comparación entre el árbol del Genoma 70 y el árbol basado en la distancia *GSS*. (a) Árbol obtenido con la distancia *GSS* y el método *neighbor-joining*. (b) Árbol basado en el Genoma 70. Árbol basado en el alineamiento concatenado 437 genes presentes en al menos el 70 % de las especies. En ambos árboles, los nombres de las especies en azul corresponden a *Bacillus* aislados de ambientes acuáticos. Los puntos en las ramas indican el apoyo basado en $p - val \leq 0{,}05$ (árbol *GSS*) y los valores de *bootstrap* ≥ 80 (árbol Genoma 70). Los nodos indicados por flechas y los rectángulos punteados muestran que los *Bacillus* acuáticos agrupan mejor en el árbol *GSS*.

Los resultados de este análisis muestran que los árboles obtenidos usando el Genoma 70 y el Genoma Núcleo presentan topologías muy similares y el mayor número de nodos con valores significativos; no obstante, la diferencia entre ellos es de sólo 5 nodos. Es de hacer notar que el árbol *GSS* ocupa un lugar intermedio entre los árboles cuando se emplean ambas medidas (Tablas 2 y 3) .

Tabla 3: Diferencia Simétrica entre árboles

	16S rRNA	Marcadores AMPHORA	Genoma Núcleo	Genoma 70	Distancia *GSS*
16S rRNA	0	-	-	-	-
Marcadores AMPHORA	112	0	-	-	-
Genoma Núcleo	102	64	0	-	-
Núcleo 70	100	70	10	0	-
Distancia *GSS*	102	72	32	24	0

Análisis del contenido funcional

El agrupamiento obtenido mediante la distancia *GSS* sugirió un contenido de genes compartido más alto entre los organismos acuáticos de lo que pudiera esperarse de sus orígenes polifiléticos (provenientes de distintos linajes evolutivos). Para probar si el contenido de genes era el responsable del mejor agrupamiento de los *Bacillus* acuáticos observado en el árbol obtenido con la distancia *GSS*, decidimos llevar a cabo el análisis de agrupamiento jerárquico empleando únicamente el contenido de genes representados por familias de proteínas anotadas funcionalmente. Así, el contenido funcional definido como el conjunto de funciones asignadas a través de cualquiera de las categorías COG (Clusters of Orthologous Groups) [55] y/o Figfams [56] fue determinado para cada genoma. La comparación del contenido funcional se realizó empleando la distancia de Jaccard (Jd) como medida de similitud, evaluando diferentes métodos de agrupamiento jerárquico. La calidad del agrupamiento fue evaluado empleando los coeficientes de aglomeración. En esencia, entre más cercano el valor del coeficiente de aglomeración a la unidad mejor es la calidad del agrupamiento (ver Materiales y Métodos). La tabla 4 muestra los valores del coeficiente de aglomeración obtenidos con la distancia Jd basadas en COGs o Figfams y diferentes métodos de agrupamiento. Para los análisis basados en COG y Figfam, el mejor dendrograma se obtuvo usando el método de agrupamiento Ward (COG $coef.agl. = 0{,}91$; Figfam $coef.agl. = 0{,}92$, respectivamente).

Tabla 4: Coeficientes de aglomeración obtenidos empleando la distancia de Jaccard (Jd) y diferentes métodos de agrupamiento.

Método de Agrupamiento	Coeficiente de Aglomeración [a]	
	COG	Figfam
average	0.73	0.69
single	0.71	0.67
complete	0.77	0.71
ward	**0.91**	**0.92**
weighted	0.75	0.69

[a] En negritas se indica el método de agrupamiento con el mejor coeficiente de aglomeración.

Los dendrogramas obtenidos empleando las categorías COG y Figfam arrojaron tres grupos principales, dos de los cuales fueron similares a los observados en el árbol filogenómico y que corresponden a los grupos de *B. cereus* y *B. subtilis*. El tercer grupo comprende al resto de los *Bacillus* (Figuras 8 y 9). Para establecer el número de grupos (k) en las jerárquias seleccionadas, los dendrogramas fueron cortados a diferentes umbrales y evaluados por sus valores de *Silhouette* [60]. Los valores de *Silhouette* reflejan la distancia de cada elemento de un grupo a los otros elementos del mismo grupo, comparadas a las distancias en contra de los miembros de los otros grupos.

Para el agrupamiento jerárquico basado en COGs, los grupos de *B. cereus* y *B. subtilis* mostraron valores positivos de *Silhouette*, mientras que el tercer grupo presentó valores bajos de *Silhouette* e incluso valores negativos (Figura 8). Para establecer el número de grupos en la jerarquía, iniciamos calculando los valores de Silhouette para los tres primeros grupos ($k = 3$) tomando como límite los grupos claramente definidos *B. cereus* y *B. subtilis*. Los análisis alcanzaron 10 grupos ($k = 10$) antes de la disrupción del grupo limitante *B. subtilis* (Figura 8). Interesantemente, uno de los 10 grupos incluyó 7 *Bacillus* acuáticos que en la reconstrucción filogenética fueron ubicados en clados diferentes (Figura 4).

Por otra parte, el agrupamiento jerárquico usando la distancia *Jd* basada en Figfams también resultó en dos grupos con buenos valores *Silhouette*, mientras que el tercer grupo presentó valores positivos pero bajos de *Silhouette* (Figura 9). Cuando cortamos el dendrograma para maximizar el valor de *Silhouette*, el resultado arrojó seis grupos antes de la disrupción de los dos grupos principales iniciales. En este punto, 10 *Bacillus* acuáticos fueron agrupados formando un grupo con sólo dos *Bacillus* no acuáticos. Este resultado es notable dado que los *Bacillus* acuáticos parecen tener orígenes polifiléticos (Figura 4) y algunos de ellos permanecieron aún separados en el agrupamiento jerárquico obtenido con la distancia *Jd* utilizando los COGs (Figura 8). Similar a lo observado en la reconstrucción utilizando la distancia *GSS*, el análisis de agrupamiento jerárquico empleando sólo el contenido funcional por medio de COGs y/o Figfams, sugiere también que los *Bacillus* acuáticos comparten entre ellos una mayor proporción de su contenido funcional.

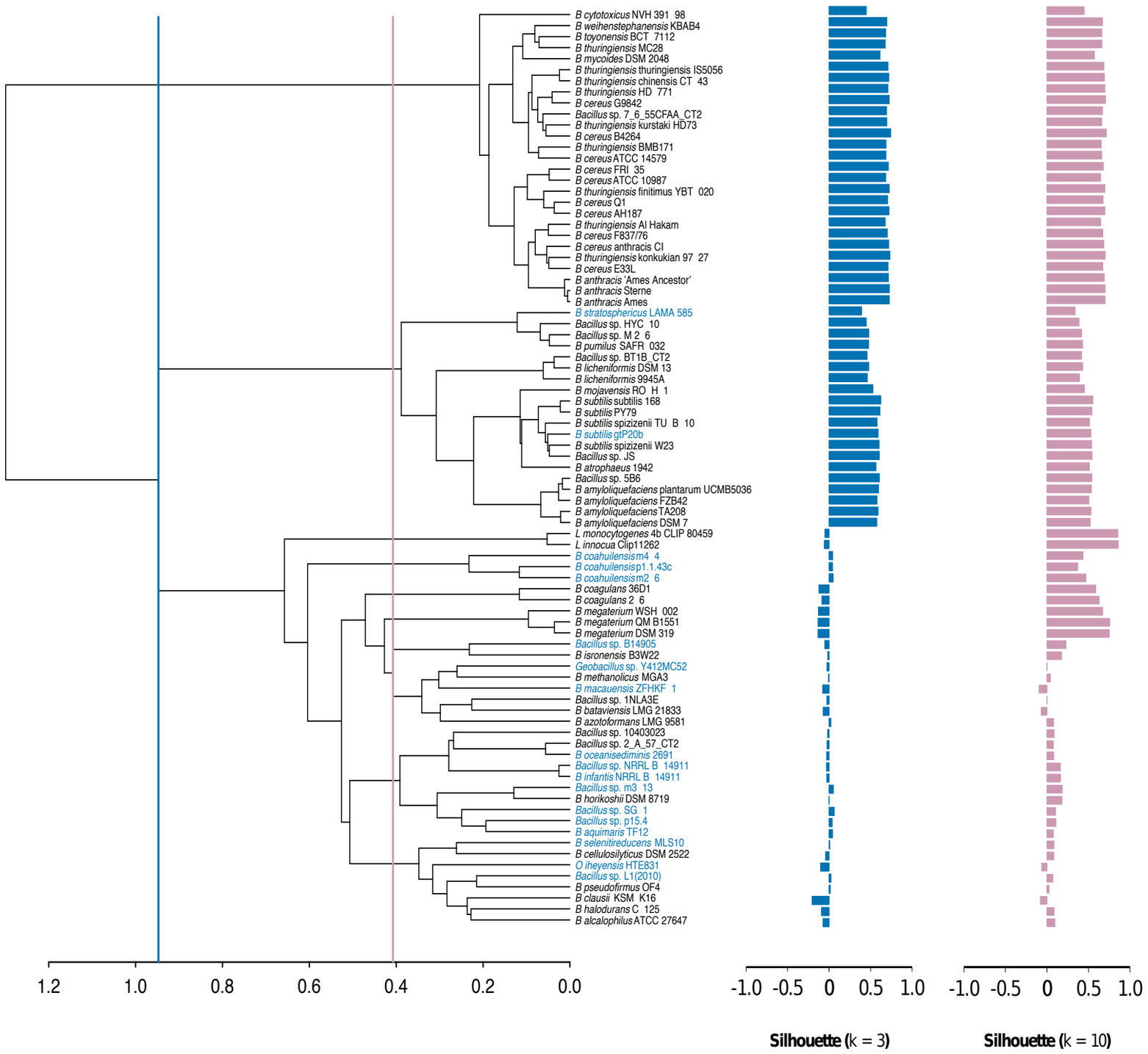

Figura 8: Agrupamiento jerárquico empleando la distancia de *Jd* basada en COGs. El agrupamiento jerárquico se obtuvo usando el método de agrupamiento Ward. La línea rosa en el dendograma señala la altura a la que el dendrograma se cortó para obtener los diferentes grupos ($k = 9$). Las barras en color azul y rosa indican los valores del índice Silhouette a $k = 3$ y $k = 9$, respectivamente. El nombre de las especies aisladas de ambientes acuáticos se muestra en color azul claro.

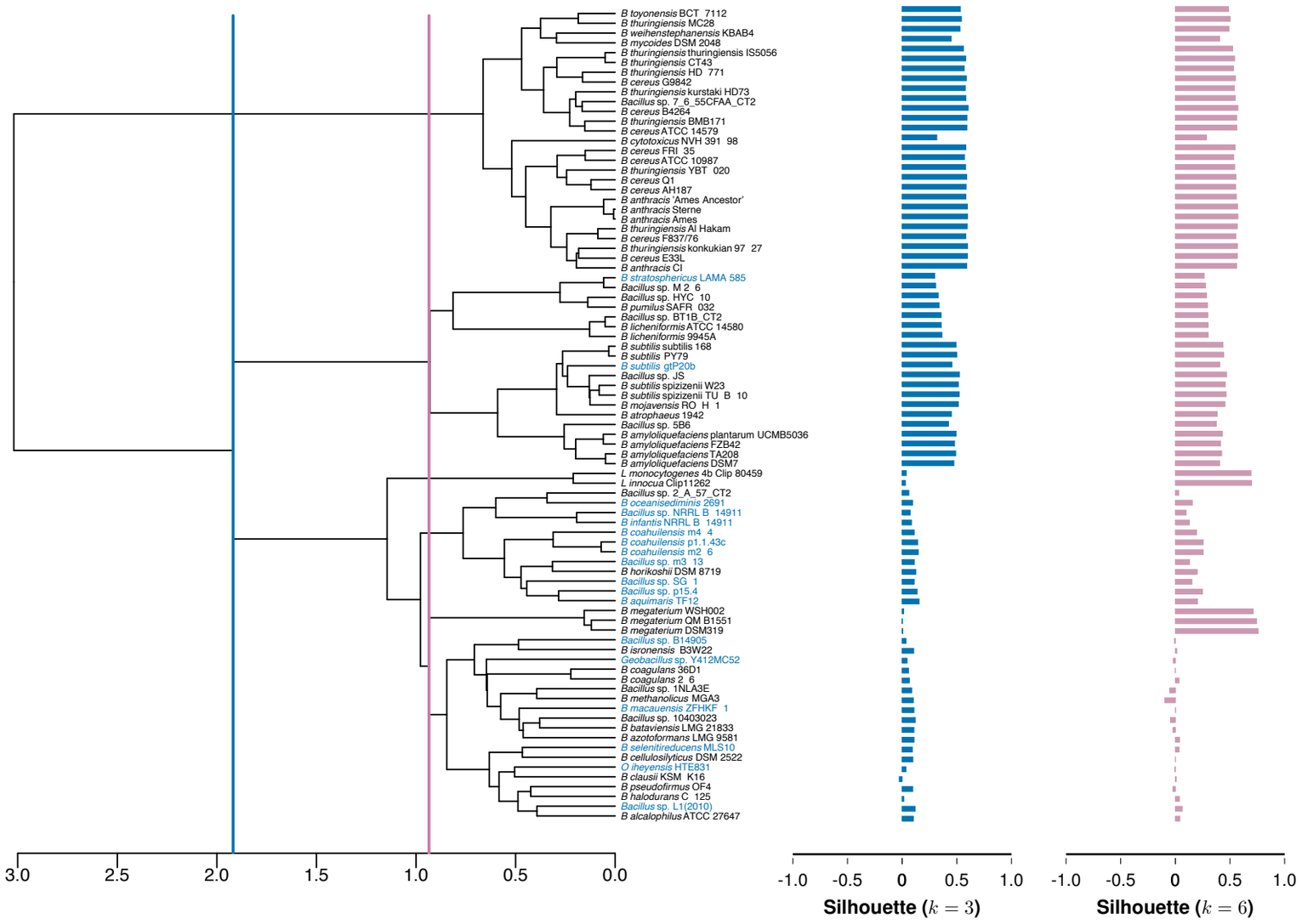

Figura 9: Agrupamiento jerárquico empleando la distancia de Jaccard (Jd) basada en Figfams. El dendograma se obtuvo mediante el método de agrupamiento Ward. Las barras en color azul y rosa indican el valor del índice de *Silhouette* a $k = 3$ y $k = 6$, respectivamente. Las líneas en color azul y rosa en el dendograma señalan la altura de corte a $k = 3$ y $k = 6$, respectivamente. Los nombres de las especies aisladas de ambientes acuáticos se indican en color azul claro.

