UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE QUÍMICA

ANÁLISIS DEL GENOMA NÚCLEO Y PANGENOMA DEL GÉNERO *STREPTOCOCCUS*

TESIS

QUE PARA OBTENER EL TÍTULO DE

QUÍMICO FARMACÉUTICO BIÓLOGO

PRESENTA

HUGO RAFAEL BARAJAS DE LA TORRE

MÉXICO, D.F. **AÑO 2014**

JURADO ASIGNADO:

PRESIDENTE: Profesor: Samuel Canizales Quinteros
VOCAL: Profesor: Mónica Berenice Heras Chavarría
SECRETARIO: Profesor: Luis David Alcaraz Peraza
1er. SUPLENTE: Profesor: Beatriz Ruíz Villafan
2° SUPLENTE: Profesor: Javier Fernández Torres

SITIO DONDE SE DESARROLLÓ EL TEMA: Instituto de Ecología, UNAM.

ASESOR DEL TEMA: Luis David Alcaraz Peraza

SUSTENTANTE: Hugo Rafael Barajas de la Torre

Índice de contenido

Índice de tablas

Índice de figuras

Resumen

Gracias al avance en las tecnologías de secuenciación, hoy (2014) contamos con 7,411 genomas de especies y cepas bacterianas completamente secuenciadas con esta información es posible avanzar en cuestiones taxonómicas, como las definiciones de especies bacterianas; problemas de salud, como el diseño dirigido de vacunas; o conocer las funciones que llevan a cabo las comunidades bacterianas en ambientes complejos. Este tipo de problemáticas se pueden abordar mediante el análisis comparativo de pangenomas. Los pangenomas pueden definirse a distintos niveles taxonómicos y nos provee de la información necesaria para conocer los elementos genéticos y las funciones que comparten los grupos bacterianos que se deseen analizar.

En este trabajo hicimos genómica comparada de los 108 proteomas disponibles del género Streptococcus, el cuál comprende un gran número de especies patógenas importantes de humanos y animales, así como especies comensales de las membranas mucosas de la cavidad oral, tracto respiratorio e intestinales. La búsqueda de ortólogos se hizo mediante 12,100 alineamientos con BLAST de forma bidireccional. La caracterización, anotación y el análisis del el genoma núcleo y pangenoma de los Streptococci, hecha en este trabajo, nos da un impactante total de 33,039 familias de proteínas que componen el pangenoma y tan solo 405 proteínas se encuentran universalmente conservadas en el género. Con el análisis funcional del pangenoma, hemos descrito las funciones conservadas (división celular; traducción y estructuras ribosomales; membrana y pared celular; y respuesta a estrés), mientras que, en el genoma accesorio existen funciones como los mecanismos de defensa que muestran variabilidad en cada especie analizada. Además con el genoma núcleo realizamos un agrupamiento de similitud genómica entre los miembros del género y contrastamos su resolución contra el gen 16S rRNA.

Introducción y antecedentes

El género *Streptococcus*

El género *Streptococcus* es un género bacteriano perteneciente al phylum de los *Firmicutes,* de la clase *Bacilli*, orden *Lactobacillales* y familia *Streptococcaceae*. Los streptococci son bacterias Gram positivas, no móviles, no esporulantes, anaerobias facultativas, catalasa negativas, y que se desarrollan generalmente en cadenas o en pares de cocos de 1µm de diámetro. Son clasificadas, típicamente, de acuerdo a sus capacidades de hemólisis (α, β o γ) y mediante los antígenos que presentan en sus paredes celulares (clasificcación de Lancefield). (Kayser, *et al.*, 2005, pp.234), aunque la clasificación actual de los streptococci se basa fuertemente en los datos de las comparaciones de las secuencias de rRNA 16S (Kawamura *et al.* 1995; Facklam, 2002). Este género incluye importantes patógenos y comensales de las membranas mucosas del tracto respiratorio superior y en algunas especies en las mucosas intestinales. El género consta de más de 40 especies y, con algunas excepciones, las especies individuales se asocian exclusivamente como patógenos o comensales del humano o algún animal en particular. El género se divide en seis grupos de especies, cada una de estas caracterizada por su potencial patogénico y otras propiedades fenotípicas resumidas en la tabla 1 (Kilian, 2007). De igual manera estos grupos han sido descritos mediante filogenias generadas con las secuencias de rRNA 16S de 34 especies de *Streptococcus* por Kawamura *et al.* (1995).

Muchas especies de los streptococci son parte de la microbiota comensal presente en las superficies mucosas de humanos y animales, las cuales generalmente no causan ningún daño pero prácticamente todas las especies comensales de *Streptococcus* son patógenos oportunistas, principalmente si logran tener acceso al flujo sanguíneo desde la cavidad oral o los intestinos (Kilian, 2007). Sin embargo, a pesar de ser portadas de manera asintomática, estas especies pueden causar una gran variedad de infecciones, variando desde las caries dentales y faringitis, hasta enfermedades que ponen en peligro la vida, como la fascitis necrozante o la meningintis (Mitchell, 2003)

Diversas clasificaciones de las especies del género.

Clasificación filogenética

El grupo *pyogenes* incluye a la mayor parte de las especies que son patógenas de humanos y animales, capaces de causar septicemia o infecciones en el tracto respiratorio(Kilian, 2007).

Los representantes del grupo *mitis*, incluye a comensales de la cavidad oral y nasofaringe humanas, aunque una de las especies, *Streptococcus pneumoniae*, es un importante patógeno humano (Kilian, 2007). Muchas de las especies de este grupo se consideran difíciles de clasificar e identificar por métodos bioquímicos ya que tienen pocas características que ayuden a su discriminación. Este es un grupo muy relacionado genéticamente y comparte una alta similitud en sus secuencias de rRNA 16S (Tagg *et al.,* 2012, pp. 136)

Especies de *Streptococcus* con importancia clínica			
Grupo filogenético	Especies	Grupo de Lancefield	Tipo de Hemólisis
Grupo pyogenes	*S. pyogenes*	A	β
	S. agalactiae	B	β
	S. equisimilis	C	β
Grupo mitis	*S. pneumoniae*	O	α
	S. mitis	O	α
	S. oralis	No identificado	α
	S. sanguinis	O	α
	S. gordonii	O	α
Grupo anginosus	*S. anginosus*	G,F y A	α
	S. intermedius		α
Grupo salivarius	*S. salivarius*	K	γ
	S. thermophilus		γ
	S. vestibularis		γ
Grupo bovis	*S. bovis*	D	α
	S. alactolyticus		α
Grupo mutans	*S. mutans*	No designado	γ
	S. sobrinus	No designado	γ

Tabla 1. Grupos de Streptococcus mediante su clasificación filogenética. Se resumen características útiles para la identificación de especies principales dentro de los grupos. (Modificada de Kilian , 2007).

Dentro del grupo *anginosus*, se encuentran bacterias que son parte de la microbiota comensal de la cavidad oral, tracto gastrointestinal y tracto genital femenino. Los integrantes de este grupo forman colonias pequeñas y en su mayoría son no hemolíticos. Su clasificación es complicada, ya que varían en el antígeno de Lancefield que presentan (Tagg *et al*, 2011, pp 135-136).

El grupo *salivarius*, cuenta con los colonizadores comensales de la mucosa humana oral.

De éstos, se ha encontrado que *S. salivarius* puede utilizarse como un probiótico debido a que esta especie es una buena productora de bacteriocinas(Kilian, 2007). Este grupo está muy relacionado con el grupo bovis, teniendo como ejemplos a *S. infantarius* y *S. alactolyticus* que eran parte del grupo bovis y ahora se encuentran dentro de este (Tagg *et al*, pp 137).

Las bacterias del grupo *bovis*, se pueden encontrar en el colon (Kilian, 2007). Sus miembros han sufrido de muchos cambios taxonómicos en los últimos años , se ha dividido a diversos biotipos de *S. bovis* en las especies *S. gallolyticus* subsp. g*allolyticus*, *S. gallolyticus* subsp. *pasteurianus* y *S. macedonicus (Poyart, Quesne, & Trieu-cuot, 2002).*

El grupo denominado *mutans*, cuenta con especies relativamente acidúricas y acidogénicas capaces de colonizar la superficie dental humana y de ciertos animales, debido a esto algunas de las especies de este grupo están involucradas en el desarrollo de las caries dentales (Tagg *et al*, pp 137).

Existe un grupo pobremente definido, en donde se encuentra a *S. suis* que es un importante patógeno de cerdos a nivel mundial y que ha emergido como causa de enfermedades zoonoticas por el contacto de los humanos con estos animales (Tagg *et al*, pp 138) y del cual se ha observado que no se relaciona filogenéticamente con ninguno de los demás grupos (Kawamura *et al*., 1995). Dentro de este grupo se encuentran especies relacionadas a *S. suis* y que causan de igual manera enfermedades a animales, como son *S. porci* (Vela *et al*., 2010) , *S. gallinaceus* y *S. ovis*, las cuales son especies patógenas de animales (Tagg *et al*, pp 138).

Clasificación por tipo de hemólisis

α-hemólisis: Las colonias, al crecer en agar sangre, se ven rodeadas de una zona verde. Esto es causado por la producción de H_2O_2, el cual reacciona con la hemoglobina, produciendo metahemoglobina, que es de color verde (Kayser *et al.*, 2005, pp. 235). Dentro de esta categoría se encuentran los integrantes del colectivo de streptococci denominado

como grupo “viridans” (del latín *virdis* = verde) , que refiere su propiedad α-hemolítica y entre los cuales se encuentran los integrantes del grupo mitis, anginosus y bovis (Kilian, 2007).

β-hemólisis: Las colonias que crecen en agar sangre se rodean de una zona hemolítica de color amarillo claro, causada por la lisis de los eritrocitos y la hemoglobina es descompuesta mediante la hemolisina bacteriana producida (Kayser *et al.*, 2005, pp. 235). Esta característica constituye uno de los principales marcadores de los streptococci potencialmente patógenos en cultivos de muestras clínicas (Kilian, 2007).

γ-hemólisis: Es un término utilizado para indicar que a nivel macroscópico no hay una zona evidente de hemólisis (Kayser *et al.*, 2005, pp. 235).

