UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

Análisis Bioinformático de genes núcleo y genes firma
en el phylum Firmicutes. Una aproximación pangenómica.

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

BIÓLOGO

P R E S E N T A :

TONATIUH ALVAREZ DEL CASTILLO ESTRADA

TUTORA:

PhD. GABRIELA OLMEDO ALVAREZ

FACULTAD DE CIENCIAS
UNAM

2009

A la mujer que más he amado,
a mi Nancy.
Ahora solo me resta idear
una eternidad con tu ausencia

A la mujer que amaré tanto como
a mi primer amor,
a mi Abril,
ansío ver la entrega
de tu sueño a mis brazos.
:)

Agradecimientos

A mi querida Gaby,
Los títulos nobiliarios siempre me fueron imposibles contigo,
te agradezco tu infinita paciencia, tu invaluable experiencia de vida
y los momentos llenos de alegría que me obsequiaste.

A mi muy estimado amigo Luis,
su invaluable paciencia, su incómoda impaciencia, sus desplantes alegres y sus objetivas críticas me hacen sentir satisfecho de tener un amigo que pocas veces estará de acuerdo conmigo,
Te deseo una larga vida en la que el amor y la salud
se sobrepongan a todo

A Varinia,
me mostraste la virtud del trabajo, el esfuerzo y el maldito temple que se debe tener en la vida para afrontar las vicisitudes que te regala.

A mis 3 familias:
Los Estrada y Álvarez del Castillo, que siempre estuvieron para forjar al hombre que digo ser ahora.
A mis Hernández Martínez, por todo el amor que siempre sobró en su casa para un irresponsable como yo, deseo únicamente dicha para todos ustedes, una dicha que se mida en universos de felicidad.

A mis amig@s en Irapuato, en el DF y en Hidalgo que tuvieron la paciencia para soportarme en los peores momentos.

A CONCYTEG,
pues por la beca.

A Buda,
de algún lado debió de haber llegado la paciencia, ¿o no?

Hoja de Datos del Jurado

Formato
1. Datos del alumno Álvarez del Castillo Estrada Tonatiuh 56777335 Universidad Nacional Autonoma de Mexico Facultad de Ciencias Biología 096220018
2. Datos del tutor Dra. Gabriela Olmedo Álvarez
3. Datos del sinodal 1 Dr Arturo Carlos II Becerra Bracho
4. Datos del sinodal 2 Dr René Cerritos Flores
5. Datos del sinodal 3 Biól. Germán Bonilla Rosso
6. Datos del sinodal 4 Biól. Luis David Alcaraz Peraza
7. Datos del trabajo escrito Análisis Bioinformática de genes núcleo y genes firma en el phylum Firmicutes. Una aproximación pangenómica. 78 p. 2009

Índice **Página**

1. Introducción.

Las bacterias han sido clasificadas dentro del enorme dominio conocido como Eubacteria, mediante una propuesta realizada por Woese, et. al., 1990; con el objetivo de proporcionar una nueva perspectiva del árbol de la vida que se encontraba segmentada en diversos reinos que eran considerados como “no naturales”. La presente clasificación es posible gracias a las nuevas técnicas moleculares en las que con una molécula presente en todos los individuos, la subunidad ribosomal de RNA 16s, se lleva a cabo dicho posicionamiento.

El ribosoma bacteriano es un complejo RNA-proteico encargado de sintetizar las proteínas necesarias para la célula, compuesto de dos diferentes subunidades tales como la subunidad 30S y la 50S. Cada una de estas subunidades se compone a su vez de otras subunidades menores cayendo en nuestro particular interés el RNA 16s, ya que éste fue utilizado como elemento principal de una nueva clasificación de los seres vivos.

El autor de dicha propuesta, Carl Woese fue quien eligió a la subunidad ribosomal 16S debido a su abundancia, a la codificación realizada tanto por organelos como por núcleos y genomas procariontes, por su muy especial conformación con zonas de rápida y lenta evolución y a que su conservación es casi total en todos los seres vivos (Woese, Kandler et al. 1990). Aunado a esto, podemos mencionar también su muy antigua y esencial función en la maquinaria molecular del organismo y su capacidad de interacción con por lo menos otras 100 unidades ribosomales y proteínas.

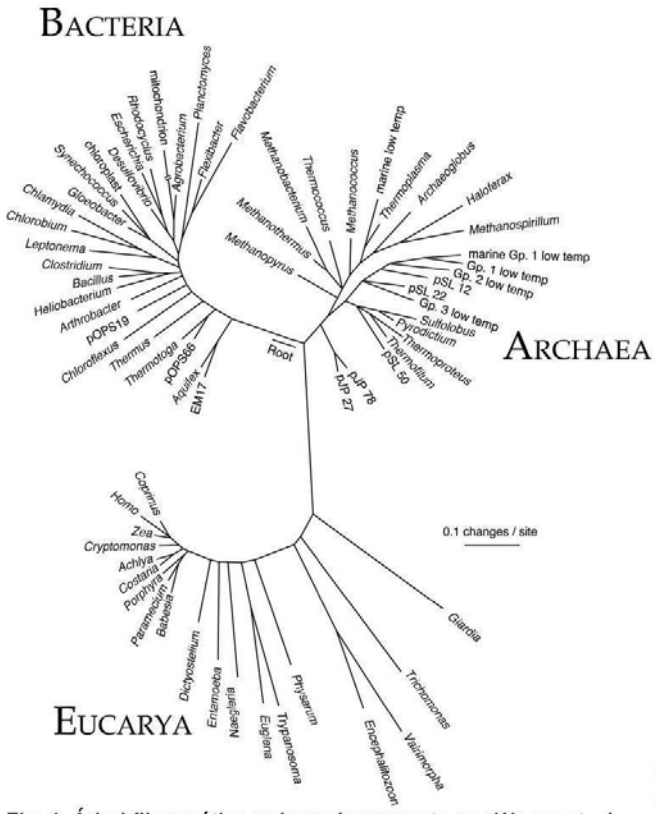

Fig. 1. Árbol filogenético universal propuesto por Woese et. al. 1990, mostrando los 3 dominios.
Tomado de http://bio.fsu.edu/~stevet/pictures/TheBigTree.jpg

El gen de la subunidad ribosomal de RNA16S se encuentra ampliamente conservado en las especies vivas, produce una molécula que, inicialmente, ha sido sumamente estudiada y es capaz de diferenciar a los 3 dominios con sencillas características: en las eubacterias presentan una estructura de “tallo-asa” en las posiciones 500 a 545 y con una protuberancia que sobresale de la estructura del tallo con una composición característica de 6 nucleótidos la cual, a su vez, presenta una diferencia con respecto a la “otra” hebra en el quinto y sexto nucleótido. Por su parte, arqueobacterias y eucariontes presentan esa misma protuberancia con ligeras diferencias, es de siete

nucleótidos, tiene una composición característica específica muy distinta a la de eubacterias y por último, la protuberancia sobresale del sexto y séptimo par. En el caso particular de los eucariontes su subunidad ribosomal carece de una estructura exclusiva localizada en la zona de los nucleótidos 585 a 655 sólo presente en arqueo y eubacterias. Mientras que en las arqueobacterias existe una estructura muy particular en las posiciones 180 a 197 y 405 a 498.

Phylum Firmicutes.

Los individuos clasificados en el phylum Firmicutes (ver Figura 2), son susceptibles a la tinción de Gram, catalogados por ello como Gram positivos (+), con una cantidad muy baja de Guanina y Citosina (contenido G+C); algunos de ellos son capaces de formar estructuras de resistencia conocidas como endosporas, la mayoría son capaces de producir enzimas extracelulares hidrolíticas que atacan a los individuos vecinos, permitiéndoles la obtención de materia prima para llevar a cabo intercambio electrónico o solo obtener una fuente de carbono y la distribución ecológica de la mayoría de éstas bacterias se reduce casi exclusivamente al suelo (Madigan, Martinko et al. 2003).
Existen varias teorías respecto a los procesos químicos que se producen al exponer a las bacterias a la tinción de Gram, sin embargo, lo importante a destacar en este punto es que debido a la constitución distinta de las paredes celulares se produce dicha coloración; en el caso de las bacterias Gram positivas (+) se tiñen a causa de la gruesa capa de peptidoglicanos (compuestos a su vez de N-acetilglucosamina y ácido N-acetilmurámico, acompañados de algunos aminoácidos tales como L-alanina, D-alanina, ácido D-glutámico e inclusive lisina o ácido diaminopimérico [DAP]) que recubren a la membrana celular, al contrario de las Gram negativas (-) cuya capa de peptidoglicanos es delgada (en ocasiones se llegan a presentar hasta 25) y se acompañan de ácido teicoico. Aunado a dicha estructura, por encima de la capa regular de peptidoglicanos y ácido teicoico existe una capa de lipopolisacáridos (LPS) con proteínas (Madigan, Martinko et al. 2003).
Las características de cada género incorporado al phylum Firmicutes son por demás diversas, variando extraordinariamente en relaciones ecológicas, metabolismo, proceso de esporulación o patogenicidad, por lo que es muy conveniente tratar a cada uno de los géneros de forma aislada así que se describirán, en dos partes, algunos de los representantes más destacados de los géneros *Bacillus, Clostridium, Mycoplasma* y *Listeria* (ver el Anexo I) y una breve descripción de los mismos y un recopilado (ver Tabla 1.1 a 1.4 en el Anexo I) donde se incluyen solo características generales de más individuos con información tal como contenido de Guanina-Citosina, tamaño del genoma, número de genes y la presencia de plásmidos.
Con el propósito de ir haciendo más ilustrativa la explicación de las bacterias con las que vamos a trabajar a continuación se muestra una figura que recopila gran cantidad de especies del phylum Firmicutes, su ancestría con parentescos y el tamaño hasta hoy registrado, cabe resaltar el hecho de que existen inconsistencias entre los valores de la siguiente figura y las Tablas 1.1 a 1.4 ya que

cotidianamente se hacen revisiones de las bases de datos donde están almacenados sus genomas y esto puede producir las variaciones entre los valores.

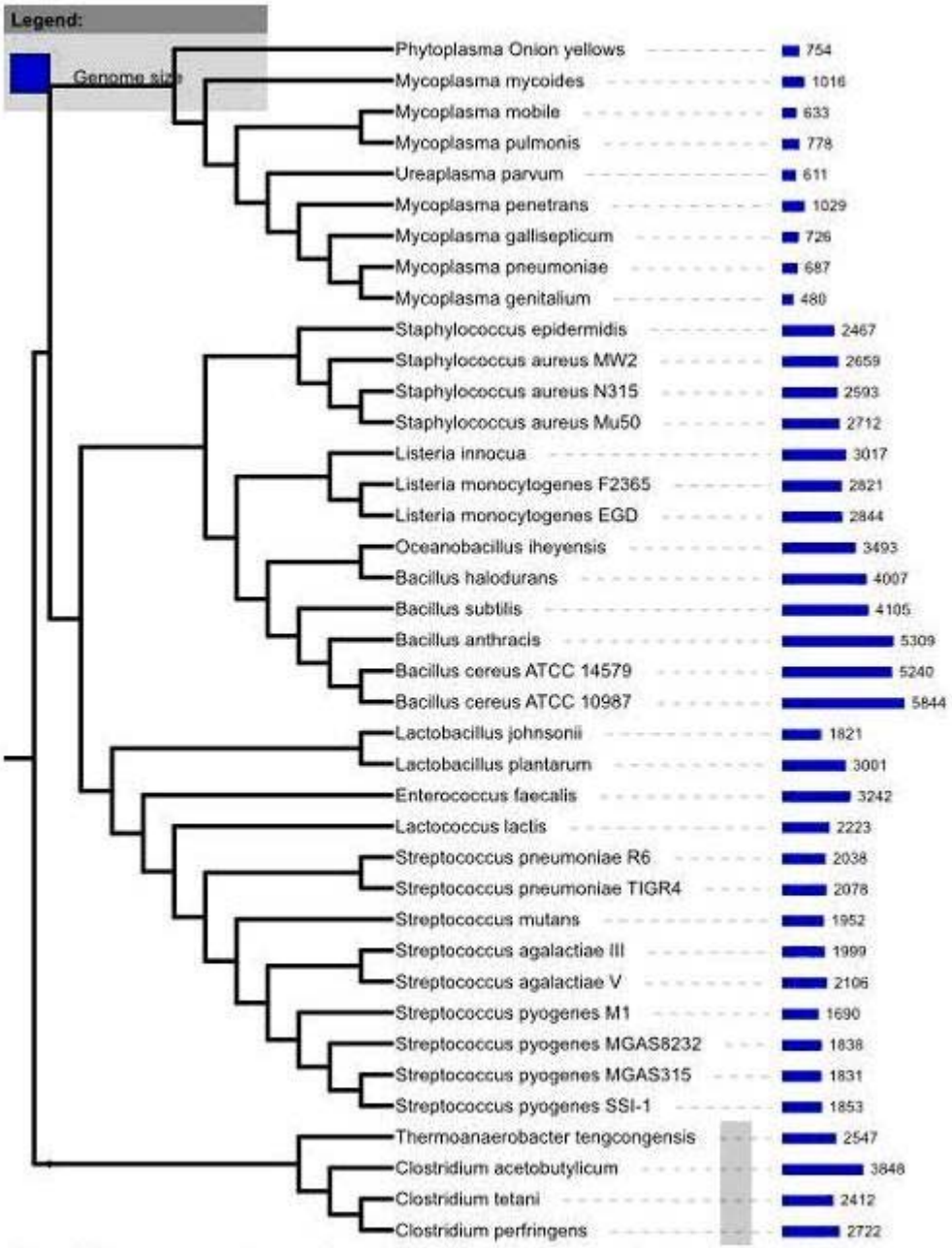

Figura 2. Imagen que muestra una filogenia del phylum Firmicute con las especies más estudiadas y características. En azul se ejemplifica el tamaño del genoma de cada una de las especies, dada en kilopares de bases (kpb).. (Filogenia realizada con ITOL en http://itol.embl.de/itol.cgi).

Endosporas y esporulación.

A lo largo de su historia evolutiva, los géneros *Bacillus* y *Clostridium*, desarrollaron una estrategia de supervivencia sumamente útil y práctica para sobreponerse a las variaciones del ambiente y la ausencia de fuentes de carbono, nitrógeno y fósforo: la formación de endosporas. Dichas estructuras

no son más que una célula hija incluida en la célula madre con diferencias estructurales importantes y con una resistencia extraordinaria al calor, la desecación, a agentes químicos tremendamente agresivos, a la radiación e inclusive al tiempo.

La formación y diferenciación de esta estructura es un proceso conocido como esporulación, que inicia con el desarrollo de la endospora en la célula vegetativa; proceso que llega a durar hasta 8 horas, para posteriormente llevar a cabo una autólisis de la célula madre y liberarla al medio con el propósito de que se propague por cualquier medio posible hasta encontrar un nuevo medio rico en nutrientes donde germinará y reiniciará el ciclo de vida normal de una bacteria (ver Tabla 2).

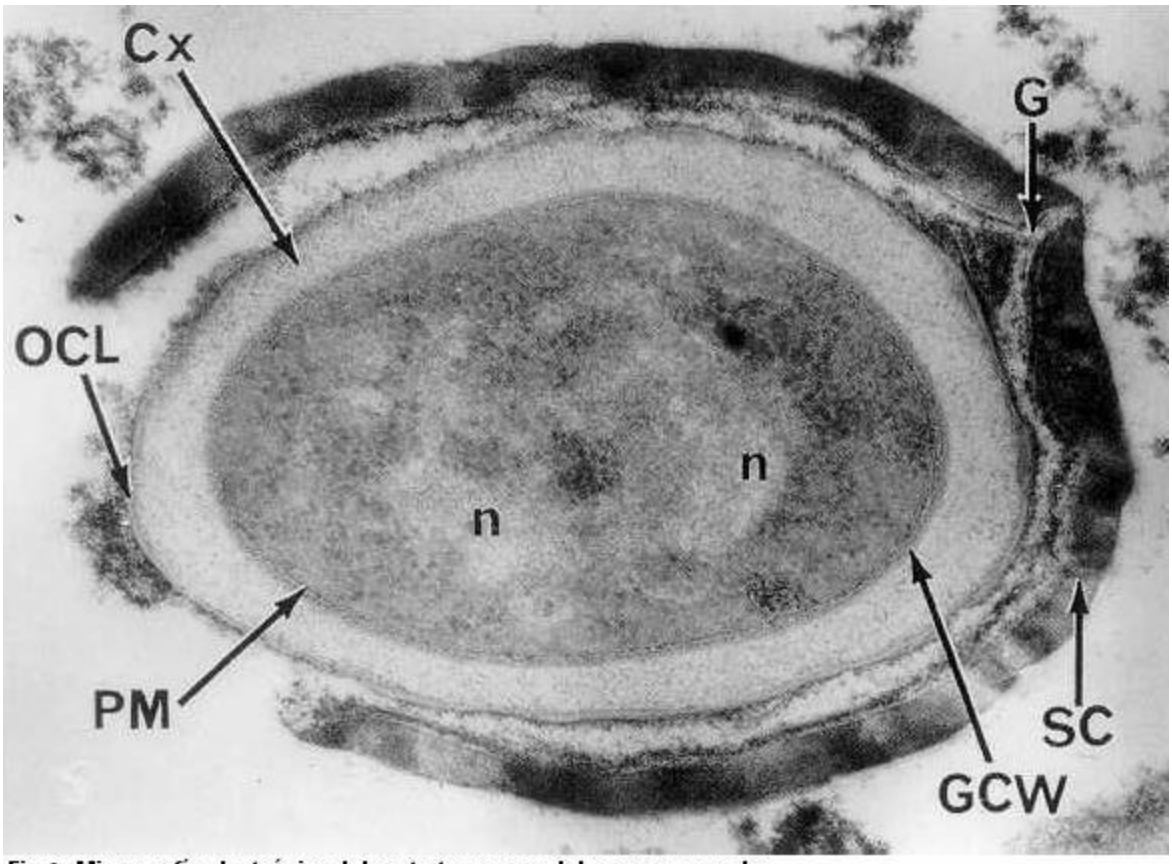

Fig. 3. Micrografía electrónica del corte transversal de una espora de B. megaterium, mostrando la cubierta de la espora (SC), la ranura germinal (G) en la cubierta de la espora, la capa exterior de la corteza (OCL), la corteza (C), la pared celular de la célula vegetativa (GCW), la membrana subyacente de la espora (PM) y regiones donde el nucleoide es visible.
Tomado de: http://www.gsbs.utmb.edu/microbook/ch002.htm

Las esporas son estructuras bacterianas hipometabólicas, de hecho poseen la tasa más baja de metabolismo de todos los organismos vivos; presentan una dormancia prácticamente inalterable y sumamente específica junto con un nivel de deshidratación muy por debajo de los niveles tolerables, con el propósito de mantener la viabilidad apropiada de la espora. En virtud de ésta característica, le es posible resistir altas temperaturas y radiaciones UV.

Núcleo de la endospora.

El núcleo de la endospora se encuentra sumamente deshidratado, conteniendo un reducido 10 a 30% del agua total de la célula, lo cual le confiere una consistencia semejante a un gel denso con una elevada termorresistencia, tolerancia a moléculas sumamente agresivas capaces de dañar el DNA como el peróxido de hidrógeno y mantiene inactivas a las enzimas presentes en su interior. Dicho gel se compone básicamente de iones Ca2+ y ácido dipicolínico, los cuales se unen en un compuesto característico y exclusivo de endosporas bacterianas llamado dipicolinato de calcio (DPC), que a su vez es la molécula responsable de disminuir la cantidad de agua en el núcleo, evitando la desnaturalización o inactivación de algunas de las enzimas del núcleo.

El pH se ve disminuido una unidad con respecto al de la célula madre, presentando grandes concentraciones de proteínas pequeñas ácido-solubles (SASP's) cuya función as adherirse al DNA y protegerlo de la acción nociva de radiaciones UV, desecación y calor seco y proporcionar una fuente de aminoácidos para la síntesis de proteínas durante la germinación.

Al igual que la bacteria madre, la espora posee su propia copia del cromosoma pero los elementos que componen el protoplasma difieren mucho de la primera; claro es el caso de la cantidad de ribosomas y algunos otros elementos de la maquinaria biosintética tales como co-factores o RNAt, cuya presencia es muy reducida. Se encuentran también algunas enzimas como la RNA polimerasa, las SASP's, nucleósidos mono y di fosfatados, ninguno trisfosfatado ya que éstos obtendrán su fuente de fósforo a partir del 3 fosfo-glicerato, que a su vez se transforma en 2 fosfo-glicerato y finalmente el donador de fósforo, el fosfo-enol-pirúvico.

Cubierta de la espora.

La composición de la cubierta puede variar dependiendo de la especia bacteriana sin embargo, es común encontrar una o varias proteínas de tipo queratina, ricas en cisteína y en aminoácidos de naturaleza hidrofóbica, que llegan a constituir hasta el 60% del peso seco de la espora. La adición de la cisteína es un proceso post-traduccional a través de una modificación de la proteína inmadura.

Las cubiertas son insolubles e impermeables que impiden la entrada de agentes químicos desnaturalizantes del DNA, enzimas de acción lítica capaces de degradar la corteza e inclusive depredación por parte de algunos protozoarios; sin embargo, a pesar de tal resistencia, es absolutamente incapaz de proferir algún tipo de resistencia al calor o las radiaciones. Gracias a la gran cantidad de puentes disulfuro, las cubiertas se presentan como estructuras compactas y químicamente muy estables.

Capa exterior de la corteza.

No es absolutamente clara la función específica de esta membrana localizada justo por encima de la corteza aunque la presencia de ésta es imperativa en la adecuada construcción de la endospora. Sin embargo, la membrana exterior pocas veces mantiene su integridad en esporas inactivas así que se sabe que no constituye una barrera de permeabilidad. De hecho, la eliminación de la capa exterior y

de la mayor parte de la corteza no tiene un efecto perceptible en la resistencia de la espora al calor, radiación o agentes químicos (Raju, Setlow et al. 2007).

Corteza.

La corteza se compone de peptidoglicano con una estructura muy similar a la de la forma vegetativa pero con algunas variaciones “esporo-específicas”. Es esencial para la formación de la espora y la reducción de la cantidad de agua en el núcleo sin embargo, dicho componente es degradado al momento de la germinación, paso elemental en la expansión del núcleo y subsecuente crecimiento. En el proceso de esporulación se involucran alrededor de 100 genes, los cuales no tienen una función exclusiva durante la esporulación sino son elementos más que tienen una utilidad normal durante la vida de la bacteria.

Germinación.

La germinación es el proceso en el que la espora termina su periodo de dormancia y regresa al estado de vida libre correspondiente al de una bacteria común. De acuerdo a Foster y Johnstone, 1990, es posible reconocer 4 etapas diferentes en este proceso:

Preactivación. Esta etapa sucede sencillamente por el proceso de degradación y envejecimiento natural de las cubiertas de la espora.

Activación. Se desencadena a causa de la presencia de un compuesto químico o bioquímico como iones inorgánicos, glucosa u otros azúcares, bases nitrogenadas o aminoácidos y es reconocido directamente por un receptor alostérico localizado en la membrana esporal interna. Al activarse, el receptor adopta una capacidad proteolítica que le permite degradar una proenzima que se hallaba unida de forma covalente al peptidoglicano de la corteza. El resultante enzimático reconoce la lactama del NAM y lleva a cabo la hidrolización del peptidoglicano cortical. Como consecuencia el agua circundante se introduce al protoplasma esporal, provocando la pérdida de las características conferidas por la deshidratación. En este punto el metabolismo aún se encuentra inactivo y la germinación es aún reversible.

Germinación. Debido a los drásticos cambios que empiezan a suceder en el interior de la espora como pérdida de Ca2+ y DPA o la hidrolización de las SASP's, el proceso germinativo es irreversible, a pesar de que el metabolismo aún sea endógeno. El calcio que se pierde comienza a migrar al córtex, neutralizando las cargas negativas, favoreciendo la rehidratación del protoplasma y su hinchamiento. Se lleva a cabo la síntesis de energía que inicia con el 3-fosfoglicerato pasa a 2-fosfoglicerato, posteriormente se transforma en PEP, el cual comienza a donar fosfatos de alta energía que servirán de cimientes en la producción de ATP. Por último, la RNA polimerasa comienza con su labor transcripcional.

Término y Crecimiento. La espora finalmente es capaz de tomar nutrientes del exterior y metabolizarlos, presenta un crecimiento pronunciado del protoplasma, la síntesis de DNA es

vigorosa y finalmente empieza a mostrar signos de ruptura de la cubierta esporal a causa de la degradación de la misma, cuyos componentes están siendo utilizados para formar la pared celular.

Bioinformática.

Actualmente se encuentran registradas más de 65 369 091 950 bases en las divisiones tradicionales del GenBank (www.ncbi.nlm.nih.gov/Genbank/index.html) por lo que la cantidad de información que existe en la ciencia Bioinformática es prácticamente imposible de administrar si no es con una computadora (Benson, Karsch-Mizrachi et al. 2006).