El agrupamiento del contenido funcional podría explicarse por algunas categorías funcionales

Con el fin de investigar qué elementos del contenido funcional podrían explicar los agrupamientos obtenidos mediante el agrupamiento jerárquico, el análisis de componentes principales (PCA – Principal Components Analysis) se llevó a cabo empleando la matriz de frecuencias de las categorías funcionales COGs y Figfams.

En el caso de los COGs, la gráfica de dispersión de puntos de los dos primeros componentes muestra que el primer componente principal separa al grupo de los *Bacillus cereus* del resto de los grupos (Figura 10). El primer componente (PC1), se describe por las categorías: [S] *Function unknown* (-0.28706416), [E] *Amino acid transport and metabolism* (-0.28657268), [R] *General function prediction only* (-0.28601708), [P] *Inorganic ion transport and metabolism* (-0.27445301), [K] *Transcription* (-0.26999934), [J] *Translation, ribosomal structure and biogenesis* (-0.25211402), [F] *Nucleotide transport and metabolism* (-0.25154681).

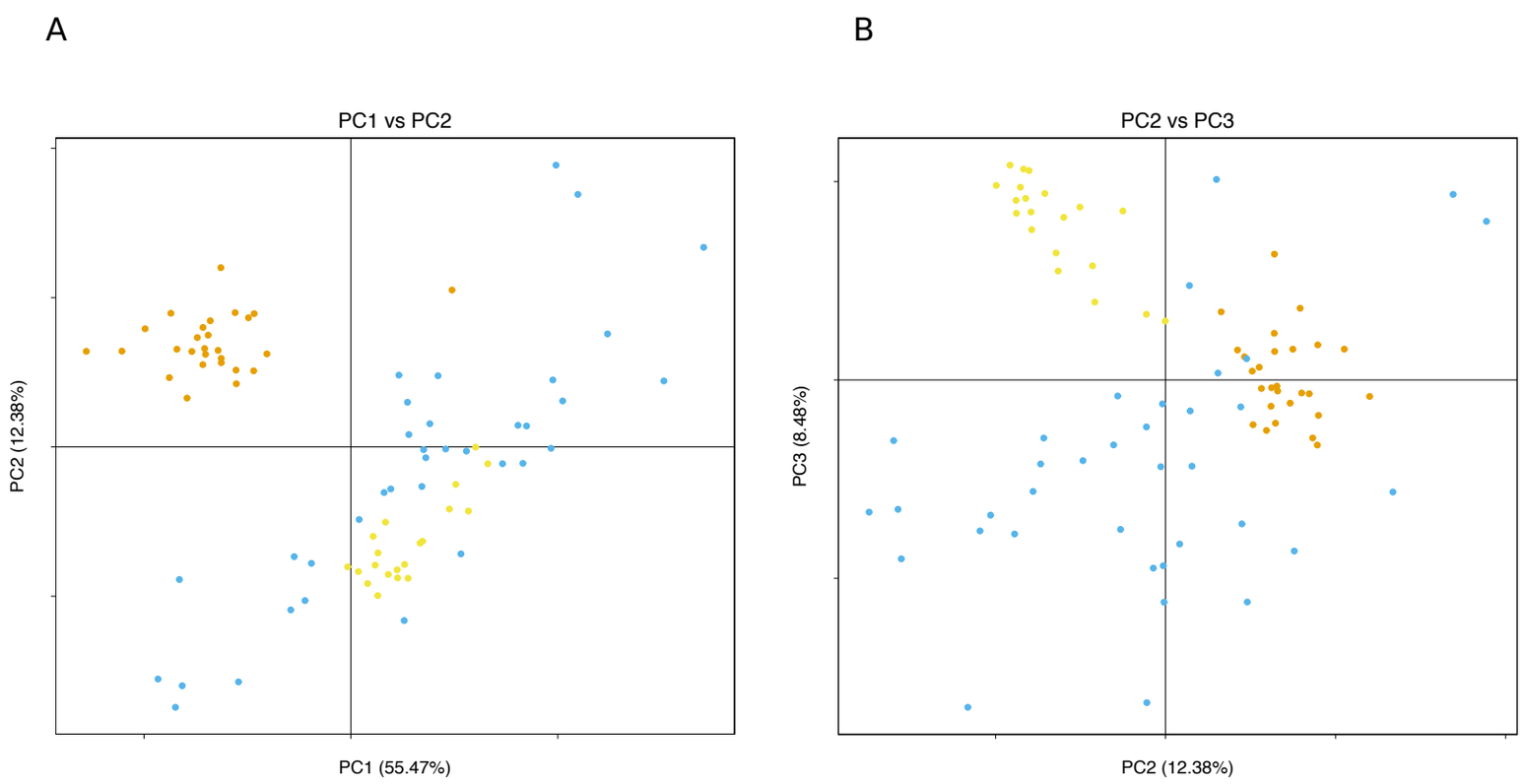

Figura 10: Análisis de Componentes Principales (PCA) coloreado por grupos. Gráfica de dispersión de puntos del PC1 vs PC2 (A) y del PC2 vs PC3 (B). Los puntos representan los miembros de los tres principales grupos obtenidos del agrupamiento jerárquico empleando las categorías COG. Los círculos naranja representan los miembros del grupo *B. cereus*, en amarillo los miembros del grupo *B. subtilis* y en azul los miembros del grupo de *Bacillus* acuáticos.

El segundo componente (PC2) se caracteriza por las categorías: [N] *Cell motility* (-0.457494066), [O] *Posttranslational modification, protein turnover, chaperones* (-0.422042059), [I] *Lipid transport and metabolism* (-0.315095407), [V] *Defense mechanisms* (0.300243449), [F] *Nucleotide transport and metabolism* (0.275410757), [G] *Carbohydrate transport and metabolism* (-0.270679026), [C] *Energy production and conversion* (-0.252397088) (Tabla 5). Lo anterior también se ve reflejado en el mapa de calor empleando las frecuencias de COGs clasificados por categoría funcional (Figura 11). Sin embargo, el grupo de los *Bacillus* acuáticos no parece estar caracterizado por alguna categoría funcional.

Tabla 5: Cargas por variables (categoriás COG)

Componente Principal 1		Componente Principal 2	
Carga	Categorías COG	Carga	Categorías COG
-0.287064	[S] Function unknown	-0.457494	[N] Cell motility
-0.286572	[E] Amino acid transport and metabolism	-0.422042	[O] Posttranslational modification, protein turnover, chaperones
-0.286017	[R] General function prediction only	-0.315095	[I] Lipid transport and metabolism
-0.274453	[P] Inorganic ion transport and metabolism	0.300243	[V] Defense mechanisms
-0.269999	[K] Transcription	0.275410	[F] Nucleotide transport and metabolism
-0.252114	[J] Translation, ribosomal structure and biogenesis	-0.270679	[G] Carbohydrate transport and metabolism
-0.251546	[F] Nucleotide transport and metabolism	-0.252397	[C] Energy production and conversion

En el caso de las categorías Figfam, una vez más el primer componente parece definir al grupo *Bacillus cereus* (Figura 12), este componente se caracteriza por las categorías: [V12] Miscellaneous (0.26121880); [V1] Amino Acids and Derivatives (0.24956872); [V20] Protein Metabolism (0.24217435); [V5] Cofactors, Vitamins, Prosthetic Groups, Pigments (0.23926786); [V2] Regulation and Cell signaling (0.23676526); [V15] Nucleosides and Nucleotides (0.23502512); [V17] Phosphorus Metabolism (0.23310239).

El segundo componente se define por las categorías: [V11] Metabolism of Aromatic Compounds (-0.373947508); [V8] Fatty Acids, Lipids, and Isoprenoids (-

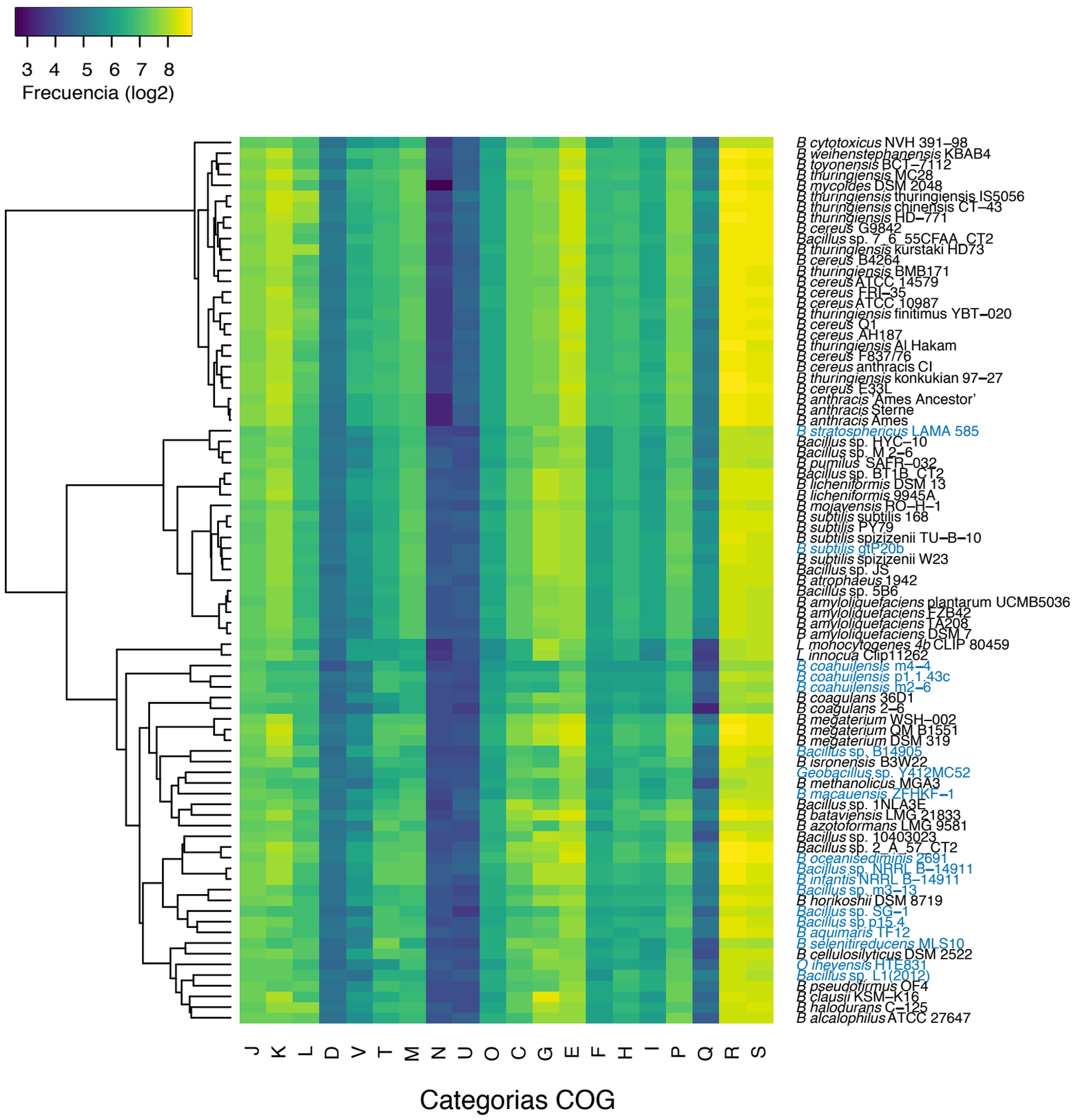

Figura 11: Mapa de calor obtenido usando las frecuencias de los COG por categoría funcional. La frecuencia de los COG por categoría funcional se muestra como el logaritmo de base 2. Categorías COG: [J] Translation, ribosomal structure and biogenesis; [K] Transcription; [L] Replication, recombination and repair; [D] Cell cycle control, cell division, chromosome partitioning; [V] Defense mechanisms; [T] Signal transduction mechanisms; [M] Cell wall/membrane/envelope biogenesis; [N] Cell motility; [U] Intracellular trafficking, secretion, and vesicular transport; [O] Posttranslational modification, protein turnover, chaperones; [C] Energy production and conversion; [G] Carbohydrate transport and metabolism; [E] Amino acid transport and metabolism; [F] Nucleotide transport and metabolism; [H] Coenzyme transport and metabolism; [I] Lipid transport and metabolism; [P] Inorganic ion transport and metabolism; [Q] Secondary metabolites biosynthesis, transport and catabolism; [R] General function prediction only; [S] Function unknown.