Clasificación de Lancefield

La clasificación de Lancefield se centra en la identificación de polisacáridos presentes en la pared celular (substancia C), llamados antígenos de Lancefield , los cuales se distinguen utilizando anticuerpos específicos y precipitando los polisacáridos (Lancefield, 1932). Mediante estos antígenos y su variación se clasifican los grupos de Lancefield del A al V. Este método es importante debido a que se ha utilizado para hacer la distintinción entre las diferentes especies del grupo pyogenes, originalmente desarrollado por Rebecca Lancefield (1932). Esta técnica serológica se ha aplicado también para la identificación de los streptococci del grupo “viridans”; así como a los enterococci,

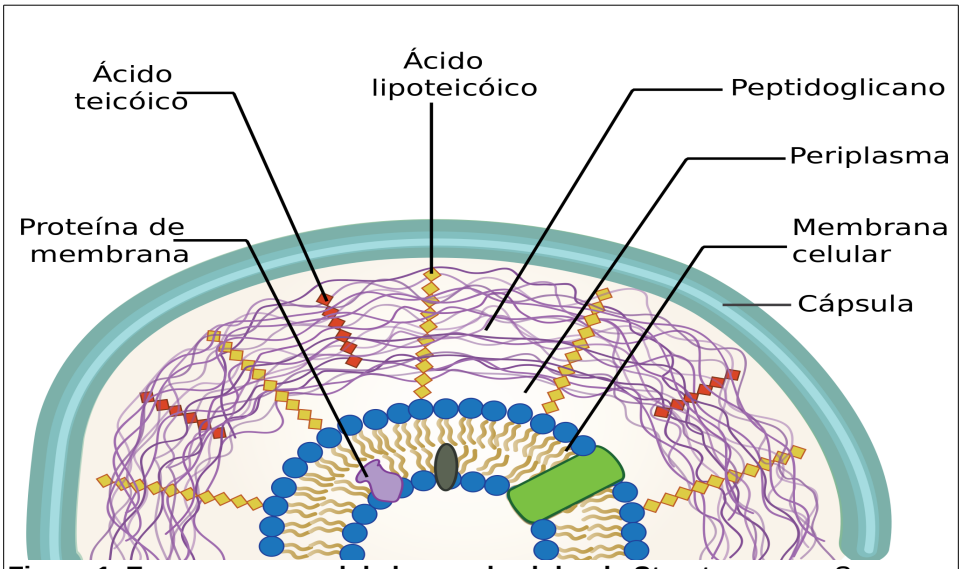

Figura 1. Esquema general de la pared celular de Streptococcus. Se muestran los componentes principales de la pared celular (Modificada de Kilian, 2007.

extendiendo el número total de serogrupos hasta 21, del A al H y del K al W (Kilian, 2007).

La estructura de la pared celular de los streptococci (ver Figura 1) se compone por una capa gruesa de peptidoglicano, ácidos teicóicos unidos covalentemente al peptidoglicano, cadenas de ácidos lipoteicóicos que cruzan por el petidoglicano y se anclan a la membrana celular y una cápsula de polisacáridos presentes en la mayor parte de las especies, la cual es conocida como subtancia C (antígeno de Lancefield) que se une covalentemente a la mureína (Kayser *et al.*, 2005, pp. 235-236).

Características nutricionales y de cultivo.

En general, los streptococci son fastidiosos en cuanto a sus requerimientos nutricionales. Su cultivo *in vitro* requiere del uso del medio de cultivo agar sangre enriquecido con peptonas, carbohidratos y vitaminas. Análisis genómicos de *S. pyogenes*, *S. agalactiae* y *S. pneumoniae* han mostrado que no presentan la capacidad de llevar a cabo el ciclo de los ácidos tricarboxílicos (Glaser et al., 2002), lo que siginifíca que son incapaces de sintetizar los precursores de la mayoría de los aminoácidos. En la naturaleza, el requerimiento de nitrógeno, se ve satisfecho por los aminoácidos excretados por otros microorganismos dentro del microbioma o por los liberados debido a la degradación de proteínas en tejidos mediante la acción de proteinasas. La temperatura óptima de la mayoría de los streptococci es de 37°C, aunque algunas especies como *S. uberis* y *S. thermophilus* pueden crecer a temperaturas bajas de 10°C o altas de hasta 45°C, respectivamente (Tagg *et al*, pp 126).

Los streptococci son anaerobios facultativos y la mayor parte de ellos pueden crecer en ausencia de oxígeno. *S. pneumoniae* y otras especies requieren concentraciones elevadas de dióxido de carbono para un crecimiento óptimo y algunas especies se desarrollan de manera óptima en condiciones de anaerobiosis. Sus requerimientos energéticos se ven satisfechos mediante la fermentación de carbohidratos. Entre las bacterias capaces de crecer aeróbicamente, los streptococci son los únicos incapaces de obtener ATP mediante el sistema de transporte de electrones y no cuentan con la habilidad de sintetizar porfirinas, citocromos o catalasa (Tagg *et al*, pp 127).

La problemática de las especies bacterianas

La más reciente definición de una especie bacteriana proviene de las era pregenómica. En 1987 (Wayne *et al.*, 1987) se propuso que las cepas bacterianas que presentaran una reasociación DNA-DNA mayor al 70 % y compartieran un cierto número de características fenotípicas podían ser consideradas una especie. Junto a esto se utiliza la comparación del gen de rRNA 16S, con la cual se considera a un microorganismo como parte de una especie cuando comparte al menos el 97% de identidad con dicha secuencia (Goris et al., 2007; Stackebrandt & Goebel, 1994).

La definición de especie utilizando esta secuencia es problemática, ya que es una molécula que evoluciona lentamente y no tiene la suficiente resolución para distinguir entre especies similares (Fox et al., 1992). De igual manera, ha recibido críticas debido a que utiliza un valor de corte arbitrario; sin embargo estas comparaciones son aceptadas universalmente para la determinación de largas escalas evolutivas. Otra complicación que tiene el uso de este estándar para la clasificación de especies bacterianas, es que el marco conceptual de la definición de especies se basa en la herencia del material genético en dirección vertical (Alcaraz, 2013); luego entonces se presenta un conflicto, ya que las bacterias son capaces de llevar a cabo la transferencia horizontal de material genético (HGT) y la recombinación de genes entre especies similares o linajes clonales (Smith et al., 1993). Debido a estas limitaciones y sesgos se han tomado distintos enfoques para intentar caracterizar a las especies bacterianas. En la actualidad contamos con una estrategia que no requiere la secuenciación de genomas, sino la comparación de alineamientos múltiples de genes, conocida como MLST (del inglés *Multi Locus Sequence Typing*). Esta estrategia se basa en el uso de 7 genes que se encuentren dispersos dentro de los genomas para evitar riesgos de ligamiento genético, los cuales son amplificados, secuenciados, alineados y concatenados en una sola secuencia artificial para maximizar la cantidad de información genética que se analiza con el modelo de substitución y finalmente proponer una hipótesis filogenética que ayude a discriminar entre especies relacionadas (Alcaraz, 2013).

Aparte de las anteriormente mencionadas, existen estrategias para definir distancias

Genómicas como el *Genomic Similarity Score* (GSS) (Alcaraz et al., 2010; Moreno-Hagelsieb & Janga, 2008). Esta métrica está basada en la suma de puntuación en bits (*bit-scores*) resultante de la comparación de las proteínas ortólogas de un organismo con las de otro organismo, las cuales son identificadas como los mejores aciertos recíprocos (RBH) en las comparaciones pareadas de genomas y normalizada contra la suma de *bit-scores* de los genes comparados contra sí mismos (Figura 2) (Moreno-Hagelsieb & Janga, 2008). Por lo tanto esta medida tiene un rango que va de 0 a 1; cuando dos proteomas son idénticos se tendrá el valor máximo de 1 y de esta manera sirve como una medida de similitud genómica. El GSS puede ser utilizado como una herramienta complementaria para aclarar las relaciones entre organismos usando genes ortólogos compartidos por pares (Alcaraz et al., 2010).

$$GSS = \frac{BBS(comp)}{BBS(self)}$$

Figura 2. Fórmula para el cálculo del GSS. GSS, *Genomic Similarity Score;* BBS (comp), *bit-score* de blast del alineamiento contra el genoma problema; BBS (self), *bit-score* de blast del alineamiento contra sí mismo.

Para poder hacer uso de esta herramienta se requiere primero de la identificación de los genes ortólogos compartidos por pares de genomas, pero ¿Qué son los genes ortólogos? Se dice que dos o más estructuras, procesos y en este caso genes son homólogos entre sí, si y

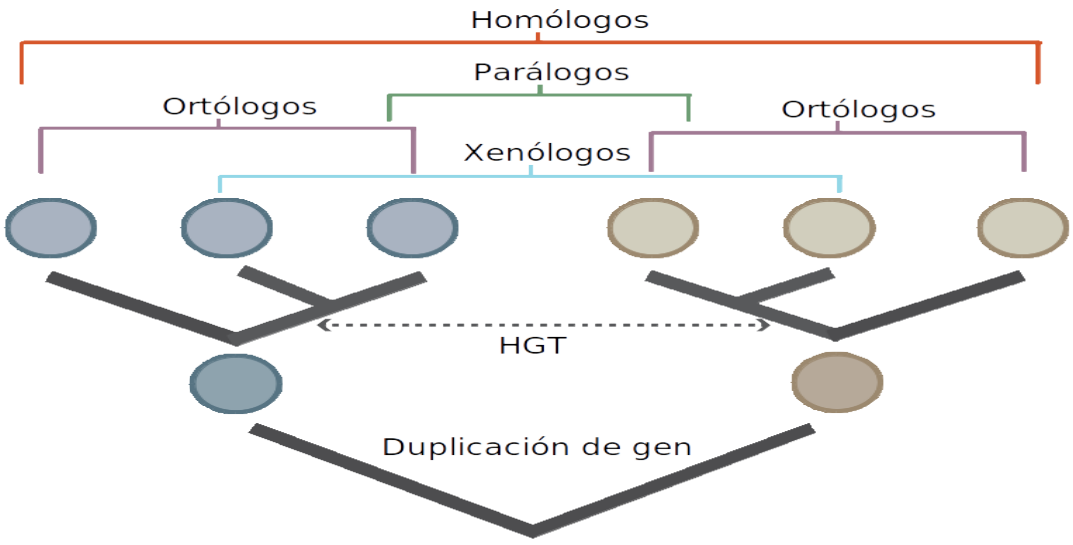

Figura 3. Relaciones de homología. Esquema simplificado de los subtipos de homología: ortología paralogía y xenología. Un caracter proveniente de un ancestro común que diverge en distintos linajes se considera un ortólogo. Un caracter presente en diferentes linajes pero que ha sido resultado de una duplicación génica en un ancestro común es considerado un parálogo. Un caracter en dos linajes distintos que ha sido adquirido vía HGT se considera un xenologo.

solo si adquirieron su estado actual por herencia directa de un ancestro común (Castilla, 2007). La homología entre dos caracteres puede estar dada por dos circunstancias; una es que dos genes de dos diferentes especies deriven de un mismo ancestro común y surjan tras un evento de especiación, esto es conocido como ortología; y la segunda es que dos genes deriven de un mismo gen que fue duplicado dentro de un genoma , esto es conocido como paralogía. Aparte de estos dos, existe un caso más de homología que es común en bacterias, el cual se denomina xenología y es debido a la transferencia horizontal de material genético (Fitch, 2000) (ver Figura 3).