De acuerdo a la información provista por el Centro Nacional para la Información Biotecnológica (NCBI, por sus siglas en inglés), en específico su departamento conocido como GenBank, a partir del año 2001 la cantidad de información generada comenzó a superar cualquier esfuerzo por parte de un investigador a mantenerse actualizado sin la ayuda de herramientas computarizadas (Gibas and Jambeck 2001).

El común denominador de los bioinformáticos se encuentra en la gran cantidad de información que deben procesar y analizar, sin importar el origen de los datos. Sin embargo, dentro de las disciplinas moleculares existen varios puntos específicos que pueden ser tratados independientemente:

Alineamientos locales BLAST.

La Herramienta de Búsqueda de Alineamientos Locales Básicos (en inglés Basic Local Alignment Search Tool, BLAST) encuentra regiones de similitud local entre secuencias. Dicho programa compara secuencias de nucleótidos o de proteínas con secuencias almacenadas en bases de datos y calcula la significancia estadística de las correspondencias. BLAST puede ser usado también para inferir funcionalidad o relaciones evolutivas entre secuencias así como ayudar a identificar miembros de familias génicas.

Ésta herramienta fue desarrollada por Eugene Myers, Stephen Altschul, Warren Gish, David J. Lipman y Webb Millar en el Instituto Nacional de Salud (Altschul, Madden et al. 1997) (NIH, por sus siglas en inglés) basándose en el algoritmo Smith-Waterman. Realizar un proceso estándar de alineación de secuencias con dicho algoritmo resultaría en un proceso demasiado lento a causa del tamaño de la base de datos (GenBank), por ello se adiciona un enfoque heurístico que reduce el tiempo de trabajo hasta 50 veces pero igualmente reduce la precisión. Sin embargo, esa pérdida de precisión no constituye un problema serio ya que la correspondencia entre secuencias comparadas es también evaluada estadísticamente para comprobar su significancia.

El algoritmo BLAST lleva a cabo el análisis de datos en 3 pasos:

Primero crea una lista de todas las secuencias cortas que superan el valor del umbral cuando se alinean con la secuencia interrogante (conocidas como “palabras”, en lenguaje de BLAST). Después,

se recorre la base de datos en busca de correspondencias con la secuencia interrogante. Dado que el tamaño de las “palabras” es demasiado corto (3 residuos por proteína u 11 residuos por nucleótido) es posible buscar en una tabla precomputada de todas las palabras y sus posiciones en las secuencias para aumentar la velocidad. Éstas palabras coincidentes son extendidas en alineamientos locales sin espacios, entre la secuencia interrogante y la secuencia de la base de datos. Las extensiones continúan hasta que el valor del alineamiento queda por debajo del umbral. Los alineamientos con valor más alto en la secuencia o pares segmentados de mayor valor (MSPs, siglas en inglés) se combinan, dentro de lo posible, en alineamientos locales. Actualmente es posible realizar alineamientos locales con fragmentos incompletos (Gibas and Jambeck 2001).
Para información más detallada sobre herramientas más específicas consulte el Anexo III.

Pangenoma

Los términos descritos a continuación, constituyen una nueva serie de vocablos utilizados dentro de las disciplinas bioinformáticas y moleculares que fueron acuñados a partir de la necesidad de delimitar ciertas características de los genomas que hasta hace algunos años no eran consideradas. Desde el momento en que se registró la primera secuencia completa de un genoma hasta la actualidad, las bases de datos que poseen tal información han visto su contenido incrementarse de forma casi exponencial, lo que ha permitido llevar a cabo análisis más detallados y obtener espectaculares resultados mediante inferencias. A partir de dichas inferencias se resolvió acuñar un término plenamente utilizado en la actualidad por biólogos moleculares y bioinformáticas, el de “pan-genoma”, que posee el prefijo “pan” que viene del griego “παν” vocablo que significa “todo” (Tettelin, 2005), determinando así la naturaleza del significado o, en otras palabras, es la suma de los genes núcleo y de los genes dispensables. Cabe destacar que los genes núcleo son todos aquellos que están presentes en todas las cepas de la especie mientras que los genes dispensables son aquellos presentes en varias pero no en todas las cepas de la especie. Dentro de los últimos podemos mencionar otra subclasificación, que incluye a los genes únicos que, como su nombre lo indica, se encuentran en una sola cepa (Tettelin, Masignani et al. 2005).
El pan-genoma varía constantemente gracias a la anotación de nuevos genomas que agregan una cantidad completamente nueva de genes, dependiendo de la especie. Hay casos particulares en los que con cada nueva secuenciación se incorporan varios genes nuevos, en el caso de *Streptococcus* se han registrado hasta 33 genes nuevos, denominando a éste como pan-genoma “abierto”, mientras que en otras especies tal número rápidamente tiende a cero, el caso de *Bacillus anthracis* es sumamente claro, dicho evento sucedió poco después de analizar cuatro genomas, designando tal genoma como un pan-genoma “cerrado”, debiéndose a causas tan diversas como nichos ecológicos

escasamente cambiantes o una mínima capacidad de adquirir material genético exógeno (Medini, Donati et al. 2005).

El incremento o estabilidad del pangenoma puede verse afectado, en el caso particular de las bacterias, por los diferentes procesos de transferencia de información genética tales como la transformación (cuando el material genético es tomado directamente del medio ambiente), la transducción (en este caso el material genético es transferido específicamente por un fago) y por conjugación (que sucede cuando las bacterias llevan a cabo un intercambio de material genético directamente entre si) (Medini, Donati et al. 2005).

La magnitud en que puede aumentar el pan-genoma bien puede verse bajo una luz distinta si tomamos en consideración lo siguiente: actualmente se sugiere que existen alrededor de 1031 fagos, responsables de más de 1023 infecciones por segundo. Considerando la cantidad de material genético que posee un fago, de entre 5,000 a 500,000 pares de bases (Madigan, Martinko et al. 2003), una increíble cifra de un orden cercano a los miles de millones de genes no se encontraría alejado de la realidad. Dicho proceso no es desconocido por la ciencia, desde la década de los 50, cuando fue descubierto, se exploró ampliamente el potencial de la transducción, que es simplemente la transferencia de DNA de una bacteria a otra teniendo como vector un fago (Madigan, Martinko et al. 2003). Sin embargo, el actor al que hacemos referencia, los fagos, no son los únicos que intervienen en un proceso que bien puede ser catalogado como normal dentro de los microecosistemas bacterianos, tales participantes pueden ser plásmidos, integrones y transposones (Szpirer, Top et al. 1999).

Típicamente el segmento del pangenoma denominado como genoma núcleo ha sido acuñado y utilizado por gran cantidad de autores como todos aquellos genes que proporcionan las características básicas de un organismo determinado, junto con sus características fenotípicas más comunes y están presentes en todos los integrantes de un taxón en particular (Rooney, Swezey et al. 2006). Tales características podrían agruparse en funciones de mantenimiento celular básico, membrana y/o pared celular, funciones regulatorias y de enlace y transporte de proteínas.

Por otra parte, todos aquellos genes que le confieren características particulares y ventajas selectivas como adaptación a diversos nichos ecológicos, resistencia a antibióticos o la colonización de un hospedero funciones que no son esenciales en el crecimiento normal de la bacteria, son reconocidos como genoma accesorio. Dichos genes se encuentran comúnmente agrupados en grandes islas genómicas, encontrándose a su vez flanqueados por fragmentos de DNA que poseen niveles de G+C sumamente diferentes.

Los procesos de selección que actúan sobre las distintas regiones difieren tremendamente entre sí, las zonas del genoma núcleo tienden a estar menos sujetas a las presiones selectivas ya que la presencia de toda la maquinaria molecular que contiene es esencial, mientras que las regiones del

genoma accesorio presentan constantes eventos de modificación a causa de una constante exposición a la transferencia lateral de genes (Medini, Donati et al. 2005).

Filogenómica.

La palabra Filogenómica fue utilizada por primera vez en 1998 en el contexto de proveer “un acercamiento a la predicción de la función del gen” para datos a escala genómica y poco después en el contexto de la inferencia filogenética (Philippe and Blanchette 2007). Recientemente la disciplina de la filogenómica, que conjunta áreas de la biología como la evolución y la biología molecular, ha visto su campo de acción ampliado gracias al progreso de tecnologías relacionadas con la secuenciación de DNA y, por consiguiente, un incremento en el número de anotaciones de genomas completos.
Su propósito inicial es el de utilizar los datos genómicos para inferir relaciones filogenéticas y ampliar el entendimiento de los procesos de evolución molecular. Por otra parte, busca inferir funciones putativas para secuencias de proteínas o de DNA mediante comparaciones de muchas especies y su filogenia (Eisen 1998). Así mismo, aunque menos socorrido por los investigadores, el análisis evolutivo es una herramienta poderosa en estudios de secuencias genómicas, proveyendo de perspectiva al análisis y facilidad de interpretación de los resultados (Eisen and Fraser 2003).
Previos a la era genómica los análisis filogenéticos que se realizaban eran muy limitados, ya que únicamente se tomaban en cuenta algunos genes o una serie de éstos, hasta que las anotaciones de genomas completos se incrementó tremendamente permitiendo darle enfoques a escala genómica, con una precisión y puntualidad mucho más finas que antes (Eisen 1998).
El enfoque genómico requiere de la compilación de gran cantidad de datos que incluyeran gran cantidad de loci de muchas especies. Tales grupos de datos son menos propensos a errores de muestreo o de sistematización y ofrecen la posibilidad de analizar cantidades enormes de características filogenéticas informativas de diferentes ubicaciones genómicas junto con la opción de corroborar dichos datos variando las especies muestreadas.
Para hacer uso de la identificación de la semejanza de secuencias entre genes, es muy útil comprender cómo surge dicha similitud. Los genes pueden hacerse idénticos en sus secuencias ya sea como resultado de una convergencia (similitudes que han surgido sin una historia evolutiva en común) o con modificación a partir de un ancestro común (también conocida como homología). Es imperativo reconocer que similitud de secuencias y homología, no son sinónimos ya que no todos los homólogos poseen secuencias idénticas (algunos divergen de tal manera que es difícil, si no imposible, reconocer las diferencias) y no todas las similitudes son debidas a homologías (Eisen 1998).
Las homologías, surgidas a partir de procesos de duplicación y una posterior divergencia, pueden dar origen a funciones distintas en un mismo organismo y existen términos para cada uno de éstos: los genes ortólogos, que se refieren a aquellos homólogos que sufrieron de un proceso de divergencia

después de una especiación; mientras que los parálogos son aquellos que divergieron justo después de un proceso de duplicación génica. Por otra parte, todos los genes que han divergido entre sí después de haber pasado por un proceso de transferencia lateral son conocidos como xenólogos (Eisen 1998).

El primer paso para llevar a cabo un estudio sobre la evolución de un gen es la identificación de homólogos mediante la comparación con otros homólogos previamente identificados que se encuentran en bases de datos.

La recopilación de datos para definir los marcadores filogenéticos requiere de criterios muy severos ya que no todos los loci poseen la señal histórica apropiada sin embargo, deben cumplir con por lo menos los siguientes tres criterios:

Los genes ortólogos deben ser fáciles de identificar y amplificar en todos los taxa de interés,
se deben de identificar todos aquellos genes que posean exones muy largos (más largos que el umbral práctico determinado actualmente por la tecnología de secuenciación de DNA, por ejemplo 800 pb) y se deben identificar genes razonablemente conservados, es decir, todos aquellos genes con bajos rangos de evolución son menos propensos a acumular homoplasias (Li, Ortí et al. 2007).

2. Justificación y Antecedentes

El género *Bacillus* es uno de los géneros más estudiados debido a la gran cantidad de representantes del grupo que se encuentran involucrados en actividades económicas de gran importancia como la industria alimentaria, química y de salud. Por consecuencia, la cantidad de estudios de cualquier índole a propósito de estas bacterias, crece cotidianamente, permitiéndonos comprender cada vez más el complejo entramado que constituye su mera existencia.

Souza y colaboradores (2006) han reportado que en la Cuenca de Cuatro Ciénegas existe un enorme sistema de pozas, arroyos y manantiales que conservan características relictuales de un mar Jurásico, con notoria escasez de compuestos de fósforo pero una generosa presencia de sales y sulfatos. La diversidad microbiológica de la zona ha sido caracterizada mediante la secuenciación de genes de rRNA 16S y demuestra que casi la mitad de los filotipos de la Cuenca de Cuatro Ciénegas están relacionadas directamente con bacterias de hábitats marinos (Souza, Espinosa-Asuar et al. 2006). Dentro de este sistema se aislaron distintas cepas del género *Bacillus*, dentro de estos aislados sobresale la cepa M4-4.

Bacillus coahuilensis M4-4 es una nueva especie descrita recientemente (Cerritos, Vinuesa et al. 2008) que a partir de la secuenciación de su genoma se descubrió que contaba con el genoma más pequeño reportado para cualquier especie del género *Bacillus*, únicamente 3,640 genes codificantes, un contenido G+C de 37% y 905 genes hipotéticos no compartidos con otros organismos en las bases de datos, además de contar con adaptaciones a su medio extremo como la síntesis de sulfolípidos y la posibilidad de censar el ambiente a través de una rodopsina, estos genes son de origen exógeno, probablemente de Cianobacterias (Alcaraz, Olmedo et al. 2008). Se plantea la hipótesis de que el tamaño del genoma de B. coahuilensis es producto de una reducción genómica mediada por secuencias de inserción (IS), transposones (Tn) y elementos repetitivos, como resultado a una adaptación y especialización de nicho

Ahora bien, el reducido número de genes de B. coahuilensis nos permite preguntarnos que genes son los que definen al género como unidad. Si se logra obtener el listado de ortólogos compartidos entre todos los *Bacillus* se puede llegar a reconstruir la historia evolutiva del género, además de entender elementos compartidos de la biología del género, con formas de vida que van desde organismos de vida libre a extremófilos y patógenos.

Por otro lado, si bien se trata de describir la cohesión en el género, mediante el análisis del genoma núcleo, y el género cuenta con tan distintos hábitats, esta diversidad de hábitats puede ser mejor comprendida mediante el inventario de la totalidad de los genes incluidos en el género, con el Pangenoma. El pangenoma del género *Bacillus* nos habla del potencial metabólico del género, siendo difícil una reconstrucción histórica porque eventos como la transferencia horizontal de genes

obscurecen la historia evolutiva, pero nos habla de adaptaciones particulares a nichos y estilos de vida de cada especie en particular.

Se han realizado trabajos tratando de averiguar la cantidad mínima de genes para que un organismo sea funcional, mediante mutagénesis se ha determinado estimaciones de genes esenciales para sobrevivir en una condición muy específica con *B. subtilis* (192 genes) y algunas otras bacterias como Mycoplasma hasta llegar a la síntesis química de este genoma mínimo por el grupo de Venter y sus colaboradores (Kobayashi, Ehrlich et al. 2003; Glass, Assad-Garcia et al. 2006; Gibson, Benders et al. 2008).

La mitad de los 192 genes esenciales de *B. subtilis* se involucran en el procesamiento de información (replicación, transcripción y traducción), una quinta parte en membrana y división celular, una décima parte a genes relacionados en requerimientos energéticos. Solo el 4% de los genes esenciales son hipotéticos (Kobayashi, Ehrlich et al. 2003). Dichos genes son para condiciones físicas específicas de crecimiento. Resulta interesante analizar que tan flexible es el núcleo de genes compartidos entre todos los *Bacillus* ya que la cantidad de genes compartidos dentro del género reportados, oscila en los 350 genes (Takami, Takaki et al. 2002), aunque los métodos de comparación no fueron extensivos para determinar ortólogos en dicho caso, pero es claro que la cantidad de genes conservados excede a los esenciales.

La clasificación actual divide al género en cepas patógenas como *B. anthracis* y *B. cereus* por un lado y aislados ambientales muchos de los cuales pueden ser de importancia industrial como *B. licheniformis* y cepas como *B. subtilis* que ha servido como organismo modelo de los Gram +. La diversidad de *Bacillus* va más allá de *B. subtilis* y si bien es el organismo mejor estudiado dentro del género resulta arriesgado hacer generalizaciones ecológicas y tratar de aplicar lo que se conoce en *B. subtilis* a todos los demás *Bacilli* (Ravel and Fraser 2005; Earl, Losick et al. 2008).

Ampliando la búsqueda de todos los ortólogos compartidos en todas las especies de *Bacillus* secuenciadas nos permitirá la identificación de genes necesarios para sustentar la vida dentro de este género además de poder definir con múltiples marcadores moleculares al género y entender sesgos funcionales que definan a los *Bacillus*.

Actualmente no existen trabajos que enlisten los genes del genoma núcleo del género *Bacillus* y, por otra parte, trabajos relacionados con el pangenoma del género solo ha sido hechos para la cepa de *Bacillus anthracis*, con indicios de ser un pangenoma cerrado (Keim, Price et al. 2000; Sacchi, Whitney et al. 2002).

3. Objetivo general.

Mediante la determinación de un genoma mínimo del género *Bacillus* realizado a partir de la comparación de los distintos genomas disponibles en bases de datos electrónicas, proponer un esquema que resalte la cantidad mínima de genes necesarios para la supervivencia de un hipotético ancestro común, así mismo la relación de ortólogos compartidos entre los integrantes del género y finalmente un desglose del potencial metabólico a partir del genoma accesorio.

Objetivos particulares.

- Definir el grupo de genes compartidos entre los integrantes del género *Bacillus* a diferentes niveles de restricción, partiendo desde el más estricto (compartición entre todos los integrantes) hasta el más laxo (compartición hasta entre 13 de las especies analizadas).
- Proponer un mapa metabólico de dicho genoma mínimo a distintos niveles de restricción para diferenciar potenciales metabólicos a distintos grados de restricción comparativa.
- Definir la composición del genoma del género *Bacillus* a partir de la adición de todos las secuencias genómicas disponibles para el género en la base de datos del NCBI.
- Proponer un mapa metabólico del genoma total del género *Bacillus* que permita identificar las diferencias genéticas existentes entre éste género y otros integrantes del phylum Firmicutes.

4. Metodología.

La descarga de los genomas se realizó durante octubre de 2007 de la base de datos de NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). Los genomas utilizados en el presente trabajo son los siguientes:

Tabla 4.1. Nombre reportado en la base de datos de NCBI para los genomas utilizados en el análisis del presente trabajo.

B. subtilis subsp. *subtilis str. 168*	*O.iheyensis* HTE 831
B.anthracis Sterne	*B. anthracis Ames* 0581
B.cereus E33L ZK	*B.thuringiensis serovar konkukian* 97-27
B. halodurans C-125	*G.kaustophilus* HTA426
B. cereus ATCC 14579	*B. anthracis* A2012
B. clausii KSM-K16 *incompleto*	*B. licheniformis* ATCC 14580
B. anthracis Ames	*Bacillus sp.* NRRL B-14911 *incompleto*
B. amyloliquefaciens FZB42	*B. pumilus* SAFR-032
B. anthracis str. *'Ames Ancestor'*	*Bacillus sp.* B14905 *incompleto*
B. anthracis str. Sterne	*Bacillus sp*. SG-1 *incompleto*
B. cereus subsp. *cytotoxis* NVH 391-98 *incompleto*	*B. thuringiensis serovar israelensis* ATCC 35646 *incompleto*
B. cereus ATCC 10987	*B. thuringiensis* str. *Al Hakam*
B. coagulans 36D1 *incompleto*	*B. weihenstephanensis* KBAB4
B. coahuilensis M4-4 *incompleto*	

La información sobre número de acceso, estatus de secuenciación, total de genes, número de proteinas, tamaño del genoma, contenido GC, tipo de hábitat, temperatura óptima y tolerancia a salinidad pueden ser consultados en la Tabla 4.2 del Anexo IV.

Obteniendo información de los genomas.

De cada genoma se obtienen las secuencias traducidas en un archivo multifasta. Dichas secuencias son formateadas para usarlas como base de datos para Blast con el siguiente comando:

```
$formatdb -$(archivo) –o T –p T
```

Con lo que se generan archivos que indexan la base de datos y nos permite hacer búsquedas en bases de datos definidas localmente.

Consecuentemente los Blasts bidireccionales se hicieron utilizando los siguientes parámetros:

```
$ blastall -p blastp -i $archivo_de_entrada -d $BLASTDB -m8 -e 1e-10 > salida.bout
```

Posteriormente fueron filtrados mediante líneas de comando de Perl las cuales se describen con detalle en el Anexo V.

Con los resultados de los mejores hits bidireccionales se analiza que genes se están compartiendo entre todos los genomas, mediante el comando diff, filtrando los resultados a una tabla de Excel y viendo qué genes son los que se comparten entre todos los *Bacillus*.

Para proporcionar un sentido de funciones globales de los genes analizados se utilizó el COG, para obtener las anotaciones COG por cada gen se utilizó el archivo ptt descargado del GenBank, una tabla de anotación del genoma que contiene información de la especie, tamaño del genoma, coordenadas de cada gen, identificadores y la anotación COG por cada gen. Un ejemplo, con las primeras dos lineas del archivo ptt del genoma de *B. subtilis* se muestra a continuación:

Bacillus subtilis subsp. subtilis str. 168, complete genome - 1..4214630
4105 proteins

Location	Strand	Length	PID	Gene	Synonym	Code	COG	Product
410..1750	+	446	1607709	dnaA	BSU00010	-	COG0593L	chromosomal replication initiation protein
1939..3075	+	378	1607700	dnaN	BSU00020	-	COG0592L	DNA polymerase III subunit beta

Cabe mencionar que la información de los archivos ptt que contienen la anotación COG solo se encuentra disponible de manera pública para los genomas completos. En el caso de los genomas en progreso utilizados en el análisis, realizamos una anotación manual de cada COG, mediante el uso de Blasts bidireccionales contra la base de datos COG (ftp://ftp.ncbi.nih.gov/pub/COG/COG/). De manera local con los siguientes parámetros:

```
$blastall -p blastp -i NOMBREDELAENTRADA.faa -d  /media/hda5/ncbi_database/COG/myva
-m 8 -e 1e-6 >NOMBREDELASALIDA.bout
```

Cabe mencionar la utilización de una categoría extra no incluida en los COGs originales que se refiere a los hipotéticos conservados, genes sin función reportada, pero conservados en distintas especies, en este trabajo se denominan la categoría “Y”.

Hecho el blast se genera una lista con los identificadores COG para cada proteína parecida a esto:

Tabla 5.1. Identificadores COG en código numérico de acuerdo a cada proteina

gi|30018746|ref|NP_830377.1|	BH3839	42.54	503	280	5	38	535	38	536	5.00E-117	420
gi|30022322|ref|NP_833953.1|	BS_comGA	55.91	347	152	1	2	347	1	347	6.00E-112	402
gi|30021671|ref|NP_833302.1|	BS_narQ	56.69	254	108	2	8	261	9	260	3.00E-079	293

Mediante los comandos de UNIX grep, sed y cat se asignó un COG general a una clase COG particular (ver Tabla 5), esto se logro usando una línea de comando del estilo siguiente:

```
$grep -lr BS_yvgX ./cog/  | sed 's/^/cat /g' |bash | grep 'COG'
>>nombredelasalida.COGS.txt
```

Con los datos de las salidas bidireccionales de Blast con respecto a *B. coahuilensis* M4-4 de 15 genomas (*B. subtilis* subsp. *subtilis* str. 168, *B.anthracis Sterne*,*B.cereus* E33L ZK, *B. halodurans* C-125, *B. cereus* ATCC 14579, B. clausii KSM-K16, B. anthracis Ames, O.iheyensis HTE 831, *B. anthracis Ames* 0581, *B.thuringiensis* serovar konkukian 97-27, *G.kaustophilus* HTA426, *B. anthracis* A2012, *B. licheniformis* ATCC 14580 y *Bacillus*. sp. NRRL B-14911) generamos una base de datos, donde eliminamos redundancia con cepas de especies como *B. anthracis*, *B. cereus* y *B. thuringiensis* que cuentan con varios representantes secuenciados sin que esto implique que existan más genes nuevos, incluso de la comparación de cepas de *B. anthracis* surge la idea del pangenoma cerrado (Keim, Price et al. 2000; Sacchi, Whitney et al. 2002) donde por más genomas nuevos que se secuencien de parientes cercanos no se encuentran más genes nuevos (a un nivel significativo). Esta base de datos se exportó a Excel para hacer una comparativa de que genes eran compartidos entre todos los *Bacillus,* mediante funciones VLOOK y usando como matriz las salidas de los blast bidireccionales se construyó una tabla dinámica. Dicha tabla se hizo con base en la anotación del genoma de *B. coahuilensis*, el más pequeño de los genomas de *Bacillus* secuenciados a la fecha, como base, comparando la presencia/ausencia de cada gen contra todos los demás genomas analizados, y utilizando además la clasificación KEGG de los genes presentes. Un extracto de dicha tabla dinámica se muestra a continuación:

Tabla 5.2. Ejemplo de tabla de contingencia con identificador de COG en código del KEGG y su compartición con el resto de las cepas del género *Bacillus.*

ID	Anotation	COG	Number of shared	Bcoah_bsu	Bcoah_bha	Bcoah_ban	Bcoah_bar	Bcoah_baa	Bcoah_bat	Bcoah_bce	Bcoah_bcz	Bcoah_btk	Bcoah_bli	Bcoah_bld	Bcoah_bcl	Bcoah_oih	Bcoah_gka	Bcoah_nrrl
M4	(3R)-hydroxymyristoyl-			K02	K02	K02	K02	K02	K02	K02	K02	K02	K02	K02	K02	K02	K02	K02
403	[acyl carrier protein]			372	372	372	372	372	372	372	372	372	372	372	372	372	372	372
340	dehydratase [H]	I	15															

Con base en la tabla anterior se pudo determinar qué genes se encuentran presentes en todos los Bacilli (columna number of shared). Si el gen se encuentra en todos los Bacilli se considera como parte del genoma núcleo, siendo este el nivel más estricto para considerar un gen parte del núcleo. El nivel de restricción va del 15 (TODOS) hasta el “corte 0” (genes únicos en B. coahuilensis).