0.331302949); [V26] Sulfur Metabolism (-0.328666220); [V18] Photosynthesis (-0.277449839); [V16] Phages, Prophages, Transposable elements, Plasmids (0.268678769). A diferencia del análsis de Componentes Principales basado en las categorías COG, las categorías Figfam sólo definen al grupo de los *Bacillus cereus*. La figura 13 muestra el mapa de calor obtenido empleando las frecuencias de las categorías Figfam.

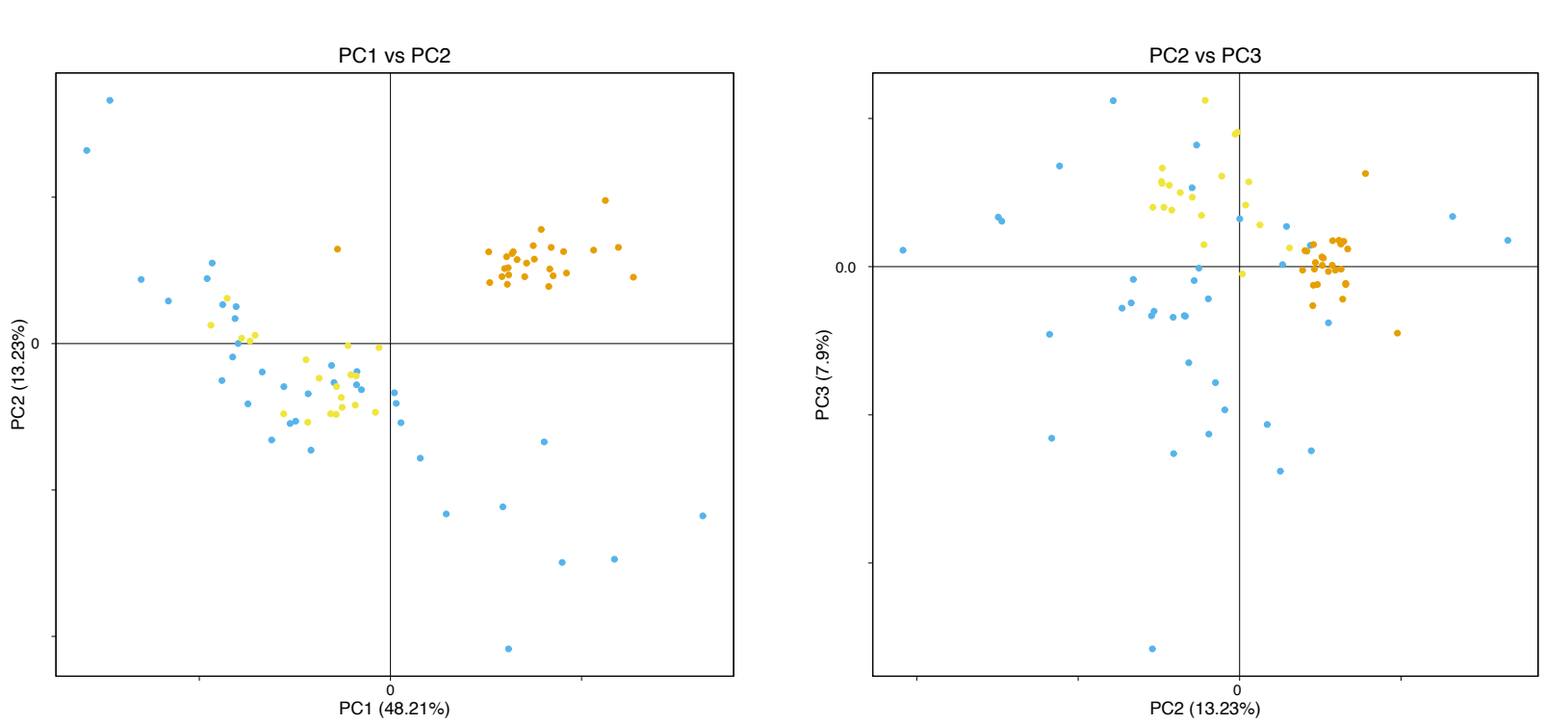

Figura 12: Análisis de Componentes Principales (PCA) empleando las categorías Figfam coloreado por grupos. Gráfica de dispersión de puntos del PC1 vs PC2 (A) y del PC2 vs PC3 (B). Los puntos representan los miembros de los tres principales grupos obtenidos del agrupamiento jerárquico empleando las categorías Figfam. Los círculos naranja representan los miembros del grupo *B. cereus*, en amarillo los miembros del grupo *B. subtilis* y en azul los miembros del grupo de *Bacillus* acuáticos.

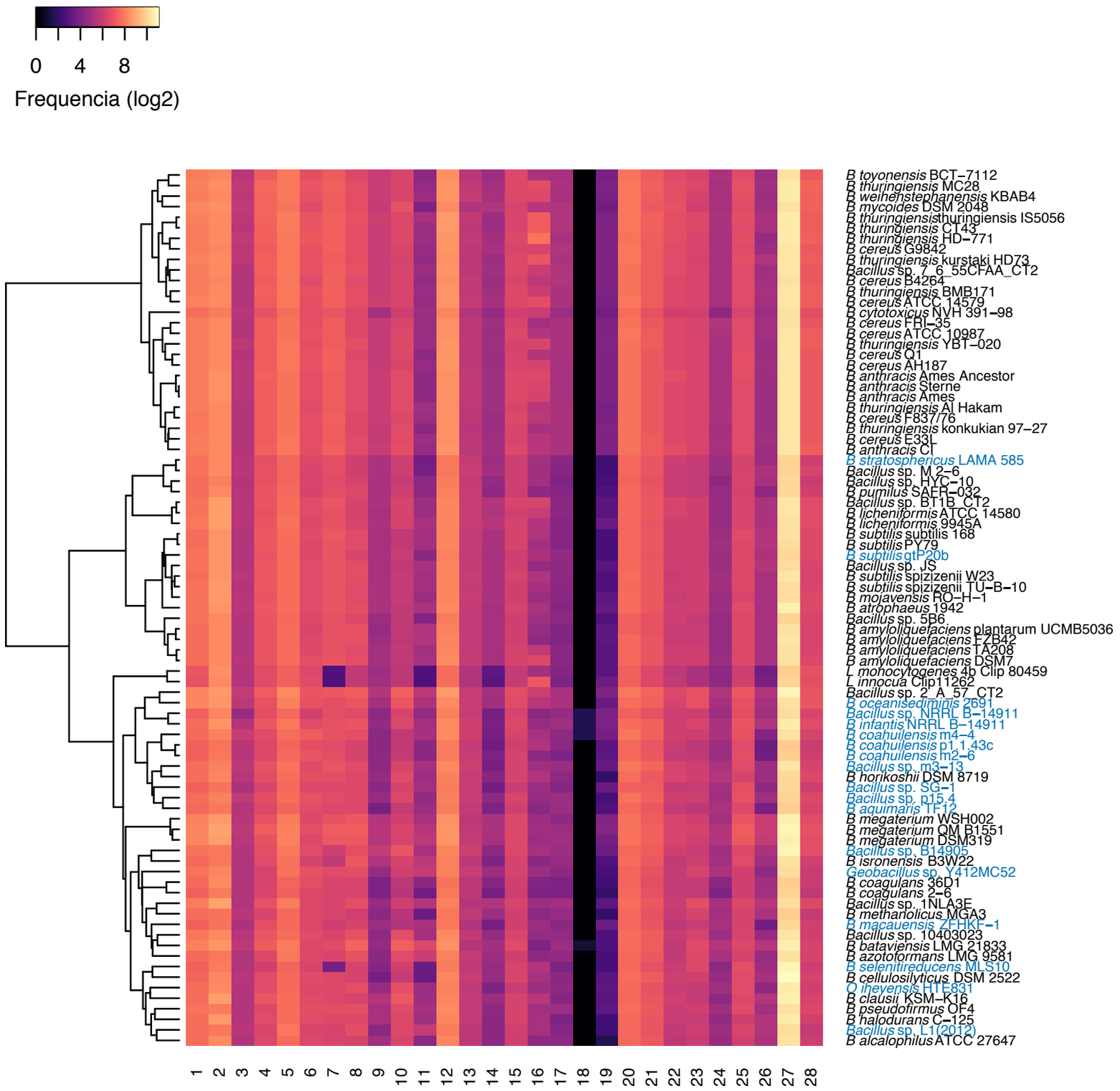

Figura 13: Mapa de calor basado en la frecuencia de Figfams clasificados por categoría funcional. El nombre de las especies aisladas de ambientes acuáticos se indica en color azul. Categorías funcionales: 1) Amino Acids and Derivatives; 2) Carbohydrates; 3) Cell Division and Cell Cycle; 4) Cell Wall and Capsule; 5) Cofactors, Vitamins, Prosthetic Groups, Pigments; 6) DNA Metabolism; 7) Dormancy and Sporulation; 8) Fatty Acids, Lipids, and Isoprenoids; 9) Iron acquisition and metabolism; 10) Membrane Transport; 11) Metabolism of Aromatic Compounds; 12) Miscellaneous; 13) Motility and Chemotaxis; 14) Nitrogen Metabolism; 15) Nucleosides and Nucleotides; 16) Phages, Prophages, Transposable elements, Plasmids; 17) Phosphorus Metabolism; 18) Photosynthesis; 19) Potassium metabolism; 20) Protein Metabolism; 21) RNA Metabolism; 22) Regulation and Cell signaling; 23) Respiration; 24) Secondary Metabolism; 25) Stress Response; 26) Sulfur Metabolism; 27) Unknown; 28) Virulence, Disease and Defense.

Asociación ambiente natural - Contenido de genes

Para investigar si existe una asociación entre el ambiente natural de los organismo bajo estudio y los agrupamientos obtenidos empleando COGs y/o Figfams, los valores P (*P-val*) fueron calculados basados en la distribución hipergeométrica (Métodos). Los resultados sugieren una asociación significativa entre el ambiente natural y dos de los tres grupos principales obtenidos con el agrupamiento jerárquico, con excepción del grupo *B. subtilis* cada grupo se asocia con un ambiente natural principal.

Para los COGs y Figfams se observó una asociación entre el ambiente natural *Facultativo* y el agrupamiento formado por el grupo *Bacillus cereus sensu lato*. Las especies comprendidas en este grupo crecen saprofíticamente bajo condiciones ricas en nutrientes, y se ha propuesto que algunos miembros de este grupo pueden establecer una relación simbiótica con hospederos invertebrados desarrollando ocasionalmente un estilo de vida patogénico [86]. Otra asociación significativa fue hallada entre el ambiente natural *Acuáticos* y el tercer grupo el cual comprende siete de los nueve clados observados en la reconstrucción filogenética (Figura 4). Es importante notar que 15 de los 18 posibles *Bacillus* acuáticos fueron ubicados en este grupo. Dentro de este grupo a su vez, se observa un subgrupo de nueve y 11 *Bacillus* acuáticos dependiendo de sí los COGs o los Figfams son usados (Figuras 14 y 15).

Dado que los organismos dependen de sus capacidades metabólicas y regulatorias para sobrevivir en ambientes específicos, es razonable esperar una relación entre el contenido funcional de genes y el ambiente. Algunos estudios han mostrado una relación entre el metabolismo de los organismos y el ambiente que ellos ocupan [87, 88], mientras que otros autores han sugerido la existencia de genes específicos al ambiente involucrados en vías metabólicas que son hipotetizados de ser responsables para la adaptación bacterial [33, 34]. Recientemente, un estudio genómico comparativo de cepas de *B. amyloliquefaciens* y *B. subtilis* asociados a plantas, en contra de cepas no asociadas a plantas, sugieren que las diferencias en sus genomas ocurren durante su adaptación a diferentes habitats [89]. La tabla en el Apéndice D muestra los ambientes naturales en los cuales los *Bacillus* empleados en este estudio han sido clasificados.