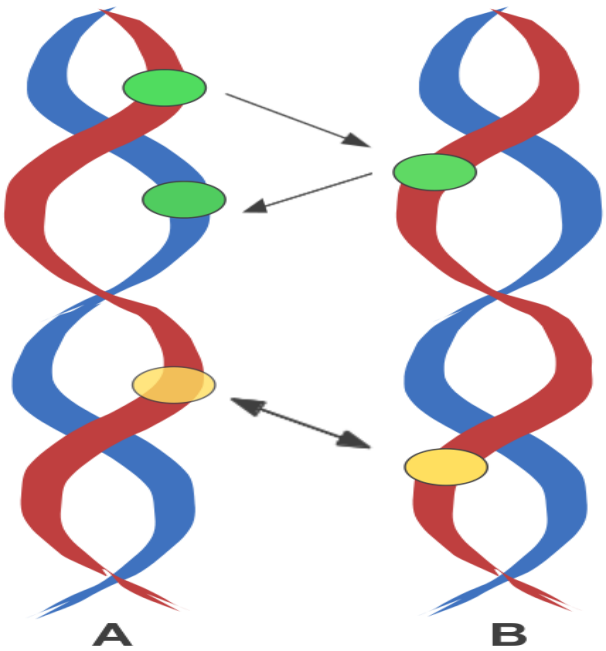

Figura 4. Ejemplificación gráfica de la identificación de ortólogos por RBH. Los circulos amarillos representan el mejor acierto recíproco de un gen entre los genomas A y B. Los circulos verdes representan un acierto que no es recíproco debido a que se encuentra a un gen del genoma A como mejor acierto en el del genoma B, per este no lo encuentra de regreso, sino a un tercero de regreso en el genoma A. Esta clase de aciertos representan una gran similitud entre dos genes y se consideran como evidencia de ortología.

Luego entonces, se tiene el problema de la identificación de los genes ortólogos. La tarea de encontrar al homólogo de una secuencia de interés dentro de una base de datos conteniendo muchas otras secuencias puede conceptualizarse como obtener el mejor alineamiento posible de la secuencia problema contra todos los blancos, dar una puntuación

a esos alineamientos y elegir a los que rebasen un umbral de corte dado o que cumplan con un cierto valor estadístico (como un valor mínimo del *E-value*). Como se mencionó anteriormente, los ortólogos pueden ser detectados utilizando el enfoque del RBH (Figura 4), donde se asumen como ortólogos a dos genes en diferentes genomas que se encuentran a sí mismos como el mejor acierto posible en el genoma opuesto, siendo esta la definición operativa más común de ortología (Moreno-Hagelsieb & Latimer, 2008).

Estas búsquedas se basan en los valores de *bit-score* resultantes de los alineamientos hechos con BLAST (*Basic Local Alignment Tool) (Altschul et al., 1990)* que es un algortimo heurístico utilizado para encontrar regiones locales de similitud entre dos secuencias de aminoácidos o DNA y ha sido diseñado para comparar una secuencia problema contra una base de datos con blancos para el alineamiento. Este programa requiere de tres partes importantes: datos de entrada (secuencia problema), una base de datos contra la cual realizar las búsquedas y un programa particular de BLAST. Desde el lanzamiento de la versión original de BLAST en 1990, una familia de programas especializados han sido desarrollados y sus diferencias radican en el tipo de datos de entrada y el tipo de base de datos que se les proporciona (Zhang, 2011).

Programa	Descripción
BLASTN	Busqueda de una secuencia de nucleótidos contra una base de datos de nucleótidos
BLASTP	Busqueda de una secuencia de aminoácidos contra una base de datos de proteínas
BLASTX	Busqueda de una secuencia de nucleótidos traducida en todos los marcos de lectura contra una base de datos de proteínas
TBLASTN	Busqueda de una secuencia de proteínas contra una base de datos de nucleótidos traducidos de forma dinámica en todos los marcos de lectura
TBLASTX	Busqueda de la traducción de los seis marcos de lectura de una secuencia de nucleótidos contra la traducción de los seis marcos de lectura de una base de datos de nucleótidos

Tabla 2. Principales programas de BLAST. Se muestran los programas más comunmente utilizados de BLAST y la descripción del tipo de datos de entrada y base de datos que utilizan.

El alineamiento es una forma de acomodar dos secuencias de DNA o aminoácidos para identificar regiones similares que se encuentren conservadas entre las especies. Cada secuencia alineada se muestra como una fila dentro de una matriz y la inserción de *gaps* entre residuos de cada secuencia se lleva a cabo, de manera que bases idénticas o similares

en cada secuencia resulten alineadas en posiciones sucesivas. Cada *gap* extiende una o más columnas dentro de la matriz del alineamiento. La puntuación de un alineamiento es calculada sumando las puntuaciones de las columnas que contienen las mismas bases y las puntuaciones de penalización por los *gaps* y las desigualdades en cada columna conteniendo diferentes bases. Un esquema de puntuaciones especifica los puntos de los aciertos y desigualdades, que forman la matriz de puntuaciones, y los puntos de los *gaps* llamados costos de *gap*. Existen dos tipos de alineamientos para las comparaciones de secuencias; los alineamientos locales y los alineamientos globales. Dado un esquema de puntuación, el cálculo del alineamiento global es una optimización global que forza al alineamiento a extender la búsqueda a lo largo de ambas secuencias, mientras los alineamientos locales solo identifican a regiones de alta similitud entre las secuencias. Para acelerar el proceso de búsqueda de homología, BLAST realiza un filtrado en el cual primero hace una búsqueda de en la base de datos aciertos de palabras con una longitud *w* y una puntuación de alineamiento de al menos *T* entre el sujeto de búsqueda y las secuencias blanco y luego extiende cada acierto hacia ambos extremos para generar el alineamiento local en las secuencias cuyos puntajes de alineamiento sean mayores a un umbral determinado. Estos aciertos son conocidos como HSP; del ingles *High Scoring Segment Pairs* (Zhang, 2011). En resumen, el proceso se divide en dos pasos: la búsqueda de la base de datos, en la cual se eligen blancos de manera rápida que puedan producir un alineamiento significativo y la fase de alineamiento, donde los blancos más prometedores se alinean y se les da una puntuación en bits. (Moreno-Hagelsieb & Janga, 2008).

La era genómica

Después de que el primer organismo de vida libre, *Haemophilus inflenzae*, fue secuenciado en 1995, la secuenciación de genomas completos se convirtió en un método estándar y rápido para el estudio de procesos biológicos en los organismos (Medini et al., 2005). Al principio de la era genómica, se pensaba que la secuenciación de un aislado representativo era suficiente para realizar la descripción de la complejidad genética de una especie (Tettelin et al., 2002), pero recientemente se han analizado aislados de las mismas especies mediante el uso de la genómica comparativa y se ha encontrado que la variación

intraespecie puede ser tan significativa como la variación interespecie (Muzzi et al., 2007; Tettelin et al., 2008). Por esta razón, el uso de datos genómicos de diferentes cepas dentro de una misma especie es necesario para resolver preguntas sobre la fisiología bacteriana, como la identificación de genes esenciales para el metabolismo o para conocer procesos fundamentales de la patogénesis, resistencia a antibióticos, adaptación al ambiente y evolución (Muzzi et al., 2007).

Gracias a los avances tecnológicos, hoy en día contamos con las tecnologías de secuenciación de próxima generación en las plataformas como Roche-454 (www.roche-applied-science.com) o Illumina (www.illumina.com) (Shendure & Ji, 2008). En las últimas décadas, el número de genomas bacterianos que han sido secuenciados ha crecido de forma exponencial de forma en que hoy se encuentran disponibles 7,411 genomas bacterianos completamente secuenciados, así como miles de proyectos de secuenciación por *Whole Genome Shotgun,* la cual es una técnica de secuenciación en donde el DNA genómico se fragmenta por completo y posteriormente es colocado en un vector para transformar a *E. coli* y secuenciar las clonas (Shendure & Ji, 2008), sumando de esta manera más de 33,000 genomas en proceso de secuenciación y disponibles (Pagani *et al*., 2012). La habilidad de secuenciar genomas completos de organismos relacionados ha abierto la posibilidad de realizar estudios a gran escala de genómica comparativa y evolutiva (Metzker, 2009), pero a pesar de la implementación de estrategias refinadas para realizar estas actividades, seguimos teniendo el problema de la gran cantidad de diversidad bacteriana que existe, tomando en cuenta que en algunos ambientes solo hemos sido capaces de cultivar el 1% de los microorganismos que los habitan (Vartoukian *et al*., 2010). Por esta razón, ha surgido el campo de la metagenómica, que es la disciplina que permite el estudio genómico de microorganismos no cultivados provenientes de muestras ambientales directamente (Wooley *et al*., 2010). Este tipo de análisis se pueden realizar prácticamente en cualquier ambiente para poder estudiar tanto la diversidad taxonómica como funcional de éstos (Lundberg *et al*., 2012). Estos ambientes y sus microbiomas son diversos y pasando por el suelo, cuerpos de agua, animales, plantas o la cavidad oral humana (Belda-Ferre *et al*., 2012; Berendsen *et al*., 2012; Lundberg *et al*., 2012). En el caso de la cavidad oral humana podemos encontrar que los streptotocci son relevantes debido a que muchos de sus

miembros son parte de la flora comensal (grupos mitis y anginosus) y *S. mutans* se encuentra asociado al desarrollo de las caries dentales aunque no es la especie dominante dentro de este ambiente (Belda-Ferre *et al*., 2012).

El progreso que se ha tenido en la secuenciación de genomas ha llevado al enriquecimiento y ampliación de las bases de datos genómicas con una gran variedad de proteínas predichas, la mayoría sin un rol funcional documentado. Con el uso de la biología computacional, se realizan esfuerzos para extraer la mayor cantidad de información posible de estas secuencias al clasificarlas de acuerdo a sus relaciones de homología, realizando de esta manera la predicción de las funciones celulares o actividad bioquímica que pueden tener. Este ha sido un reto complicado en su mayor parte porque el organismo mejor caracterizado con el que contamos es E*. coli,* del cual solo se han caracterizado experimentalmente el 40% de sus productos génicos (Koonin E. V., 1997). Por otro lado, contamos con la ventaja de que las proteínas de los procariontes, en general, se encuentran altamente conservadas, lo que hace que podamos inferir las funciones de proteínas de organismos pobremente caracterizados a partir de sus homólogos en otros organismos. Generalmente las proteínas que son ortólogas conservan la misma arquitectura de sus dominios y por ende las mismas funciones, aunque pueden existir excepciones y complicaciones para realizar esta generalización(Tatusov *et al*., 2000).