Mapeo metabólico del pangenoma

En la página web del KEGG (Kyoto Enciclopedia of Genes and Genomes): http://www.genome.jp/kegg/ se utilizó la herramienta KAAS (KEGG automatic annotation server) para efectuar la reconstrucción. Dicha herramienta únicamente requiere de que se le “suba” una secuencia o, en este caso, un archivo multifasta; se elige la opción BBH o “best bi-directional hit” para la

elección de ortólogos, se proporciona un correo electrónico para recibir la información actualizada de la solicitud de trabajo y se eligen las opciones “Representative Sets” y “For genes” ya que solo requerimos trabajaremos con el grupo específico de genes de *Bacillus.*.

El archivo multifasta utilizado para dicha reconstrucción incluyó todas las secuencias traducidas en aminoácidos de todos los *Bacillus*. De esta manera podemos crear un mapa general donde tengamos idea del potencial metabólico de todos los *Bacillus* así como del grado de redundancia en cada ruta metabólica.

5. Resultados y Discusión.

Para llevar a cabo una correcta identificación de los genes incluidos en el estudio se construyó una tabla, que se muestra a continuación:

Tabla 6. Grupos de COG´s y las características generales de cada una de las clases.

A	Procesamiento y modificación del RNA
B	Estructura y dinámica de la cromatina
C	Producción y conversión de energía
D	Control del ciclo celular, división celular y segmentación de cromosomas
E	Transporte y metabolismo de aminoácidos
F	Transporte y metabolismo de nucleótidos
G	Transporte y metabolismo de carbohidratos
H	Transporte y metabolismo de coenzimas
I	Transporte y metabolismo de lípidos
J	Traducción, estructura ribosomal y biogénesis
K	Transcripción
L	Replicación, recombinación y reparación
M	Biogénesis de pared celular/membrana/cubierta celular
N	Movilidad celular
O	Modificación post-traduccional, intercambio de proteínas y chaperoninas
P	Transporte y metabolismo de iones inorgánicos
Q	Biosíntesis, transporte y catabolismo de metabolitos secundarios
R	Función general únicamente predicha
S	Función desconocida
T	Mecanismos de transducción de señales
U	Tráfico intracelular, secreción y transporte vesicular
V	Mecanismos de Defensa
X	Hipotéticos
Y	Estructura nuclear
Z	Citoesqueleto

Cabe destacar que la tabla anterior es muy importante en la interpretación de datos ya que prácticamente todos los resultados se basarán en la interpretación de las clases COG.

Genoma núcleo. Se construyó una tabla donde todos los genes de cada una de las bacterias se comparan contra los genes de *B. coahuilensis* M4-4, con el propósito de realizar tablas de presencia y ausencia a diferentes niveles de restricción es decir, genes que se compartieran entre las 15 especies, después entre 14 y así sucesivamente hasta poder determinar tanto los genes únicos de *B. coahuilensis* como los genes que componen el genoma núcleo del género (véase la Tabla 7). Hecho esto se efectuaron re-arreglos de la tabla para contabilizar e identificar los genes de las diferentes clases COG y el nivel de restricción, obteniendo así los siguientes resultados:

Tabla 7. Tabla de los COGs compartidos en los 14 genomas incluidos en el estudio. El corte 15 corresponde al nivel más alto de restricción y se distribuyen las presencias por clase COG.

COG Corte	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	Dif 1-15	%
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
B	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100.00
C	39	53	61	68	70	73	74	75	78	80	80	81	87	92	93	54	58.06
D	15	18	20	22	22	22	22	22	22	23	23	23	23	24	25	10	40.00
E	53	82	104	113	120	126	130	133	139	143	145	147	150	156	165	112	67.88
F	38	42	47	47	47	48	48	48	50	52	53	54	55	55	56	18	32.14
G	24	39	48	52	54	62	65	66	67	70	73	80	81	86	90	66	73.33
H	34	46	53	54	56	56	56	57	57	57	57	59	60	63	65	31	47.69
I	19	34	43	45	48	53	54	59	59	60	60	62	63	73	79	60	75.95
J	68	80	86	88	98	100	102	104	106	106	107	109	114	117	118	50	42.37
K	25	34	42	45	49	53	55	55	58	58	60	61	64	65	70	45	64.29
L	49	58	66	70	74	77	82	85	88	88	88	90	92	96	111	62	55.86
M	30	37	42	49	51	54	54	56	56	58	59	64	69	77	85	55	64.71
N	13	18	20	21	24	25	26	26	30	32	33	35	36	39	43	30	69.77
O	22	30	31	40	44	45	48	48	49	51	51	51	52	54	59	37	62.71
P	13	22	34	42	45	48	52	54	55	58	61	66	68	80	88	75	85.23
Q	0	0	0	1	1	1	2	4	4	5	5	8	8	14	21	21	100.00
R	29	40	46	63	84	95	103	110	117	119	123	129	138	154	178	149	83.71
S	5	7	12	12	14	14	14	14	16	18	18	19	20	24	26	21	80.77
T	11	21	26	30	31	33	34	37	38	42	44	48	54	59	61	50	81.97
U	7	9	10	10	11	11	11	11	11	11	11	11	12	13	13	6	46.15
V	0	8	13	17	18	21	23	25	27	27	28	32	33	38	42	42	100.00
X	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	100.00
Y	16	20	29	35	37	43	45	46	49	52	53	55	58	63	72	56	77.78
Total	**510**	**699**	**834**	**925**	**999**	**1061**	**1101**	**1136**	**1178**	**1212**	**1234**	**1286**	**1339**	**1444**	**1562**		

Los resultados obtenidos en el análisis de los 15 genomas muestran la presencia de los genes en comparación con los de *B. coahuilensis* es decir, comparando el genoma completo de esta cepa y contra los genomas de las otras 14 bacterias podemos definir la cantidad de genes que se comparten, así, podemos tener la certeza de que el total del corte 15 (el nivel de restricción más alto) contiene los genes compartidos por todas las cepas. Por otra parte, el nivel de restricción 14 nos habla de un total de 189 genes más que no se comparten entre todas las cepas, solo entre catorce de ellas. Los resultados mostrados en la columna de Dif 1-15 provienen de la resta de todos los genes acumulados desde el corte 14 hasta el corte 1 menos los genes del corte 15, así mismo el porcentaje que se encuentra en la siguiente columna solo es una representación porcentual de la variación

existente entre las cifras antes mencionadas. Éste número que podría hablarnos de la flexibilidad del pool de genes de cada una de las categorías. Por consiguiente es importante resaltar que en la tabla 7, en un sombreado de gris claro, se muestran las categorías COG relacionadas directamente con el transporte de biomoléculas, específicamente aminoácidos, nucleótidos, carbohidratos, iones inorgánicos y coenzimas; prestando especial atención a la categoría “E”, que es una de las que presentan una mayor flexibilidad en su pool genético, junto con la categoría “R”.
La obtención de un genoma estricto, producto del resultado del corte 15, resulta biológicamente cuestionable, como veremos a continuación ya que una de las rutas metabólicas elementales se encuentra incompleta (en el Anexo VI podemos ver exactamente cuáles son los genes que componen cada corte, recordando que en los cortes inferiores se cuentan también los genes del nivel de restricción inmediato superior), por lo tanto es conveniente comparar los resultados de los 3 niveles de restricción más altos, los cortes 15, 14 y 13:

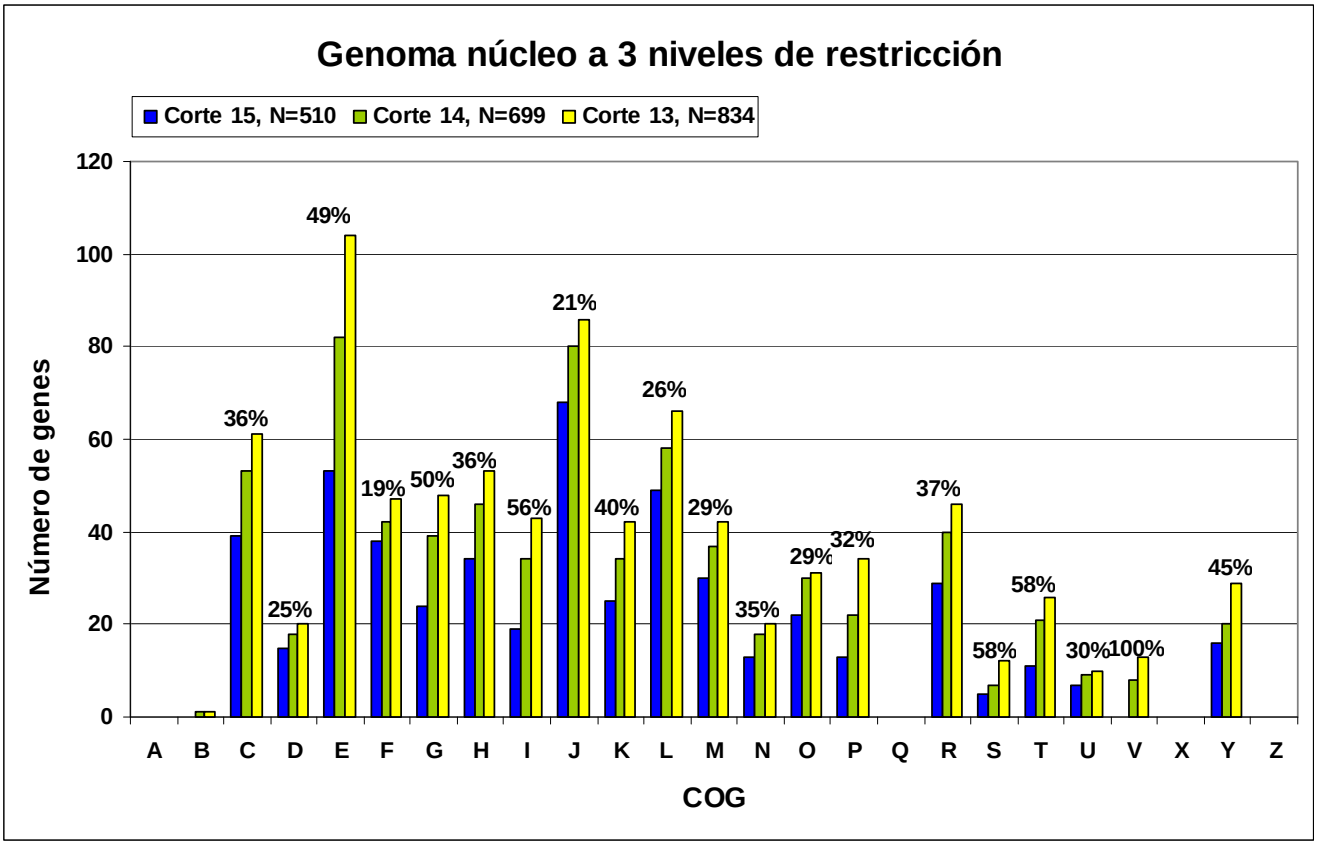

Fig 4. Diferencias entre el número de genes en el corte 15, 14 y 13 y el porcentaje de variación entre el corte con mayor restricción y el de menor.

En la gráfica se notan las diferencias drásticas entre niveles de restricción, siendo muy notorios aquellos que rebasan el 45%, éste porcentaje es la representación de la variación que existe entre los compartidos en el corte 15 y el valor de los compartidos del corte 13. Tal como las clases E, G, I, S y T, cuyos cambios son esperados ya que son categorías COG que estan involucradas en transporte y

metabolismo de aminoácidos, azúcares y lípidos, así como todos aquellos genes involucrados en trasnsucción de señales. Por otro lado es de esperarse que genes involucrados en metabolismo secundario (Q) se encuentren totalmente ausentes del núcleo, ya que éstos son respuestas particulares de cada organismo a su medio ambiente.

Mientras que, en lo opuesto, las clases que presentan una menor variación son aquellas con un porcentaje menor a 25, como son las clases F, J y L, cuya conservación se espera ya que son las categorías COG involucradas en funciones celulares básicas como el metabolismo de nucleótidos, replicación, recombinación y reparación, así como todas las proteinas que se encuentran relacionadas con biogénesis ribosomal y traducción.

La clase T corresponde a los genes involucrados en mecanismos de transducción de señales, resaltando la presencia de genes involucrados en la esporulación (ver Tabla 8) que son:

Proteina E de esporulación en Fase II. Posee 2 funciones específicas, cambiar de la división celular de media a polar y activar el factor de transcripción específico □F.

Proteína A de esporulación en Fase 0. Regulador maestro de entrada al proceso de esporulación.

Fosfotransferasa F de incio en la esporulación. Mensajero secundario del proceso de fosftransferencia durante la esporulación.

La clase S, relacionada con aquellos genes cuya función resulta desconocida, cuenta con varios relacionados con la esporulación:

Proteína M de esporulación en Fase II. Requerida para deshacerse de la pared celular del septo de la espora.

Proteína B de maduración de la espora. Probablemente desarrolla un papel preponderante en la maduración de la cubierta esporal.

Proteína R de esporulación en Fase V. Probablemente involucrada en la formación de la corteza esporal.

Proteína S de esporulación en Fase V. Parece jugar un papel de regulación positiva en permitir a las células progresar más allá de la Fase V mediante la desecación de la pared esporal.

El transporte y metabolismo de lípidos y aminoácidos, representado por la clase I y E, respectivamente, carece de genes involucrados en cualquier función de la esporulación. Aunque por otra parte, en la clase G, transporte y metabolismos de carbohidratos, encontramos un gen asociado a la esporulación:

Proteína AE de esporulación en Fase III. Permite la esporulación después de embeber a la espora dentro de la célula.

Sin embargo, al ser más precisos en el análisis de la tabla anterior, notamos que las clases K y Y sobrepasan el 40% de variación, así que resulta interesante averiguar si presentan genes involucrados en la esporulación.

El proceso de transcripción está designado a la clase K, donde encontramos los siguientes genes:

Proteína T de esporulación en Fase V. Regulador transcripcional tanto positivo como negativo de los genes Sigma-G dependientes.

Proteína J de esporulación en Fase 0. Necesaria para iniciar el proceso esporulativo mediante la correcta segregación de los cromosomas.

Proteína D de esporulación en Fase III. Reguladora de la transcripción de sigK, que codifica para la RNA polimerasa factor Sigma que se encuentra en la cámara de la célula madre.

Una de las características distintivas dentro del género es la capacidad de esporular y de estos genes encontramos a tan solo 22 genes que se encuentran conservados (Ver Tabla 8) en todos los *Bacillus*. De esperarse son los factores sigmas responsables de este proceso, y spo0A que es un sensor de dos componentes necesario para el inicio de la esporulación. Resulta curioso que se encuentren genes presentes para las etapas 0, II, III, IV y V de la esporulación, lo cual quiere decir que los detonantes para iniciar la diferenciación hacia endospora son muy variables y la integración inicial de estímulos para esporular responde al hábitat/estrés particulares de cada especie:

Tabla 8. 22 genes núcleo del proceso de esporulación presentes en todas las cepas de *Bacillus.*

M4401790	Anti-sigma F factor antagonist	K06378
M4403330	Stage II sporulation protein D	K06381
M4400073	Stage II sporulation protein E	K06382
M4401329	Sporulation sigma-E factor processing peptidase	K06383
M4401782	Stage II sporulation protein M	K06384
M4402279	Stage II sporulation protein P	K06385
M4403333	Stage II Sporulation Protein	K06386
M4403297	Stage II sporulation protein R	K06387
M4402070	Stage III sporulation protein AA	K06390
M4402071	Stage III sporulation protein AB	K06391
M4402072	Stage III sporulation protein AD	K06393
M4402073	Stage III sporulation protein AE	K06394
M4402076	Stage III sporulation protein AG	K06396
M4402077	Stage III sporulation protein AH	K06397
M4402095	Stage IV sporulation protein B	K06399
M4402368	Stage IV sporulation protein FA	K06401
M4402369	Stage IV sporulation protein FB	K06402
M4402312	Stage V sporulation protein B	K06409
M4401141	Stage V sporulation protein D	K08384
M4402335	Stage VI sporulation protein D	K06417
M4400132	RNA polymerase sigma-H factor	K03091
M4402096	Stage 0 sporulation protein A	K07699
M4401141	Stage V sporulation protein D	K08384
M4400073	Stage II sporulation protein E	K06382
M4400082	Stage V sporulation protein T	K04769

M4400016 Sporulation initiation inhibitor protein soj K03496

Metabolismo.

Los datos obtenidos, al ser tratados con la herramienta KASS del KEGG (http://www.genome.jp/kegg/kaas/) de reconstrucción metabólica, arrojaron mapas para cada uno de los cortes aunque, con fines prácticos se optó por el sobrelapamiento de los 3 niveles y notar las ganancias metabólicas conforme se relaja el nivel de restricción, la figura 5 muestra dichos resultados.

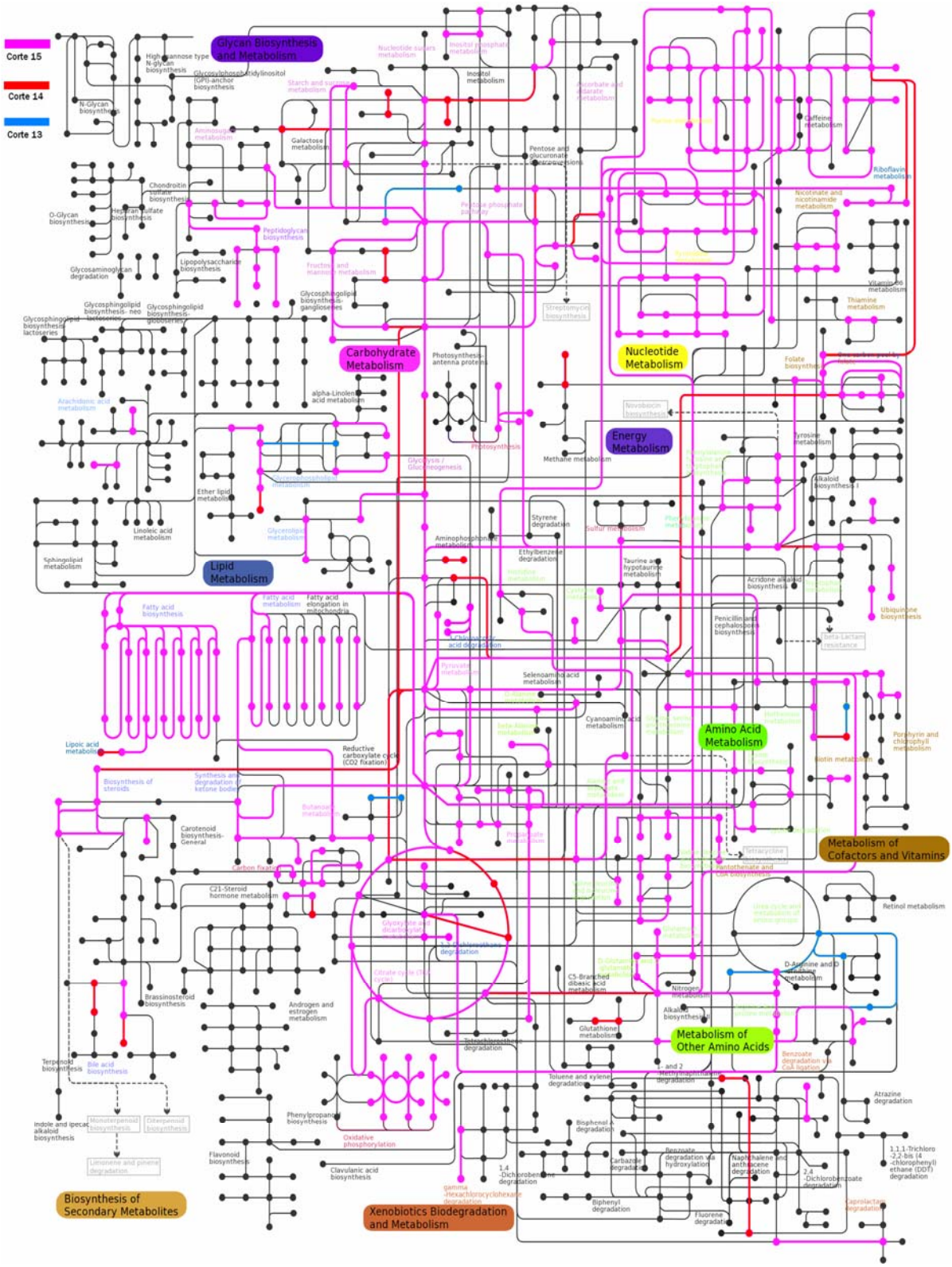

Fig 5. Reconstrucción metabólica del genoma núcleo con los cortes diferenciados por color: corte 15(rosa), corte 14(rojo) y corte 13 (azul).

El corte 15 se encuentra representado por las líneas en color rosa, donde es notable reconocer que rutas tan elementales como el ciclo de Krebs, el ciclo de la carboxilación reductiva o la fosforilación

oxidativa no se encuentran completos, lo cual representaría una imposibilidad a nivel biológico. Los 510 genes que constituyen el genoma núcleo se encuentran distribuidos en todas las posibles categorías de componentes metabólicos. Para tener certidumbre de la realidad biológica hace falta cuantificar los reemplazos no ortólogos de genes, éstos son genes que, con un origen evolutivo distinto, cubren la misma función celular/metabólica, este trabajo se encuentra en proceso en el laboratorio actualmente.

Cuando se baja el nivel de restricción permitiendo que uno de los organismos no cuente con un ortólogo estricto a 14 (líneas rojas; Fig. 5) los genes compartidos por todos empiezan a mostrar más categorías funcionales, algunas de ellas pudiendo ser consideradas como funciones secundarias, como se muestra a continuación:

Biodegradación y metabolismo de Xenobióticos.

Solo se presenta ganancia de intermediarios en la degradación del ácido 3-Cloroacrílico, naftaleno, antraceno, 1 y 2 metil-naftaleno.

Metabolismo de aminoácidos.

Nuevos intermediarios se suman en la biosíntesis de la fenilalanina, tirosina y triptofano, metabolismo de taurina, hipotaurina, propanoato y citrato.

Metabolismo de carbohidratos.

Nuevos intermediarios para la glucólisis/gluconeogénesis, el metabolismo de galactosa, inositol, pentosa fosfato y del ácido dibásico ramificado C5.

Metabolismo de cofactores y vitaminas.

Existe ganancia en el metabolismo de riboflavina, biotina, ácido lipóico y biosíntesis del folato.

Metabolismo de energía.

Intermediarios adicionales para el ciclo de la carboxilación reductiva, fijación de CO_2, metabolismo del metano,

Metabolismo de lípidos.

Ganancia de intermediarios en el metabolismo de glicerofosfolípidos, biosíntesis de esteroides y ácido biliar, síntesis y degradación de cuerpos cetónicos.

Biosíntesis de metabolitos secundarios.

Incorporación de intermediarios de la biosíntesis de alcaloides

Metabolismo de nucleótidos.

Se adquieren intermediarios para el metabolismo de pirimidinas,

Metabolismo de otros aminoácidos.

Ganancia de intermediarios en el metabolismo del aminofosfonato y del glutatión.

Por otra parte, las ganancias reportadas para el genoma núcleo al corte 13 son menores y se centran exclusivamente en la ganancia de intemediarios metabólicos, como se muestra a continuación:

Metabolismo de carbohidratos.

Intermediarios de la ruta de las pentosas fosfato y del metabolismo del butanoato.

Metabolismo de lípidos.

Nuevos intermediarios en el metabolismo de glicerofosfolípidos.

Metabolismo de cofactores y vitaminas.

Se presentan intermediarios involucrados en el metabolismo de la biotina, de la D-arginina y de la D-ornitina.

Metabolismo de aminoácidos.

Parte del ciclo de la urea y del metabolismo de grupos amino se completa.