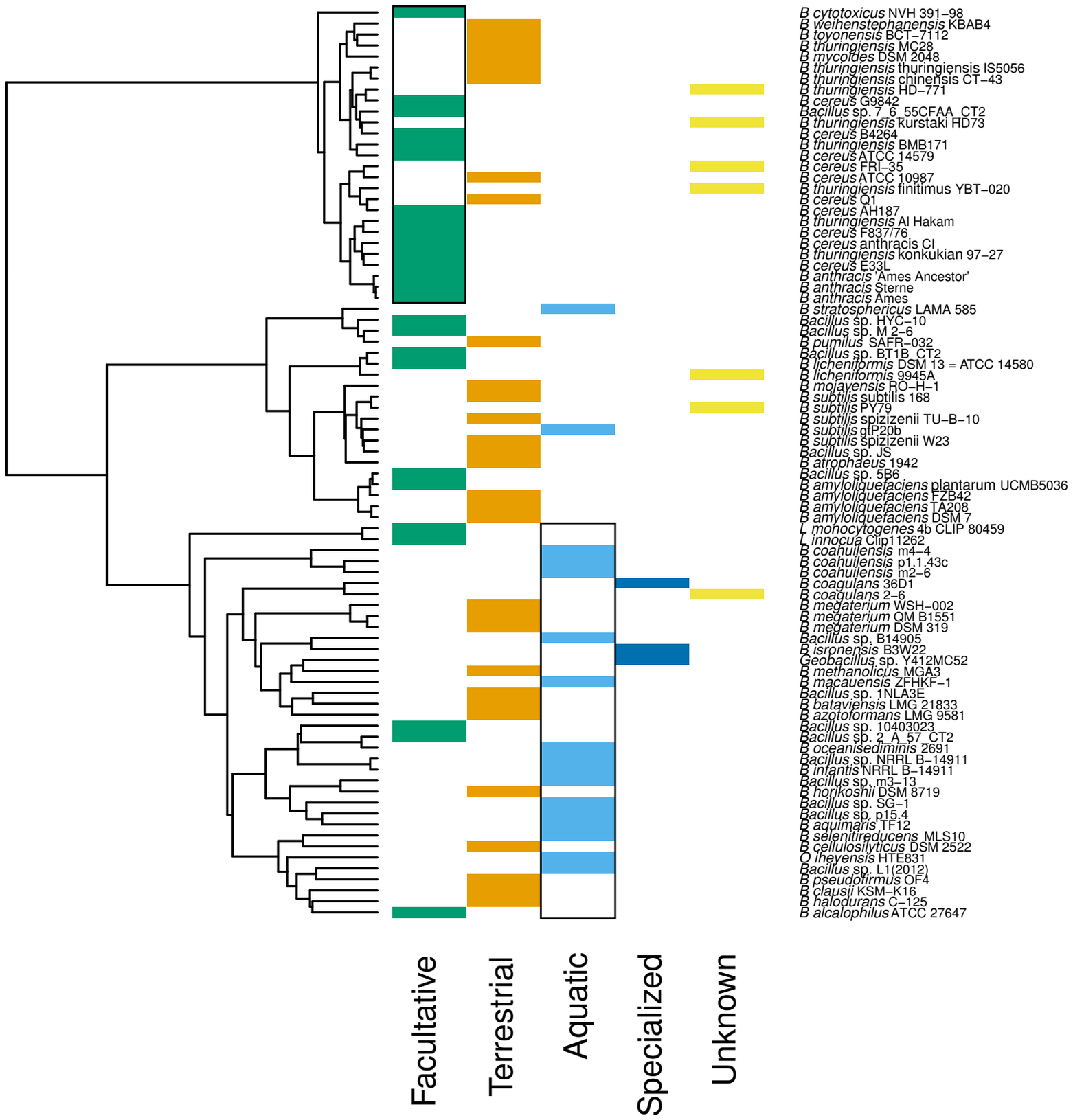

Figura 14: Asociación entre el ambiente natural y los grupos obtenidos del agrupamiento jerárquico utilizando COGs. Los valores P fueron calculados basados en la distribución hipergeométrica empleando los grupos obtenidos con el agrupamiento jerárquico con COGs. Las asociaciones significativas son indicadas por los recuadros negros. Facultativo P-val corregido $P = 1{,}7 \times 10^{-03}$; Acuático P-val corregido $P = 5{,}6 \times 10^{-04}$.

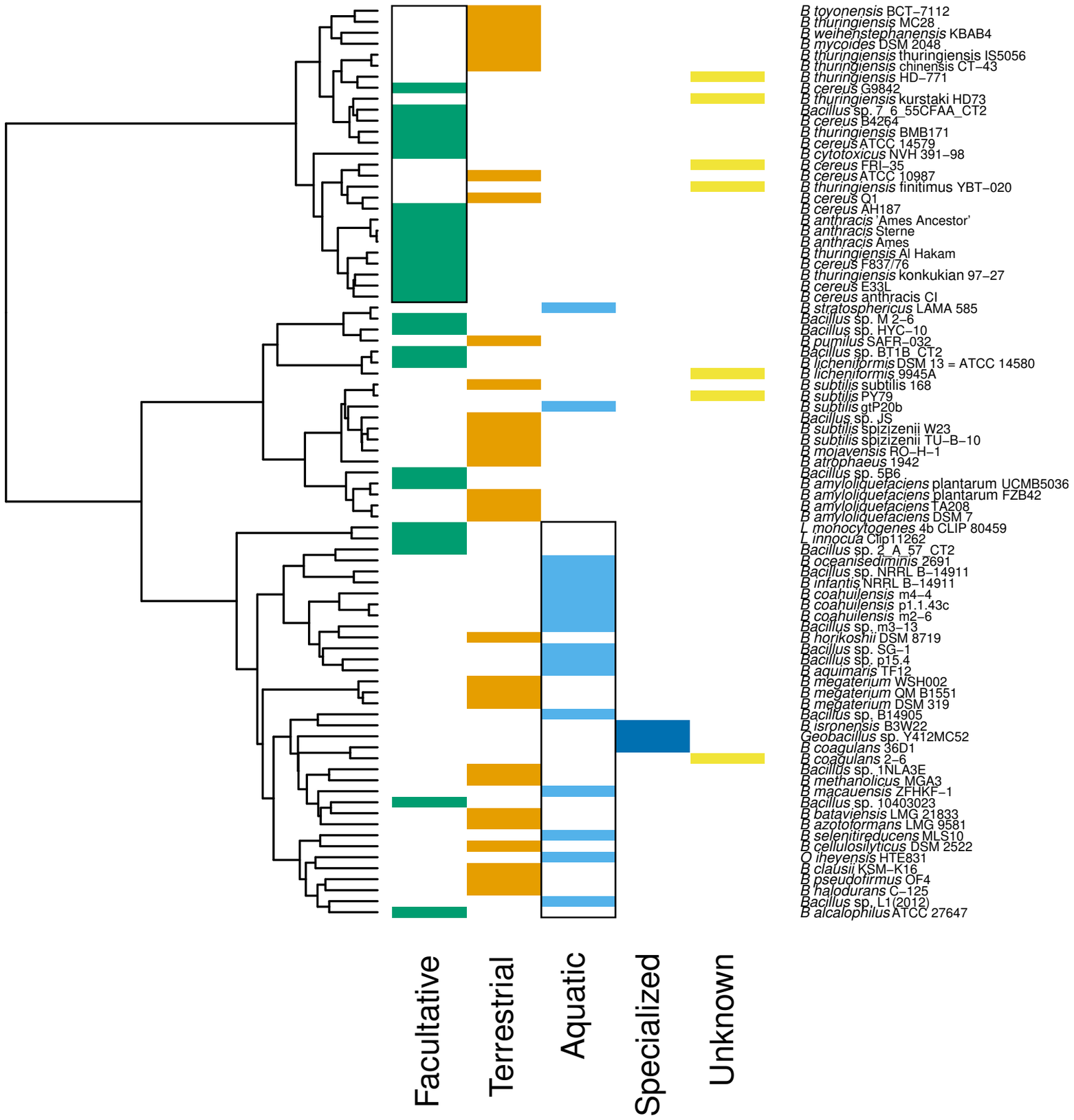

Figura 15: Asociación entre el ambiente natural y los grupos obtenidos del agrupamiento jerárquico utilizando Figfams. Los valores P fueron calculados basados en la distribución hipergeométrica empleando los grupos obtenidos con el agrupamiento jerárquico con Figfams. Las asociaciones significativas son indicadas por los recuadros negros. Facultativo P-val corregido $P = 1{,}8 \times 10^{-03}$; Acuático P-val corregido $P = 9{,}0 \times 10^{-05}$.

Funciones ambiente específicas

Con el objeto de investigar la existencia de funciones específicas en los grupos obtenidos del agrupamiento jerárquico las funciones núcleo de cada grupo fueron determinadas y sus funciones potenciales comparadas con las observadas en los otros grupos. Para este análisis los tres principales grupos del dendrograma obtenido con la distancia de Jaccard (Jd) y la matriz con la frecuencia de COGs fueron empleados (Materiales y Métodos). Los resultados muestran que el grupo de *Bacillus cereus* contiene el mayor número de posibles COGs adaptativos con 196, seguido por el grupo de *Bacillus subtilis* con 106 y el grupo de los *Bacillus* Acuáticos con 29 (Apénd. C, Apénd. B y tabla 6).

Dentro del grupo de los *Bacillus cereus* los COGs específicos pertenecen a 20 categorías funcionales, de las cuales 7 comprenden el 75 % de los COGs específicos. Las categorías con más COG específicos fueron las categorías [R] *General function prediction only* y [S] *Function unknown* con 36 y 54 COGs, respectivamente, seguidas por las categorías: [E] *Amino acid transport and metabolism*, [P] *Inorganic ion transport and metabolism*, [L] *Replication, recombination and repair*, [K] *Transcription* y [M] Cell wall/membrane/envelope biogenesis. Se ha sugerido que la abundancia de enzimas proteolíticas y de transportadores de péptidos y aminoácidos en *Bacillus cereus* y *Bacillus anthracis* indicaria que proteínas, péptidos y aminoácidos podrían ser un nutriente preferido por estas bacterias [23]. La presencia de COGs específicos pertenecientes a la categoría [V] *Defense mechanisms* en este grupo podría estar relacionada con los *Bacillus* patogénicos incluidos en este grupo, ya que esta categoría involucra principalmente genes de resistencia a antibióticos (Apéndice C).

Por otra parte, el grupo *Bacillus subtilis* presentó 106 COGs específicos distribuidos en 16 categorías siendo la categoría [G] *Carbohydrate transport and metabolism* la más representada con 24 COGs específicos, las categorías [S] *Function unknown*, [E] *Amino acid transport and metabolism* y [M] *Cell wall/membrane/envelope biogenesis* con 12 COGs cada una, y la categoría [R] *General function prediction only* con 10 COGs. El hecho de que la categoría [G] haya sido la más representada podría reflejar la especialización del metabolismo de carbohidratos en este grupo. La presencia de múltiples genes involucrados en el metabolismo de carbohidratos en *Bacillus subtilis* ha sido interpretada como un rasgo característico de las bacterias de suelo, donde el material derivado de plantas es la principal fuente de nutrientes [23]. Es de llamar la atención el hecho de que tanto el grupo *Bacillus cereus* y el de *Bacillus subtilis* presentan COGs específicos en la categoría [M] *Cell wall/membrane/envelope biogenesis*, dado que los COGs en el grupo de *Bacillus subtilis* corresponden a genes involucrados en la formación de la esdoespora. Estas proteínas podrían ser usadas como marcadores de las especies que forman el grupo *Bacillus subtilis*. La pérdida de conservación en los genes que participan en el pro-

Tabla 6: Posibles COGs adaptativos hallados en el grupo *Bacillus* acuáticos.

EP	COG0601	ABC-type dipeptideoligopeptidenickel transport systems, permease components
E	COG0624	Acetylornithine deacetylaseSuccinyl-diaminopimelate desuccinylase and related deacylases
E	COG0747	ABC-type dipeptide transport system, periplasmic component
E	COG1703	Putative periplasmic protein kinase ArgK and related GTPases of G3E family
G	COG0395	ABC-type sugar transport system, permease component
G	COG1175	ABC-type sugar transport systems, permease components
G	COG1653	ABC-type sugar transport system, periplasmic component
HE	COG0111	Phosphoglycerate dehydrogenase and related dehydrogenases
I	COG0183	Acetyl-CoA acetyltransferase
I	COG1250	3-hydroxyacyl-CoA dehydrogenase
I	COG1884	Methylmalonyl-CoA mutase, N-terminal domainsubunit
I	COG1960	Acyl-CoA dehydrogenases
I	COG2185	Methylmalonyl-CoA mutase, C-terminal domainsubunit (cobalamin-binding)
J	COG4108	Peptide chain release factor RF-3
M	COG1215	Glycosyltransferases, probably involved in cell wall biogenesis
O	COG0695	Glutaredoxin and related proteins
O	COG1765	Predicted redox protein, regulator of disulfide bond formation
P	COG0607	Rhodanese-related sulfurtransferase
Q	COG0179	2-keto-4-pentenoate hydratase2-oxohepta-3-ene-1,7-dioic acid hydratase (catechol pathway)
R	COG0388	Predicted amidohydrolase
R	COG0673	Predicted dehydrogenases and related proteins
R	COG1647	Esteraselipase
S	COG2966	Uncharacterized conserved protein
S	COG3610	Uncharacterized conserved protein
T	COG0784	FOG: CheY-like receiver
T	COG2199	FOG: GGDEF domain
T	COG2200	FOG: EAL domain
T	COG2202	FOG: PASPAC domain
V	COG0841	Cationmultidrug efflux pump

ceso de formación de la endoespora habría sido descrito previamente por Alcaraz y colaboradores [27].

El grupo de los *Bacillus* acuáticos presentó 29 COGs específicos, algunos de los COGs específicos pertenecientes a las categorías [G] *Carbohydrate transport and metabolism* y algunos en la categoría [E] *Amino acid transport and metabolism* consisten de secuencias relacionadas al transporte de carbón orgánico disuelto (DOC - Dissolved Organic Carbon). Estas funciones han sido halladas como sobrerepresentadas en un estudio metagenómico de genes expresados por bacterioplancton [90]. Los tres COGs hallados en la categoría [G] (COG1653, COG1175, COG0395) son transportadores generales de carbohidratos, y los COGs en la categoría [E] (COG747 y COG0601) son transportadores de oligopéptidos. Los COGs sobrerepresentados en la categoría [I] mostraron posibles diferencias metabólicas entre grupos. Por ejemplo, las especies de *Bacillus* acuáticos contienen los COG0183 (acetyl-CoA acetyltransferase), COG1250 (3-hydroxyacyl-CoA dehydrogenase) y COG1960 (Acyl-CoA dehydrogenase), los cuales están involucrados en la degradación de lípidos. La sobrerepresentación de algunos COGs involucrados en la degradación de lípidos ha sido previamente observada en la bacteria marina oligotrófica *Sphingopyxis alaskensis* RB2256 [91]. En contraste, los COGs sobrerepresentados en la categoría [I] dentro de los grupos *B. cereus* y *B. subtilis*, están solamente involucrados en la biosíntesis de lípidos.

El hecho de que el contenido funcional de genes representados por las categorías COG o Figfam, sean conservados dentro de los miembros de un grupo y no compartidos con los miembros en los otros grupos, podría sugerir la especialización vía el genoma accesorio. Esta especialización podría estar relacionada a las condiciones ambientales que estos organismos enfrentan. Más importante, los grupos aislados principalmente de ambientes terrestres, contienen el mayor número de COGs grupo específicos. Nosotros especulamos que, mientras los organismos que viven en ambientes complejos como el terrestre, donde las condiciones pueden cambiar drásticamente, podrían necesitar una batería de funciones para ayudarlos a sobrevivir, los organismos acuáticos podrían encarar un ambiente menos heterogéneo.