Existen bases de datos que han sido diseñadas para intentar clasificar a las proteínas de organismos secuenciados basados en el concepto de ortología; una de ellas es el *Cluster of Orthologous Groups* (COG) (http://www.ncbi.nlm.nih.gov/COG/). Esta base de datos representa las relaciones de ortología de uno a muchos y muchos entre muchos genes, y es por esto que una de sus utilidades es la asignación de funciones conocidas a grupos de proteínas . La base de datos original, en el año 2000, contenía a 2091 COGs con proteínas de 21 genomas completos (Tatusov *et al*., 2000). En el 2003, la cantidad de COGs había aumentado hasta el número de 4873 COGs de las proteínas de 66 genomas secuenciados (Tatusov *et al*., 2003) Hoy en día, la base de datos eggNOG contiene información tanto sobre procariontes como eucariontes, sumando un total de 2031 genomas base (Powell *et al.,* 2013) .

La base de datos de SEED (http://www.theseed.org/wiki/Main_Page) contaba inicialmente con 180,177 proteínas con 2133 roles funcionales. Esto datos provienen de 383 organismos distintos y se colocan en 173 subsistemas; entendiéndose por subsistema a un conjunto de roles funcionales que juntos son la implementación de un proceso biológico específico o un complejo estructural. Un subsitema puede pensarse de manera general como una ruta. Así como la glucólisis se compone de roles funcionales (glucocinasa, glucosa-6-fosfato isomerasa, fosfofructocinasa, etc.), un complejo ribosomal o un sistema de transporte pueden verse como colecciones de roles funcionales. (Overbeek *et al*., 2005)

El pangenoma bacteriano

El concepto del pangenoma (pan, de la palabra griega παν, que significa todo) surgió a raíz de análisis de genómica comparativa realizados con cepas patógenas de *S. agalactiae* por Tettelin *et al.* (2005) para conocer cuál era el número de genomas necesario para poder describir por completo a una especie bacteriana. El pangenoma es la suma de dos componentes; el genoma núcleo, que es el conjunto total de genes que comparten entre

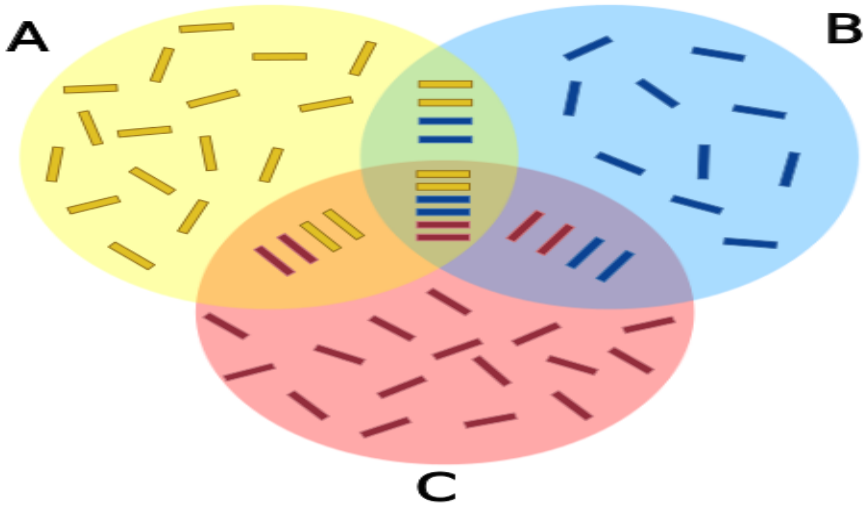

Figura 5. Representación gráfica del pangenoma. Diagrama de Venn representando los diferentes genes ortólogos compartidos entre tres genomas. La intersección de los conjuntos A, B y C representa al genoma núcleo. La suma del genoma núcleo con los genes específicos de cada cepa o cepas componen el pangenoma.

todas las cepas y el genoma accesorio, compuesto por los genes que se encuentran ausentes en una o más cepas y los genes únicos de cada cepa (ver Figura 5) (Tettelin et al., 2005). El genoma accesorio es donde se encuentra la diversidad de las especies y provee de funciones que no son esenciales pero que si confieren ventajas selectivas, entre las cuales se encuentran la adaptación al nicho, resistencia a antibióticos y las habilidades para colonizar a nuevos huéspedes (Tettelin et al., 2008). Las especies pueden tener lo que se denomina un pangenoma abierto o cerrado, dependiendo de su comportamiento al añadir más genomas en la construcción del mismo. Un pagenoma abierto es aquel que al incorporar más genomas continúa incrementando la cantidad de genes que contiene, mientras que el cerrado deja de acumular genes nuevos con cada genoma añadido (Medini et al., 2005) .

El análisis filogenético del genoma núcleo, puede ser considerado como el siguiente nivel del MSLT dentro del ámbito filogenético, ya que compara la información genómica de todos los ortólogos que se comparten entre los grupos que se estén trabajando y pueden ser definidos a distintos niveles taxonómicos para determinar que características genéticas se comparten dentro de las especies al phylum. Una de las ventajas de realizar análisis del genoma núcleo de un grupo de taxa es que puede revelar el grupo de genes que se encuentran conservados dentro de un rango taxonómico y de esta manera encontrar funciones que sean las que determinen la cohesión de un grupo o que sean características fenotípicas compartidas por todos como la resistencia a antibióticos o la resistencia al calor. A nivel evolutivo se pueden hacer hallazgos sobre la conservación de funciones en distintos niveles taxonómicos (Alcaraz, 2013).

Objetivo General

Realizar la caracterización, descripción y análisis del genoma núcleo y pangenoma del género *Streptococccus* y especies selectas mediante perfiles funcionales y relacionar estas funciones con metagenomas orales humanos mediante el análisis de la abundancia taxonómica del género en este ambiente.

Objetivos particulares

1. Construcción de una base de datos local con los proteomas y genomas (CDS) del género *Streptococcus*, que se encuentren completamente secuenciados.

2. Realizar alineamientos locales de los proteomas del género de forma pareada y alineamientos múltiples de sus secuencias de rRNA 16S.

3. Llevar a cabo la reconstrucción filogenética del género mediante las secuencias de rRNA 16S.

4. Generar una matriz de distancias genómicas entre todas las cepas utilizadas.

5. Estimar la abundancia del género en metagenomas orales humanos y determinar su perfil funcional dentro de este ambiente.

Hipótesis

El genoma núcleo del género S*treptococcus* se compone de una serie de genes/funciones conservados. Mediante la construcción de perfiles de abundancia de estos genes en grupos filogenéticamente relacionados, se puede obtener información de distancias genéticas que discriminen similitudes de composición de genes y que puedan ser correlacionadas con funciones de estos organismos en sus distintos ambientes.

Metodología

Construcción inicial de la base de datos

Se descargaron los archivos FASTA de aminoácidos, DNA y de RNA de todas las cepas del género *Streptococcus* cuyos genomas estuvieran completamente secuenciados y disponibles en la base de datos de genomas del Genbank a través del servidor FTP (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) en la fecha 08/07/2013. De igual manera fueron descargados los archivos de *Bacillus subtilis* str. 168 y *Bacillus licheniformis* 9945A, los cuales se decidieron utilizar como grupos externos, en este trabajo, con base en la localización de dichas especies en el árbol filogenético propuesto por (Ciccarelli et al., 2006)

Los números de acceso de los genomas que se descargaron son los siguientes: NC_002737, NC_003028, NC_003098, NC_003485, NC_004070, NC_004116, NC_004350, NC_004368, NC_004606, NC_006086, NC_006448, NC_006449, NC_007296, NC_007297, NC_007432, NC_008021, NC_008022, NC_008023, NC_008024, NC_008500, NC_008501, NC_008532, NC_008533, NC_009009, NC_009332, NC_009442, NC_009443, NC_009785, NC_010380, NC_010582, NC_011072, NC_011134, NC_011375, NC_011900, NC_012004, NC_012466, NC_012467, NC_012468, NC_012469, NC_012470, NC_012471, NC_012891, NC_012923, NC_012924, NC_012925, NC_012926, NC_013798, NC_013853, NC_013928, NC_014251, NC_014494, NC_014498, NC_015215, NC_015219, NC_015291, NC_015433, NC_015558, NC_015600, NC_015678, NC_015760, NC_015875, NC_015876, NC_016749, NC_016750, NC_016826, NC_016837, NC_017040, NC_017053, NC_017563, NC_017567, NC_017576, NC_017581, NC_017582, NC_017591, NC_017592, NC_017593, NC_017594, NC_017595, NC_017596, NC_017617, NC_017618, NC_017619, NC_017620, NC_017621, NC_017622, NC_017768, NC_017769, NC_017905, NC_017927, NC_017950, NC_018073, NC_018089, NC_018526, NC_018594, NC_018630, NC_018646, NC_018712, NC_018936, NC_019042, NC_019048, NC_020526, NC_020540, NC_021005, NC_021006, NC_021026, NC_021028, NC_021175, NC_021195, NC_021213, NC_021314, NC_021485, NC_021486, NC_021507, NC_021807, NC_021900, NC_000964 y NC_021362.

Para la descarga de los archivos se utilizó el comando:

```
$rsync -avh ftp.ncbi.nih.gov::genomes/Bacteria/Streptococcus*/*.faa .
```

Se utilizó la extensión .faa para designar archivos con secuencias de aminoácidos, la extensión .ffn para designar archivos con secuencias de nucleótidos de regiones codificantes, la extensión .frn para designar archivos de secuencias de RNA no codificante y .gff para los archivos de información general (General Feature Format).

Se organizó la información en subcarpetas como se muestra en el diagrama de la Figura 6.

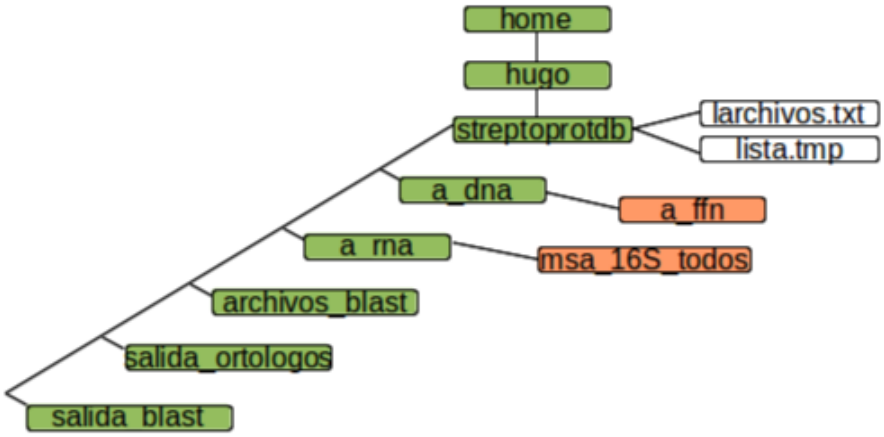

Figura 6. Diagrama de organización de información. El directorio a_dna contiene a los archivos de nucletidos (.ffn y .gff). El directorio a_rna contiene los archivos de RNA no codficante. El directorio archivos_blast contiene a los proteomas. El directorio salida_blast contiene a los archivos de salida de los alineamientos en blast. El directorio salida_ortologos contiene a los archivos de salida de la búsqueda de ortólogos.