Otra característica de algunos miembros de la especie es la competencia (la habilidad de tomar y procesar DNA exógeno bajo ciertas condiciones fisiológicas). Aunque la conservación de los genes involucrados en la competencia es poco conocido (Rey, Ramaiya et al. 2004), y parece ser un proceso más bien conservado a nivel intra-específico (Earl, Losick et al. 2008) y con poco nivel de conservación en estados iniciales (Rey, Ramaiya et al. 2004). Dentro de los genes involucrados en competencia conservados en todos los *Bacillus* solo tenemos a: comEA, comEB, comEC, comFA, comFC y comGA. Esto confirma que la variabilidad en el fenotipo de competencia no es conservada a nivel genético y que si bien algunos genes de este proceso se encuentran universalmente conservados en los *Bacillus* no son suficientes para explicar qué es exactamente lo que inicia el proceso en cada especie.

Genoma núcleo de hábitats acuáticos.

Si bien *Bacillus* ha sido estudiado desde finales de 1,800 (Ravel and Fraser 2005) el papel ecológico que desempeña este género, con toda la información genómica dentro del género, apenas es posible empezar a abordar el problema. Unas cuantas comparaciones existen a nivel de las especies patógenas (*B. anthracis y B. cereus*) con fines de diagnóstico y de entender diferencias polimórficas entre los distintos aislados (Ivanova, Vysotskii et al. 1999; Xu and Cote 2003; Helgason, Tourasse et al. 2004; Priest, Barker et al. 2004; Rasko, Ravel et al. 2004; Anderson, Sorokin et al. 2005). Sin embargo, el rol ecológico de las cepas no patógenas no es claro y nos dimos a la tarea de comparar *Bacillus* de hábitats acuáticos, para describir que genes conservan estas cepas con respecto a los ortólogos compartidos con todos los demás genomas del género.

B. coahuilensis M4-4, *Geobacillus kaustophilus* HTA426, *Oceanobacillus iheyensis* HTE831 y *Bacillus sp*. NRRL B-14911, son las bacterias que fueron seleccionadas para hacer dicha comparación, todas han sido aislados de ambientes acuáticos, comparten otras características generales como que generalmente tienen un tamaño reducido de genoma (3.2 – 3.5 Mb) a comparación de los demás *Bacillus* “generalistas” con la única excepción de *B.*. *sp*. NRRLB-14911 (cerca de 5 Mb), aunque las diferencias en los modos de vida originales de las bacterias son evidentes: con organismos aislados a gran profundidad (*G. kaustophilus* - >3000 m; *O. iheyensis* -

>2000m), temperaturas elevadas (*G. kaustophilus*), de zonas fóticas (*B. coahuilensis* y *B. sp*. NRRLB-14911).

La diferencia en el tamaño de genoma puede responder a presiones de selección similares como vivir en ambientes oligotróficos y quizá encontrarse especializado en el medio ambiente, en el caso de *B. coahuilensis* (3.2 Mb) dicho efecto se observa con la adquisición de estrategias para sobreponerse a la limitación de Fósforo en el medio ambiente como la síntesis de sulfolípidos y un reducido tamaño en el genoma con múltiples carencias cuando se le compara contra otros *Bacillus* (i.e. en enzimas clave para sobrevivir, como las del ciclo del Nitrógeno) (Alcaraz, Olmedo et al. 2008). En el caso de *O. iheyensis* también con un genoma reducido (3.6 Mb) (Takami, Takaki et al. 2002) su genoma está adaptado a vivir en alta salinidad y medios alcalinos, además de resistir una presión atmósferica mayor a 0.1 MPa, esta cepa fue aislada a 1,050 m de profundidad. *G. kaustophilus* (3.5 Mb) (Takami, Takaki et al. 2004), también un organismo marino fue aislado a 3,000 m de profundidad en una ventila hidrotermal, con un óptimo de crecimiento de 60°C. Finalmente B. sp. NRLB14911 es el del genoma más grande (~5 Mb) de todos los *Bacillus* aislados del agua, este aumento del tamaño se traduce en un organismo robusto y redundante, este organismo fue aislado de la zona fótica (=> 10 m) en el Golfo de México (Martin, Siefert et al. 2003) que puede parecerse más a un organismo generalista como *B. subtilis* y de tamaño solo similar a *B. anthracis/cereus*.

Cuando se comparan los genes núcleo de todas las especies contra los acuáticos observamos una diferencia de 125 genes más en los acuáticos (ver Fig. 8). Los *Bacillus* acuáticos comparten características como una sobre-representación de transportadores/importadores de aminoácidos (COG E) y un aumento ligero en la cantidad de todos los demas transportadores dentro de los cuales podemos encontrar a transportadores de tipo ABC, permeasas, bombas H+/Na+, solutos orgánicos como glicin-betaina, colina, carnitina, prolina y pantotenato. Todas estas estrategias les permiten conservar la homeostasis cuando viven en ambientes salinos y alcalinos (Takami, Takaki et al. 2002). Una característica común a *B. coahuilensis*, *O. iheyensis* y *G. kaustophilus* es que presentan una elevada cantidad de secuencias de Inserción (IS) y transposones, por lo que es posible pensar que este ha sido el mecanismo mediante el cual se ha reducido el tamaño de su genoma. Este efecto se ha observado en genomas que han sufrido una especialización de nicho, pseudogenizando funciones no vitales, y posteriormente reduciendo el tamaño de genoma debido a un sesgo erocional en las bacterias (Pushker, Mira et al. 2004). Esto mismo no se ve en genomas generalistas como *B. subtilis* que no tienen una sola IS en su genoma (Kunst, Ogasawara et al. 1997).

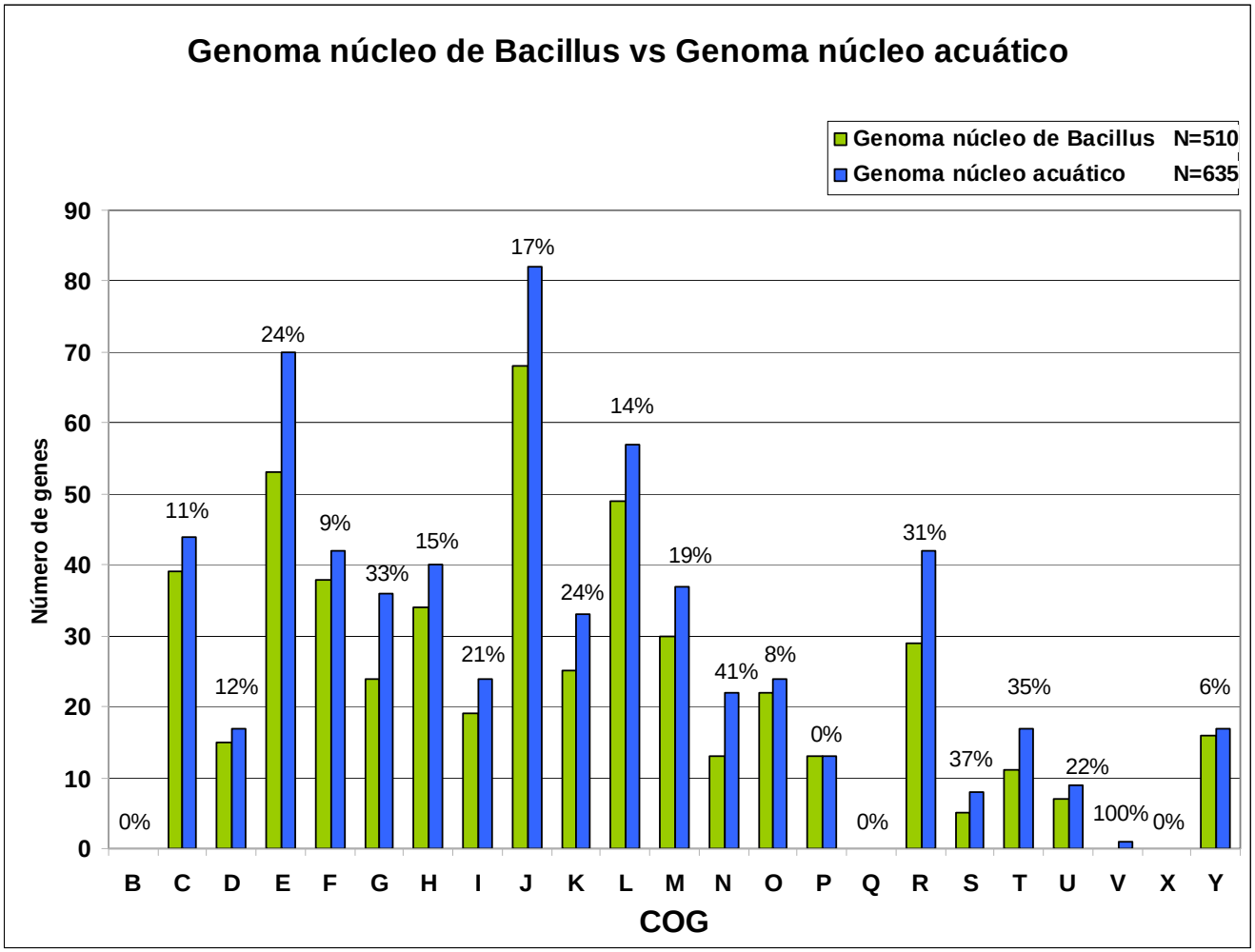

Fig. 6. Comparativa de presencia de genes entre los genomas mínimos del género contra bacterias de hábitat acuático.

La comparativa del genoma núcleo contra el núcleo de los genomas acuáticos (Fig. 6) muestra más representados los COGs G, N, R, S y T, de donde podemos sustraer las clases R, S y V que corresponden a funciones predichas y desconocidas mientras que la última solamente muestra una variación de un solo gen que difícilmente podríamos considerar como representativo. A pesar de que notemos una diferencia clara entre las clases es necesario verificar si estadísticamente es significativa la variación, por lo que se llevó un análisis de ji-cuadrada que se muestra a continuación:

Chi-cuadrada	4.637746
G. de L.	22
p	0.000032
Tabla 0.05	12.33802
Tabla 0.01	9.542492

El análisis, entonces, determina que existe una diferencia significativa entre el genoma núcleo del género *Bacillus* y el genoma de bacterias que habitan ambientes acuáticos, lo que significaría que las funciones correspondientes a los COGs son muy importantes para la supervivencia de las bacterias

en ambientes acuáticos. Estas clases corresponden al transporte y metabolismo de carbohidratos, la movilidad celular y la transducción de señales. Más no por poseer los porcentajes más altos son las únicas clases a tomar en cuenta, están presentes otras con variaciones moderadas involucradas en el transporte y metabolismo de aminoácidos (clase E), transporte y metabolismo de lípidos (clase I), transcripción (clase K) y finalmente la clase U relacionada con el tráfico intracelular, secreción y transporte vesicular.

Pangenoma.

La determinación del pangenoma se llevó a cabo con aquellos genomas disponibles en la base de datos del NCBI (ver Tabla 3), en este caso el propósito es describir el potencial metabólico absoluto del grupo por lo que la redundancia es aceptada en este análisis.
Un panorama general de las funciones metabólicas representadas en el total de los 103,695 genes analizados se presenta en la figura 7 que, como era de esperarse, la función más representada dentro de este análisis es la que corresponde a los genes hipotéticos (COG X), con cerca de 23 mil genes, luego a los del COG R y S con cerca de 17 mil genes, estas categorías corresponden a genes hipotéticos o predichos también. Esto es de esperarse ya que después de secuenciar y anotar un genoma tenemos un estimado actual del 30% de genes hipotéticos. Esto sirve para ver la falta de estrategias de genómica funcional masivas, pero más aún también habla de un potencial de funciones desconocidas dentro de estos organismos.

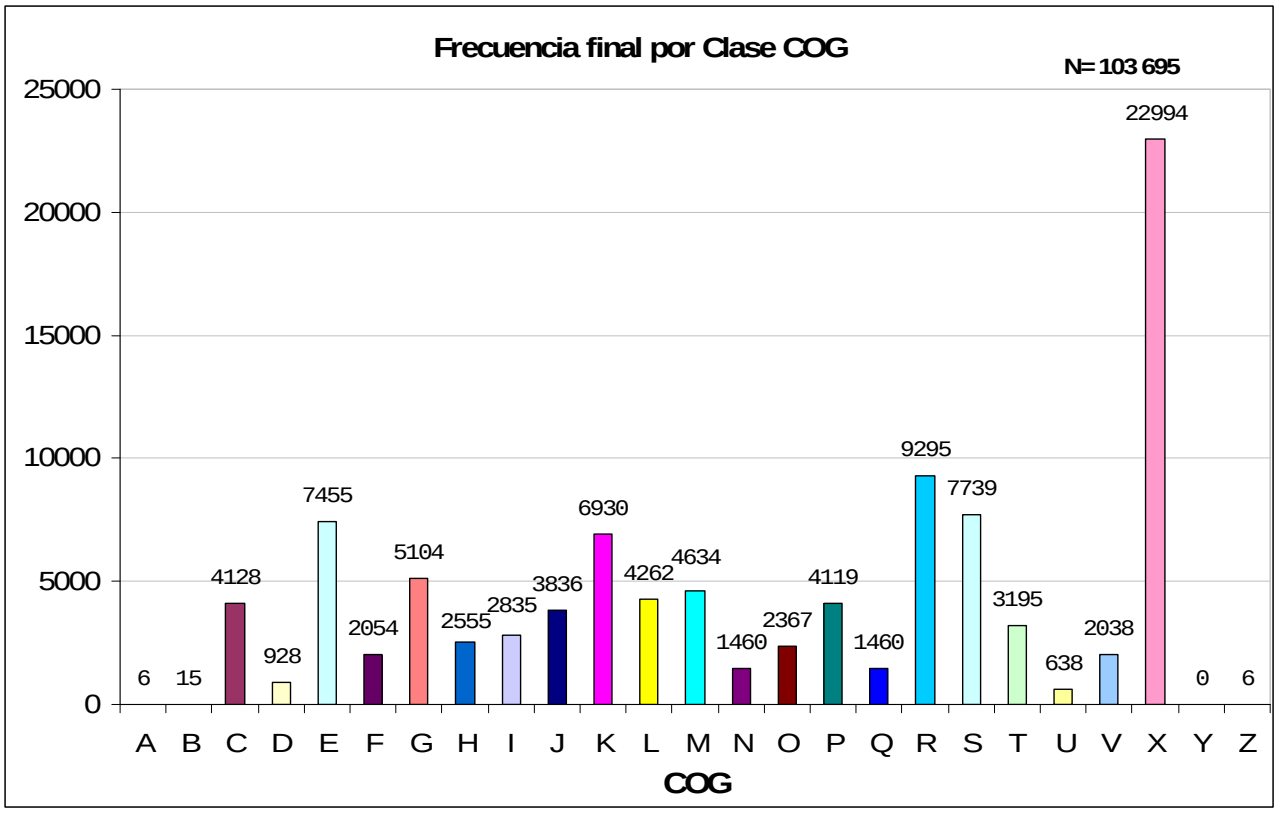

Figura 7. Distribución de indicadores COG del pangenoma del género *Bacillus*.

Categorías a considerar son la K (transcripción) con cerca de 7 mil genes relacionados a la regulación de la expresión genética, la E (transporte y metabolismo de aminoácidos) con cerca de 7,500 genes. Hay categorías con un mínimo nivel de redundancia si las comparamos con las demás como la C, D, F, N, O que tienen que ver con mecanismos de obtención/conversión de energía, división celular, metabolismo y transporte de nucleótidos, mobilidad celular y chaperonas respectivamente, en estas categorías los mecanismos de acción son tan conservados que quizá no haya lugar a tener redundancia en estos genes. Por otro lado el arsenal químico de *Bacillus* se reduce a 1,460 genes relacionados al metabolismo secundario (COG Q) y a 2,038 genes que tienen que ver en mecanismos de defensa y producción de antibióticos, si se compara este mismo resultado con el genoma núcleo se puede ver un incremento dramático, cero genes de las categorías hipotéticas en el genoma núcleo. Sin embargo, quizá muchos de los genes incluidos en las categorías de hipotéticos pudieran ser genes involucrados en metabolismo secundario, esto es especulativo pero al menos no parecen ser parte de las funciones celulares fundamentales en su mayoría, esto si se hacen las búsquedas por homología de función. Con estos datos se puede tener una cuantificación (37% del total de genes) del potencial biotecnológico de los *Bacillus* cuando decimos que si lo que interesa obtener son metabolitos secundarios, antibióticos, etc.

Cabe mencionar que cuando se considera al pangenoma completo se incluyen todos los elementos producto de transferencia horizontal de genes (HGT). Esto no tiene implicaciones ya que lo que

interesa aquí es ver el potencial metabólico como género. Otra tarea es analizar los “huecos” metabólicos que no han sido cubiertos por genes endógenos o HGT. Para visualizar esto se mapearon todos los genes en un mapa metabólico (ver Figura 8). Los huecos podrían ser representantes de funciones limitadas en movilidad o que el género no cubre en la naturaleza, un ejemplo obvio es la fotosíntesis y el metabolismo de porfirinas y clorofila, que en este caso no se dan, sin embargo otras formas de sensar la luz como las rodopsinas han sido descritas en al menos una especie (Alcaraz, Olmedo et al. 2008), y este tipo de eventos son imposibles de visualizar en este mapa general.

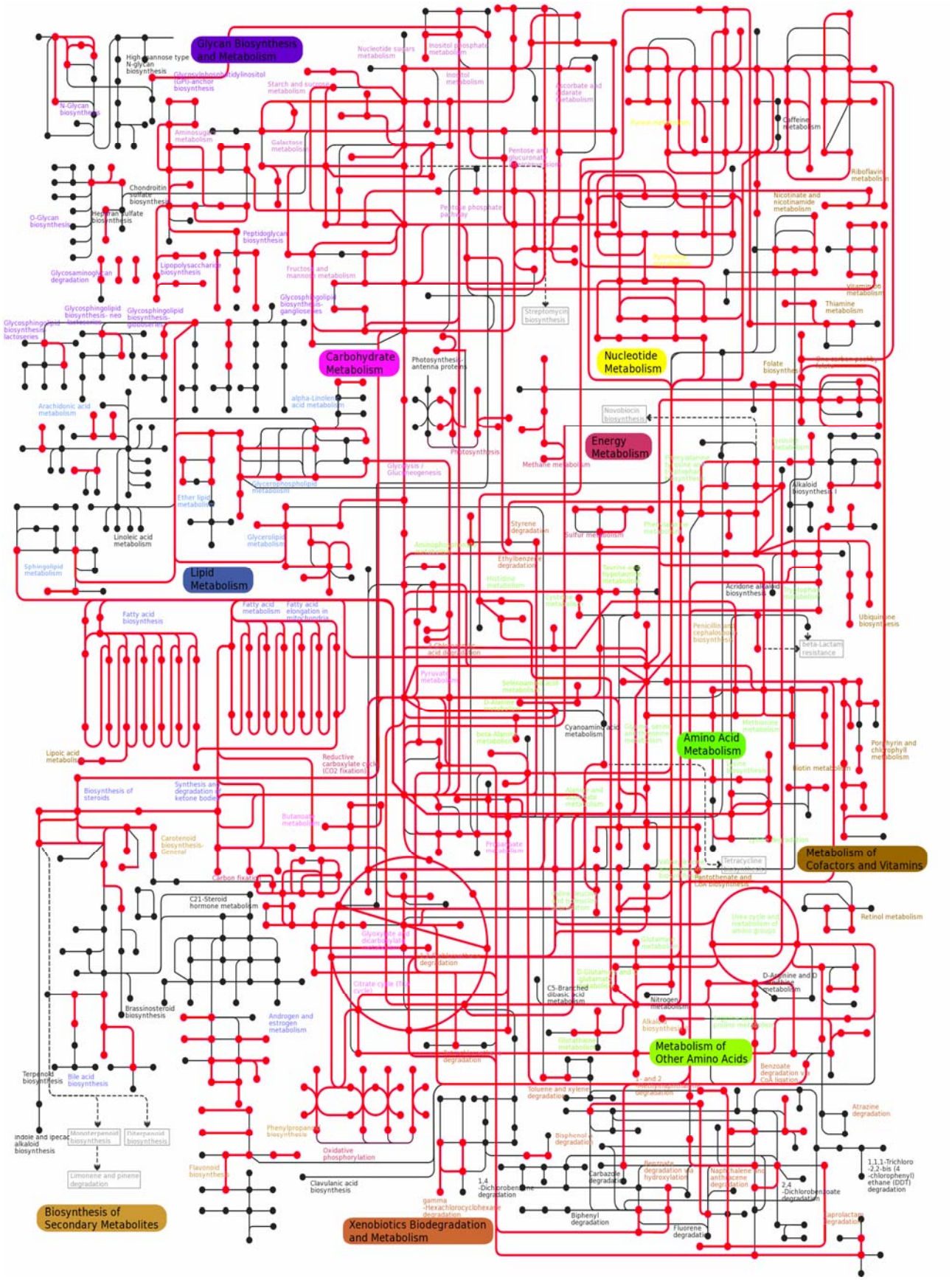

Figura 8. Reconstrucción metabólica del pangenoma del género *Bacillus*, los genes presentes son aquellos marcados en rojo.

6. Perspectivas

Como se ha demostrado anteriormente (Tettelin, Masignani et al. 2005), la concepción de definir especies bacterianas con los criterios tradicionales (16S rRNA, MLST, CGH) fallan ya que desprecian una gran cantidad de información del genoma accesorio. Con los enfoques actuales de secuenciación como el 454 (www.454.com), Solid (www.appliedbiosystems.com) y Solexa (www.illumina.com) es posible pensar en secuenciar más de un par de cepas de la misma especie y ganar información sobre su diversidad intraespecífica y tener idea primero de que genes son los indispensables en estos genomas y descubrir su potencial metabólico mediante la descripción del genoma accesorio. Dentro de la misma especie se han hecho intentos en B. cereus/anthracis y B. subtilis para describir la variabilidad intraespecífica (Ivanova, Vysotskii et al. 1999; Helgason, Tourasse et al. 2004; Priest, Barker et al. 2004; Rasko, Ravel et al. 2004; Anderson, Sorokin et al. 2005; Earl, Losick et al. 2008). Este tipo de comparaciones pueden ser aplicados para el desarrollo de vacunas, diseñadas a partir del conocimiento del genoma núcleo y el accesorio. Yo creo que no tiene sentido hablar de la variación intraespecifica. Es un tema muy amplio de el área de genética de poblaciones y por supuesto que si se tiene información al respecto.

Ahora bien, para el análisis de la funcionalidad de un género tan ubicuo como *Bacillus* este tipo de análisis sirven como un punto de partida para describir primero los genes que son informativos para reconstruir la historia evolutiva del género y de esta manera dar nuevas alternativas más allá del 16S y el MLST (Bentley and Parkhill 2004; Gevers, Cohan et al. 2005). Además de dar una idea funcional completa de que es necesario para ser un integrante de este género. No entiendo esta frase Con el genoma accesorio del género podemos definir, como se hizo en este trabajo, algunas funciones ambiente-específicas como el caso de los *Bacillus* acuáticos cuando se compara el núcleo de ortólogos de todos contra el núcleo de los acuáticos.

El siguiente paso es pasar del carácter cualitativo y descriptivo de este estudio a poder utilizar la información generada para cuantificar a nivel global las diferencias entre todos los genomas utilizado, mediante índices que consideren un alineamiento gen por gen y algún tipo de normalización para el tamaño del genoma / número de genes. En el grupo de trabajo de la Dra. Olmedo, Cinvestav Guanajuato, se tiene otra cepa de *Bacillus*, aislada de Cuatro Ciénegas como B. coahuilensis, que servirá para confirmar hipótesis sobre la factibilidad de tener más genes compartidos de los esperados a nivel filogenético si se comparte el nicho, también se tiene contemplada la secuenciación a mediana cobertura de distintas cepas de *Bacillus* aislados de distintas partes del mundo, con énfasis en organismos acuáticos. El análisis de cepas secuenciadas a baja cobertura se ha probado exitosamente en otros estudios (Goo, Roach et al. 2004). Por otro lado la información generada del genoma núcleo puede ser utilizada por el laboratorio de la Dra. Souza, en el Instituto de Ecología,

UNAM, para tener una colección de marcadores de *Bacillus* y profundizar en los estudios de genética poblacional de Cuatro Ciénegas.

7. Conclusiones.

En el presente trabajo se investigarón características distintivas del género *Bacillus*, así como profundizar en el conocimiento de su biología, para con esta información tratar de hacer un sentido biológico de todos los genes ortólogos compartidos por todos los *Bacillus*. Para generar las listas de genes ortólogos se utilizaron técnicas de bioinformática que surgieron a partir de los mejores alineamientos locales bidireccionales (mediante BLAST). Las listas de ortólogos compartidas fueron evaluadas en cuestión de presencia/ausencia de un ortólogo y se clasificaron en base al COG. De un promedio de 4 mil genes en cada genoma de *Bacillus* se tiene que es cerca de un 10% (510 genes) los que se encuentran compartidos en todos los organismos del género.