Conclusiones

El análisis evolutivo de los *Bacillus* sugiere que las especies de *Bacillus* acuáticos tienen un orígen polifilético. Sin embargo, aún cuando el grupo de *Bacillus* acuáticos descrito por Alcaraz *et al.* no se mantiene en nuestros análisis, es posible observar un grupo compuesto exclusivamente por *Bacillus* aislados de ambientes acuáticos. Esta observación es apoyada a través de la reconstrucción filogenómica empleando el Genoma Núcleo y el Genoma 70. Además, el análisis del contenido de genes empleando el índice de similitud genómica (*GSS*) sugiere que los *Bacillus* aislados de ambientes acuáticos pueden compartir una proporción mayor de su contenido de genes más allá de lo que podría esperarse a partir de sus orígenes polifiléticos, sugiriendo un proceso de evolución convergente a nivel del contenido de genes. Por otra parte, la comparación del contenido funcional entre los grupos obtenidos mediante el análisis de agrupamiento jerárquico, nos permitió determinar la existencia de funciones específicas de cada grupo, las cuales podrían estar relacionadas a la adaptación al ambiente del cual han sido aislados.

Perspectivas

- La falta de información sobre la ecología de los organismos impide analizar las relaciones Evolución - Contenido de genes - Ambiente. Actualmente se han lanzado algunas iniciativas para que los metadatos relacionados a la ecología de cada genoma reportado en las bases de datos sean incluídos. La disponibilidad de esta información hace necesario el desarrollo de herramientas infórmáticas que permitan obtener esta información de forma rápida y confiable.

- El creciente número de genomas depositados en las bases de datos permitirá en un futuro incluir especies de *Bacillus* descritas recientemente, extendiendo así el análisis a grupos poco representados o incluso grupos nuevos.

- Por otra parte, la alta redundancia que existe en las bases de datos hace necesario el diseño de nuevas estrategias para la selección de los genomas preliminares que no tiene un genoma completo de referencia.

Apéndices

Apéndice A

Genomas completos no – Redundantes

Tabla A.1: Genomas completos seleccionados usando el índice GSS

Grupo	Especies
Group 1	***B. amyloliquefaciens* DSM 7, *B. amyloliquefaciens* TA208**, *B. amyloliquefaciens* XH7, *B. amyloliquefaciens* LL3
Group 2	***B. thuringiensis* serovar chinensis CT-43, *B. thuringiensis* serovar thuringiensis str. IS5056**, *B. thuringiensis* Bt407, *B. thuringiensis* YBT-1518
Group 3	***B. megaterium* DSM 319, *B. megaterium* QM B1551, *B. megaterium* WSH-002**
Group 4	***B. thuringiensis* serovar finitimus YBT-020**
Group 5	***B. subtilis* subsp. subtilis str. 168**, *B. subtilis* subsp. subtilis str. BSP1, *B. subtilis* XF-1, *B. subtilis* QB928, *B. subtilis* subsp. subtilis 6051-HGW, *B. subtilis* subsp. subtilis str. BAB-1, *B. subtilis* BSn5, ***B. subtilis* PY79**, *B. subtilis* subsp. subtilis str. RO-NN-1, *B. subtilis* subsp. natto BEST195
Group 6	***B. amyloliquefaciens* subsp. plantarum UCMB5036, *B. amyloliquefaciens* FZB42**, *B. amyloliquefaciens* subsp. plantarum UCMB5113, *B. amyloliquefaciens* subsp. plantarum CAU B946, *B. amyloliquefaciens* subsp. plantarum YAU B9601-Y2, *B. amyloliquefaciens* subsp. plantarum NAU-B3, *B. amyloliquefaciens* subsp. plantarum UCMB5033, *B. amyloliquefaciens* subsp. plantarum AS43.3, *B. amyloliquefaciens* IT-45, *B. amyloliquefaciens* CC178, *B. amyloliquefaciens* Y2
Group 7	***B. anthracis* str. Sterne, *B. anthracis* str. Ames, *B. anthracis* str. Ames Ancestor**, *B. anthracis* str. A0248, *B. cereus* AH820, *B. anthracis* str. H9401, *B. anthracis* str. CDC 684
Group 8	***Bacillus* sp. JS**
Group 9	***B. cereus* G9842, *B. thuringiensis* HD-771**, *B. thuringiensis* HD-789
Group 10	***B. cereus* FRI-35, *B. cereus* ATCC 10987**
Group 11	***B. toyonensis* BCT-7112, *B. thuringiensis* MC28**
Group 12	***B. thuringiensis* serovar kurstaki str. HD73**
Group 13	***B. licheniformis* 9945A**
Group 14	***B. infantis* NRRL B-1491**

continua en la siguiente página

Tabla A.1 – *Continuación de la página previa*

Grupo	**Especies**
Group 15	***B. cellulosilyticus* DSM 2522**
Group 16	***B. pseudofirmus* OF4**
Group 17	***B. cereus* ATCC 14579, *B. thuringiensis* BMB171, *B. cereus* B4264**
Group 18	***B. selenitireducens* MLS10**
Group 19	***B. cereus* biovar anthracis str. CI**
Group 20	***B. subtilis* subsp. spizizenii str. W23, *B. subtilis* subsp. spizizenii TU-B-10**
Group 21	***B. coagulans* 36D**
Group 22	***B. halodurans* C-125**
Group 23	***B. thuringiensis* serovar konkukian str. 97-27, *B. cereus* E33L**
Group 24	***B. licheniformis* DSM 13 = ATCC 14580**, *B. licheniformis* DSM 13 = ATCC 14580
Group 25	***B. clausii* KSM-K16**
Group 26	***B. weihenstephanensis* KBAB4**
Group 27	***B. cytotoxicus* NVH 391-98**
Group 28	***B. cereus* Q1, *B. cereus* AH187**, *B. cereus* NC7401
Group 29	***B. thuringiensis* str. Al Hakam, *B. cereus* F837/76**, *B. cereus* 03BB102
Group 30	***B. pumilus* SAFR-032**
Group 31	***B. atrophaeus* 1942**
Group 32	***B. coagulans* 2-6**
Group 33	***Bacillus* sp. 1NLA3E**

Los grupos de genomas completos no-Redundantes fueron obtenidos usando un valor de $GSS = 0{,}95$ el cual corresponde al umbral de 97 % de identidad empleado para definir la unidad taxonómica operacional (OTU). Las especies seleccionadas de cada grupo se indican en negritas.

Apéndice B

Posibles COGs específicos en el grupo *Bacillus subtilis*

Tabla B.1: Lista de los posibles COGs específicos en el grupo *Bacillus subtilis*

C	COG0371	Glycerol dehydrogenase and related enzymes
C	COG0372	Citrate synthase
C	COG1069	Ribulose kinase
C	COG1251	NAD(P)H-nitrite reductase
C	COG2141	Coenzyme F420-dependent N5,N10-methylene tetrahydromethanopterin reductase and related flavin-dependent oxidoreductases
C	COG2851	H+/citrate symporter
D	COG2385	Sporulation protein and related proteins
ER	COG0591	Na+/proline symporter
ER	COG1063	Threonine dehydrogenase and related Zn-dependent dehydrogenases
E	COG0263	Glutamate 5-kinase
E	COG0620	Methionine synthase II (cobalamin-independent)
E	COG0710	3-dehydroquinate dehydratase
E	COG0765	ABC-type amino acid transport system, permease component
E	COG1113	Gamma-aminobutyrate permease and related permeases
E	COG1174	ABC-type proline/glycine betaine transport systems, permease component
E	COG2040	Homocysteine/selenocysteine methylase (S-methylmethionine-dependent)
E	COG2755	Lysophospholipase L1 and related esterases
E	COG2866	Predicted carboxypeptidase
E	COG3591	V8-like Glu-specific endopeptidase
GM	COG5039	Exopolysaccharide biosynthesis protein
GR	COG2140	Thermophilic glucose-6-phosphate isomerase and related metalloenzymes
G	COG0153	Galactokinase
G	COG0246	Mannitol-1-phosphate/altronate dehydrogenases

continua en la siguiente página

Tabla B.1 – *Continuación de la página previa*

G	COG1263	Phosphotransferase system IIC components, glucose/maltose/N-acetylglucosamine-specific
G	COG1264	Phosphotransferase system IIB components
G	COG1445	Phosphotransferase system fructose-specific component IIB
G	COG1482	Phosphomannose isomerase
G	COG1486	Alpha-galactosidases/6-phospho-beta-glucosidases, family 4 of glycosyl hydrolases
G	COG1621	Beta-fructosidases (levanase/invertase)
G	COG2160	L-arabinose isomerase
G	COG2190	Phosphotransferase system IIA components
G	COG2211	Na+/melibiose symporter and related transporters
G	COG2271	Sugar phosphate permease
G	COG2721	Altronate dehydratase
G	COG2723	Beta-glucosidase/6-phospho-beta-glucosidase/beta-galactosidase
G	COG2814	Arabinose efflux permease
G	COG3507	Beta-xylosidase
G	COG3534	Alpha-L-arabinofuranosidase
G	COG3866	Pectate lyase
G	COG3867	Arabinogalactan endo-1,4-beta-galactosidase
G	COG3936	Protein involved in polysaccharide intercellular adhesin (PIA) synthesis/biofilm formation
G	COG4468	Galactose-1-phosphate uridyltransferase
H	COG0611	Thiamine monophosphate kinase
H	COG1072	Panthothenate kinase
H	COG1424	Pimeloyl-CoA synthetase
IQ	COG0304	3-oxoacyl-(acyl-carrier-protein) synthase
I	COG0331	(acyl-carrier-protein) S-malonyltransferase
I	COG4247	3-phytase (myo-inositol-hexaphosphate 3-phosphohydrolase)
KG	COG1349	Transcriptional regulators of sugar metabolism
K	COG1510	Predicted transcriptional regulators
K	COG1846	Transcriptional regulators
K	COG2732	Barstar, RNAse (barnase) inhibitor
L	COG1525	Micrococcal nuclease (thermonuclease) homologs
MI	COG0615	Cytidylyltransferase
M	COG0399	Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis
M	COG0463	Glycosyltransferases involved in cell wall biogenesis
M	COG1442	Lipopolysaccharide biosynthesis proteins, LPS:glycosyltransferases
M	COG1861	Spore coat polysaccharide biosynthesis protein F, CMP-KDO synthetase homolog
M	COG1887	Putative glycosyl/glycerophosphate transferases involved in teichoic acid biosynthesis TagF/TagB/EpsJ/RodC
M	COG2089	Sialic acid synthase
M	COG3773	Cell wall hydrolyses involved in spore germination
M	COG3980	Spore coat polysaccharide biosynthesis protein, predicted glycosyltransferase

continua en la siguiente página

Tabla B.1 – *Continuación de la página previa*

M	COG5520	O-Glycosyl hydrolase
M	COG5577	Spore coat protein
N	COG3144	Flagellar hook-length control protein
OC	COG0526	Thiol-disulfide isomerase and thioredoxins
O	COG0265	Trypsin-like serine proteases, typically periplasmic, contain C-terminal PDZ domain
O	COG1331	Highly conserved protein containing a thioredoxin domain
O	COG1404	Subtilisin-like serine proteases
O	COG1651	Protein-disulfide isomerase
O	COG1764	Predicted redox protein, regulator of disulfide bond formation
P	COG0306	Phosphate/sulphate permeases
P	COG0369	Sulfite reductase, alpha subunit (flavoprotein)
P	COG1117	ABC-type phosphate transport system, ATPase component
P	COG1914	Mn2+ and Fe2+ transporters of the NRAMP family
P	COG3540	Phosphodiesterase/alkaline phosphatase D
P	COG4594	ABC-type Fe3+-citrate transport system, periplasmic component
Q	COG1020	Non-ribosomal peptide synthetase modules and related proteins
Q	COG2124	Cytochrome P450
Q	COG3208	Predicted thioesterase involved in non-ribosomal peptide biosynthesis
Q	COG3319	Thioesterase domains of type I polyketide synthases or non-ribosomal peptide synthetases
Q	COG3458	Acetyl esterase (deacetylase)
Q	COG3479	Phenolic acid decarboxylase
R	COG1783	Phage terminase large subunit
R	COG2079	Uncharacterized protein involved in propionate catabolism
R	COG3318	Predicted metal-binding protein related to the C-terminal domain of SecA
R	COG3953	SLT domain proteins
R	COG4112	Predicted phosphoesterase (MutT family)
R	COG4195	Phage-related replication protein
R	COG5518	Bacteriophage capsid portal protein
S	COG1652	Uncharacterized protein containing LysM domain
S	COG1714	Predicted membrane protein/domain
S	COG2246	Predicted membrane protein
S	COG3299	Uncharacterized homolog of phage Mu protein gp47
S	COG3619	Predicted membrane protein
S	COG3665	Uncharacterized conserved protein
S	COG3876	Uncharacterized protein conserved in bacteria
S	COG4336	Uncharacterized conserved protein
S	COG4485	Predicted membrane protein
S	COG4721	Predicted membrane protein
S	COG4894	Uncharacterized conserved protein
S	COG5444	Uncharacterized conserved protein
TK	COG2197	Response regulator containing a CheY-like receiver domain and an HTH DNA-binding domain
TQ	COG2508	Regulator of polyketide synthase expression

continua en la siguiente página

Tabla B.1 – *Continuación de la página previa*

T	COG1734	DnaK suppressor protein
T	COG3434	Predicted signal transduction protein containing EAL and modified HD-GYP domains