Dado que cada organismo puede tener almacenada su información genética en más de un cromosoma y plásmidos, cada conjunto de proteomas predichos fueron organizados en un sólo archivo que contenía la información de cromosomas y elementos extracromosomales; por lo tanto, los archivos con los números de acceso: NC_015219, NC_016837, NC_016750, NC_015876, NC_012923, NC_008500 y NC_008501 fueron concatenados a sus especies correspondientes de la base de datos, usando el número de acceso del cromosoma como identificador. La concatenación de las secuencias en dichos archivos con sus proteomas correspondientes se logra con el siguiente comando:

```
$cat NC_015215.faa NC_015219.faa >NC_015215.faa
```

Posteriormente cada archivo de aminoácidos fue formateado con el programa *formatdb* de la paquetería de Blast v. 2.2.26 ((Altschul et al., 1990) indicando que las secuencias son proteínas, generando así bases de datos de cada proteoma las cuales son necesarias para hacer búsquedas mediante el programa BLASTP. El comando utilizado para realizar esta acción fue:

```
$formatdb -i NC_002737 -p T -o T
```

Donde:
-i es el archivo de entrada
-p es la indicación del tipo de secuencias, ya sean proteínas (T) o nucleótidos (F).

Finalmente se determinó el número de proteínas con las que cuenta cada proteoma utilizando una lista con los nombres de los archivos de los proteomas, denominada “lista.tmp” y el siguiente bucle de comandos:

```
#!/bin/bash
for ARCHIVO in `cat lista.tmp`
 do
grep “>” $ARCHIVO | wc -l >>no_prots.txt
 done
```

Donde el comando grep busca la expresión regular del símbolo “mayor que” (>) dentro de cada archivo y esa búsqueda se direcciona hacia el comando wc con la opción -l que cuenta el número de veces que aparece por cada línea el caracter especificado. Esta cuenta se entiende como el número de secuencias (proteínas) contenidas en cada archivo, ya que en los archivos FASTA cada secuencia tiene un encabezado el cual comienza con este símbolo seguido de una descripción.

Búsqueda de genes ortólogos

Para la búsqueda de los ortólogos por pares se siguió la estrategia de Mejores Aciertos Recíprocos (*Reciprocal Best Hit;* RBH), donde se consideran como genes ortólogos a dos genes que se encuentran a sí mismos como el mejor acierto (*hit)* posible, basados en el puntaje de bits entre la secuencia problema y la base de datos (Moreno-Hagelsieb & Janga, 2008).

Se realizaron 12,100 alineamientos locales de los 108 proteomas completos de los

Streptotocci, mediante Legacy Blast (v. 2.2.26), de los proteomas completos de cada una de las 108 cepas con genomas completos secuenciados del género *Streptococcus* y de las cepas *Bacillus subtilis* str. 168 y *Bacillus licheniformis* 9945A, utilizando BLASTP con un valor de corte e- de $1x10^{-6}$, utilizando un formato de salida tabular y para las demás opciones se utilizaron los valores predeterminados del programa. El comando utilizado fue el siguiente:

```
blastall -p blastp -i NC_002737.faa -d NC_003028.faa -m 8 -e 1e-6 -v 10 -b 10 -a 3
> NC_002737.faa- NC_003028.faa.bout
```

Donde:

-p es el programa a utilizar.

-i es el archivo de secuencias problema.

-d es la base de datos.

-m es el formato de salida.

-e es el valor de corte.

-v es el número de secuencias de la base de datos de las que se muestran descripciones de una línea.

-b es el número de secuencias de bases de datos de las que se muestran alineamientos.

-a es el número de procesadores a utilizar.

Se utilizó la lista denominada “lista.tmp” dentro del directorio “streptoprotdb”, de los archivos que contienen a cada proteoma y un script de bash donde se combinaron todos los proteomas predichos, el script es el siguiente:

```
#!/bin/bash

DIRECTORIO=~/streptoprotdb/archivos_blast
cd ~/streptoprotdb/archivos_blast

for VARIABLEQ in `cat ~/streptoprotdb/lista.tmp` #Se define la primer variable que
es cada uno de los archivos denotados en lista.tmp
 do

for VARIABLEDB in `cat ~/streptoprotdb/lista.tmp`
 do

echo "blastall -p blastp -i $DIRECTORIO/$VARIABLEQ".faa" -d $DIRECTORIO/
$VARIABLEDB".faa" -m 8 -e 1e-6 -v 10 -b 10 -a 3 >>
~/streptoprotdb/salida_blast/"$VARIABLEQ-$VARIABLEDB".bout"
```

```
done
done
```

Un ejemplo de la salida tabular se muestra en la Tabla 3

gi|15674251|ref|NP_268424.1|	gi|15899950|ref|NP_344554.1|	61.1	455	165	5	5	451	3	453	0	550
gi|15674252|ref|NP_268425.1|	gi|15899951|ref|NP_344555.1|	73.02	378	102	0	1	378	1	378	0	539
gi|15674253|ref|NP_268426.1|	gi|15899952|ref|NP_344556.1|	70.31	64	19	0	1	64	19	82	5.00E-034	110
gi|15674254|ref|NP_268427.1|	gi|15899953|ref|NP_344557.1|	94.34	371	21	0	1	371	1	371	0	689
gi|15674254|ref|NP_268427.1|	gi|15900948|ref|NP_345552.1|	35.21	142	60	4	6	147	162	271	5.00E-016	75.1

Tabla 3. Ejemplo de la salida tabular de BLAST. Se muestran 12 columnas de datos. La primera es el identificador global y el númro de acceso de la secuencia problema, la segunda es el identificador global y número de acceso de la base de datos, la tercera es el valor de porcentaje de identidad, la cuarta es la longitud del alineamiento, la quinta son los mismatches, la sexta es el númro de gaps, la séptima y octava; novena y décima son los sitios de inicio y fin del alineaminto de la secuencia problema y base de datos respectivamente, la onceava es el e-value y la doceava es el valor de bit-score.

Los alineamientos fueron realizados de manera bidireccional, comparando a un proteoma “A” contra un proteoma “B”, utilizando primero al proteoma “A” como secuencia problema y al proteoma “B” como la base de datos contra la cual se realiza la búsqueda y después siendo el proteoma “B” la secuencia problema y el proteoma “A” la base de datos.

Mediante un script se compararon los alineamientos de Blast en dos vías. Primero se remueven los aciertos duplicados (DUPL) Posteriormente se obtienen los mejores valores de alineamiento (*bit-scores*) de cada uno de los 2 archivos a comparar (BEST). donde después se utiliza la comparación de identificadores para unir dos salidas de blast en una misma tabla (MERGE). Se hacen las comparaciones por medio de los valores de bit-scores y los identificadores de cada secuencia (COMPARE). A partir del archivo de comparación se genera una lista donde si la secuencia encuentra a el mismo acierto siendo problema y base de datos, comparación por nombres, se imprime la salida con el identificador de la secuencia, de la base de datos y el valor de bit-score (ORTHOLOGS).

El script de perl se llama bbh.sh y se muestra a continuación:

```
#!/bin/bash

#Script para obtener lista de genes ortólogos mediante blast bidireccionales
#Se necesitan dos archivos de salida de blast con salida tabular (-m8) para poder
#comparar las listas. La forma de uso del script es la siguiente:


#$bash bbh.sh nombre_archivo_1 nombre_archivo_2
#Se generan varios archivos, el resultado final se almacena en los que tienen la terminación .orthologs

#Luis David Alcaraz, última actualización 2013-04-09
```

```
i=0
argv=()
for arg in "$@";
do argv[$i]="$arg"
i=$((i + 1))
done

i=0
arg1="${argv[$i]}"
i=$((i + 1))
arg2="${argv[$i]}"


#Se remueven duplicados del mismo valor, para evitar tener múltiples hits

echo "perl -ne 'BEGIN {\$column = 0}' -e 'BEGIN {\$unique=0}; s/\r?\\n//; @F=split /\t/, \$_; if (!(\
$save{\$F[\$column]}++)) {print \"\$_\\n\"; \$unique++} END' $1 >$1.dupl" >$1-$2.bash

echo "perl -ne 'BEGIN {\$column = 0}' -e 'BEGIN {\$unique=0}; s/\r?\\n//; @F=split /\t/, \$_; if (!(\
$save{\$F[\$column]}++)) {print \"\$_\\n\"; \$unique++} END' $2 >$2.dupl" >>$1-$2.bash

#Se obtienen los mejores valores de alineamiento para cada uno de los archivos a comparar
echo "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})) {push @names, \$n}; 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}}}' $1.dupl > $1.best" >>$1-$2.bash

echo "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})) {push @names, \$n}; 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}}}' $2.dupl > $2.best"  >>$1-$2.bash

#Se Unen las listas con los valores máximos

echo "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}}'
$1.best $2.best >$1-$2.merge" >>$1-$2.bash

#Se comparan las listas unidas por columnas
echo "perl -ne 'BEGIN {\$colm=0; \$coln=13;}' -e 's/\r?\\n//; @F=split /\t/, \$_; if (\$F[\$colm] eq \
$F[\$coln]) {print \"\$_\\n\"}' $1-$2.merge > $1-$2.compare" >>$1-$2.bash

#Se genera la lista de ortólogos por pares
echo "perl -ne 'BEGIN {@cols=(0, 1, 11)}' -e 's/\r?\\n//; @F=split /\t/, \$_; print join(\"\t\",
@F[@cols]), \"\\n\"' $1-$2.compare >$1-$2.orthologs" >>$1-$2.bash



bash $arg1-$arg2.bash

#EOF
```

La forma de ejecutar el script es:

```
$./bbh.sh ARCHIVO1 ARCHIVO2
```

Un ejemplo de las salidas de los archivos es:

NC_012469-NC_017769.bout-NC_017769-NC_012469.bout.bash

NC_012469-NC_017769.bout-NC_017769-NC_012469.bout.compare

NC_012469-NC_017769.bout-NC_017769-NC_012469.bout.merge

NC_012469-NC_017769.bout-NC_017769-NC_012469.bout.orthologs

Donde el archivo con la terminación .orthologs es el que contiene la lista de proteínas ortólogas entre los dos genomas comparados, siendo en este caso los proteomas de NC_012469 y NC_017769. Se presenta a continuación una muestra de la salida de dicho archivo.

gi|225860013|ref|YP_002741522.1|	gi|387787131|ref|YP_006252199.1|	884
gi|225860014|ref|YP_002741523.1|	gi|387787132|ref|YP_006252200.1|	754
gi|225860015|ref|YP_002741524.1|	gi|387787133|ref|YP_006252201.1|	134
gi|225860016|ref|YP_002741525.1|	gi|387787134|ref|YP_006252202.1|	275

Tabla 4. Ejemplo de la salida de BLAST. Se muestra el ejemplo de los aciertos identificados como ortólogos. La primera columna corresponde a los identificadores del primer proteoma comparado, la segunda contiene a los identificadores del segundo proteoma y la tercera con los valores de *bit score*.