Con los datos generados en el presente trabajo se tiene una lista de genes núcleo del género *Bacillus*. Estando plenamente consciente que no es un dato cerrado y que está a disposición de la nueva información producto de la secuenciación de otros genomas. Sin embargo, se aporta un listado de ortólogos que pueden ser utilizados en posteriores trabajos como punto de partida cuando, con genomas nuevos, se tenga que redefinir el genoma núcleo y el accesorio. A nivel biológico la relevancia de estos datos radican en que los métodos de comparación filogenética que usan información de todo el genoma describen de manera fina la diversidad genética en su conjunto. Correspondiendo más bien aun continuo de diversidad genética, más que a entidades aisladas (como las especies biológicas), esto es debido a que la conservación de secuencia de los genes rRNA es tan larga que ocultan importantes niveles de diferenciación fenotípica y ecológica.

Comparar el genoma núcleo contra bacterias que tengan un común denominador ambiental abre posibilidades a descubrir funciones sobre o subrepresentadas compartidas por los organismos que tengan nichos similares.

También se abren posibilidades a hacer genómica funcional en genes del núcleo que pertenecen a una categoría conocida como hipotéticos conservados, si son genes que están en todos los *Bacillus* pero se desconoce su función se trata de un punto de inicio para tratar de descubrir las funciones del universo de genes hipotéticos.

Bibliografía

Alcaraz, L. D., G. Olmedo, et al. (2008). "The genome of *Bacillus coahuilensis* reveals adaptations essential dor survival in the relic of an ancient marine environment." *Proceedings of the National Academy of Sciences PNAS* **105**: 5803-8.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Anderson, I., A. Sorokin, et al. (2005). "Comparative genome analysis of Bacillus cereus group genomes with Bacillus subtilis." FEMS Microbiol Lett **250**(2): 175-84.

Baweja, R. B., M. S. Zaman, et al. (2007). "Properties of Bacillus anthracis spores prepared under various environmental conditions." Archives of Microbiology **(In print)**.

Benson, D. A., I. Karsch-Mizrachi, et al. (2006). "GenBank." Nucleic Acids Research **35**: D21-D-25.

Bentley, S. D. and J. Parkhill (2004). "Comparative genomic structure of prokaryotes." Annu Rev Genet **38**: 771-92.

Cerritos, R., P. Vinuesa, et al. (2008). "Bacillus coahuilensis sp. nov. a new moderately halophilic species from different pozas in the Cuatro Ciénegas Valley in Coahuila, México." International Journal of Systematic and Evolutionary Microbiology **In press**.

Dupuy, B. and A. L. Sonenshein (1998). "Regulated transcription of *Clostridium* difficile toxin genes." Molecular Microbiology **27**(1): 107–120.

Earl, A. M., R. Losick, et al. (2008). "Ecology and genomics of *Bacillus subtilis*." Trends Microbiol **16**(6): 269-75.

Eisen, J. A. (1998). "Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis." Genome Research **8**: 163-167.

Eisen, J. A. and C. M. Fraser (2003). "Phylogenomics: Intersection of Evolution and Genomics." Science **300**(1706): 1706-1707.

Gevers, D., F. M. Cohan, et al. (2005). "Opinion: Re-evaluating prokaryotic species." Nat Rev Microbiol **3**(9): 733-9.

Gibas, C. and P. Jambeck (2001). Developing Bioinformatics Computer Skills. Sebastopol, CA, O'Reilly & Associates Inc.

Gibson, D. G., G. A. Benders, et al. (2008). "Complete Chemical Synthesis, Assembly, and Cloning of a Mycoplasma genitalium Genome." Science **319**(5867): 1215-1220.

Glass, J. I., N. Assad-Garcia, et al. (2006). "Essential genes of a minimal bacterium." Proc Natl Acad Sci U S A **103**(2): 425-30.

Goo, Y. A., J. Roach, et al. (2004). "Low-pass sequencing for microbial comparative genomics." BMC Genomics **5**(1): 3.

Helgason, E., N. J. Tourasse, et al. (2004). "Multilocus Sequence Typing Scheme for Bacteria of the Bacillus cereus Group." Appl. Environ. Microbiol. **70**(1): 191-201.

Hennequin, C., F. Porcheray, et al. (2001). "GroEL (Hsp60) of Clostridium difficile is involved in cell adherence." Microbiology **147**: 87–96.

Hiratsuka, Y., M. Miyata, et al. (2006). "A microrotary motor powered by bacteria." Proceedings of the National Academy of Sciences PNAS **103**(37): 13618–13623.

Ivanova, E. P., M. V. Vysotskii, et al. (1999). "Characterization of Bacillus strains of marine origin." Int Microbiol **2**(4): 267-71.

Keim, P., L. B. Price, et al. (2000). "Multiple-Locus Variable-Number Tandem Repeat Analysis Reveals Genetic Relationships within *Bacillus anthracis*." Journal of Bacteriology **182**(10): 2928–2936.

Kobayashi, K., S. D. Ehrlich, et al. (2003). "Essential Bacillus subtilis genes." Proceedings of the National Academy of Sciences PNAS **100**(8): 4678-83.

Kunst, F., N. Ogasawara, et al. (1997). "The complete genome sequence of the gram-positive bacterium Bacillus subtilis." Nature **390**(6657): 249-56.
Larkin, M. A., G. Blackshields, et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics Applications Note **23**(21): 2947–2948.
Li, C., G. Ortí, et al. (2007). "A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study." BMC Evolutionary Biology **7**(44).
Li, Q., J. S. Sherwood, et al. (2006). "Antimicrobial resistance of Listeria spp. recovered from processed bison." Letters in Applied Microbiology **44**: 86-91.
Madigan, M. T., J. M. Martinko, et al. (2003). Brock. Biología de los Microorganismos. Englewood Cliffs, New Jersey.
Martin, K. A., J. L. Siefert, et al. (2003). "Cyanobacterial signature genes." Photosynthesis Research **75**: 211–221.
Medini, D., C. Donati, et al. (2005). "The microbial pan-genome." Current Opinion in Genetics & Development **15**: 589-584.
Morikawa, M. (2006). "Beneficial Biofilm Formation by Industrial Bacteria Bacillus subtilis and Related Species." Journal of Bioscience and Bioengineering **101**(1): 1–8.
Murray, P. R., K. S. Rosenthal, et al. (2005). Medical Microbiology. Philadelphia, PA, USA, Elsevier Mosby.
Nakano, M. M. and P. Zuber (1998). "Anaerobic Growth of a "Strict Aerobe" (*Bacillus subtilis).*" Annual Review of Microbiology **52**: 165–190.
Nölling, J., G. Breton, et al. (2001). "Genome Sequence and Comparative Analysis of the Solvent-Producing Bacterium Clostridium acetobutylicum." Journal of Bacteriology **183**(16): 4823–4838.
Peake, R. C., D. A. James, et al. (1996). Medical Microbiology. S. Baron. Galveston, Texas, USA, CIP.
Philippe, H. and M. Blanchette (2007). "Overview of the First Phylogenomics Conference." BMC Evolutionary Biology **7**(Suppl 1): S1-S4.
Portnoy, D. A., V. Auerbuch, et al. (2002). "The cell biology of Listeria monocytogenes infection: the intersection of bacterial pathogenesis and cell-mediated immunity." The Journal of Cell Biology **158**(3): 409–414.
Priest, F. G., M. Barker, et al. (2004). "Population Structure and Evolution of the Bacillus cereus Group." J. Bacteriol. **186**(23): 7959-7970.
Pushker, R., A. Mira, et al. (2004). "Comparative genomics of gene-family size in closely related bacteria." Genome Biology **5**(4): R27.
Raju, D., P. Setlow, et al. (2007). "Antisense-RNA-Mediated Decreased Synthesis of Small, Acid-Soluble Spore Proteins Leads to Decreased Resistance of Clostridium perfringens Spores to Moist Heat and UV Radiation." Applied and Environmental Microbiology **73**(7): 2048-2053.
Rasko, D. A., J. Ravel, et al. (2004). "The genome sequence of Bacillus cereus ATCC 10987 reveals metabolic adaptations and a large plasmid related to Bacillus anthracis pXO1." Nucl. Acids Res. **32**(3): 977-988.
Ravel, J. and C. M. Fraser (2005). "Genomics at the genus scale." Trends Microbiol **13**(3): 95-7.
Razin, S., D. Yogev, et al. (1998). "Molecular Biology and Pathogenicity of Mycoplasmas." Microbiology and Molecular Biology Reviews **62**(4): 1094-1156.
Rey, M. W., P. Ramaiya, et al. (2004). "Complete genome sequence of the industrial bacterium Bacillus licheniformis and comparisons with closely related Bacillus species." Genome Biol **5**(10): R77.
Rooney, A. P., J. L. Swezey, et al. (2006). "Analysis of Core Housekeeping and Virulence Genes Reveals Cryptic Lineages of Clostridium perfringens That Are Associated With Distinct Disease Presentations." Genetics **172**: 2081-2091.
Sacchi, C. T., A. M. Whitney, et al. (2002). "Sequencing of 16S rRNA Gene:A Rapid Tool for Identification of Bacillus anthracis." Emerging Infectious Diseases **8**(10): 1117-1123.

Souza, V., L. Espinosa-Asuar, et al. (2006). "An endangered oasis of aquatic microbial biodiversity in the Chihuahuan deset." Proceedings of the National Academy of Sciences PNAS **103**(17): 6565-6570.

Szpirer, C., E. Top, et al. (1999). "Retrotransfer or gene capture: a feature of conjugative plasmids, with ecological and evolutionary significance." Microbiology **145**: 3321–3329.

Takami, H., Y. Takaki, et al. (2004). "Thermoadaptation trait revealed by the genome sequence of thermophilic Geobacillus kaustophilus." Nucleic Acids Res **32**(21): 6292-303.

Takami, H., Y. Takaki, et al. (2002). "Genome sequence of Oceanobacillus iheyensis isolated from the Iheya Ridge and its unexpected adaptive capabilities to extreme environments." Nucleic Acids Res **30**(18): 3927-35.

Tettelin, H., V. Masignani, et al. (2005). "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial ‘‘pan-genome’’." PNAS Proceedings of the National Academy of Sciences **102**(39): 13950–13955.

Todar, K. (2006). Todar's Online Textbook of Bacteriology. K. Todar, University of Wisconsin-Madison, Department of Bacteriology.

Woese, C. R., O. Kandler, et al. (1990). "Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya." PNAS Proceedings of the National Academy of Sciences **87**: 4576-4579.

Xu, D. and J. C. Cote (2003). "Phylogenetic relationships between Bacillus species and related genera inferred from comparison of 3' end 16S rDNA and 5' end 16S-23S ITS nucleotide sequences." Int J Syst Evol Microbiol **53**(Pt 3): 695-704.

ANEXOS

I

Bacillus sp.

Las especies de éste género poseen, bajo el microscopio, una forma muy característica al microscopio similar a una barra, son capaces de llevar a cabo esporulación, aerobios principalmente y sin embargo, si las condiciones no son favorables pueden desempeñarse como aerobios facultativos. En su gran mayoría son saprófitos y quimio-heterótrofos, pudiéndoseles encontrar en nichos tan extremos como desiertos, arena, aguas termales o suelos del Ártico. Presentan movimiento gracias a una serie de flagelos peritricos, con respuesta positiva o negativa a los diferentes factores y pueden ser termófilos, psicrófilos, acidófilos, alcalófilos, halotolerantes e inclusive halófilos

La pared celular de éste género se caracteriza por poseer un peptidoglicano que contiene ácido meso-diaminopimérico (DAP, siglas en inglés), mismo polímero de pared celular que es prácticamente universal en las bacterias Gram -. Así mismo, todas las especies de *Bacillus* presentan gran cantidad de ácido teicoico unido a residuos de ácido murámico; tal hecho no es sino una generalización poco consistente, ya que el contenido de la pared varía tremendamente en y entre las especies (Todar 2006).

Bacillus anthracis. Es un bacilo esporoformador, de 1-1.2 µm de ancho x 3-5 µm de largo, es capaz de infectar a humanos y vertebrados pequeños ya sean de vida libre o domésticos. Los animales se contagian al ponerse en contacto con las esporas del bacilo ya sea ingiriéndolo, respirándolo o tocándolo. De igual manera el ser humano puede contagiarse al entrar en contacto con la carne de animales enfermos o muertos.

La cápsula bacteroide de ésta especie se compone de ácido poli-D-glutámico, el cuál es un factor determinante de virulencia pues por sí misma no es nociva pero confiere resistencia a factores bactericidas del suero y a digestión fagocítica, es producido por el plásmido pX02, el cual es a su vez transferido a los individuos de B. anthracis sin cápsula por medio de transducción, cambiando así el fenotipo; dicha cápsula no es de radical importancia al final de la infección, donde está involucrada casi exclusivamente la toxina de ántrax sino al principio, en el establecimiento de las colonias bacterianas en los tejidos (Todar 2006). La toxina de ántrax se produce por el plásmido pX01 y consta de tres segmentos termolábiles con un peso molecular aproximado de 80kDa.

Históricamente B. anthracis es reconocido como una plaga que afecta principalmente al ganado bovino y, recientemente ha sido catalogado como organismo riesgoso ya que por sus mecanismos de infección bien puede ser utilizado como arma biológica.

Bacillus subtilis. Es un bacilo esporoformador mesófilo, aerobio estricto, aunque recientemente se ha reportado su anaerobiosis facultativa (Nakano and Zuber 1998), que ha tomado como hábitat la rizósfera, punto intermedio entre las raíces de las plantas y el suelo circundante, es capaz de aumentar la densidad de las cosechas promoviendo el desarrollo de otras bacterias como *Rhizobacterium* gracias a la síntesis de algunos péptidos antifúngicos (Todar 2006).

Recientemente se ha determinado que B. subtilis, con el propósito de sobrevivir en un ambiente que ya no es favorable libera antibióticos al medio para destruir a sus vecinos y de ellos, obtener la materia prima necesaria para un proceso energéticamente costoso que es el formar esporas (Kobayashi, Ehrlich et al. 2003). Ha sido constantemente utilizado en la investigación científica y en la producción industrial por su gran adaptabilidad a diferentes medios, escasez de factores patógenos y facilidad de formación de bio-películas que permite a la especie comenzar a producir diferentes antibióticos o metabolitos (Morikawa 2006).

Tabla 1.1 Algunos de los integrantes más representativos de 4 de los principales géneros que caracteriza al género *Bacillus*.

Bacillus					
Especie	**Tamaño del genoma**	**Porcentaje de G+C**	**Número de Genes**	**Plásmidos**	**Observaciones**
B. anthracis Ames	5,227,293	35%	5,630		
B. anthracis Ames 0581	5,228,310	35%	5,635	pX01 pX02	Patógeno de animales y del hombre por exposición a esporas, produce daño cutáneo o pulmonar. Patogenicidad contenida en plásmidos.
B. anthracis Sterne	5,228,663	35%	5,415		
B. cereus ATCC 10987	5,224,283	35%	5,772	pBc10987	Patógeno por intoxicación, produce enterotoxinas necrotizantes y hemolisinas. Patogenicidad contenida en plásmidos.
B. cereus ATCC 14579	5,411,809	35%	5,481	pBClin15	
B. cereus E33L (ZK)	5,300,915	35%	5,268	pE33L466 pE33L54 pE33L9 pE33L8 pE33L5	

B. clausii KSM-K16	4,303,871	44%	4,204		Bacilo alcalinotolerante con presencia ubicua
B. halodurans C125	4,202,353	43%	4,171		
B.licheniformis ATCC 14580	4,222,645	46%	4,289		Relacionado ocasionalmente a bacteremia o septicemia.
B. subtilis 168	4,214,814	43%	4,225		Bacteria de gran valor médico, científico e industrial por su facilidad de crecimiento en cultivos, gran producción de enzimas y antibióticos.
B. thuringiensis 97-27	5,237,682	35%	5,263	pBT9727	Especie patógena de insectos abundantemente utilizada en el control de plagas como bioinsecticida.
B. thuringiensis Al Hakam	5,257,091	35%	4,883	pALH1	

Clostridium sp.

Los integrantes del género *Clostridium* se caracterizan, a diferencia de sus parientes cercanos, *Bacillus*, por carecer completamente de un sistema citocrómico y mecanismos de fosforilación de transporte de electrones por lo que se les denomina fermentadores. Tanto pueden responder a la tinción de Gram como no hacerlo sin embargo, la mayoría de los individuos son Gram +.

Muchos son incapaces de desarrollarse en ambientes aerobios aunque sus esporas pueden pasar años en ambientes aerobios y conservar su viabilidad. Debido a la extensa ocupación de ambientes anaerobios, han desarrollado procesos de fermentación de muchas moléculas orgánicas como carbohidratos, aminoácidos, purinas, etanol e inclusive ácidos grasos dando como resultado la producción de otros compuestos de uso industrial como etanol, butanol, acetona, ácido butírico y acético, junto con grandes cantidades de CO2 y H2, compuestos que nos permiten inferir que éste género está involucrado directamente con la biodegradación y el ciclo del carbono (Todar 2006).

Dado que son saprófitos, únicamente podemos catalogarlos como patógenos oportunistas cuyas mismas enzimas involucradas en la degradación de los antes mencionados compuestos tienen un rol preponderante en su patogenicidad (Peake, James et al. 1996).

Clostridium acetobutilicum. Este es un clostridio anaerobio esporoformador quimio-organótrofo distribuido ampliamente en suelos, con un rango de tolerancia de temperatura partiendo desde los 10 hasta los 65ºC; es capaz de fermentar azúcares y produce muchos compuestos de alto valor industrial (ver tabla 2) que si bien dejó de ser abundantemente explotada por el aumento en la accesibilidad de procesos petroleros, recientemente ha visto un resurgimiento a causa del alza de los precios de crudo en el mundo (Madigan, Martinko et al. 2003).

Al igual que todas las bacterias posee un genoma circular más un plásmido, reconocido como megaplásmido (pSOL1) con 192,000 pares de bases, codificantes para un polipéptido de 178 aminoácidos (Nölling, Breton et al. 2001). En dicho plásmido se encuentran los genes involucrados en la producción de solventes, aunque en modelos experimentales se ha visto que conforme aumentan las generaciones, la actividad del plásmido se pierde, ya que las mutantes son incapaces de adquirir dichos plásmidos y sobreviene una degeneración de la cepa.

Poseen movilidad gracias a flagelos peritricos, notándose que dicha característica es capaz de aumentar la producción de solventes, desplazándose directamente a cualquier molécula de azúcar o de ácido butírico. En esta parte de su ciclo de vida el metabolismo es intenso, aunque aumenta el ritmo de éste cuando la bacteria forma una endospora, proceso que fácilmente puede ser disparado por una presencia excesiva de oxígeno, aumento de la cantidad de productos de desecho o simplemente la pérdida de intensidad del gradiente de protones del exterior (Todar 2006).

Clostridium difficile. Este es un clostridio esporoformador anaerobio, integrante de la microbiota colónica en el ser humano y capaz de utilizar los aminoácidos como fuente de energía, aunque en

ocasiones utiliza carbohidratos. Especie íntimamente relacionada a infecciones hospitalarias debido a la tremenda tolerancia a antibióticos que posee, tal resistencia es provista por transposones conjugativos. Los tratamientos prolongados con antibióticos eliminan la microbiota competidora, permitiéndole un desarrollo libre de competidores, su crecimiento describe una curva típica mientras el alimento se encuentre disponible. Durante esta fase no existe una expresión de toxinas; sin embargo, en cuanto el alimento escasea y la población alcanza la meseta poblacional, los genes codificantes de las toxinas A y B se expresan, lo que indica que se encuentran bajo algún tipo de represión catabólica (Dupuy and Sonenshein 1998).

La toxina A ha sido clasificada como una enterotoxina ya que producen una acumulación de líquido en el intestino mientras que la toxina B es una toxina citopática extremadamente letal, ambas son termolábiles a temperatura ambiental por lo que cualquier diagnóstico realizado a partir de presencia de toxina debiera ser efectuado a la brevedad (Todar 2006).

Sus esporas se producen durante un periodo de estrés relacionado con bajos niveles de pH, exposición a antibióticos o altas temperaturas y mantienen su viabilidad hasta por dos años; sin embargo no es la única estrategia de supervivencia que poseen, cuando las mismas situaciones se presentan pero a un nivel menor, las bacterias pueden aumentar su capacidad de adhesión sobre expresando genes relacionados a proteínas exteriores, facilitando la colonización (Hennequin, Porcheray et al. 2001).

Clostridium tetani. Es un clostridio anaerobio esporoformador y patógeno de humanos, distribuido ampliamente en suelos, en particular suelos con gran cantidad de abono, dado que su espora se desarrolla en la zona terminal, esto le otorga una muy característica apariencia de baqueta. Ha sido clasificada como bacteria Gram +, sin embargo conforme las cepas envejecen pueden teñir como Gram – o Gram variable (Todar 2006).

Las infecciones por C. tetani se producen a causa de la exposición de alguna herida a la bacteria y el ambiente anaeróbico del tejido permite la replicación y secreción de exotoxinas. Una toxina espasmogénica en particular, la tetanospamina, se sintetiza a partir de la información contenida en el plásmido pE88 durante el crecimiento y esporulación celular, posteriormente, durante la lisis, ésta se libera al torrente mostrando una muy especial afinidad por el tejido nervioso (Todar 2006). La toxina se fija exactamente en neuronas periféricas y se transporta a lo largo de los axones y dendritas hasta el Sistema Nervioso Central, donde se une a las terminales nerviosas motoras inhibidoras presinápticas y bloquea la liberación de los neurotransmisores glicina y ácido gamma-aminobutírico (Peake, James et al. 1996). En consecuencia, al no poder registrar el impulso nervioso, la estimulación es permanente y se producen los característicos espasmos musculares generalizados del tétanos (Todar 2006).

La toxina ha sido clasificada como una neurotoxina, y es un polipéptido precursor, con un peso alrededor de los 150kDa aunque posteriormente gracias a la acción de otra proteína se corta,

resultando en dos fragmentos de diferente peso, uno de 100kDa (Fragmento B) y otro de 50kDa (Fragmento A) manteniéndose unidos por un puente bisulfuro. Exclusivamente el fragmento B se adhiere a los gangliósidos GT y GD1b, por el lado del fragmento A se tienen una actividad enzimática tóxica. Al parecer el evento de unión es irreversible y solo puede eliminarse al desarrollar una terminal nerviosa nueva.

Tabla 1.2 Algunos de los integrantes más representativos de 4 de los principales géneros que caracteriza al género *Clostridium*.

Clostridium					
C. acetobutylicum ATCC 824	3,940,880	30%	3,844	pSOL1	Bacteria sacarolítica que produce compuestos de alto valor industrial como etanol, butanol y acetona.
C. difficile 630	4,290,252	29%	3,971	pCD630	Integrante de microbiota intestinal capaz de provocar Diarrea Asociada a Antibióticos y Colitis Pseudomembranosa Asociada a Antibióticos.
C. novyi NT	2,547,720	28%	2,437		Patógeno de animales de corral y en ocasiones de humanos. Potencial uso médico por sintetizar potentes efectos antitumorales.
C. perfringens 13	3,031,430	23%	2,786	pCP13	Patógeno capaz de producir gangrena gaseosa, infecciones quirúrgicas e infecciones uterinas por las invasinas y exotoxinas que es capaz de producir.
C. perfringens ATCC 13124	3,256,683	28%	3,017		
C. perfringens SM101	2,897,393	28%	2,701	Plasmid 1 Plasmid 2 Plasmid 3	
C. tetani E88	2,799,251	28%	2,445	pE88	Patógeno causante del tétanos cuya toxina es catalogada como aberración ya que no posee función alguna en su tipo de vida.
C. thermocellum ATCC 27405	3,843,301	38%	3,307		Único clostridio capaz de degradar celulosa y sus productos fermentativos primarios como alfa-celulosa y celobiosa.

Mycoplasma sp.

Género taxonómico perteneciente a la clase Mollicutes, que posee poco más de 100 especies y que han sido ampliamente estudiadas en años recientes debido a su participación en patologías crónicas tanto en animales como en seres humanos.

Presentan características metabólicas y morfológicas muy diversas tales como la completa ausencia de una pared celular, la presencia de esteroles en la membrana celular (Murray, Rosenthal et al. 2005), una total incapacidad de síntesis de purinas y hasta una seria variación en su código genético, específicamente en el codón de término (Li, Sherwood et al.), que en vez de ser utilizado para finalizar la síntesis de proteínas, fue modificado a uno adicional para el triptófano.

Dichas variaciones son debidas a una historia evolutiva sumamente vigorosa caracterizada por un parasitismo o comensalismo estricto, así como un promedio de desarrollo evolutivo inusualmente alto (Razin, Yogev et al. 1998). Se cree que su historia evolutiva está estrechamente vinculada a la del ser humano, en forma de una degeneración evolutiva que le hizo perder gran cantidad de sus genes ancestrales y una consecuente reducción de su genoma. Gracias a ésta característica en particular, varios integrantes del género se han convertido en modelos de estudios relacionados con el tamaño mínimo de genoma o genes esenciales para la vida.