Apéndice C

Posibles COGs específicos en el grupo *Bacillus cereus*

Tabla C.1: Lista de los posibles COGs específicos en el grupo *Bacillus cereus*

CR	COG0604	NADPH:quinone reductase and related Zn-dependent oxidoreductases
C	COG0039	Malate/lactate dehydrogenases
C	COG0277	FAD/FMN-containing dehydrogenases
C	COG1151	6Fe-6S prismane cluster-containing protein
C	COG4781	Membrane domain of membrane-anchored glycerophosphoryl diester phosphodiesterase
D	COG4942	Membrane-bound metallopeptidase
E	COG0460	Homoserine dehydrogenase
E	COG0754	Glutathionylspermidine synthase
E	COG0814	Amino acid permeases
E	COG1114	Branched-chain amino acid permeases
E	COG1115	Na+/alanine symporter
E	COG1231	Monoamine oxidase
E	COG1280	Putative threonine efflux protein
E	COG2502	Asparagine synthetase A
E	COG2515	1-aminocyclopropane-1-carboxylate deaminase
E	COG3104	Dipeptide/tripeptide permease
E	COG3186	Phenylalanine-4-hydroxylase
E	COG3227	Zinc metalloprotease (elastase)
E	COG3616	Predicted amino acid aldolase or racemase
E	COG3938	Proline racemase
E	COG4166	ABC-type oligopeptide transport system, periplasmic component
F	COG0295	Cytidine deaminase
F	COG0572	Uridine kinase
F	COG0737	5'-nucleotidase/2',3'-cyclic phosphodiesterase and related esterases
F	COG0775	Nucleoside phosphorylase

continua en la siguiente página

Tabla C.1 – *Continuación de la página previa*

F	COG1051	ADP-ribose pyrophosphatase
F	COG1328	Oxygen-sensitive ribonucleoside-triphosphate reductase
F	COG1957	Inosine-uridine nucleoside N-ribohydrolase
F	COG1972	Nucleoside permease
GER	COG0697	Permeases of the drug/metabolite transporter (DMT) superfamily
G	COG0406	Fructose-2,6-bisphosphatase
G	COG0588	Phosphoglycerate mutase 1
G	COG1172	Ribose/xylose/arabinose/galactoside ABC-type transport systems, permease components
G	COG1830	DhnA-type fructose-1,6-bisphosphate aldolase and related enzymes
G	COG2376	Dihydroxyacetone kinase
G	COG3469	Chitinase
H	COG0303	Molybdopterin biosynthesis enzyme
H	COG0314	Molybdopterin converting factor, large subunit
H	COG0476	Dinucleotide-utilizing enzymes involved in molybdopterin and thiamine biosynthesis family 2
H	COG1977	Molybdopterin converting factor, small subunit
H	COG2154	Pterin-4a-carbinolamine dehydratase
H	COG2896	Molybdenum cofactor biosynthesis enzyme
I	COG3243	Poly(3-hydroxyalkanoate) synthetase
I	COG3963	Phospholipid N-methyltransferase
J	COG1236	Predicted exonuclease of the beta-lactamase fold involved in RNA processing
J	COG1670	Acetyltransferases, including N-acetylases of ribosomal proteins
K	COG0640	Predicted transcriptional regulators
K	COG1309	Transcriptional regulator
K	COG1316	Transcriptional regulator
K	COG1329	Transcriptional regulators, similar to M. xanthus CarD
K	COG1476	Predicted transcriptional regulators
K	COG1695	Predicted transcriptional regulators
K	COG2345	Predicted transcriptional regulator
K	COG2378	Predicted transcriptional regulator
K	COG3682	Predicted transcriptional regulator
LR	COG0494	NTP pyrophosphohydrolases including oxidative damage repair enzymes
L	COG0675	Transposase and inactivated derivatives
L	COG0776	Bacterial nucleoid DNA-binding protein
L	COG1041	Predicted DNA modification methylase
L	COG1484	DNA replication protein
L	COG1573	Uracil-DNA glycosylase
L	COG1961	Site-specific recombinases, DNA invertase Pin homologs
L	COG2801	Transposase and inactivated derivatives
L	COG3449	DNA gyrase inhibitor
L	COG4912	Predicted DNA alkylation repair enzyme
L	COG4974	Site-specific recombinase XerD
MG	COG0451	Nucleoside-diphosphate-sugar epimerases
M	COG0768	Cell division protein FtsI/penicillin-binding protein 2

continua en la siguiente página

Tabla C.1 – *Continuación de la página previa*

M	COG1247	Sortase and related acyltransferases
M	COG1686	D-alanyl-D-alanine carboxypeptidase
M	COG1696	Predicted membrane protein involved in D-alanine export
M	COG1794	Aspartate racemase
M	COG1876	D-alanyl-D-alanine carboxypeptidase
M	COG3757	Lyzozyme M1 (1,4-beta-N-acetylmuramidase)
M	COG5386	Cell surface protein
N	COG1749	Flagellar hook protein FlgE
OU	COG1585	Membrane protein implicated in regulation of membrane protease activity
O	COG2377	Predicted molecular chaperone distantly related to HSP70-fold metalloproteases
O	COG4846	Membrane protein involved in cytochrome C biogenesis
P	COG0370	Fe2+ transport system protein B
P	COG0474	Cation transport ATPase
P	COG0475	Kef-type K+ transport systems, membrane components
P	COG0798	Arsenite efflux pump ACR3 and related permeases
P	COG1055	Na+/H+ antiporter NhaD and related arsenite permeases
P	COG1178	ABC-type Fe3+ transport system, permease component
P	COG1840	ABC-type Fe3+ transport system, periplasmic component
P	COG1918	Fe2+ transport system protein A
P	COG2060	K+-transporting ATPase, A chain
P	COG2116	Formate/nitrite family of transporters
P	COG2156	K+-transporting ATPase, c chain
P	COG2193	Bacterioferritin (cytochrome b1)
P	COG2216	High-affinity K+ transport system, ATPase chain B
P	COG2824	Uncharacterized Zn-ribbon-containing protein involved in phosphonate metabolism
QR	COG0500	SAM-dependent methyltransferases
Q	COG1335	Amidases related to nicotinamidase
Q	COG2162	Arylamine N-acetyltransferase
Q	COG3508	Homogentisate 1,2-dioxygenase
R	COG0110	Acetyltransferase (isoleucine patch superfamily)
R	COG0446	Uncharacterized NAD(FAD)-dependent dehydrogenases
R	COG0491	Zn-dependent hydrolases, including glyoxylases
R	COG0546	Predicted phosphatases
R	COG0596	Predicted hydrolases or acyltransferases (alpha/beta hydrolase superfamily)
R	COG0666	FOG: Ankyrin repeat
R	COG0693	Putative intracellular protease/amidase
R	COG0730	Predicted permeases
R	COG0733	Na+-dependent transporters of the SNF family
R	COG1011	Predicted hydrolase (HAD superfamily)
R	COG1159	GTPase
R	COG1266	Predicted metal-dependent membrane protease
R	COG1409	Predicted phosphohydrolases

continua en la siguiente página

Tabla C.1 – *Continuación de la página previa*

R	COG1923	Uncharacterized host factor I protein
R	COG2374	Predicted extracellular nuclease
R	COG2389	Uncharacterized metal-binding protein
R	COG2503	Predicted secreted acid phosphatase
R	COG2514	Predicted ring-cleavage extradiol dioxygenase
R	COG2704	Anaerobic C4-dicarboxylate transporter
R	COG2872	Predicted metal-dependent hydrolases related to alanyl-tRNA synthetase HxxxH domain
R	COG2936	Predicted acyl esterases
R	COG3173	Predicted aminoglycoside phosphotransferase
R	COG3560	Predicted oxidoreductase related to nitroreductase
R	COG3568	Metal-dependent hydrolase
R	COG3631	Ketosteroid isomerase-related protein
R	COG3964	Predicted amidohydrolase
R	COG3979	Uncharacterized protein contain chitin-binding domain type 3
R	COG3981	Predicted acetyltransferase
R	COG4533	ABC-type uncharacterized transport system, periplasmic component
R	COG4667	Predicted esterase of the alpha-beta hydrolase superfamily
R	COG5018	Inhibitor of the KinA pathway to sporulation, predicted exonuclease
R	COG5485	Predicted ester cyclase
S	COG0327	Uncharacterized conserved protein
S	COG0393	Uncharacterized conserved protein
S	COG0401	Uncharacterized homolog of Blt101
S	COG0586	Uncharacterized membrane-associated protein
S	COG1285	Uncharacterized membrane protein
S	COG1434	Uncharacterized conserved protein
S	COG1720	Uncharacterized conserved protein
S	COG1944	Uncharacterized conserved protein
S	COG2035	Predicted membrane protein
S	COG2259	Predicted membrane protein
S	COG2261	Predicted membrane protein
S	COG2306	Uncharacterized conserved protein
S	COG2311	Predicted membrane protein
S	COG2320	Uncharacterized conserved protein
S	COG2326	Uncharacterized conserved protein
S	COG2510	Predicted membrane protein
S	COG2733	Predicted membrane protein
S	COG2879	Uncharacterized small protein
S	COG3223	Predicted membrane protein
S	COG3238	Uncharacterized protein conserved in bacteria
S	COG3272	Uncharacterized conserved protein
S	COG3274	Uncharacterized protein conserved in bacteria
S	COG3339	Uncharacterized conserved protein
S	COG3384	Uncharacterized conserved protein
S	COG3397	Uncharacterized protein conserved in bacteria
S	COG3506	Uncharacterized conserved protein

continua en la siguiente página

Tabla C.1 – *Continuación de la página previa*

S	COG3575	Uncharacterized protein conserved in bacteria
S	COG3584	Uncharacterized protein conserved in bacteria
S	COG3589	Uncharacterized conserved protein
S	COG3760	Uncharacterized conserved protein
S	COG3797	Uncharacterized protein conserved in bacteria
S	COG3817	Predicted membrane protein
S	COG3819	Predicted membrane protein
S	COG3874	Uncharacterized conserved protein
S	COG3878	Uncharacterized protein conserved in bacteria
S	COG4043	Uncharacterized conserved protein
S	COG4198	Uncharacterized conserved protein
S	COG4377	Predicted membrane protein
S	COG4412	Uncharacterized protein conserved in bacteria
S	COG4427	Uncharacterized protein conserved in bacteria
S	COG4478	Predicted membrane protein
S	COG4509	Uncharacterized protein conserved in bacteria
S	COG4627	Uncharacterized protein conserved in bacteria
S	COG4696	Uncharacterized protein conserved in bacteria
S	COG4702	Uncharacterized conserved protein
S	COG4709	Predicted membrane protein
S	COG4720	Predicted membrane protein
S	COG4760	Predicted membrane protein
S	COG4815	Uncharacterized protein conserved in bacteria
S	COG4842	Uncharacterized protein conserved in bacteria
S	COG4877	Uncharacterized protein conserved in bacteria
S	COG4990	Uncharacterized protein conserved in bacteria
S	COG5294	Uncharacterized protein conserved in bacteria
S	COG5523	Predicted integral membrane protein
TK	COG0745	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain
T	COG0642	Signal transduction histidine kinase
T	COG2365	Protein tyrosine/serine phosphatase
T	COG3103	SH3 domain protein
T	COG5278	Predicted periplasmic ligand-binding sensor domain
U	COG0681	Signal peptidase I
V	COG0842	ABC-type multidrug transport system, permease component
V	COG1131	ABC-type multidrug transport system, ATPase component
V	COG1136	ABC-type antimicrobial peptide transport system, ATPase component
V	COG1680	Beta-lactamase class C and other penicillin binding proteins
V	COG1968	Uncharacterized bacitracin resistance protein
V	COG3510	Cephalosporin hydroxylase
V	COG4767	Glycopeptide antibiotics resistance protein

Apéndice D

Clasificación del ambiente natural de los genomas de *Bacillus* usadas en este estudio

Tabla D.1: Clasificación del ambiente natural de los *Bacillus* empleados en este estudio