Caracterización del genoma núcleo por género y especie.

Para determinar las proteínas que componen el genoma núcleo del género *Streptococcus* se tomó como referencia al integrante con el genoma más pequeño (*S. Agalactiae* 2-22; No. de acceso: NC_021195; No. de proteínas: 1548). Con las listas de proteínas codificadas en este genoma se hizo la búsqueda de ortólogos contra todos los demás proteomas del género y se obtuvieron patrones de ausencia, presencia de cada una de las proteínas con respecto a los demás integrantes del género.

Esta búsqueda se hizo con los identificadores globales (GI, del inglés *Global Identifier*) los cuales están asociados a las claves de acceso de cada una de las proteínas de NC_021195 en cada una de las listas de ortólogos, mediante la función VLOOKUP en LibreOffice Calc (v 4.0.2.2). Un ejemplo de dichas búsquedas se presenta en la Tabla 5.

GI S.agalactiae	matriz de búsqueda		Fórmula de búsqueda
494702402	494702402	1	1
494702403	494702403	1	1
494702404	494702405	1	0
494702405	494702406	1	1
494702406	494702407	1	1

Tabla 5. Ejemplo de las búsquedas de proteínas del genoma núcleo en LibreOffice Calc.

Donde la fórmula de búsqueda es:
=VLOOKUP(T3,$U$3:$V$1101,2,0)

La fórmula busca el identificador de la referencia, presente en la primera columna, dentro de la lista de identificadores en la segunda columna de la Tabla 5. Si existe el identificador en la lista de la segunda columna entonces se imprime el número 1, si no existe se imprime el número 0.

Por cada proteína encontrada en todas las especies analizadas se hace una suma de presencia y cuando ésta es igual al total de cepas analizadas (108) se considera a dicho gen como parte del genoma núcleo del género.

Las secuencias correspondientes al genoma núcleo fueron extraídas del genoma de referencia con el programa *seqret* del paquete EMBOSS (v. 6.4.0.0) (Rice et al., 2000) utilizando los identificadores USA (del inglés *Uniform Sequence Address*), los cuales utilizan a las claves de acceso de las proteínas y se nombran de la siguiente manera: fasta::NC_021195.faa:YP_007967863.1

Se hizo una lista con los USA, denominada “listaUSA_core.txt” y mediante un bucle de comandos en bash se realizó la extracción de las secuencias de manera automática, generándose así un archivo multifasta conteniendo a todas las secuencias del genoma núcleo. El bucle utilizado fue el siguiente:

```
#!/bin/bash
for i in `cat listaUSA_core.txt`
 do
seqret -auto -stdout $i >mf_core.faa
 done
```

donde:
-auto es la opción para desactivar el modo interactivo.
-stdout indica que se escriba el archivo en salida estándar.

Utilizando la estrategia anteriormente descrita, se caracterizaron a los genomas núcleo de

las especies con más de tres genomas completamente secuenciados. Los genomas que se utilizaron como referencia para cada especie son los siguientes: *S.agalactiae* 2-22 (NC_021195), *S.dysgalactiae* subsp. Esquimilis RE37 (NC_018712), *S. equii* subsp. Zooepidemicus (NC_012470), *S. gallolyticus* UCN 34 (NC_013798), *S. mutans* UA159 (NC_004350), *S. pneumoniae* SPN034183 (NC_021028), *S. pyogenes* M1 476 (NC_020540), *S. salivarius* 57.I (NC_017594), *S. suis* P1/7 (NC_012925) y *S. thermophilus* LMD-9 (NC_008532). El número de genomas utilizados para la construcción de cada genoma núcleo se encuentra en la tabla 2.

Caracterización del pangenoma.

En este trabajo se utilizó como definición de pangenoma al conjunto total de genes que forman parte del género *Streptococcus*. Para realizar la construcción del pangenoma, se concatenaron los archivos de aminoácidos de todos los genomas disponibles en un archivo multifasta utilizando una lista de dichos archivos, denominada “lista_aa_pangenoma.txt” y el siguiente bucle:

```
#!/bin/bash
for $i in `cat lista_aa_pangenoma.txt`
 do
cat $i >>mf_pangenoma.fas
 done
```

Con este archivo se utilizaron diferentes *scripts* de la paquetería QIIME (Caporaso et al., 2010), iniciando por el script *pick_otus.py* (http://qiime.org/1.3.0/scripts/pick_otus.html) con el cual se asignan secuencias similares a familias de proteínas, mediante el algoritmo de clustering CD-HIT (Huang et al., 2010), utilizando un valor de corte del 80% de similitud. A continuación se utilizó el script *make_otu_table.py* (http://qiime.org/1.3.0/scripts/make_otu_table.html) con el cual se tabulan el número de veces que aparece un OTU en cada muestra, siendo cada muestra un genoma. Esta tabla se convirtió a un formato tabular mediante el script del proyecto BIOM (McDonald et al., 2012) *convert_biom.py*. Los comandos utilizados fueron los siguientes:

```
$pick_otus.py -i mf_pangenoma.fas -m cdhit -o strepto-CDHIT/ -M 2000 -T 3 -s 0.8
```

```
$make_otu_table.py -i strepto-CDHIT/seq_otus.txt -o strepto.biom
$convert_biom.py -i strepto.biom -o strepto.biom.tab
```

Donde:

-i es el archivo de entrada.

-m es el método que se utiliza para seleccionar a los OTU's.

-o es el nombre que se le da al directorio donde se guardan los archivos de salida.

-M es el máximo de memoria que se le da a CD-HIT.

-s es el valor de corte de similitud.

Con la tabla generada mediante el script *convert_biom.py* se calculó el número de familias génicas formadas por el agrupamiento con CD-HIT. Esta tabla incluye la descripción de la proteína, el número de acceso de la secuencia y el organismo del cual proviene; así como el conteo de presencia de cada secuencia por genoma. Una vez generada esta tabla, se procedió a generar una gráfica de acumulación de genes por genoma añadido en el programa R (v. 3.0.2) utilizando la función *specaccum* de la librería “vegan” (Oksanen *et al.,* 2013) mediante los siguientes comandos:

```
$strepto_tabla_transpuesta <- t(strepto_data)   #Se transpone la tabla para cambiar las columnas por filas.
$strepto_acum2 <-specaccum(strepto_tabla_transpuesta, "random", permutations = 100) #se realiza con specaccum la acumulación de familias, añadiendo las muestras de mnera aleatoria.
$plot(strepto_acum2,     ci=2,     ci.type="polygon",     col="blue",     ci.lty=0, ci.col="lightblue")  #Se genera el gáfico con un polígono indicando la desviación estandar de los datos.
```

Para calcular la estadística descriptiva básica del pangenoma (máximos, mínimos, medias y medianas) se utilizó el script de QIIME *per_library_stats.py (*http://qiime.org/1.3.0/scripts/per_library_stats.html)*,* con el comando:

```
$per_library_stats.py -i strepto.biom >stats_pg.txt
```

Donde:

-i indica el archivo de entrada

>stats_pg.txt indica que la salida del script se guarde en un archivo llamado stats_pg.txt

Anotación de los genomas núcleo y pangenoma.

Se generaron archivos multifasta de las secuencias codificantes de nucleótidos tanto del pangenoma como del genoma núcleo mediante dos estrategias distintas. Para el genoma núcleo se realizaron las siguientes acciones:

Se utilizó el comando *infoseq* sobre el archivo FASTA de regiones codificantes de nucleótidos del genoma de referencia (S. agalactiae 2-22, NC_021195) para obtener los USA de cada secuencia y colocarlos en una lista.

Posteriormente, se utilizó el archivo correspondiente a este genoma con extensión .gff, denominado General Feature Format (GFF), el cual contiene información general sobre las secuencias contenidas en el archivo de regiones codificantes de nucleótidos con extensión .ffn, como son el nombre de la secuencia, su número de acceso de genbank, atributos, productos, etc. Se tomaron los números de acceso de cada secuencia del genoma a partir de este archivo y se hizo una lista, la cual fue asociada a la lista de identificadores USA's generada anteriormente. Mediante la función VLOOKUP en LibreOffice Calc (v 4.0.2.2) se comparó la lista de identificadores de las proteínas contenidas en el genoma núcleo contra la lista de identificadores-anotaciones extraídos del archivo GFF y se imprimen los aciertos concordantes entre las listas. Un ejemplo de la búsqueda se muestra en la Tabla 6.

No. de acceso .faa	No. de acceso .gff	Lista USA's	Fórmula de búsqueda
YP_007967863.1	YP_007967863.1	fasta::NC_021195.ffn:207-1568	fasta::NC_021195.ffn:207-1568
YP_007967867.1	YP_007967864.1	fasta::NC_021195.ffn:1723-2859	fasta::NC_021195.ffn:4580-5695
YP_007967868.1	YP_007967865.1	fasta::NC_021195.ffn:2929-3789	fasta::NC_021195.ffn:5779-6354
YP_007967873.1	YP_007967866.1	fasta::NC_021195.ffn:3820-4017	fasta::NC_021195.ffn:10901-12187
YP_007967874.1	YP_007967867.1	fasta::NC_021195.ffn:4580-5695	fasta::NC_021195.ffn:12189-13463

Tabla 6. Ejemplo de las búsquedas de los USAs asociados a cada secuencia de proteínas.

Donde la fórmula de búsqueda es:

=VLOOKUP(A2,$Q$3:$R$1550,2,0)

La fórmula busca el número de acceso (A2) de la secuencia de aminoácidos (primera columna) dentro del rango establecido ($Q$3:$R$1550) el número de acceso de las

secuencias de nucleótidos (B) y si existe dentro de de dicha lista, se imprime en la cuarta columna (Fórmula de búsqueda) el identificador USA correspondiente.

Se generó una lista de identificadores USAs, denominada “listaUSA_core.txt” y se separaron aquellos cuyo marco de lectura se encontraba inverso utilizando el comando grep, el cual realiza la búsqueda de expresiones regulares en archivos de texto. Los siguientes comandos fueron utilizados:

```
grep “:c” listaUSA_core.txt >>sqrt_coreinv.txt
grep “:c” -v listaUSA_core.txt >>sqrt_core.txt
```

Donde:
“:c” es la expresión que se pretende buscar en los archivos. Esta expresión está incluida solo en las secuencias cuyo marco de lectura es invertido.
-v indica que se busque todo aquello que no es la expresión que se está indicando.
sqrt_coreinv.txt es el archivo con la lista de las secuencias invertidas.
sqrt_core.txt es el archivo con la lista de las secuencias en sentido regular.