La morfología del género *Mycoplasma* se presenta sumamente reducida, es decir, solo poseen una membrana, ribosomas y un cromosoma altamente condensado. De igual manera cabe destacar que los cromosomas de estos organismos son sumamente reducidos, con un contenido de G-C pequeño y uniforme, un tamaño genómico que varía dentro del rango de las 577 a 2220 kpb y, por consecuencia, la diversidad genética o proteica pocas veces supera las 1000 secuencias codificantes.

A pesar de la perdida de los genes relacionados con funciones vitales como el metabolismo oxidativo, gluconeogénesis, catalasas, peroxidasas o cualquier otra proteína de protección, aún conserva algunos de biosíntesis de ácidos nucleicos, vitaminas y ácidos grasos, aunque su actividad parasitaria es obligatoria y debe obtenerlo casi todo de su hospedero. Conservan genes involucrados en la replicación, transcripción y traducción del DNA, aunque son pocos los representantes del grupo de los RNA ribosomales y de transferencia.

Los genes asociados a la membrana son prácticamente inexistentes, en particular aquellos asociados al metabolismo de ácidos grasos, por lo que toman de su hospedero los fosfolípidos y glucolípidos necesarios y sintetizan, así, su propia membrana. El costo de haber ahorrado en genes lo paga en este punto en particular ya que son incapaces de controlar la fluidez de la membrana y están obligados a incorporar una enorme cantidad de colesterol, provista por el hospedero.

No tienen una forma definida ya que carecen de pared celular sin embargo son capaces de desplazarse suavemente sobre superficies sólidas y han desarrollado estructuras de anclaje a

membranas celulares. Tal situación sugiere que la membrana de *Mycoplasma* posee un citoesqueleto rudimentario que le permite extender estructuras tales como puntas o filamentos o adoptar distintas formas en el espacio intracelular (Razin, Yogev et al. 1998).

Sin variar, lo que ha provocado una alta especificidad con su hospedero. Los individuos portadores de cualquiera de las presentes bacterias llegan a desarrollar una cierta relación de comensalismo, aunque tal interacción no excluye la posibilidad de volverse infeccioso (Razin, Yogev et al. 1998)

Mycoplasma genitalium. Bacteria parásita de seres humanos cuyo genoma es reconocido como el más pequeño de todas las bacterias de vida libre, solo 580 074 pares de bases y 479 secuencias codificantes para proteínas. Generalmente se encuentra en el tejido urogenital y se contagia a través del contacto sexual, causando la inflamación de tejidos debido a residuos metabólicos desechados tales como peróxido de hidrógeno y metabolitos super-oxidantes. Recientemente fue localizado, también, en el tracto respiratorio de humanos (Razin, Yogev et al. 1998).

El bajo contenido G+C es una característica común entre el género *Mycoplasma* y en el caso particular de M. genitalium, es solo de 32%. En caso contrario, las zonas donde se encuentran los genes que codifican para RNAr y RNAt presentan un sólido 42 y 52%, respectivamente, esto a causa de la importancia de dichos genes y la imposibilidad práctica de perderlos por una mutación. Otros genes que igualmente se encuentran altamente conservados son todos aquellos relacionados con el transporte de nutrientes como fructosa y glucosa, así como los especializados en la formación de estructuras de anclaje, adhesinas y variación antigénica para evadir el sistema inmune de los hospederos. Se estima, incluso, que hasta el 5% del total de su DNA ha sido asignado para conservar fragmentos repetidos de adhesinas que serán usados para efectuar recombinación antigénica.

Es capaz de desplazarse sobre superficies a una velocidad de 0.1µm/s con el propósito de adherirse a las células eucariontes y penetrar en ellas, esto debido a la casi nula presencia de genes vinculados al metabolismo energético y la vital necesidad de apropiarse del ATP del hospedero.

Mycoplasma mobile. Esta bacteria es encontrada comúnmente en cuerpos de agua corriente y que se han adaptado a vivir como patógenos de algunas especies de peces. Se le reconoce como la única especie capaz de desplazarse sobre superficies a una velocidad de 7µm/s y poder arrastrar macrófagos 10 veces más grandes sin significativamente su velocidad; es capaz de desarrollar una fuerza de arrastre estimada en 27 pN (Hiratsuka, Miyata et al. 2006).

Su cromosoma tiene 770 079 pb, con un contenido de G+C de tan solo 24.9% y escasas 635 proteínas, de las cuales el 88% son expresadas. Presenta exclusivamente una copia del DNA de cada una de las subunidades ribosomales 16s, 23s y 5s, sin embargo la subunidad 5s no se localiza dentro del operón 16s-23s-5s sino que se ubica a <180° con respecto al cromosoma circular, sugiriendo esto que sucedió, en algún momento de su historia evolutiva, un re-arreglo genómico serio sobre este punto. Así mismo, posee una secuencia repetida 5 veces de 2 435 pb que codifican para proteínas muy parecidas, evidenciando un claro ejemplo de transferencia lateral de genes.

Sus estructuras de adherencia no son comunes dentro de eubacterias o eucariontes y no poseen genes que codifiquen para mecanismos de desplazamiento celular comunes al resto. Por el contrario, se ha identificado un proceso ATP-dependiente con grandes proteínas ultraestructurales involucradas en la adhesión celular. Las posibilidades industriales y médicas que ésta bacteria es capaz de ofrecer varían alrededor de su mecanismo de desplazamiento, el cual no tiene par en el mundo bacteriano y abriría las puertas a sistemas de locomoción que utilizaran azúcares para desplazar objetos.

Tabla 1.3 Algunos de los integrantes más representativos de 4 de los principales géneros que caracteriza al género Micoplasma.

Mycoplasma					
M. capricolum subsp. Capricolum ATCC 27343	1,010,023	23%	867	Probable	Agente causal de mastitis, artritis y enfermedades respiratorias de ganado caprino con una tasa de mortalidad de 60 hasta 70%.
M. gallisepticum R	996,422	31%	781		Agente causal de enfermedades respiratorias en aves de vida libre y de corral.
M. genitalium G37	580,074	31%	525		Agente causal de enfermedades del tracto urogenital de humanos. Bacteria con el genoma más pequeño conocido.
M. hyopneumoniae 232	892,758	28%	727		Agente causal de neumonía porcina.
M. hyopneumoniae 7448	920,079	28%	711		
M. hyopneumoniae J	897,405	28%	709		
M. mobile 163K	777,079	24%	667		Agente patógeno de peces. Posee la mayor movilidad de todos los *Mycoplasma*s con 7µm/s.
M. mycoides SC	1,211,703	23%	1,052		Agente causal de pleuroneumonía bovina contagiosa. Posee el mayor porcentaje de secuencias de inserción, 13% del total del genoma
M. penetrans HF-2	1,358,633	25%	1,069		Patógeno de humanos, específicamente de tracto urogenital y pulmonar. Asociado a VIH aunque puede invadir sin la presencia del virus.
M. pneumonia M129	816,394	40%	733		Agente causal de neumonía ligera en humanos, en un 10% de los casos es fatal.
M. pulmonis UAB CTIP	963,879	26%	815		Agente causal de micoplasmosis murínica respiratoria en ratones y ratas.

M. synoviae 53	799,476	28%	728		Patógeno de pollos y pavos, causando sinovitis y enfermedades del tracto respiratorio.

Listeria sp.

Perteneciente al gran grupo de bacterias Gram +, el género *Listeria* se caracteriza por su forma bacilar, carente del proceso de esporulación, capaz de multiplicarse en temperaturas cercanas a los 0°, un contenido bajo de G+C en su genoma, en la franja del 36 al 38%, y con presencia de 1 a 5 flagelos periféricos capaces de conferir movilidad a los 28°.

El género posee 6 especies: *Listeria monocytogenes, Listeria seeligeri, Listeria ivanovii, Listeria innocua, Listeria welshimeri* y *Listeria grayi*, de las cuales la primera es la única que genera interés en los servicios de salud ya que es capaz de desarrollar una serie de peligrosas patologías. Se distribuyen principalmente en el suelo y en materia orgánica en descomposición, con una tremenda resistencia a condiciones tan adversas como altas concentraciones de sales, altas temperaturas y niveles elevados de pH. Puede, además, encontrársele en sitios como drenajes, agua, carnes procesadas, leche cruda, quesos y humanos, aunque ha sido aislada de otros organismos vivos como son los peces, crustáceos, ostras, pulgas y garrapatas. Se estima que entre el 5 y 10% de la población mundial es portadora de éstas bacterias sin presentar síntomas de enfermedad alguna. Cuando las bacterias actúan como patógenas han sido definitivamente transportadas mediante comida contaminada y una vez que han ingresado en el organismo hospedero, utilizan filamentos de actina para desplazarse intracelularmente y alojarse en el sistema digestivo (Todar 2006).

La habilidad de *Listeria* para habitar esta diversidad de ambientes y condiciones se debe a la particularidad de poseer 331 genes codificantes a diferentes proteínas de transporte y un repertorio sumamente extenso de mecanismos de regulación. Además de ser una bacteria microaerofílica, es un anaerobio facultativo, por lo que es puede producir ATP mediante diferentes rutas de fermentación y también a través de una cadena respiratoria completa (Portnoy, Auerbuch et al. 2002).

Listeria monocytogenes. Agente causal de diferentes patologías tales como septicemia, meningitis, meningoencefalitis y listeriosis en individuos inmunodeprimidos, recién nacidos y ancianos. También es capaz de inducir abortos y partos prematuros. Su rango de mortalidad varía entre un 20 hasta un 40% (Benson, Karsch-Mizrachi et al. 2006). Es la única especie del género capaz de producir patologías en el humano, ya que presenta una amplia resistencia a los factores humorales de defensa y gracias a un gen de hemolisina presente en su cromosoma, puede escapar al citoplasma de los macrófagos y comenzar la infección.

El área donde se localiza al antes mencionado gen, es un operón muy particular ya que se constituye de diversos genes que se encuentran asociados a virulencia. Posee otro gen cuyo producto promueve la polimerización de actina, existente también en la superficie de las células del organismo hospedero, obligando a la célula hospedera a empujar al grupo de bacterias hacia el exterior, las cuales se han agrupado y recubierto con una lámina de actina. Una vez que arribaron a la superficie de la célula hospedera, ésta comienza a proyectar protuberancias saturadas de bacterias vivas, que

son reconocidas por las células adyacentes y engullidas. Tal proceso permite a *Listeria* invadir órganos completos sin necesidad de pasar por una fase extracelular (Peake, James et al. 1996) Podría deducirse, entonces, que aquellas bacterias que no produzcan hemolisinas son medianamente patógenas, ya que presentan mecanismos de adhesión a las células hospederas pero no son capaces de producir lisis en las mismas para cerrar su ciclo de virulencia.

Tabla 1.4 Algunos de los integrantes más representativos de 4 de los principales géneros que caracteriza al género *Listeria*.

Listeria					
L. innocua Clip11262	3,011,208	37%	3,065	pLI100	Bacteria capaz de producir bacteriocinas, la cuales inhiben el desarrollo de *L. monocytogenes*.
L. monocytogenes 4b F2365	2,905,310	38%	2,934		Agente causal de listeriosis en humanos y ganado vacuno. Pueden existir infecciones uterinas provocando abortos
L. monocytogenes EGD-e	2,944,528	37%	2,940		
L. welshimeri serovar 6b str. SLCC5334	2,814,130	36%	2,864		Bacteria no patógena.

II

Tabla 2. Descripción generalizada de las fases de la esporulación.

Fase 0		
La célula madre posee dos cromosomas al final del periodo de crecimiento exponencial		
FASE 1		
El material genético se condensa constituyendo un filamento axial ancho, que ocupa el centro de la célula, según su eje longitudinal. Cada nucleoide está unido a uno de los extremos de la célula.	Se forman dos espículas en la pared celular cercana a los polos por el interior, con su correspondiente invaginación de la membrana citoplasmática.	Se sintetizan y liberan al medio exoenzimas junto con antibióticos. El peptidoglicano sintetizado durante la septación es digerido para utilizar los aminoácidos y proceder a la Fase II.
FASE 2		
Se termina por formar un septo transversal acéntrico, cerca de un polo de la célula, por invaginación de la membrana citoplásmica, y deposición de nuevo peptidoglicano entre las dos membranas adyacentes.	Cada nucleoide queda segregado en uno de los dos compartimentos que se han formado, repartiéndose entre la endospora y la célula madre.	En esta fase continúa la síntesis de los antibióticos bacilisina y surfactina y de las exoenzimas serina-proteasas, metalo-proteasas, ribonucleasas, a-amilasa, etc.
FASE 3		
Independización del protoplasto de la pre-espora respecto a la célula madre mediante la degradación selectiva del peptidoglicano del septo que se había depositado en la fase I. Entonces, la membrana citoplásmica de la célula madre va creciendo unidireccionalmente alrededor de la pre-espora, hasta que ésta queda libre en el citoplasma del esporangio.	El citoplasma de la pre-espora queda rodeado por dos membranas de polaridad opuesta: la interior tiene la polaridad normal, pero la exterior, derivada del crecimiento de la membrana de la célula madre, tiene polaridad invertida.	Síntesis e incremento de ácido tricarboxílico y enzimas asociadas al ciclo del glioxilato.
FASE 4		
Se forma casi por completo la corteza de la espora, por deposición de peptidoglicano entre las dos membranas.	Se deposita el peptidoglicano de la pared celular, preliminar al de la pared celular de la futura célula.	Comienza la síntesis del ácido dipicolínico (Baweja, Zaman et al.), así como la acumulación de Ca2+.
FASE 5		
Se lleva a cabo el proceso de maduración de la pre-espora hasta espora, junto con cubiertas y corteza.	La espora adquiere resistencia a diversos agentes desnaturalizantes y mutagénicos del DNA como radiaciones UV, calor, lisozimas y solventes orgánicos.	El citoplasma esporal pierde agua y comienza a hacerse más denso adquiriendo una constitución similar a la de un gel.
FASE 6		
Culmina el proceso de esporulación con la liberación de la endospora al medio mediante la autólisis de la célula madre.		

Tabla 5. Todos los genes reconocidos hasta la actualidad que participan en la esporulación. (Elaborado a partir de KEGG Orthology http://www.genome.jp/kegg/)

bofA	Inhibitor of the pro-sigma K processing machinery
bofC	Forespore regulator of the sigma-K checkpoint
cgeA	Spore maturation protein CgeA

cgeB	Spore maturation protein CgeB
cgeC	Spore maturation protein CgeC
cgeD	Spore maturation protein CgeD
cgeE	Spore maturation protein CgeE
cotA	Spore coat protein A
cotB	Spore coat protein B
cotC	Spore coat protein C
cotD	Spore coat protein D
cotE	Spore coat protein E
cotF	Spore coat protein F
cotH	Spore coat protein H
cotI	Spore coat protein I
cotJA	Spore coat protein JA
cotJB	Spore coat protein JB
cotJC	Spore coat proteinJC
cotM	Spore coat protein M
cotN	Spore coat-associated protein N
cotS	Spore coat-associated protein S
cotSA	Spore coat protein SA
cotT	Spore coat protein T
cotV	Spore coat protein V
cotW	Spore coat protein W
cotX	Spore coat protein X
cotY	Spore coat protein Y
cotZ	Spore coat protein Z
yrbB, coxA	Spore cortex protein
jag	spoIIIJ-associated protein
kapB	Kinase-associated protein B
kapD	Sporulation inhibitor KapD
kbaA	KinB signaling pathway activation protein
kipA	Antagonist of KipI
kipI	Inhibitor of KinA
phrA	Phosphatase RapA inhibitor
phrC	Phosphatase RapC regulator
phrE	Phosphatase RapE regulator
phrF	Phosphatase RapF regulator
phrG	Phosphatase RapG regulator
phrI	Phosphatase RapI regulator
phrK	Phosphatase RapK regulator
rapA, spo0L	Response regulator aspartate phosphatase A (stage 0 sporulation protein L)
rapB	Response regulator aspartatephosphatase B
rapC	Response regulator aspartatephosphatase C
rapD	Response regulator aspartatephosphatase D
rapE	Response regulator aspartatephosphatase E
rapF	Response regulator aspartatephosphatase F
rapG	Response regulator aspartatephosphatase G
rapH	Response regulator aspartatephosphatase H
rapI	Response regulator aspartatephosphatase I
rapJ	Response regulator aspartatephosphatase J
rapK	Response regulator aspartatephosphatase K
safA	Morphogenetic protein associated with SpoVID
sda	Developmental checkpoint coupling sporulation initiation to replication initiation
sinI	Antagonist of SinR

spmA	Spore maturation protein A
spmB	Spore maturation protein B
spo0B	Stage 0 sporulation protein B (sporulation initiation phosphotransferase)
spo0E	Stage 0 sporulation regulatory protein
spo0M	Sporulation-control protein
spoIIAA	Stage II sporulation protein AA (anti-sigma F factor antagonist)
spoIIAB	Stage II sporulation protein AB (anti-sigma F factor)
spoIIB	Stage II sporulation protein B
spoIID	Stage II sporulation protein D
spoIIE	Stage II sporulation protein E
spoIIGA	Stage II sporulation protein GA (sporulation sigma-E factor processing peptidase)
spoIIM	Stage II sporulation protein M
spoIIP	Stage II sporulation protein P
spoIIQ	Stage II sporulation protein Q
spoIIR	Stage II sporulation protein R
spoIISA	Stage II sporulation protein SA
spoIISB	Stage II sporulation protein SB
spoIIIAA	Stage III sporulation protein AA
spoIIIAB	Stage III sporulation protein AB
spoIIIAC	Stage III sporulation protein AC
spoIIIAD	Stage III sporulation protein AD
spoIIIAE	Stage III sporulation protein AE
spoIIIAF	Stage III sporulation protein AF
spoIIIAG	Stage III sporulation protein AG
spoIIIAH	Stage III sporulation protein AH
spoIVA	Stage IV sporulation protein A
spoIVB	Stage IV sporulation protein B
spoIVCA	Site-specific DNA recombinase
spoIVFA	Stage IV sporulation protein FA
spoIVFB	Stage IV sporulation protein FB
spoVAA	Stage V sporulation protein AA
spoVAB	Stage V sporulation protein AB
spoVAC	Stage V sporulation protein AC
spoVAD	Stage V sporulation protein AD
spoVAE	Stage V sporulation protein AE
spoVAF	Stage V sporulation protein AF
spoVB	Stage V sporulation protein B
spoVD	Stage V sporulation protein D (sporulation-specific penicillin binding protein)
spoVFA	Dipicolinate synthase subunit A
spoVFB	Dipicolinate synthase subunit B
spoVG	Stage V sporulation protein G
spoVK	Stage V sporulation protein K
spoVM	Stage V sporulation protein M
spoVR	Stage V sporulation protein R
spoVS	Stage V sporulation protein S
spoVID	Stage VI sporulation protein D
spsF	Spore coat polysaccharide biosynthesis protein SpsF
sasP-A, sspA	Small acid-soluble spore protein A (major alpha-type SASP)
sasP-B, sspB	Small acid-soluble spore protein B (major beta-type SASP)
sspC	Small acid-soluble spore protein C (minor alpha/beta-type SASP)
sspD	Small acid-soluble spore protein D (minor alpha/beta-type SASP)
sspE	Small acid-soluble spore protein E (minor gamma-type SASP)
sspF	Small acid-soluble spore protein F (minor alpha/beta-type SASP)

sspG	Small acid-soluble spore protein G (minor)
sspH	Small acid-soluble spore protein H (minor)
sspI	Small acid-soluble spore protein I (minor)
sspJ	Small acid-soluble spore protein J (minor)
sspK	Small acid-soluble spore protein K (minor)
sspL	Small acid-soluble spore protein L (minor)
sspM	Small acid-soluble spore protein M (minor)
sspN	Small acid-soluble spore protein N (minor)
sspO, cotK	Small acid-soluble spore protein O (minor)
sspP, cotL	Small acid-soluble spore protein P (minor)
tgl	Protein-glutamine gamma-glutamyltransferase
tlp	Small acid-soluble spore protein (thioredoxin-like protein)
usd	Required for translation of spoIIID
yabG	Spore coat assemly protein
yknT	Sigma-E controlled sporulation protein
yqfD	Similar to stage IV sporulation protein
yraD	Similar to spore coat protein
yraG	Similar to spore coat protein

III

Fredj Tekaia, investigador del instituto Pasteur, propone una definición sobre la Bioinformática: son los métodos estadísticos, matemáticos y computacionales orientados a la resolución de problemas biológicos utilizando secuencias de DNA, aminoácidos e información relacionada. La incorporación de esta herramienta a las ciencias biológico-moleculares sucedió de forma gradual, de la mano con el avance en potencia y miniaturización de las computadoras que se hicieron accesibles al público en general. De esta manera, los análisis de datos comenzaron a hacerse en computadoras de escritorio, sin necesidad de grandes módulos de procesamiento o supercomputadoras. Durante la década de los 90's, con la explosión de la Internet, se pudo entonces acceder de forma inmediata a la información generada por otros investigadores, que poco a poco fue acumulándose en grandes bases de datos que ahora son administradas por grandes instituciones tanto privadas como de salud y permiten el libre acceso a cualquier individuo interesado.

Estas bases de datos se encuentran asociadas con programas computacionales diseñados específicamente para actualizar, buscar y obtener componentes de la información almacenada dentro del sistema. Los registros de los datos que se encuentran en las bases de datos están categorizados por nombre y proveen de información relacionada con la secuencia de entrada con la descripción del tipo de molécula, el nombre científico del organismo fuente e inclusive proporciona sugerencias de literatura. Es obligatorio que las bases de datos provean de un sencillo acceso a la información y que las herramientas proporcionadas estén diseñadas para proporcionar la información exacta para resolver una pregunta biológica específica. Algunos ejemplos de bases de datos de nucleótidos, proteínas y genomas se muestran en la Tabla 3.

Tabla 3. Algunas de las bases de datos más importantes del mundo con acceso abierto al público en general.

Nucleótidos		
EMBL Nucleotide Sequence Database Fuente principal de acceso a secuencias nucleotídicas en Europa http://www.ebi.ac.uk/embl	**NCBI GenBank** Colección de secuencias de diferentes orígenes como GenBank, RefSeq y PDB http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide	**DDBJ DNA Data Bank of Japan** Una de las 3 bases de datos cumbre de secuencias nucleotídicas. http://www.ddbj.nig.ac.jp
Proteínas		
SWISSPROT Contiene secuencias anotadas o comentadas es decir, cada secuencia ha sido revisa, documentada y enlazada a otras bases de datos http://www.ebi.ac.uk/swissprot/access.html	**UniProtKB/TrEMBL** Contiene la traducción de todas las secuencias codificantes presentes en el EMBL, GenBank y DDBJ. Igualmente posee secuencias proteicas extraídas de la literatura o enviadas a la UniProtKB/Swiss-Prot. http://www.ebi.ac.uk/trembl/	**PIR (Protein Resource Information)** Provee a la comunidad científica de una fuente centralizada, sencilla y de autoridad para secuencias de proteínas e información funcional. http://pir.georgetown.edu
RCSB PDB Colaboración de investigación para la Bioinformática Estructural con	**INTERPRO** Base de datos de familias proteicas, dominios y sitios funcionales en los que	**PROSITE** Base de datos de dominios

herramientas de visualización en 3D de estructuraas terciarias de proteínas http://www.rcsb.org/pdb/home/home.do	características identificables de proteínas conocidas pueden aplicarse a secuencias proteicas desconocidas http://www.ebi.ac.uk/interpro/index.html	proteicos, familias y sitios funcionales. http://us.expasy.org/prosite/
Genomas		
ENSEMBL Proyecto conjunto entre el EMBL-EBI y el Instituto Wellcome Trust Sanger que intenta desarrollar un sistema que mantenga la anotación automática de genomas de eucariontes. http://www.ebi.ac.uk/ensembl/index.html	**NCBI Entrez Genome** Base de datos que provee de perspectivas para una variedad de genomas, cromosomas completos, mapas de secuenciación y físicos. http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome	**J. Craig Venter Institute** Instituto de reciente creación producto de la fusión de 5 instituciones en el cual es posible acceder a los genomas completos de todos los organismos secuenciados hasta la actualidad. http://www.tigr.org/index.shtml

BLAST.

Existen varios programas diseñados a partir del mismo logaritmo pero poseen diferentes usos, a continuación se detalla cada uno de ellos:

BLASTp: Compara una secuencia de aminoácidos contra una base de datos de secuencias proteicas.

BLASTn: Compara una secuencia de nucleótidos contra una base de datos con secuencias de nucleótidos.