Organismo	**Bioproject**	**Ecosistema**	**Categoría**	**Tipo**
B. alcalophilus ATCC 27647	PRJNA171835	Host-associated	Human	Facultative
B. amyloliquefaciens DSM 7	PRJEA41719	Environmental	Terrestrial	Terrestrial
B. amyloliquefaciens TA208	PRJNA64581	Engineered	Industrial production	Terrestrial
B. amyloliquefaciens plantarum FZB42	PRJNA13403	Environmental	Terrestrial	Terrestrial
B. amyloliquefaciens plantarum UCMB5036	PRJEA82107	Host-associated	Plants	Facultative
B. anthracis Ames	PRJNA309	Derivated		Facultative
B. anthracis Ames Áncestor'	PRJNA10784	Host-associated	Animal	Facultative
B. anthracis Sterne	PRJNA10878	Host-associated	Unclassified	Facultative
B. aquimaris TF12	PRJNA71387	Environmental	Aquatic	Aquatic
B. atrophaeus 1942	PRJNA46075	Environmental	Terrestrial	Terrestrial
B. azotoformans LMG 9581	PRJNA80827	Environmental	Terrestrial	Terrestrial
B. bataviensis LMG 21833	PRJNA77725	Environmental	Terrestrial	Terrestrial
B. cellulosilyticus DSM 2522	PRJNA38423	Environmental	Terrestrial	Terrestrial
B. cereus AH187	PRJNA17715	Host-associated	Human	Facultative
B. cereus ATCC 10987	PRJNA74	Engineered	Food production	Terrestrial
B. cereus ATCC 14579	PRJNA384	Host-associated		Facultative
B. cereus B4264	PRJNA17731	Host-associated	Human	Facultative
B. cereus E33L	PRJNA12468	Host-associated	Mammals	Facultative
B. cereus F837/76	PRJNA15716	Host-associated	Unclassified	Facultative

continua en la siguiente página

Tabla D.1 – *Continuación de la página previa*

Organismo	**Bioproject**	**Ecosistema**	**Categoría**	**Tipo**
B. cereus FRI-35	PRJNA171769			Unknown
B. cereus G9842	PRJNA17733	Host-associated	Animal	Facultative
B. cereus Q1	PRJNA16220	Environmental	Terrestrial	Terrestrial
B. cereus anthracis CI	PRJNA36309	Host-associated	Chimpanzee	Facultative
B. clausii KSM-K16	PRJNA13291	Environmental	Terrestrial	Terrestrial
B. coagulans 2-6	PRJNA61501			Unknown
B. coagulans 36D1	PRJNA15679	Environmental	Aquatic	Specialized
B. coahuilensis m2-6	PRJNA84423	Environmental	Aquatic	Aquatic
B. coahuilensis m4-4	PRJNA19551	Environmental	Aquatic	Aquatic
B. coahuilensis p1.1.43c	PRJNA84425	Environmental	Aquatic	Aquatic
B. cytotoxicus NVH 391-98	PRJNA13624	Host-associated	Plants	Facultative
B. halodurans C-125	PRJNA235	Environmental	Terrestrial	Terrestrial
B. horikoshii DSM 8719	PRJNA72395	Environmental	Terrestrial	Terrestrial
B. infantis NRRL B-14911	PRJNA212797	Environmental	Aquatic	Aquatic
B. isronensis B3W22	PRJNA173035	Environmental	Air	Specialized
B. licheniformis 9945A	PRJNA49115			Unknown
B. licheniformis DSM 13 = ATCC 14580	PRJNA13082	Host-associated		Facultative
B. macauensis ZFHKF-1	PRJNA167832	Environmental	Aquatic	Aquatic
B. megaterium DSM 319	PRJNA42425	Environmental	Soil	Terrestrial
B. megaterium QM B1551	PRJNA30165	Environmental	Soil	Terrestrial
B. megaterium WSH002	PRJNA71447	Environmental	Soil	Terrestrial
B. methanolicus MGA3	PRJNA49595	Environmental	Terrestrial	Terrestrial
B. mojavensis RO-H-1	PRJNA68567	Environmental	Terrestrial	Terrestrial
B. mycoides DSM 2048	PRJNA29701	Environmental	Terrestrial	Terrestrial
B. oceanisediminis 2691	PRJNA167766	Environmental	Aquatic	Aquatic
B. pseudofirmus OF4	PRJNA28811	Environmental	Terrestrial	Terrestrial
B. pumilus SAFR-032	PRJNA20391	Engineered	Built environment	Terrestrial
B. selenitireducens MLS10	PRJNA13376	Environmental	Aquatic	Aquatic
B. stratosphericus LAMA 585	PRJNA176166	Environmental	Aquatic	Aquatic
B. subtilis PY79	PRJNA225627			Unknown
B. subtilis gtP20b	PRJNA53249	Environmental	Aquatic	Aquatic
B. subtilis spizizenii TU-B-10	PRJNA68561	Environmental	Terrestrial	Terrestrial
B. subtilis spizizenii W23	PRJNA38713	Environmental	Terrestrial	Terrestrial
B. subtilis subtilis 168	PRJNA29889	Environmental	Terrestrial	Terrestrial
B. thuringiensis Al Hakam	PRJNA18255	Engineered	Industrial production	Facultative
B. thuringiensis BMB171	PRJNA43631	Engineered		Facultative
B. thuringiensis HD-771	PRJNA171845			Unknown
B. thuringiensis MC28	PRJNA167562	Environmental	Terrestrial	Terrestrial
B. thuringiensis chinensis CT-43	PRJNA43737	Environmental	Terrestrial	Terrestrial
B. thuringiensis finitimus YBT-020	PRJNA60447			Unknown
B. thuringiensis konkukian 97-27	PRJNA10877	Host-associated	Human	Facultative

continua en la siguiente página

Tabla D.1 – *Continuación de la página previa*

Organismo	**Bioproject**	**Ecosistema**	**Categoría**	**Tipo**
B. thuringiensis kurstaki HD73	PRJNA185468			Unknown
B. thuringiensis thuringiensis IS5056	PRJNA187142	Environmental	Terrestrial	Terrestrial
B. toyonensis BCT-7112	PRJNA225857	Environmental	Terrestrial	Terrestrial
B. weihenstephanensis KBAB4	PRJNA13623	Environmental	Terrestrial	Terrestrial
Bacillus sp. 10403023	PRJEA70827	Host-associated	Human	Facultative
Bacillus sp. 1NLA3E	PRJNA53255	Engineered	Bioremediation	Terrestrial
Bacillus sp. 2 A 57 CT2	PRJNA40003	Host-associated	Human	Facultative
Bacillus sp. 5B6	PRJNA79215	Host-associated	Plants	Facultative
Bacillus sp. 7 6 55CFAA CT2	PRJNA40005	Host-associated	Human	Facultative
Bacillus sp. B14905	PRJNA18949	Environmental	Aquatic	Aquatic
Bacillus sp. BT1B CT2	PRJNA40001	Host-associated	Human	Facultative
Bacillus sp. HYC-10	PRJNA162763	Host-associated	Fish	Facultative
Bacillus sp. JS	PRJNA79217	Environmental	Terrestrial	Terrestrial
Bacillus sp. L1(2012)	PRJNA182346	Environmental	Aquatic	Aquatic
Bacillus sp. M 2-6	PRJNA161543	Host-associated	Plants	Facultative
Bacillus sp. NRRL B-14911	PRJNA13545	Environmental	Aquatic	Aquatic
Bacillus sp. SG-1	PRJNA19283	Environmental	Aquatic	Aquatic
Bacillus sp. m3-13	PRJNA38237	Environmental	Aquatic	Aquatic
Bacillus sp. P15.4	PRJNA384653	Environmental	Aquatic	Aquatic
Geobacillus sp. Y412MC52	PRJNA30797	Environmental	Aquatic	Specialized
L. innocua Clip11262	PRJNA86	Engineered	Food production	Facultative
L. monocytogenes 4b Clip 80459	PRJEA32207	Host-associated		Facultative
O. iheyensis HTE831	PRJNA284	Environmental	Aquatic	Aquatic

Referencias

1. Cohan, F. M. and Koeppel, A. F. (nov, 2008) The origins of ecological diversity in prokaryotes.. *Current Biology,* **18**(21):R1024–34.

2. Abby, S. and Daubin, V. 2007. Comparative genomics and the evolution of prokaryotes. *Trends in Microbiology,* **15**(3):135–141.

3. Ochman, H. and Davalos, L. M. 2006. The Nature and Dynamics of Bacterial Genomes. *Science,* **311**:1730–1733.

4. Eisen, J. A. and Fraser, C. M. 2003. Phylogenomics : Intersection of Evolution and Genomics. *science,* **300**:1706–1707.

5. Snel, B., Bork, P., and Huynen, M. a. 1999. Genome phylogeny based on gene content.. *Nature genetics,* **21**(1):108–110.

6. Konstantinidis, K. T. and Tiedje, J. M. 2005. Genomic insights that advance the species definition for prokaryotes.. *Proceedings of the National Academy of Sciences of the United States of America,* **102**(7):2567–72.

7. Ochman, H., Lawrence, J. G., and Groisman, E. A. (May, 2000) Lateral gene transfer and the nature of bacterial innovation. *Nature,* **405**(6784):299–304.

8. Wiedenbeck, J. and Cohan, F. 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev,* **35**(5):957–76.

9. Fitch, W. M. 2000. Homology a personal view on some of the problems. *Trends Genet,* **16**(5):227–231.

10. Tettelin, H., Masignani, V., Cieslewicz, M., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A., Deboy, R., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J., Hauser, C., Sundaram, J., Nelson, W., Madupu, R., Brinkac, L., Dodson, R., Rosovitz, M., Sullivan, S., Daugherty, S., Haft, D., Selengut, J., Gwinn, M., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K., Smith, S., Utterback, T., White, O., Rubens, C., Grandi, G., Madoff, L., Kasper, D., Telford, J., Wessels, M., Rappuoli, R., and Fraser, C. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A,* **102**(39):13950–5.

11. Collins, R. E. and Higgs, P. G. 2012. Testing the Infinitely Many Genes Model for the Evolution of the Bacterial Core Genome and Pangenome. *Molecular biology and evolution,* **29**(11):3413–3425.

12. van de Guchte, M. (September, 2017) Horizontal Gene Transfer and Ecosystem Function Dynamics. *Trends in Microbiology,* **25**(9):699–700.

13. Tamames, J., Sánchez, P. D., Nikel, P. I., and Pedrós-alió, C. 2016. Quantifying the Relative Importance of Phylogeny and Environmental Preferences As Drivers of Gene Content in Prokaryotic Microorganisms. *Frontiers in Microbiology,* **7**(433):1–12.

14. Mandic-mulec, I., Stefanic, P., and Elsas, J. 2015. Ecology of Bacillaceae. *Microbioloy Spectrum,* **3**(2):1–24.

15. Ravel, J. and Fraser, C. M. 2005. Genomics at the genus scale. *Trends in Microbiology,* **13**(3):95–7.

16. Earl, A. M., Losick, R., and Kolter, R. 2008. Ecology and genomics of Bacillus subtilis.. *Trends in microbiology,* **16**(6):269–75.

17. Slepecky, R. A. and Ernest, H. H. 2006. The Prokaryotes, Springer, New York, NY. USA third edit edition.

18. Fajardo-cavazos, P., Maughan, H., and Nicholson, W. L. 2014. Evolution in the Bacillaceae. *Microbiology Spectrum,* **2**(5):1–32.

19. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessières, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., Bron, S., Brouillet, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S. K., Codani, J. J., Connerton, I. F., and Danchin, A. 1997. The complete genome sequence of the gram-positive bacterium Bacillus subtilis.. *Nature,* **390**(6657):249–256.

20. Takami, H., Nakasone, K., Takaki, Y., Maeno, G., Sasaki, R., Masui, N., Fuji, F., Hirama, C., Nakamura, Y., Ogasawara, N., Kuhara, S., and Horikoshi, K. 2000. Complete genome sequence of the alkaliphilic bacterium Bacillus halodurans and genomic sequence comparison with Bacillus subtilis. *Nucleic Acids Research,* **28**(21):4317–4331.

21. Takami, H., Takaki, Y., and Uchiyama, I. 2002. Genome sequence of Oceanobacillus iheyensis isolated from the Iheya Ridge and its unexpected adaptive capabilities to extreme environments. *Nucleic Acids Research,* **30**(18):3927–3935.

22. Read, T. D., Peterson, S. N., and Tourasse, N. 2003. The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. *Nature,* **423**:81–86.

23. Ivanova, N., Sorokin, A., Anderson, I., Galleron, N., Candelon, B., Kapatral, V., Bhattacharyya, A., Reznik, G., Mikhailova, N., Lapidus, A., Chu, L., Mazur, M., Goltsman, E., Larsen, N., Souza, M. D., Walunas, T., Grechkin, Y., Pusch, G., Haselkorn, R., Fonstein, M., Ehrlich, S. D., Overbeek, R., and Kyrpides, N. 2003. Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis. *Nature,* **423**:87–91.

24. Rasko, D. A., Ravel, J., Andreas, O., Helgason, E., Cer, R. Z., Jiang, L., Shores, K. A., Fouts, D. E., Tourasse, N. J., Angiuoli, S. V., Kolonay, J., Nelson, W. C., Kolstù, A.-b., Fraser, C. M., and Read, T. D. 2004. The genome sequence of Bacillus cereus ATCC 10987 reveals metabolic adaptations and a large plasmid related to Bacillus anthracis pXO1. *Nucleic Acids Research,* **32**(3):977–988.

25. Rey, M. W., Ramaiya, P., Nelson, B. A., Brody-karpin, S. D., Zaretsky, E. J., Tang, M., Leon, A. L. D., Xiang, H., Gusti, V., Clausen, I. G., Olsen, P. B., Rasmussen, M. D., Andersen, J. T., Jørgensen, P. L., Larsen, T. S., Sorokin, A., Bolotin, A., Lapidus, A., Galleron, N., Ehrlich, S. D., and Berka, R. M. 2004. Complete genome sequence of the industrial bacterium Bacillus licheniformis and comparisons with closely related Bacillus species. *Genome biology,* **5**(10):R77.

26. Veith, B., Herzberg, C., Steckel, S., Feesche, J., Heinz Maurer, K., Ehrenreich, P., Baumer, S., Henne, A., Liesegang, H., Merkl, R., Ehrenreich, A., and Gottschalk, G. 2004. The Complete Genome Sequence of Bacillus licheniformis DSM13 , an Organism with great industrial potential. *Journal of molecular microbiology and biotechnology,* **7**:204–211.

27. Alcaraz, L. D., Moreno-Hagelsieb, G., Eguiarte, L. E., Souza, V., Herrera-Estrella, L., and Olmedo, G. 2010. Understanding the evolutionary relationships and major traits of Bacillus through comparative genomics. *BMC Genomics,* **11**:332.

28. Ivanova, E., Vysotskii, M., Svetashev, V., Nedashkovskaya, O., Gorshkova, N., Mikhailov, V., Yumoto, N., Shigeri, Y., Taguchi, T., and Yoshikawa, S. 1999. Characterization of Bacillus strains of marine origin. *Interntional Microbiology,* **2**(4):267–271.