Se prosiguió a extraer las secuencias de nucleótidos del genoma núcleo del género a partir del archivo del genoma de referencia (S.agalactiae 2-22; NC_021195) utilizando las listas generadas en el punto anterior mediante los siguientes bucles de comandos:

```
#!/bin/bash
for i in `cat sqrt_core.txt`
 do
seqret -auto -stdout $i >>mf_core_genero.ffn
 done

for i in `cat sqrt_coreinv.txt`
 do
seqret -auto -stdout -srev $i >>mf_core_genero.ffn
 done
```

Donde:
-srev indica que imprima el complemento reverso.
mf_core_genero.ffn es el archivo multifasta generado que contiene las secuencias de nucleótidos.

Para la anotación del pangenoma se generó un archivo multifasta que contuviera todos los genes de *Streptococcus,* utilizando una lista con los nombres de todos los archivos de nucleótidos de regiones codificantes con extensión .ffn, denominada “lista_archivos_nuc.ffn” y mediante el siguiente bucle de comandos:

```
for i in `cat lista_archivos_nuc.txt”
 do
cat $i >>mf_strepto_seqall.ffn
 done
```

Se utilizaron los archivos multifasta de nucleótidos con los genes del pangenoma, genoma núcleo del género y especies selectas para realizar su anotación mediante el servidor MG-RAST v. 3.0 (Meyer et al., 2008), en una clasificación jerárquica por subsistemas y por COGs (del inglés *Cluster of Orthologous Groups*) utilizando la base de datos M5NR, con valores de corte e-value de $1e^{-5}$, identidad mínima del 60% y una longitud mínima de alineamiento de 15. Los archivos fueron subidos al servidor indicando las opciones de dereplicación, la cual elimina secuencias duplicadas; filtro de longitud y bases ambiguas, indicando un factor de 2.0 desviaciones estándar y un máximo de 5 pares de bases ambiguas. Estas últimas opciones remueven las secuencias que difieren del promedio de longitud por más del número de desviaciones estándar especificada y las secuencias que contienen más del número de bases ambiguas especificadas. Posteriormente, se descargaron los datos de abundancia de las secuencias anotadas por subsistemas y COGs y se procedió a normalizar los valores en R (v. 3.01) (R Core Team, 2013), generándose un *heat map* que utiliza los valores normalizados de las abundancias por columna (Z-*score*), utilizando la librería “gplots” (Warnes *et al.,* 2013). Se presenta un ejemplo de los comandos utilizados para generar uno de los gráficos:

```
library(gplots)
library(RColorBrewer)      #Se cargan las librerías

heatmap_sbs <- read.table (tabla_hmp_sbs.csv, header=TRUE, row.names, sep=”t”)
#Se lee la tabla con los datos de abundancia de funciones

colores <- colorRampPalette(brewer.pal(9, “PuOr”)    #Se genera una paleta de
colores

heatmap.2 (as.matrix(heatmap_sbs), key=T, symkey=F, trace=”none”, scale=”column”,
col=colores, Rowv=T, keysize=1.5, cexRow=0.5, cexCol=0.5)
#Se genera el gráfico
```

Reconstrucción filogenética.

Se extrajeron los identificadores de las secuencias de RNA ribosomal 16S de cada uno de los 108 genomas y de los genomas de *Bacillus subtilis* str. 168 y *Bacillus licheniformis* 9945A, a partir de los archivos fasta de RNA no codificantes con extensión .frn, utilizando una lista con los nombres de los 110 archivos denominada “listaRNA.txt” y mediante el uso del siguiente bucle de comandos:

```
for i in `cat listaRNA.txt`
 do
infoseq $i | grep "16S" >> lista16s.txt
 done
```

Se revisó y editó la lista de USA generada para solo contuviera una secuencia de RNA 16S por cada genoma debido a que existen múltiples copias de este gen. Una vez editada esta lista, se procedió a extraer las secuencias de los archivos y concatenarlas en un archivo multifasta con el siguiente bucle de comandos:

```
#!/bin/bash
for i in `cat lista16s.txt`
 do
seqret -auto -stdout $i  >>mf_16s_todos.fas
 done
```

El archivo multifasta generado se utilizó para realizar un alineamiento múltiple con el programa SSU-align v. 0.1 (Nawrocki *et al*., 2013) con el siguiente comando:

```
$ssu-align mf_16S_todos.fas msa_16S_todos.fas
```

Donde el primer término del comando es el programa a utilizar, el segundo es el nombre del archivo de entrada y el tercero es el nombre de la carpeta que se generará donde se guardan los archivos de salida.

El programa SSU-align genera un alineamiento múltiple en formato Stockholm, denominado “msa_16S_todos.fas.bacteria.stk”, el cual es un formato utilizado para incluir información y características dentro del archivo del alineamiento múltiple. Este archivo se convirtió a formato FASTA con este mismo programa mediante el comando:

```
$ssu-mask –stk2afa -a msa_16S_todos.fas.bacteria.stk
```

Donde:
ssu-mask es el nombre del programa a utilizar.
--stk2afa es la opción que indica que se convierta el formato Stockholm a FASTA.
-a indica cual es el archivo a convertir.

El alineamiento múltiple en formato FASTA fue sometido a una prueba de modelos de sustitución en el programa MEGA 5.2 (Tamura *et al*., 2011) para conocer cuál era el más adecuado para el conjunto de datos. Posteriormente se utilizó el modelo de dos parámetros de Kimura para generar un árbol filogenético de máxima verosimilitud con cien repeticiones de *bootstrap* en MEGA 5.2 .

Distancias genómicas.

Para comparar la similitud entre genomas (GSS, del inglés *Genomic Similarity Score*) (Alcaraz et al., 2010; Moreno-Hagelsieb & Janga, 2008), se sumaron los valores de *bit scores* de BLAST de las listas de genes ortólogos resultantes de las comparaciones entre los 108 genomas de los streptococci y los 2 bacilli, detectados como RBH, por cada par de proteomas utilizando una lista con los nombres de los archivos, denominada “larchivos.txt” con un script en bash que se presenta a continuación:

```
#!/bin/bash
#Script para hacer sumas de bitscores en cada archivo de las salidas de ortólogos

# i = nombre del archivo (NC_XXX)
# ARCHIVO = salida .orthologs

DIRECTORIO=/home/hugo/streptoprotdb/salida_ortologos/

 cd $DIRECTORIO
for i in `cat /home/hugo/streptoprotdb/larchivos.txt`
 do
  cd $DIRECTORIO/$i/
  ls * >lista.tmp
done

for i in `cat /home/hugo/streptoprotdb/larchivos.txt`
 do
  cd $DIRECTORIO/$i/
for ARCHIVO in `cat lista.tmp`
 do
```

```
  awk '{ SUM += $3} END { print SUM}' $ARCHIVO >> sumabs”$i”.tmp
done
done
```

Este script genera, primero, listas con los nombres de cada archivo conteniendo a las listas de ortólogos, denominadas “lista.tmp” y posteriormente utiliza éstas para indicar, con el comando awk, que se haga la suma de la tercera columna de los archivos que contienen las listas de ortólogos como se muestra en el ejemplo de la Tabla 4.

Una vez obtenidas las sumas de *bit scores* por pares de genomas se procedió a calcular el GSS y generar una matriz de distancias genómicas de todos los organismos utilizados en este estudio, donde cada valor fue normalizado contra el valor de la comparación entre el mismo genoma (*auto score*), para normalizar. La fórmula para calcular la distancia genómica se encuentra en la Figura 2.

La matriz de distancias GSS se graficó en un árbol por el método de *Neighbor-Joining* en MEGA 5.2.

Abundancia y funciones de *Streptococcus* en metagenomas selectos

Se analizaron los ocho metagenomas orales humanos disponibles en el servidor MG-RAST en búsqueda de la abundancia de las especies del género *Streptococcus.* Los metagenomas explorados fueron los siguientes: CA_04P, CA1_01P, CA1_02P, CA_06_1.6, CA_05_4.6, CA_06P, NOCA_03P y NOCA_01P (Belda-Ferre et al., 2012).

Para lograr esto se ingresó en cada metagenoma y se pidió el análisis de clasificación por mejor acierto, utilizando la base de datos M5NR, con valores de corte e-value de $1e^{-5}$, identidad mínima del 60% y una longitud mínima de alineamiento de 15. Se solicitó que los datos se presentaran como gráficas de barras y se exploraron los datos de abundancias taxonómicas con valores crudos hasta llegar a los datos del género *Streptococcus,* donde se desplegaron en cada caso los datos de abundancia por especie y cepa. En esta ventana se descargaron los datos que generaron las gráficas de barras y se procedió a generar tablas en LibreOffice Calc (v 4.0.2.2) con estos datos para poder generar gráficas comparativas de las abundancias entre los metagenomas. Posteriormente se enviaron a la “mesa de trabajo”

de MG-RAST los features correspondientes a los streptococci de cada metagenoma para realizar el análisis funcional de éstos mediante la clasificación por subsistemas y por COG de la misma manera que se describe en el apartado de la anotación de los genomas núcleo y pangenoma. Se descargaron las anotaciones realizadas por cada clasificación y se hizo una tabla para generar un gráfico de calor (*heatmap*) en R (v.3.0.2). Un ejemplo de los comandos utilizados se muestra en el apartado de la anotación de genomas núcleo y pangenoma.

Resultados

Construcción de la base de datos.

En este estudio se generó una base de datos que contiene a los archivos de secuencias de aminoácidos, nucleótidos de regiones codificantes y RNA no codificante de los 108 genomas completos de *Streptococcus* y a los proteomas de las cepas *Bacillus subtilis* str. 168 y *Bacillus licheniformis* 9945A. Esta base de datos ha sido formateada para realizar alineamientos con BLAST y se encuentra disponible en la dirección URL (http://dx.doi.org/10.6084/m9.figshare.974600).

Se contó el número de proteínas predichas para cada organismo utilizado en este trabajo y se reviso la bibliografía asociada a cada uno para reportar su sitio de aislamiento y estilo de vida, mostrados en la Tabla 7. A partir del conteo del número de proteínas de cada cepa, se ha calculado el número promedio con las que cuenta el género *Streptococcus* (108 cepas utilizadas); resultando en un promedio de 1969 ± 163 proteínas para todo el género.

Caracterización del pangenoma, genoma núcleo del género y especies.

Se realizaron 12,100 alineamientos bidireccionales entre los 108 proteomas de los streptococci y los 2 bacilli anteriormente mencionados mediante BLASTP, identificándose a los ortólogos por pares como RBH. Se adiciona la tabla S1, como material suplementario, donde se muestra el número de genes ortólogos que presenta cada par de genomas en la dirección URL (http://dx.doi.org/10.6084/m9.figshare.956201). Utilizando la información

resultante de la búsqueda de genes ortólogos por pares, se procedió a calcular el número de genes codificantes compartidos por las 108 cepas de *Streptococcus* utilizadas en este estudio (genoma núcleo del género). Debido a que *S. agalactiae* 2-22 (NC_021195), cuenta con el genoma más pequeño de entre todas las cepas utilizadas en este estudio, resulta ser una buena referencia para poder conocer el número de proteínas compartidas entre todos los genomas (genoma núcleo del género). Al comparar las listas de ortólogos de cada cepa contra el proteoma predicho de *S. agalactiae* 2-22 y detectar la ausencia o presencia de éstos en cada genoma comparado, se encontró que el genoma núcleo del género se conforma por un total de 404 genes, representando apenas una quinta parte del promedio de genes calculado para el género.