BLASTx: Compara una secuencia de nucleótidos traducida en todos sus posibles marcos de lectura contra una base de datos de secuencias proteicas. Es posible usar esta opción para encontrar posibles productos de traducción de una secuencia de nucleótidos desconocida.

tBLASTn: Compara una secuencia proteica contra una base de datos de nucleótidos dinámicamente traducida en todos los posibles marcos de lectura.

tBLASTx: Compara las traducciones de seis marcos de una secuencia de nucleótidos contra las traducciones de seis marcos de una base de datos de secuencias de nucleótidos.

Alineamientos globales Clustal. Es un programa computacional de alineamientos de secuencias múltiples, que viene en dos presentaciones distintas:

ClustalW: Interfaz de líneas de comando

ClustalX: Interfaz gráfica para usuario. Disponible para Windows, Unix/Linux y Mac.

El heurístico utilizado en ClustalW se basa en el análisis filogenético, realizando en primer lugar una serie de alineamientos por pares, comparando cada secuencia con todas las demás para construir una matriz de distancias, en la que se refleja la relación de cada secuencia con las demás. Esta matriz sirve a continuación, para calcular un árbol filogenético que es utilizado en los pasos posteriores para generar el alineamiento, empezando por el par de secuencias más relacionadas y añadiendo las demás sucesivamente. A medida que se añaden secuencias más divergentes, será necesario introducir huecos en el alineamiento (Larkin, Blackshields et al. 2007).

Dicha estrategia produce alineamientos razonables dentro de una serie de condiciones. No es a prueba de errores; para secuencias relativamente distanciadas, puede construirse en las imprecisiones del alineamiento por pares y el análisis filogenético. Pero para grupos de secuencias

con algunos pares relacionados, se construye en las fortalezas de éstos métodos. El análisis filogenético es poco ambiguo para secuencias relacionadas muy cercanas. Utilizando múltiples secuencias para crear perfiles incrementa la precisión del alineamiento por pares para las secuencias relacionadas más distanciadas (Gibas and Jambeck 2001).

Recientemente ha sido liberada al mundo informático una versión nueva de ClustalW, la versión 2.0, que ha sido re-escrita completamente en el lenguaje C++ con el objetivo de eliminar problemas de compatibilidad en las diferentes plataformas en las cuales es posible correr el programa, hacer más sencillo el proceso de mantenimiento del código y, en caso de ser necesario, modificar sencilla y rápidamente los algoritmos de alineamiento utilizados en el programa (Larkin, Blackshields et al. 2007).

IV.

Tabla 4.2 Información general de las bacterias cuyos genomas fueron utilizados para el presente trabajo.

Bacillus amyloliquefaciens FZB42		
Refseq: NC_009725	Genes: 3813	Hábitat: Suelo
GenBank: CP000560	Proteínas codificantes: 3693	Temp: Mesófilo
Status de secuencia: Completo	Longitud: 3,918,589 nt	Salinidad: No halófilo
	Contenido de GC: 46%	
	% Codificante: 88%	

Bacillus anthracis str. 'Ames Ancestor'		
Refseq: NC_007530	Genes: 5635	Hábitat: Suelo
GenBank: AE017334	Proteínas codificantes: 5309	Temp: Mesófilo
Status de secuencia: Completo	Longitud: 5,227,419 nt	Salinidad: No halófilo
	Contenido de GC: 35%	
	% Codificante: 80%	

Bacillus anthracis str. Sterne		
Refseq: NC_005945	Genes: 5415	Hábitat: Suelo
GenBank: AE017225	Proteínas codificantes: 5287	Temp: Mesófilo
Status de secuencia: Completo	Longitud: 5,228,663 nt	Salinidad: No halófilo
	Contenido de GC: 35%	
	% Codificante: 83%	

Bacillus anthracis str. Ames		
Refseq: NC_003997	Genes: 5630	Hábitat: Suelo
GenBank: AE016879	Proteínas codificantes: 5311	Temp: Mesófilo
Status de secuencia: Completo	Longitud: 5,227,293 nt	Salinidad: No halófilo
	Contenido de GC: 35%	
	% Codificante: 80%	

Bacillus cereus ATCC 10987		
Refseq: NC_003909	Genes: 5772	Hábitat: Suelo
GenBank: AE017194	Proteínas codificantes: 5603	Temp: Mesófilo
Status de secuencia: Completo	Longitud: 5,224,283 nt	Salinidad: No halófilo
	Contenido de GC: 35%	
	% Codificante: 84%	

Bacillus cereus ATCC 14579		
Refseq: NC_004722	Genes: 5481	Hábitat: Suelo
GenBank: AE016877	Proteínas codificantes: 5234	Temp: Mesófilo
Status de secuencia: Completo	Longitud: 5,411,809 nt	Salinidad: No halófilo
	Contenido de GC: 35%	
	% Codificante: 80%	

Bacillus cereus E33L		
Refseq: NC_006274	Genes: 5269	Hábitat: Suelo
GenBank: CP000001	Proteínas codificantes: 5134	Temp: Mesófilo
Status de secuencia: Completo	Longitud: 5,300,915 nt	Salinidad: No halófilo
	Contenido de GC: 35%	

	% Codificante: 83%	

Bacillus cereus subsp. cytotoxis NVH 391-98		
Refseq: NC_009674	Genes: 4165	Hábitat: Suelo
GenBank: CP000764	Proteínas codificantes: 3833	Temp: Mesófilo
Status de secuencia: Incompleta	Longitud: 4,087,024 nt	Salinidad: No halófilo
	Contenido de GC: 35%	
	% Codificante: 79%	

Bacillus clausii KSM-K16		
Refseq: NC_006582	Genes: 4204	Hábitat: Biota intestinal
GenBank: AP006627	Proteínas codificantes: 4096	Temp: Mesófilo
Status de secuencia: Incompleta	Longitud: 4,303,871 nt	Salinidad: No Halófilo
	Contenido de GC: 44%	pH: Alcalófila
	% Codificante: 85%	

Bacillus coagulans 36D1		
Refseq: NZ_AAWV00000000	Genes: 2721	Hábitat: Biota intestinal
GenBank: AAWV00000000	Proteínas codificantes: 2675	Temp: Termófila
Status de secuencia: Incompleta	Longitud: 2,941,017 nt	Salinidad: No Halófilo
	Contenido de GC: 46%	pH: Acidófilo
	% Codificante: 79%	

Bacillus coahuilensis M4-4		
Refseq: N/A	Genes: 3640	Hábitat: Acuático
GenBank: N/A	Proteínas codificantes:	Temp: Mesófila
Status de secuencia: Incompleta	Longitud: 3 358 093 nt	Salinidad: Halófila
	Contenido de GC:	
	% Codificante:	

Bacillus halodurans C-125		
Refseq: NC_002570	Genes: 4171	Hábitat: Halófilo-Marino
GenBank: BA000004	Proteínas codificantes: 4066	Temp: Mesófila
Status de secuencia: Completa	Longitud: 4,202,352 nt	Salinidad: Halófila
	Contenido de GC: 43%	pH: Alcalófila
	% Codificante: 85%	

Bacillus licheniformis ATCC 14580		
Refseq: NC_006322	Genes: 4289	Hábitat: Suelo
GenBank: AE017333	Proteínas codificantes: 4196	Temp: Mesófilo
Status de secuencia: Completa	Longitud: 4,222,645 nt	Salinidad: No halófilo
	Contenido de GC: 46%	
	% Codificante: 86%	

Bacillus pumilus SAFR-032		
Refseq: NC_009848	Genes: 3825	Hábitat: Suelo
GenBank: CP000813	Proteínas codificantes: 3681	Temp: Mesófilo
Status de secuencia: Completa	Longitud: 3,704,465 nt	Salinidad: No halófilo
	Contenido de GC: 41%	
	% Codificante: 86%	

Bacillus sp. B14905		
Refseq: NZ_AAXV00000000	Genes: 4750	Hábitat: Marino
GenBank: AAXV00000000	Proteínas codificantes: 4624	Temp: Mesófilo
Status de secuencia: Incompleta	Longitud: 4,497,271 nt	Salinidad: Halófila
	Contenido de GC: 37%	
	% Codificante: 85%	

Bacillus sp. NRRL B-14911		
Refseq: NZ_AAOX00000000	Genes: 5797	Hábitat: Marino
GenBank: AAOX00000000	Proteínas codificantes: 5691	Temp: Mesófilo
Status de secuencia: Incompleta	Longitud: 5,085,825 nt	Salinidad: Halófila
	Contenido de GC: 45%	
	% Codificante: 86%	

Bacillus sp. SG-1		
Refseq: NZ_ABCF00000000	Genes: 4426	Hábitat: Marino
GenBank: ABCF00000000	Proteínas codificantes: 4337	Temp: Mesófilo
Status de secuencia: Incompleta	Longitud: 3,948,965 nt	Salinidad: Halófila
	Contenido de GC: 42%	
	% Codificante: 80%	

Bacillus subtilis subsp. subtilis str. 168		
Refseq: NC_000964	Genes: 4225	Hábitat: Suelo
GenBank: AL009126	Proteínas codificantes: 4105	Temp: Mesófilo
Status de secuencia: Completa	Longitud: 4,214,630 nt	Salinidad: No halófila
	Contenido de GC: 43%	
	% Codificante: 87%	

Bacillus thuringiensis serovar konkukian str. 97-27		
Refseq: NC_005957	Genes: 5263	Hábitat: Suelo
GenBank: AE017355	Proteínas codificantes: 5117	Temp: Mesófilo
Status de secuencia: Completa	Longitud: 5,237,682 nt	Salinidad: No halófila
	Contenido de GC: 35%	
	% Codificante: 83%	

Bacillus thuringiensis serovar israelensis ATCC 35646		
Refseq: NZ_AAJM00000000	Genes: 6229	Hábitat: Suelo
GenBank: AAJM00000000	Proteínas codificantes: 6132	Temp: Mesófilo
Status de secuencia: Incompleta	Longitud: 5,880,839 nt	Salinidad: No halófila
	Contenido de GC: 35%	
	% Codificante: 76%	

Bacillus thuringiensis str. Al Hakam		
Refseq: NC_008600	Genes: 4883	Hábitat: Suelo
GenBank: CP000485	Proteínas codificantes: 4736	Temp: Mesófilo
Status de secuencia: Completa	Longitud: 5,257,091 nt	Salinidad: No halófila
	Contenido de GC: 35%	pH: Alcalófilo
	% Codificante: 82%	

Bacillus weihenstephanensis KBAB4

Refseq: NZ_AAOY00000000	Genes: 5629	Hábitat: Suelo
GenBank: AAOY00000000	Proteínas codificantes: 5532	Temp: Psicrotolerante
Status de secuencia: Incompleta	Longitud: 5,602,503 nt	Salinidad: No halófila
	Contenido de GC: 35%	
	% Codificante: 81%	

Geobacillus kaustophilus HTA426		
Refseq: NC_006510	Genes: 3612	Hábitat: Marino
GenBank: BA000043	Proteínas codificantes: 3498	Temp: Termófila
Status de secuencia: Completa	Longitud: 3,544,776 nt	Salinidad: Halófila
	Contenido de GC: 52%	
	% Codificante: 84%	

Oceanobacillus iheyensis HTE831		
Refseq: NC_004193	Genes: 3594	Hábitat: Marino
GenBank: BA000028	Proteínas codificantes: 3500	Temp: Mesófila
Status de secuencia: Completa	Longitud: 3,630,528 nt	Salinidad: Halófila
	Contenido de GC: 35%	pH: Alcalófilo
	% Codificante: 84%	

V

Las siguientes son las líneas de commando utilizadas de Perl para llevar a cabo el filtrado de los resultados obtenidos mediante el análisis BLAST bidireccional.

```
perl -ne 'BEGIN {$column = 0}' -e 'BEGIN {$unique=0}; s/\r?\n//; @F=split /\t/, $_;
if (!($saveñppñp-p`p´-++)) {print "$_\n"; $unique++} END {warn "Chose $unique
unique lines out of $. total lines\nRemoved duplicates in column $column.\n"}'
$exit.bout > unique.out

perl -ne 'BEGIN {$name_col=0; $score_col=11;}' -e 's/\r?\n//; @F=split /\t/, $_;
($n, $s) = @F[$name_col, $score_col]; if (! exists($max{$n})) (Pushker, Mira et
al.); if (! exists($max{$n}) || $s >$max{$n}) {$max{$n} = $s; $best{$n} = ()}; if
($s == $max{$n}) {$best{$n} .= "$_\n"}; END {for $n (@names) {print $best{$n}}}'
unique.out > best.out

perl -e '$col1=1; $col2=0;' -e '($f1,$f2)=@ARGV; open(F1,$f1); while (<F1>)
{s/\r?\n//; @F=split /\t/, $_; $line1{$F[$col1]} .= "$_\n"} open(F2,$f2); while
(<F2>) {s/\r?\n//;@F=split /\t/, $_; if ($x = $line1{$F[$col2]}) {$x =~
s/\n/\t$_\n/g; print $x}}' best0.out best1.out > two_genomes_merged.out

perl -ne 'BEGIN {$colm=0; $coln=13;}' -e 's/\r?\n//; @F=split /\t/, $_; if
($F[$colm] eq $F[$coln]) {print "$_\n"}' two_genomes_merged.out >
reciprocal_best_hits
```

VI

Tabla 9. Genes que componen al genoma núcleo al corte 15 con una N de 510.

Anotation	COG
Putative L-lactate permease	C
Aldehyde dehydrogenase, mitochondrial precursor	C
Fumarate hydratase class II	C
Pyruvate dehydrogenase E1 component, alpha subunit	C
Pyruvate dehydrogenase E1 component, beta subunit	C
Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex	C
Dihydrolipoyl dehydrogenase	C
Cytochrome c oxidase polypeptide II precursor	C
Cytochrome c oxidase polypeptide I	C
Cytochrome c oxidase polypeptide III	C
2-oxoglutarate dehydrogenase E1 component	C
Dihydrolipoyllysine-residue succinyltransferase component of 2- oxoglutarate dehydrogenase complex	C
Succinyl-CoA synthetase beta chain	C
Succinyl-CoA synthetase alpha chain	C
Aerobic glycerol-3-phosphate dehydrogenase	C
Glycerol kinase	C
Menaquinol-cytochrome c reductase iron-sulfur subunit	C
Menaquinol-cytochrome c reductase cytochrome b subunit	C
Menaquinol-cytochrome c reductase cytochrome b/c subunit	C
Glycerophosphoryl Diester Phosphodiesterase	C
2-oxoisovalerate dehydrogenase alpha subunit	C
2-oxoisovalerate dehydrogenase beta subunit	C
Succinate dehydrogenase cytochrome b558 subunit	C
Electron transfer flavoprotein beta-subunit	C
Acetate kinase	C
Citrate synthase 2	C
Isocitrate dehydrogenase [NADP]	C
Malate dehydrogenase	C
Phosphoenolpyruvate carboxykinase [ATP]	C
1-pyrroline-5-carboxylate dehydrogenase	C
Na(+)/H(+) antiporter subunit A	C
Na(+)/H(+) antiporter subunit D	C
NifU-like protein	C
ATP synthase a chain	C
ATP synthase B chain	C
ATP synthase delta chain	C
ATP synthase gamma chain	C
Phosphate acetyltransferase	C
L-lactate dehydrogenase P	C
tRNA uridine 5-carboxymethylaminomethyl modification enzyme gidA	D
Sporulation initiation inhibitor protein soj	D
tRNA(Ile)-lysidine synthase	D
Protein mrp homolog salA	D
Anthranilate phosphoribosyltransferase	D

Tryptophan synthase alpha chain	D
Cell division protein ftsA	D
Cell division protein ftsZ	D
Cell-Division Initiation Protein	D
Chromosome partition protein smc	D
Rod shape-determining protein mreB	D
Probable septum site-determining protein minC	D
Septum site-determining protein minD	D
Septation ring formation regulator ezrA	D
Stage II sporulation protein D	D
Para-aminobenzoate synthase component I	E
Cysteine synthase	E
Serine acetyltransferase	E
Diaminopimelate decarboxylase	E
Aminopeptidase ampS	E
Transaminase mtnE	E
Asparagine synthetase [glutamine-hydrolyzing] 3	E
Oligoendopeptidase F homolog	E
2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase	E
Indole-3-glycerol phosphate synthase	E
Tryptophan synthase beta chain	E
Carbamoyl-phosphate synthase pyrimidine-specific small chain	E
Probable L-serine dehydratase, beta chain	E
Dihydrodipicolinate synthase	E
Aspartokinase 1	E
Aspartate-semialdehyde dehydrogenase	E
Pyrroline-5-carboxylate reductase 1	E
Hydroxymethylglutaryl-CoA lyase, mitochondrial precursor	E
Probable L-asparaginase 4 precursor	E
Chorismate synthase	E
3-dehydroquinate synthase	E
Histidinol-phosphate aminotransferase	E
Prephenate dehydrogenase	E
3-phosphoshikimate 1-carboxyvinyltransferase	E
Dihydrodipicolinate reductase	E
Homoserine dehydrogenase	E
Threonine synthase	E
Homoserine kinase	E
Shikimate kinase	E
Aminomethyltransferase	E
Probable glycine dehydrogenase [decarboxylating] subunit 1	E
Probable glycine dehydrogenase [decarboxylating] subunit 2	E
Putative peptidase yqhT	E
Leucine dehydrogenase	E
Probable cysteine desulfurase	E
Cystathionine beta-lyase	E
Shikimate dehydrogenase	E
Prephenate dehydratase	E
AroA(G) protein [Includes: Phospho-2-dehydro-3-deoxyheptonate aldolase	E
Alanine dehydrogenase	E
Ornithine aminotransferase	E
Putative aminotransferase B	E
D-alanine aminotransferase	E

Glutaminase 1	E
ATP phosphoribosyltransferase regulatory subunit	E
ATP phosphoribosyltransferase	E
Histidinol dehydrogenase	E
Imidazoleglycerol-phosphate dehydratase	E
Imidazole glycerol phosphate synthase subunit hisH	E
1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino] imidazole-4-carboxamide isomerase	E
Imidazole glycerol phosphate synthase subunit hisF	E
Glycine cleavage system H protein	E
Spermidine synthase 1	E
Inosine-5'-monophosphate dehydrogenase	F
Hypoxanthine-guanine phosphoribosyltransferase	F
Thymidylate kinase	F
Adenylate kinase	F
GMP synthase [glutamine-hydrolyzing]	F
Phosphoribosylaminoimidazole carboxylase catalytic subunit	F
Phosphoribosylaminoimidazole carboxylase ATPase subunit	F
Adenylosuccinate lyase	F
Phosphoribosylaminoimidazole-succinocarboxamide synthase	F
Phosphoribosylformylglycinamidine synthase I	F
Amidophosphoribosyltransferase precursor	F
Phosphoribosylformylglycinamidine cyclo-ligase	F
Phosphoribosylglycinamide formyltransferase	F
Bifunctional purine biosynthesis protein purH [Includes: Phosphoribosylaminoimidazolecarboxamide formyltransferase	F
Phosphoribosylamine--glycine ligase	F
Putative adenine deaminase BH0637	F
Thymidylate synthase	F
PyrR bifunctional protein [Includes: Pyrimidine operon regulatory protein; Uracil phosphoribosyltransferase	F
Uracil permease	F
Aspartate carbamoyltransferase	F
Dihydroorotase	F
Orotidine 5'-phosphate decarboxylase	F
Orotate phosphoribosyltransferase	F
Guanylate kinase	F
Uridylate kinase	F
Purine nucleoside phosphorylase I	F
Pyrimidine-nucleoside phosphorylase	F
Cytidylate kinase	F
Nucleoside diphosphate kinase	F
Deoxyribose-phosphate aldolase	F
Cytidine deaminase	F
Uridine kinase	F
ComE operon protein 2	F
HAM1 protein homolog	F
Adenylosuccinate synthetase	F
CTP synthase	F
Thymidine kinase	F
Uracil phosphoribosyltransferase	F
Phosphoglucosamine mutase	G
Fructokinase	G
Probable multiple sugar-binding transport ATP-binding protein msmX	G

Maltodextrin transport system permease protein malC	G
Maltodextrin transport system permease protein malD	G
Probable inorganic polyphosphate/ATP-NAD kinase	G
Phosphoenolpyruvate-protein phosphotransferase	G
Inositol-1-monophosphatase	G
1-phosphofructokinase	G
Transketolase	G
Phosphopentomutase	G
6-phosphogluconate dehydrogenase, decarboxylating 2	G
Stage III sporulation protein AE	G
6-phosphofructokinase	G
Pyruvate kinase	G
N-acetylglucosamine-6-phosphate deacetylase	G
PTS system trehalose-specific EIIBC component	G
Glyceraldehyde-3-phosphate dehydrogenase	G
2,3-bisphosphoglycerate-independent phosphoglycerate mutase	G
Enolase	G
Glucosamine-6-phosphate deaminase	G
Probable fructose-bisphosphate aldolase 1	G
Transaldolase	G
Glucose-6-phosphate isomerase A	G
Glutamine amidotransferase subunit pdxT	H
Pyridoxal biosynthesis lyase pdxS	H
2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	H
Dihydropteroate synthase	H
NH(3)-dependent NAD(+) synthetase	H
Glutamate-1-semialdehyde 2,1-aminomutase 2	H
Uroporphyrinogen decarboxylase	H
Ferrochelatase	H
Protoporphyrinogen oxidase	H
Phosphopantetheine adenylyltransferase	H
Riboflavin biosynthesis protein ribAB [Includes: GTP cyclohydrolase-2	H
6,7-dimethyl-8-ribityllumazine synthase	H
Dihydroorotate dehydrogenase electron transfer subunit	H
Dihydroorotate dehydrogenase, catalytic subunit	H
Coenzyme A biosynthesis bifunctional protein coaBC	H
Riboflavin biosynthesis protein ribC [Includes: Riboflavin kinase	H
Phosphomethylpyrimidine kinase	H
Menaquinone biosynthesis methyltransferase ubiE	H
Heptaprenyl diphosphate synthase component II	H
BirA bifunctional protein [Includes: Biotin operon repressor; Biotin-- [acetyl-CoA-carboxylase] synthetase	H
3-methyl-2-oxobutanoate hydroxymethyltransferase	H
Pantoate--beta-alanine ligase	H
FolD bifunctional protein [Includes: Methylenetetrahydrofolate dehydrogenase	H
Geranyltranstransferase	H
Probable nicotinate-nucleotide adenylyltransferase	H
Glutamyl-tRNA reductase	H
Porphobilinogen deaminase	H
Uroporphyrinogen-III synthase	H
Delta-aminolevulinic acid dehydratase	H
Folylpolyglutamate synthase	H
Probable thiamine biosynthesis protein thiI	H

Dephospho-CoA kinase	H
S-adenosylmethionine synthetase	H
Lipoyl synthase	H
4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	I
3-oxoacyl-[acyl-carrier-protein] synthase III protein 1	I
3-oxoacyl-[acyl-carrier-protein] synthase II	I
Fatty acid/phospholipid synthesis protein plsX	I
Malonyl CoA-acyl carrier protein transacylase	I
Acyl carrier protein	I
Phosphatidate cytidylyltransferase	I
Undecaprenyl pyrophosphate synthetase	I
CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase	I
Enoyl-[acyl-carrier-protein] reductase [NADH]	I
Biotin carboxyl carrier protein of acetyl-CoA carboxylase	I
3-hydroxybutyryl-CoA dehydratase	I
3-Hydroxyisobutyrate Dehydrogenase	I
Acetyl-coenzyme A synthetase	I
Acetyl-coenzyme A carboxylase carboxyl transferase subunit beta	I
Acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha	I
3-ketoacyl-CoA thiolase 2, peroxisomal precursor	I
(3R)-hydroxymyristoyl-[acyl carrier protein] dehydratase	I
Probable 3-hydroxybutyryl-CoA dehydrogenase	I
Ribonuclease P protein component	J
GTP-dependent nucleic acid-binding protein engD	J
50S ribosomal protein L9	J
Seryl-tRNA synthetase	J
Lysyl-tRNA synthetase	J
Peptidyl-tRNA hydrolase	J
Dimethyladenosine transferase	J
Methionyl-tRNA synthetase	J
Glutamyl-tRNA synthetase	J
Cysteinyl-tRNA synthetase	J
TRNA/RRNA Methyltransferase	J
50S ribosomal protein L11	J
50S ribosomal protein L10	J
30S ribosomal protein S7	J
Elongation factor G	J
Elongation factor Tu	J
50S ribosomal protein L2	J
30S ribosomal protein S3	J
50S ribosomal protein L5	J
30S ribosomal protein S8	J
50S ribosomal protein L6	J
50S ribosomal protein L18	J
30S ribosomal protein S5	J
50S ribosomal protein L15	J
Methionine aminopeptidase	J
30S ribosomal protein S11	J
Glutamyl-tRNA(Gln) amidotransferase subunit C	J
Glutamyl-tRNA(Gln) amidotransferase subunit A	J
Aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B	J
23S RRNA-Methyltransferase	J
Tryptophanyl-tRNA synthetase	J