29. Siefert, J., Larios-Sanz, M., Nakamura, L., Slepecky, R., Paul, J., Moore, E., Fox, G., and Jurtshuk, P. J. 2000. Characterization of Bacillus strains of marine origin. *Current Microbiology,* **41**(2):84–88.

30. Ettoumi, B., Raddadi, N., Borin, S., Daffonchio, D., Boudabous, A., and Cherif, A. 2009. Diversity and phylogeny of culturable spore-forming Bacilli isolated from marine sediments. *Journal Basic Microbiology,* **49**:S13–S23.

31. Cerritos, R., Luis, E., Morena, A., Janet, S., Michael, T., Alejandra, R.-V., and Valeria, S. 2011. Diversity of culturable thermo-resistant aquatic bacteria along an environmental gradient in Cuatro Cie ť xico Coahuila , Me. *Antonie van Leeuwenhoek„* **99**:303–318.

32. Jensen, P. R. and Fenical, W. 1994. Strategies for the discovery of secondary metabolites from marine bacteria: ecological perspectives. *Annual review of microbiology,* **48**:559–584.

33. Martiny, A. C., Huang, Y., and Li, W. 2009. Occurrence of phosphate acquisition genes in Prochlorococcus cells from different ocean regions. *Environmental Microbiology,* **11**(6):1340–1347.

34. Luo, C., Walk, S. T., Gordon, D. M., Feldgarden, M., Tiedje, J. M., and Konstantinidis, K. T. 2011. Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proc Natl Acad Sci USA,* **108**(17):7200–7205.

35. O'Sullivan, O., Callaghan, J. O., Sangrador-vegas, A., Mcauliffe, O., Slattery, L., Kaleta, P., Callanan, M., Fitzgerald, G. F., Ross, R. P., and Beresford, T. 2009. Comparative genomics of lactic acid bacteria reveals a niche-specific gene set. *BMC Microbiology,* **9**(50):1–9.

36. Wu, X., Monchy, S., Taghavi, S., Zhu, W., Ramos, J., and van der Lelie, D. 2011. Comparative genomics and functional analysis of niche-specific adaptation in Pseudomonas putida. *FEMS Microbiol Rev.,* **35**:299–323.

37. Alcaraz, L. D., Olmedo, G., Bonilla, G., Cerritos, R., Hernandez, G., Cruz, A., Alcaraz, L. D., Olmedo, G., Putonti, C., Jime, B., Mart, E., Lo, V., Arvizu, J. L., Ayala, F., Razo, F., Caballero, J., Siefert, J., Eguiarte, L., Vielle, J.-p., Souza, V., Herrera-estrella, A., and Herrera-estrella, L. 2008. The genome of Bacillus coahuilensis reveals adaptations essential for survival in the relic of an ancient marine environment. *Proceedings of the National Academy of Sciences,* **105**(15):5803–5808.

38. Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., OÑeill, K., Li, W., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zheng, C., Thibaud-Nissen, F., Geer, L. Y., Marchler-Bauer, A., and Pruitt, K. D. 2018. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res,* **46**(D1):D851–D860.

39. Moreno-Hagelsieb, G., Wang, Z., Walsh, S., and ElSherbiny, A. 2013. Phylogenomic clustering for selecting non-redundant genomes for comparative genomics.. *Bioinformatics,* **29**(7):947–949.

40. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics,* **28**(23):3150–3152.

41. Nawrocki, E. P. and Eddy, S. R. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics,* **29**:2933–2935.

42. Wu, M., Chatterji, S., and Eisen, J. A. 2012. Accounting For Alignment Uncertainty in Phylogenomics. *PLoS ONE,* **7**(1):e30288.

43. Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods,* **9**(8):772.

44. Guindon, S., F, D. J., V, L., M, A., W, H., and Gascuel, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology,* **59**(3):307–321.

45. Wu, M. and Eisen, J. E. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biology,* **9**:R151.

46. Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics,* **27**:1164–1165.

47. Moreno-hagelsieb, G. and Latimer, K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics,* **24**(3):319–324.

48. Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* **32**(5):1792–1797.

49. Miller, M. A., Pfeiffer, W., and Schwartz, T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gateway Computing Environments Workshop, GCE 2010,*.

50. Kunin, V., Ahren, D., Goldovsky, L., Janssen, P., and Ouzounis, C. A. 2005. Measuring genome conservation across taxa : divided strains and united kingdoms. *Nucleic Acids Research,* **33**(2):616–621.

51. Moreno-Hagelsieb, G. and Janga, S. C. 2008. Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. *Proteins,* **70**(2):344–352.

52. Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics,* **5**:164–166.

53. Sanjuán, R. and Wróbel, B. 2005. Weighted least-squares likelihood ratio test for branch testing in phylogenies reconstructed from distance measures. *Syst Biol,* **54**(2):218–229.

54. Robinson, D. and Foulds, L. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences,* **53**:131–147.

55. Tatusov, R. L., Natale, D. a., Garkavtsev, I. V., Tatusova, T. a., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes.. *Nucleic Acids Research,* **29**(1):22–28.

56. Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics,* **9**:75.

57. Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Lu, F., Marchler, G. H., Mullokandov, M., Omelchenko, M. V., Robertson, C. L., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., Zhang, N., Zheng, C., and Bryant, S. H. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research,* **39**:D225–D229.

58. Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. a., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., and Stevens, R. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).. *Nucleic acids research,* **42**(Database issue):D206–14.

59. Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E., Vonstein, V., Warren, A., Will, R., Wilson, M. J. C., Yoo, H. S., Zhang, C., Zhang, Y., and Sobral, B. W. (jan, 2014) PATRIC, the bacterial bioinformatics database and analysis resource.. *Nucleic acids research,* **42**(Database issue):D581–91.

60. Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* **20**(C):53–65.

61. Coordinators, N. R. 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research,* **44**(D1):7–19.

62. Pedruzzi, I., Rivoire, C., Auchincloss, A. H., Coudert, E., Keller, G., Castro, E. D., Baratin, D., Cuche, A., Bougueleret, L., Poux, S., Redaschi, N., Xenarios, I., and Bridge, A. 2015. HAMAP in 2015 : updates to the protein family classification and annotation system. *Nucleic Acids Research,* **43**(D1):1064–1070.

63. Parter, M., Kashtan, N., and Alon, U. 2007. Environmental variability and modularity of bacterial metabolic networks. *BMC Evolutionary Biology,* **7**(1):169.

64. R Development Core Team, R. R: A Language and Environment for Statistical Computing. 2011.

65. Kristiansson, E., Hugenholtz, P., and Dalevi, D. 2009. ShotgunFunctionalizeR: An R-package for functional comparison of metagenomes. *Bioinformatics,* **25**(20):2737–2738.

66. Delsuc, F., Brinkmann, H., and Philippe, H. 2005. Phylogenomics and the reconstruction of the tree of life.. *Nature reviews. Genetics,* **6**(5):361–375.

67. Medini, D., Donati, C., Tettelin, H., Masignani, V., and R., R. 2005. The microbial pan-genome. *Curr Opin Genet Dev,* **15**:6.

68. Lapierre, P. and Gogarten, J. P. 2009. Estimating the size of the bacterial pan-genome. *Trends in Genetics,* **25**(3):107–110.

69. Bazinet, A. L. 2017. Pan-genome and phylogeny of Bacillus cereus sensu lato. *BMC Evol Biol,* **17**(1):176.

70. Helgason, E., Okstad, O. a., Caugant, D. a., Johansen, H. a., Fouet, a., Mock, M., Hegna, I., and Kolstø, a. B. 2000. Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis–one species on the basis of genetic evidence.. *Applied and environmental microbiology,* **66**(6):2627–2630.

71. Tourasse, N. J., Helgason, E., Økstad, O. a., Hegna, I. K., and Kolstø, a. B. 2006. The Bacillus cereus group: Novel aspects of population structure and genome dynamics. *Journal of Applied Microbiology,* **101**(3):579–593.

72. Zwick, M., Joseph, S., Didelot, X., Chen, P., Bishop-Lilly, K., Stewart, A., Willner, K., Nolan, N., Lentz, S., Thomason, M., Sozhamannan, S., Mateczun, A., Du, L., and Read, T. 2012. Genomic characterization of the Bacillus cereus sensu lato species: backdrop to the evolution of Bacillus anthracis. *Genome Res.,* **22**(8):1512–24.

73. Helgason, E., Okstad, O. A., Caugant, D. A., Johansen, H. A., Fouet, A., Mock, M., Hegna, I., and Kolstø, A. B. 2000. Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis–one species on the basis of genetic evidence. *Appl Environ Microbiol,* **66**(6):2627–2630.

74. Priest, F. G., Barker, M., Baillie, L. W. J., Holmes, E. C., and Maiden, M. C. J. 2004. Population structure and evolution of the Bacillus cereus groups. *Journal of Bacteriology,* **186**(23):7959–7970.

75. Tourasse, N. J., Helgason, E., Okstad, O. A., Hegna, I. K., and Kolstø, A. B. 2006. The Bacillus cereus group: novel aspects of population structure and genome dynamics. *J Appl Microbiol,* **101**(3):579–593.

76. Rasko, D. a., Altherr, M. R., Han, C. S., and Ravel, J. 2005. Genomics of the Bacillus cereus group of organisms. *FEMS Microbiol Rev,* **29**(2):303–329.

77. Lai, Q., Liu, Y., and Shao, Z. 2014. Bacillus xiamenensis sp . nov ., isolated from intestinal tract contents of a flathead mullet (Mugil cephalus). *Antonie van Leeuwenhoek,* **105**:99–107.

78. Liu, Y., Lai, Q., Dong, C., Sun, F., Wang, L., Li, G., and Shao, Z. 2013. Phylogenetic diversity of the Bacillus pumilus group and the marine ecotype revealed by multilocus sequence analysis. *PLoS ONE,* **8**(11):1–11.

79. Branquinho, R., Meirinhos-Soares, L., Carrico, J. A., Pintado, M., and V, P. L. 2014. Phylogenetic and clonality analysis of Bacillus pumilus isolates uncovered a highly heterogeneous population of different closely related species and clones. *FEMS microbiol Ecol,* **90**:689–698.

80. Bhandari, V., Ahmod, N. Z., Shah, H. N., and Gupta, R. S. 2013. Molecular signatures for Bacillus species: Demarcation of the Bacillus subtilis and Bacillus cereus clades in molecular terms and proposal to limit the placement of new species into the genus Bacillus. *International Journal of Systematic and Evolutionary Microbiology,* **63**(7):2712–2726.

81. Bezuidt, O. K., Pierneef, R., Gomri, A. M., Adesioye, F., Makhalanyane, T. P., Kharroub, K., and Cowan, D. A. 2016. The Geobacillus Pan-Genome: Implications for the Evolution of the Genus. *Frontiers in Microbiology,* **7**(May):1–9.

82. Schmidt, T. R., Scott, E. J., and Dyer, D. W. 2011. Whole-genome phylogenies of the family Bacillaceae and expansion of the sigma factor gene family in the Bacillus cereus species-group.. *BMC genomics,* **12**:430–445.

83. Brown, A. M. V., Howe, D. K., Wasala, S. K., Peetz, A. B., Zasada, I. A., and Denver, D. R. 2015. Comparative Genomics of a Plant-Parasitic Nematode Endosymbiont Suggest a Role in Nutritional Symbiosis.. *Genome biology and evolution,* **7**(9):2727–46.

84. Manzano-Marín, A., Oceguera-Figueroa, A., Latorre, A., Jiménez-García, L. F., and Moya, A. 2015. Solving a Bloody Mess: B-Vitamin Independent Metabolic Convergence among Gammaproteobacterial Obligate Endosymbionts from Blood-Feeding Arthropods and the Leech Haementeria officinalis. *Genome Biology and Evolution,* **7**(10):2871–2884.

85. Paul, S., Bhardwaj, A., Bag, S. K., Sokurenko, E. V., and Chattopadhyay, S. 2015. PanCoreGen - Profiling, detecting, annotating protein-coding genes in microbial genomes. *Genomics,* **106**(6):367–372.

86. Jensen, G., Hansen, B., Eilenberg, J., and Mahillon, J. 2003. The hidden lifestyles of Bacillus cereus and relatives. *Environ Microbiol.,* **5**(8):631–40.

87. DeLong, E. F. and Karl, D. M. 2005. Genomic perspectives in microbial oceanography. *Nature,* **437**(7057):336–42.

88. Pedrós-Alió, C. and Pedrós-Alió, C. 2006. Genomics and marine microbial ecology. *International Microbiol,* **9**(3):191–197.

89. Zhang, N., Yang, D., Kendall, J. R. A., Borriss, R., Druzhinina, I. S., Kubicek, C. P., Shen, Q., and Zhang, R. 2016. Comparative Genomic Analysis of Bacillus amyloliquefaciens and Bacillus subtilis Reveals Evolutional Traits for Adaptation to Plant-Associated Habitats. *Front Microbiol,* **7**:2039.

90. Poretsky, R. S., Sun, S., Mou, X., and Moran, M. A. 2010. Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ Microbiol,* **12**(3):616–627.

91. Lauro, F. M., McDougald, D., Thomas, T., Williams, T. J., Egan, S., Rice, S., DeMaere, M. Z., Ting, L., Ertan, H., Johnson, J., Ferriera, S., Lapidus, A., Anderson, I., Kyrpides, N., Munk, A. C., Detter, C., Han, C. S., Brown, M. V., Robb, F. T., Kjelleberg, S., and Cavicchioli, R. 2009. The genomic basis of trophic strategy in marine bacteria.. *Proc Natl Acad Sci USA,* **106**(37):15527–15533.