Isoleucyl-tRNA synthetase 1	J
Methionyl-tRNA formyltransferase	J
Ribosomal RNA small subunit methyltransferase B	J
30S ribosomal protein S16	J
Probable 16S rRNA processing protein rimM	J
tRNA	J
Polyribonucleotide nucleotidyltransferase	J
Translation initiation factor IF-2	J
Ribosome recycling factor	J
Elongation factor Ts	J
30S ribosomal protein S2	J
tRNA delta(2)-isopentenylpyrophosphate transferase	J
30S ribosomal protein S1 homolog	J
Asparaginyl-tRNA synthetase	J
Elongation factor P	J
Hemolysin A	J
Ribosomal protein L11 methyltransferase	J
Histidyl-tRNA synthetase 2	J
Aspartyl-tRNA synthetase	J
Probable tRNA	J
Alanyl-tRNA synthetase	J
50S ribosomal protein L21	J
S-adenosylmethionine:tRNA ribosyltransferase-isomerase	J
Queuine tRNA-ribosyltransferase	J
Valyl-tRNA synthetase	J
Ribonuclease PH	J
TRNA/RRNA Methyltransferase	J
Phenylalanyl-tRNA synthetase alpha chain	J
Phenylalanyl-tRNA synthetase beta chain	J
Tyrosyl-tRNA synthetase 1	J
30S ribosomal protein S4	J
Threonyl-tRNA synthetase 1	J
Translation initiation factor IF-3	J
Leucyl-tRNA synthetase	J
Peptide chain release factor 1	J
Arginyl-tRNA synthetase	J
50S ribosomal protein L31 type B	J
Protein yyaA	K
Transcriptional Regulator	K
Stage V sporulation protein T	K
RNA polymerase sigma-H factor	K
Transcription antitermination protein nusG	K
DNA-directed RNA polymerase beta chain	K
DNA-directed RNA polymerase beta' chain	K
DNA-directed RNA polymerase alpha chain	K
RNA polymerase sigma factor sigW	K
Transcription antiterminator licT	K
Transcriptional Regulator DeoR Family	K
Ribonuclease III	K
Transcription elongation protein nusA	K
Glycerol uptake operon antiterminator regulatory protein	K
LexA repressor	K
Segregation and condensation protein B	K

N utilization substance protein B homolog	K
Arginine repressor	K
RNA polymerase sigma factor rpoD	K
Heat-inducible transcription repressor hrcA	K
Transcription elongation factor greA	K
Catabolite control protein A	K
Trehalose operon transcriptional repressor	K
Central glycolytic genes regulator	K
DNA-directed RNA polymerase delta subunit	K
DNA gyrase subunit A	L
DNA gyrase subunit B	L
DNA replication and repair protein recF	L
DNA polymerase III beta subunit	L
Chromosomal replication initiator protein dnaA	L
Single-strand binding protein 1	L
Replicative DNA helicase	L
Recombination protein recR	L
DNA polymerase III gamma/tau subunit	L
Transcription-repair coupling factor	L
Putative deoxyribonuclease yabD	L
DNA polymerase III delta' subunit	L
ATP-dependent DNA helicase pcrA	L
A/G-specific adenine DNA glycosylase	L
Methylated-DNA--protein-cysteine methyltransferase	L
Primosomal protein N'	L
ATP-dependent DNA helicase recG	L
Ribonuclease HII	L
DNA topoisomerase I	L
Tyrosine recombinase xerC	L
DNA polymerase III polC-type	L
Protein recA	L
DNA mismatch repair protein mutS	L
DNA mismatch repair protein mutL	L
Topoisomerase IV subunit A	L
Topoisomerase IV subunit B	L
ADP-ribose pyrophosphatase	L
Tyrosine recombinase xerD	L
DNA-binding protein HU 1	L
Probable endonuclease III	L
Probable exodeoxyribonuclease VII large subunit	L
DNA primase	L
Probable endonuclease IV	L
DNA repair protein recO	L
Single-stranded-DNA-specific exonuclease recJ	L
Putative Holliday junction resolvase	L
ComE operon protein 1	L
DNA Polymerase III Subunit Delta	L
Holliday junction DNA helicase ruvB	L
DNA repair protein radC homolog	L
UvrABC system protein C	L
DNA polymerase III alpha subunit	L
DNA polymerase I	L
Formamidopyrimidine-DNA glycosylase	L

Primosomal protein dnaI	L
ComF operon protein 1	L
UvrABC system protein B	L
UvrABC system protein A	L
Uracil-DNA glycosylase	L
Methyltransferase gidB	M
D-alanyl-D-alanine carboxypeptidase precursor	M
Bifunctional gcaD protein	M
Germination-specific N-acetylmuramoyl-L-alanine amidase	M
Glucosamine--fructose-6-phosphate aminotransferase [isomerizing]	M
D-alanine--D-alanine ligase B	M
Alanine racemase	M
Penicillin-binding protein 1F	M
UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase	M
S-adenosyl-methyltransferase mraW	M
Penicillin-binding protein 2B	M
Stage V sporulation protein D	M
Phospho-N-acetylmuramoyl-pentapeptide-transferase	M
UDP-N-acetylmuramoylalanine--D-glutamate ligase	M
UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase 2	M
Lipoprotein signal peptidase	M
Penicillin-binding protein dacF precursor	M
Spore cortex-lytic enzyme precursor	M
Stage IV sporulation protein B	M
GTP-binding protein lepA	M
Rod shape-determining protein mreC	M
Stage IV sporulation protein FA	M
Glutamate racemase	M
UDP-N-acetylmuramate--L-alanine ligase	M
UTP--glucose-1-phosphate uridylyltransferase	M
Carboxy-terminal-processing protease precursor	M
Prolipoprotein diacylglyceryl transferase	M
D-alanyl-D-alanine carboxypeptidase	M
UDP-N-acetylglucosamine 1-carboxyvinyltransferase 2	M
Stage II Sporulation Protein	M
Flagellar basal-body rod protein flgC	N
Flagellar motor switch protein fliG	N
Flagellum-specific ATP synthase	N
Flagellar hook protein flgE	N
Flagellar motor switch protein fliY	N
Flagellar biosynthetic protein fliR	N
Chemotaxis protein methyltransferase	N
ComG operon protein 1	N
Flagellar Motor Protein	N
Flagellar Motor Protein	N
Flagellar hook-associated protein 1	N
Flagellin	N
Flagellar hook-basal body complex protein flhO	N
Glutathione peroxidase homolog bsaA	O
33 kDa chaperonin	O
Negative regulator of genetic competence clpC/mecB	O
DNA repair protein radA homolog	O

Probable O-sialoglycoprotein endopeptidase	O
10 kDa chaperonin	O
Putative peroxiredoxin bcp	O
Cytochrome aa3-controlling protein	O
Protoheme IX farnesyltransferase	O
ATP-dependent hsl protease ATP-binding subunit hslU	O
Peptidyl-prolyl cis-trans isomerase B	O
Protein resB	O
Protein grpE	O
Chaperone protein dnaJ	O
Trigger factor	O
ATP-dependent Clp protease ATP-binding subunit clpX	O
ATP-dependent protease La 1	O
Protein hemX	O
Peptide methionine sulfoxide reductase msrA/msrB	O
ATP-dependent Clp protease proteolytic subunit 2	O
SsrA-binding protein	O
Foldase protein prsA precursor	O
Cobalt import ATP-binding protein cbiO 2	P
Cobalt import ATP-binding/permease protein cbiO	P
Ferric uptake regulation protein	P
Superoxide dismutase [Mn]	P
Protein sphX precursor	P
Probable ABC transporter permease protein yqgH	P
Phosphate transport system permease protein pstA-1	P
Phosphate import ATP-binding protein pstB	P
Zinc uptake system ATP-binding protein zurA	P
Zinc-specific metalloregulatory protein	P
Ferric enterobactin transport ATP-binding protein fepC	P
Copper-transporting P-type ATPase copA	P
Alkaline phosphatase III precursor	P
Ribosomal-protein-alanine acetyltransferase	R
Redox-sensing transcriptional repressor rex	R
Protein moxR	R
Regulatory protein recX	R
Sugar Nucleotide Epimerase	R
3'-5' exoribonuclease yhaM	R
Hemoglobin-like protein yjbI	R
Peptidase	R
Probable GTPase engC	R
GTPase	R
Peptidase	R
CinA-like protein	R
Protein hfq	R
GTP-binding protein hflX	R
GTP-binding protein engA	R
Recombination protein U	R
Ribonuclease Z	R
GTP-binding protein era homolog	R
ComE operon protein 3	R
Spo0B-associated GTP-binding protein	R
Stage V sporulation protein B	R
Probable GTP-binding protein engB	R

Stage IV sporulation protein FB	R
tRNA	R
Acetoin utilization acuB protein	R
Protein rarD	R
ComF operon protein 3	R
Pyrophosphatase ppaX	R
Carboxylesterase precursor	R
GTP Pyrophosphokinase	S
N-(5'-phosphoribosyl)anthranilate isomerase	S
Hypothetical UPF0124 protein ylmD	S
Stage II sporulation protein M	S
Segregation and condensation protein A	S
Stage II sporulation protein E	T
GTP-binding protein typA/bipA homolog	T
Protein ccdB	T
Anti-sigma F factor antagonist	T
Transcriptional regulatory protein resD	T
Sensor protein resE	T
Stage 0 sporulation protein A	T
PhoH-like protein	T
Alkaline phosphatase synthesis sensor protein phoR	T
Arsenate reductase	T
HPr kinase/phosphorylase	T
Membrane protein oxaA 1 precursor	U
Preprotein translocase secY subunit	U
Cell division protein ftsY homolog	U
Signal recognition particle protein	U
Preprotein Translocase Subunit YajC	U
Protein-export membrane protein secD	U
Preprotein translocase secA subunit	U
MazG Family Protein	Y
Nitric oxide synthase oxygenase	Y
Sporulation sigma-E factor processing peptidase	Y
DNA replication protein dnaD	Y
Putative propionyl-CoA carboxylase beta chain	Y
Stage III sporulation protein AA	Y
Stage III sporulation protein AB	Y
Stage III sporulation protein AD	Y
Stage III sporulation protein AG	Y
Stage III sporulation protein AH	Y
Germination protease precursor	Y
Stage II sporulation protein P	Y
Stage VI sporulation protein D	Y
Germination protein gerM	Y
Acetoin utilization protein acuA	Y
Stage II sporulation protein R	Y

Tabla 10. Genes que componen al genoma núcleo al corte 14 con una N total de 699 genes, (199 genes nuevos).

Anotation	COG
Acetoin utilization protein acuC	B
ABC Transporter Permease Protein	C
Aconitate hydratase	C
ATP synthase alpha chain	C
ATP synthase beta chain	C
Dihydrolipoyl dehydrogenase	C
Electron transfer flavoprotein alpha-subunit	C
Glycerol-3-phosphate dehydrogenase [NAD(P)+]	C
Isocitrate lyase	C
Nitro/flavin reductase	C
Probable NAD-dependent malic enzyme 4	C
Quinone oxidoreductase	C
Succinate dehydrogenase flavoprotein subunit	C
Succinate dehydrogenase iron-sulfur protein	C
Succinate-semialdehyde dehydrogenase [NADP+]	C
Cell division ATP-binding protein ftsE	D
Cell division protein ftsX homolog	D
DNA translocase ftsK	D
Agmatinase	E
Amino acid carrier protein alsT	E
Aminodeoxychorismate lyase	E
Aspartate aminotransferase	E
Carbamoyl-phosphate synthase pyrimidine-specific large chain	E
Carboxypeptidase	E
Choline transport ATP-binding protein opuBA	E
ComE operon protein 4	E
Diaminopimelate epimerase	E
Dipeptide transport system permease protein dppB	E
Dipeptide transport system permease protein dppC	E
Glutamate 5-kinase 2	E
Glutamate synthase [NADPH] large chain precursor	E
NAD-specific glutamate dehydrogenase	E
NifS/icsS protein homolog	E
Oligopeptide transport ATP-binding protein appD	E
Oligopeptide transport ATP-binding protein oppF	E
Para-aminobenzoate/anthranilate synthase glutamine amidotransferase component II [Includes: Para-aminobenzoate synthase glutamine amidotransferase component II	E
Peptidase T	E
Probable amino-acid ABC transporter ATP-binding protein yckI	E
Probable amino-acid ABC transporter ATP-binding protein yqiZ	E
Probable amino-acid ABC transporter extracellular-binding protein yckK precursor	E
Probable amino-acid ABC transporter permease protein yckJ	E
Probable cytosol aminopeptidase	E
Probable L-serine dehydratase, alpha chain	E
Serine hydroxymethyltransferase	E
Sodium/proline symporter	E
Spermidine/putrescine import ATP-binding protein potA	E

Xaa-Pro dipeptidase	E
MTA/SAH nucleosidase	F
Phosphoribosylformylglycinamidine synthase II	F
Protein hit	F
Purine nucleoside phosphorylase II	F
Facilitator Superfamily Protein	G
Fructose 1 6-Bisphosphatase II	G
Glucose-specific phosphotransferase enzyme IIA component	G
Glyceraldehyde-3-phosphate dehydrogenase 2	G
L-arabinose transport ATP-binding protein araG	G
Maltose/maltodextrin-binding protein precursor	G
N-acetylglucosamine-6-phosphate deacetylase	G
Phosphoglycerate kinase	G
Probable phosphomannomutase	G
PTS system fructose-specific EIIBBC component	G
PTS system glucose-specific EIICBA component	G
Ribulose-phosphate 3-epimerase	G
Trehalose-6-phosphate hydrolase	G
Triosephosphate isomerase	G
Xylose transport system permease protein xylH	G
1-deoxy-D-xylulose-5-phosphate synthase	H
2-amino-3-ketobutyrate coenzyme A ligase	H
5-Formyltetrahydrofolate Cyclo-Ligase	H
Glutamate-1-semialdehyde 2,1-aminomutase	H
GTP cyclohydrolase I	H
L-aspartate oxidase	H
Oxygen-independent coproporphyrinogen III oxidase 1	H
Probable 2-dehydropantoate 2-reductase	H
Probable lipoate-protein ligase A	H
Probable nicotinate-nucleotide pyrophosphorylase [carboxylating]	H
Quinolinate synthetase A	H
Riboflavin biosynthesis protein ribD [Includes: Diaminohydroxyphosphoribosylaminopyrimidine deaminase	H
1-deoxy-D-xylulose 5-phosphate reductoisomerase 2	I
2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	I
2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	I
3-oxoacyl-[acyl-carrier-protein] reductase	I
4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase	I
Acetyl-CoA acetyltransferase	I
Acetyl-CoA acetyltransferase	I
Acetyl-coenzyme A synthetase	I
Acyl-CoA dehydrogenase	I
Acyl-CoA dehydrogenase	I
Biotin carboxylase	I
Cardiolipin Synthetase	I
Cardiolipin synthetase	I
Long-chain-fatty-acid--CoA ligase	I
Short chain 3-hydroxyacyl-CoA dehydrogenase, mitochondrial precursor	I
50S ribosomal protein L1	J
50S ribosomal protein L16	J
50S ribosomal protein L3	J
50S ribosomal protein L32	J
50S ribosomal protein L4	J

50S ribosomal protein L7/L12	J
CCA-adding enzyme	J
D-tyrosyl-tRNA(Tyr) deacylase	J
Peptide deformylase 2	J
Prolyl-tRNA synthetase	J
Protein hemK homolog	J
rRNA Methylase	J
Cold shock protein cspB	K
Cold shock protein cspD	K
Glucokinase	K
Ribonuclease R	K
RNA polymerase sigma-28 factor precursor	K
RNA polymerase sigma-F factor	K
Stage 0 sporulation protein J	K
Thiaminase-2	K
Transcriptional repressor mprA	K
DNA ligase	L
DNA polymerase beta	L
DNA polymerase III gamma/tau subunit	L
DNA polymerase IV 1	L
DNA repair protein recN	L
Exodeoxyribonuclease V alpha chain	L
Recombination protein recR	L
Replication initiation and membrane attachment protein	L
Spore photoproduct lyase	L
Penicillin-binding protein 1F	M
Penicillin-binding protein 5* precursor	M
Probable undecaprenyl-phosphate N-acetylglucosaminyl 1-phosphate transferase	M
Putative UDP-glucose 4-epimerase	M
UDP-glucose 6-dehydrogenase	M
UDP-N-acetylglucosamine 1-carboxyvinyltransferase 1	M
UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase	M
ComG operon protein 2 homolog	N
Flagellar biosynthesis protein flhA	N
Flagellar hook-associated protein 2	N
Methyl-accepting chemotaxis protein mcpC	N
Type 4 prepilin-like proteins leader peptide processing enzyme	N
ATP-dependent protease La homolog	O
Cell division protein ftsH homolog	O
Chaperone protein dnaK	O
FeS Assembly Protein SufB	O
Ribose-phosphate pyrophosphokinase	O
Serine Protease	O
Thioredoxin reductase	O
Thioredoxin reductase	O
Achromobactin transport system permease protein cbrB	P
Arsenical pump membrane protein	P
Catalase X	P
D-methionine transport ATP-binding protein metN	P
D-methionine-binding lipoprotein metQ precursor	P
Ferrichrome transport system permease protein fhuG	P
Probable D-methionine transport system permease protein metI	P
Probable superoxide dismutase [Fe]	P

Zinc-binding protein adcA precursor	P
ABC Transporter	R
Aluminum Resistance Protein	R
Biotin Synthesis BioY Protein	R
Hypothetical protein HI0933	R
Hypothetical UPF0001 protein ylmE	R
Probable serine/threonine-protein kinase yloP	R
Probable tRNA modification GTPase trmE	R
Protein jag	R
Protein pcrB homolog	R
Sodium-Dependent Transporter	R
Xanthine/Uracil Permease Family Protein	R
Protein ctaG	S
Ribonuclease BN	S
Alkaline phosphatase synthesis transcriptional regulatory protein phoP	T
Arsenate reductase	T
GTP pyrophosphokinase	T
Probable C4-dicarboxylate sensor kinase	T
Protein prkA	T
Putative protein phosphatase	T
Sensor protein ydfH	T
Sensor protein yycG	T
Ser/Thr Protein Phosphatase Family Protein	T
Serine-protein kinase rsbW	T
Sec-independent protein translocase protein tatCy	U
Signal peptidase I	U
ABC Transporter	V
ABC Transporter	V
ABC Transporter Permease Protein	V
ABC-type transporter ATP-binding protein ecsA	V
Bacitracin transport ATP-binding protein bcrA	V
Inner membrane transport permease ybhS	V
Probable multidrug resistance protein norM	V
Putative ABC transporter ATP-binding protein exp8	V
ATPGuanido Phosphotransferase	Y
GTP-sensing transcriptional pleiotropic repressor codY	Y
Protein ecsB	Y
Pyruvate carboxylase	Y

Tabla 11. Genes que componen al genoma núcleo al corte 13 con una N total de 834 genes, (135 genes nuevos).

Anotation	COG
Cytochrome c oxidase polypeptide IVB	C
Cytochrome d ubiquinol oxidase subunit I	C
Cytochrome D Ubiquinol Oxidase Subunit II	C
Ferredoxin	C
Probable aldehyde dehydrogenase ywdH	C
Probable butyrate kinase	C
Probable manganese-dependent inorganic pyrophosphatase	C
Probable NADH-dependent butanol dehydrogenase 1	C
Protein FtsW	D
Stage V sporulation protein E	D
Arginase	E
Arginine decarboxylase	E
Aspartate aminotransferase	E
Branched-chain amino acid transport system carrier protein braB	E
Cysteine synthase	E
Dipeptide-binding protein dppE precursor	E
Gamma-glutamyl phosphate reductase 2	E
Glycine betaine transport ATP-binding protein opuAA	E
Glycine betaine-binding protein precursor	E
Histidine biosynthesis bifunctional protein hisIE [Includes: Phosphoribosyl-AMP cyclohydrolase	E
Oligoendopeptidase F, chromosomal	E
Oligopeptide transport ATP-binding protein appF	E
Oligopeptide transport ATP-binding protein oppD	E
Oligopeptide transport system permease protein oppB	E
Oligopeptide transport system permease protein oppC	E
Periplasmic dipeptide transport protein precursor	E
Probable cysteine desulfurase	E
Probable cysteine synthase	E
Probable succinyl-diaminopimelate desuccinylase	E
Sodium/Alanine Symporter Family Protein	E
Spermidine/putrescine transport system permease protein potB	E
Spermidine/putrescine-binding periplasmic protein 2 precursor	E
5'-nucleotidase precursor	F
Adenine phosphoribosyltransferase	F
Deoxyribose-phosphate aldolase	F
GMP reductase	F
Nucleoside Permease	F
2,3-diketo-5-methylthiopentyl-1-phosphate enolase	G
Glucose-1-phosphate adenylyltransferase	G
Glucose-6-phosphate 1-dehydrogenase	G
Glycogen synthase	G
Multidrug resistance protein 2	G
Probable inorganic polyphosphate/ATP-NAD kinase 2	G
PTS system N-acetylglucosamine-specific EIICBA component	G
Pullulanase precursor	G
Ribose-5-Phosphate Isomerase B	G
Dipicolinate synthase, B chain	H

Menaquinone biosynthesis protein menD [Includes: 2-succinyl-6-hydroxy- 2,4-cyclohexadiene-1-carboxylate synthase	H
Menaquinone-specific isochorismate synthase	H
Naphthoate synthase	H
Nicotinate Phosphoribosyltransferase	H
Nicotinate Phosphoribosyltransferase	H
Probable aminotransferase yhxA	H
3-oxoacyl-[acyl-carrier-protein] reductase	I
4-hydroxy-3-methylbut-2-enyl diphosphate reductase	I
Acyl-CoA dehydrogenase	I
Acyl-CoA dehydrogenase	I
Long-chain-fatty-acid--CoA ligase	I
Omega-6 fatty acid desaturase, chloroplast precursor	I
Probable succinyl-CoA:3-ketoacid-coenzyme A transferase subunit A	I
Putative acyl carrier protein phosphodiesterase 2	I
Succinyl-CoA:3-ketoacid-coenzyme A transferase subunit B	I
Peptide chain release factor 2	J
Ribosomal Large Subunit Pseudouridine Synthase D	J
Sua5/YciO/YrdC/YwlC Family Protein	J
tRNA pseudouridine synthase A	J
tRNA pseudouridine synthase B	J
tRNA uridine 5-carboxymethylaminomethyl modification enzyme gid	J
HTH-type transcriptional regulator glnR	K
Probable ATP-dependent helicase dinG homolog	K
RNA polymerase sigma-B factor	K
Sigma Factor SgiI	K
Stage III sporulation protein D	K
Transcription termination factor rho	K
Transcriptional Regulator	K
Transcriptional Regulator GntR Family	K
ATP-dependent DNA helicase recQ	L
DNA topoisomerase III	L
Exonuclease sbcD homolog	L
Primase-Related Protein	L
Probable exodeoxyribonuclease VII small subunit	L
Protein smf	L
Putative 3-methyladenine DNA glycosylase	L
Ribonuclease HIII	L
Diacylglycerol kinase	M
Glycine betaine/carnitine/choline-binding protein precursor	M
Mechanosensitive Ion Channel Family Protein	M
Membrane-Associated Zinc Metalloprotease	M
Putative septation protein spoVG	M
Flagellar biosynthesis protein flhF	N
Flagellar hook-associated protein 3	N
Putative protease yegQ	O
Achromobactin-binding periplasmic protein precursor	P
Arsenical pump membrane protein	P
Arsenical pump membrane protein	P
Chromate Transporter	P
Cobalt import ATP-binding protein cbiO 1	P
General stress protein 20U	P
Lipoprotein	P

Potassium Uptake Protein	P
Probable 3-mercaptopyruvate sulfurtransferase	P
Putative aliphatic sulfonates transport permease protein ssuC	P
Taurine import ATP-binding protein tauB	P
Vitamin B12 binding protein precursor	P
Exopolyphosphatase-Related Protein	R
Hypothetical protein yqfN	R
Lysine Decarboxylase	R
Phenylalanyl-tRNA synthetase beta chain	R
Putative esterase ytxM	R
Spore maturation protein A	R
ABC Transporter Permease Protein	S
Flotillin-Like Protein	S
Spore maturation protein B	S
Stage V sporulation protein R	S
Stage V sporulation protein S	S
Anti-sigma-B factor antagonist	T
PhoH-like protein	T
Sporulation initiation phosphotransferase F	T
S-ribosylhomocysteine lyase	T
Transcriptional regulatory protein yycF	T
Signal peptidase I W	U
ABC Transporter	V
ABC Transporter	V
Lipid A export ATP-binding/permease protein msbA	V
Multidrug resistance-like ATP-binding protein mdlA	V
Undecaprenyl-diphosphatase 2	V
Dipicolinate synthase, A chain	Y
Small, acid-soluble spore protein I	Y
Spore coat protein E	Y
Sporulation initiation phosphotransferase B	Y
Sporulation protein ypeB	Y
Stage IV sporulation protein A	Y
Stage V sporulation protein AC	Y
Stage V sporulation protein AD	Y
Stage V sporulation protein AF	Y