Así mismo, se hicieron búsquedas para determinar el genoma núcleo de especies con más de tres genomas secuenciados, presentadas en la tabla 8.

Nombre	CDS	No. de acceso	Sitio de asilamiento	Estilo de vida	Referencia
Streptococcus agalactiae 2-22	1548	NC_021195	Truchas en Israel	Patógeno	Rosinski-Chupin *et al*., 2013
Streptococcus agalactiae 09mas018883	2089	NC_021485	Leche de vaca con mastitis clínica en Suiza	Patógeno	Zubair *et al*., 2013
Streptococcus agalactiae 2603V/R	2124	NC_004116	*Homo sapiens*: aislado clínico	Patógeno	Tettelin *et al*., 2001
Streptococcus agalactiae A909	1996	NC_007432	*Homo sapiens*: neonato en estado de sepsis	Patógeno	Tettelin *et al*., 2005
Streptococcus agalactiae GD201008-001	1964	NC_018646	Tilapia: meningoencefalitis	Patógeno	Liu *et al*., 2012
Streptococcus agalactiae ILRI005	2155	NC_021486	*Camelus dormedarius*	Patógeno	Zubair *et al*., 2013
Streptococcus agalactiae ILRI112	2073	NC_021507	*Camelus dormedarius*	Patógeno	Zubair *et al*., 2013
Streptococcus agalactiae NEM316	2094	NC_004368	*Homo sapiens*: septicemia fatal	Patógeno	Glaser *et al*., 2002
Streptococcus agalactiae SA20-06	1710	NC_019048	*Tilapia*	Patógeno	Pereira *et al*., 2013
Streptococcus dysgalactiae Subsp. equisimilis AC-2713	2215	NC_019042	*Homo sapiens*: hemocultivo	Patógeno	Watanabe *et al*., 2013
Streptococcus dysgalactiae Subsp. equisimilis ATCC 12394	2056	NC_017567	Ubre bovina	Patógeno	Zuzuki *et al*., 2011
Streptococcus dysgalactiae Subsp. equisimilis GGS_124	2094	NC_012891	*Homo sapiens*: STSS	Patógeno	Kirikae *et al*., 2011
Streptococcus dysgalactiae Subsp. equisimilis RE378	1877	NC_018712	*Homo sapiens*: infecciones invasivas	Patógeno	Okumura *et al*., 2012
Streptococcus equi subsp. Equi 4047	2000	NC_012471	Caballo	Patógeno	Holden *et al*., 2009
Streptococcus equi subsp. Zooepidemicus	1869	NC_012470	Caballo	Patógeno	Holden *et al*., 2009
Streptococcus equi subsp. Zooepidemicus ATCC 35246	2087	NC_017582	*Sus scrofa*	Patógeno	Ma *et al*., 2011
Streptococcus equi subsp. Zooepidemicus MGCS10565	1893	NC_011134	*Homo sapiens*: garganta	Patógeno	Beres *et al*., 2008
Streptococcus gallolyticus subsp. Gallolyticus ATCC 43143	2246	NC_017576	Homo sapiens: hemocultivo	Patógeno	Lin et al., 2011
Streptococcus gallolyticus subsp. Gallolyticus ATCC BAA-2069 *	2329	NC_015215	Homo sapiens: hemocultivo	Patógeno	Camilli *et al*., 2008
Streptococcus gallolyticus UCN34	2223	NC_013798	*Homo sapiens*	Patógeno	Rusniok *et al*., 2010
Streptococcus gordonii str. Challis substr. CH1	2051	NC_009785	*Homo sapiens*	Patógeno	Vickerman *et al*., 2007
Streptococcus infantarius Subsp. infantarius CJ18 *	1906	NC_016826	Leche de camello	No patógeno	Jans *et al*., 2012
Streptococcus iniae SF1	2125	NC_021314	N/D	N/D	N/D
Streptococcus intermedius JTH08	1702	NC_018073	*Homo sapiens:* absceso hepático	Patógeno	Tomoyasu *et al*., 2010
Streptococcus lutetiensis 033	1890	NC_021900	*Homo sapiens*: materia fecal	Patógeno	Jin *et al*., 2013
Streptococcus macedonicus ACA-DC 198 *	1994	NC_016749	Queso kasseri	No patógeno	Papadimitriou *et al*., 2012
Streptococcus mitis B6	2004	NC_013853	*Homo sapiens*	Patógeno	Denapaite *et al*., 2010
Streptococcus mutans GS-5	1878	NC_018089	Homo sapiens: caries dental	Patógeno	Biswas *et al*., 2012
Streptococcus mutans LJ23	1921	NC_017768	*Homo sapiens*: cavidad oral	Patógeno	Aikawa *et al*., 2012
Streptococcus mutans NN2025	1895	NC_013928	*Homo sapiens*: caries dental	Patógeno	Maruyama *et al*., 2009
Streptococcus mutans UA159	1960	NC_004350	*Homo sapiens*: caries dental activa	Patógeno	Ajdić *et al*., 2002
Streptococcus oligofermentans AS 1.3089	2069	NC_021175	*Homo sapiens*: Dientes sanos	No patógeno	Tong *et al*., 2013
Streptococcus oralis Uo5	1907	NC_015291	Homo sapiens: cavidad oral	Patógeno	Reichmann *et al*., 2011
Streptococcus parasanguinis ATCC 15912	2022	NC_015678	*Homo sapiens*: garganta	Patógeno	http://genomesonline.org/cgi-bin/GOLD/GOLDCards.cgi?Goldstamp=Gc01842
Streptococcus parasanguinis FW213	2019	NC_017905	*Homo sapiens*: placa dental	No patógeno	Geng *et al*., 2012
Streptococcus parauberis KCTC 11537	1868	NC_015558	Peces enfermos (*Paralichthys sp.)*	Patógeno	Nho *et al*., 2011
Streptococcus pasteurianus ATCC 43144	1869	NC_015600	*Homo sapiens*: hemocultivo	Patógeno	Lin *et al*., 2011
Streptococcus pneumoniae 670-6B	2352	NC_014498	*Homo sapiens*	Patógeno	Donati et al., 2010
Streptococcus pneumoniae 70585	2202	NC_012468	*Homo sapiens*	Patógeno	http://genomesonline.org/Cgi-bin/GOLD/GOLDCards.cgi?Goldstamp=Gc00969

Tabla 7. Información de los genomas utilizados en este estudio. Se muestra la lista de especies y cepas de Streptococcus utilizadas, el número de proteínas que presenta cada una, el sitio de aislamiento y su estilo de vida. (*Genomas que contienen plásmidos; STSS=Streptococcal Toxic Shock Syndrome).

Nombre	CDS	No. de acceso	Sitio de asilamiento	Estilo de vida	Referencia
Streptococcus pneumoniae AP200	2216	NC_014494	*Homo sapiens*: LCE: meningitis	Patógeno	Camilli *et al*., 2008
Streptococcus pneumoniae ATCC 700669	1990	NC_011900	*Homo sapiens*	Patógeno	Croucher *et al*., 2009
Streptococcus pneumoniae CGSP14	2206	NC_010582	*Homo sapiens*: pneumonía Necrozante	Patógeno	Ding *et al*., 2009
Streptococcus pneumoniae D39	1914	NC_008533	*Homo sapiens*	Patógeno	Lanie *et al*., 2007
Streptococcus pneumoniae G54	2114	NC_011072	*Homo sapiens*: vías respiratorias	Patógeno	Dopazo *et al*., 2001
Streptococcus pneumoniae GamPNI0373	2119	NC_018630	*Homo sapiens*: nasofaringe	Patógeno	http://genomesonline.org/ cgi-bin/GOLD/GOLDCards.cgi? Goldstamp=Gc02735
Streptococcus pneumoniae Hungary19A-6	2155	NC_010380	*Homo sapiens*: oído	Patógeno	http://genomesonline.org/ cgi-bin/GOLD/GOLDCards.cgi? goldstamp=Gc00735
Streptococcus pneumoniae INV104	1820	NC_017591	*Homo sapiens*	Patógeno	Donati *et al*., 2010
Streptococcus pneumoniae INV200	1929	NC_017593	*Homo sapiens*	Patógeno	Donati *et al*., 2010
Streptococcus pneumoniae JJA	2123	NC_012466	*Homo sapiens*	Patógeno	http://genomesonline.org/ cgi-bin/GOLD/GOLDCards.cgi? Goldstamp=Gc00973
Streptococcus pneumoniae OXC141	1823	NC_017592	*Homo sapiens*	Patógeno	Donati *et al*., 2010
Streptococcus pneumoniae P1031	2073	NC_012467	*Homo sapiens*	Patógeno	http://genomesonline.org/ cgi-bin/GOLD/GOLDCards.cgi? goldstamp=Gc00972
Streptococcus pneumoniae R6	2042	NC_003098	*Homo sapiens*	Patógeno	Hoskins *et al*., 2001
Streptococcus pneumoniae SPN034156	1799	NC_021006	*Homo sapiens*	Patógeno	Donati *et al*., 2010
Streptococcus pneumoniae SPN034183	1819	NC_021028	*Homo sapiens*	Patógeno	Donati *et al*., 2010
Streptococcus pneumoniae SPN994038	1819	NC_021026	*Homo sapiens*	Patógeno	Donati *et al*., 2010
Streptococcus pneumoniae SPN994039	1819	NC_021005	*Homo sapiens*	Patógeno	Donati *et al*., 2010
Streptococcus pneumoniae SPNA45	1926	NC_018594	N/D	N/D	Donati *et al*., 2010
Streptococcus pneumoniae ST556	2148	NC_017769	*Homo sapiens*	Patógeno	Li *et al*., 2012
Streptococcus pneumoniae Taiwan19F-14	2044	NC_012469	Homo sapiens: líquido cerebroespinal	Patógeno	http://genomesonline.org/ cgi-bin/GOLD/GOLDCards.cgi? goldstamp=Gc00974
Streptococcus pneumoniae TCH8431/19A	2275	NC_014251	Homo sapiens: tracto respiratorio	Patógeno	http://genomesonline.org/ cgi-bin/GOLD/GOLDCards.cgi? goldstamp=Gc01351
Streptococcus pneumoniae TIGR4	2105	NC_003028	*Homo sapiens*: hemocultivo	Patógeno	Tettelin *et al*., 2001
Streptococcus pseudopneumoniae IS7493 *	2236	NC_015875	*Homo sapiens*: esputo	Patógeno	Shahinas *et al*., 2011
Streptococcus pyogenes SF370	1696	NC_002737	Homo sapiens: herida infectada	Patógeno	Ferreti *et al*., 2001
Streptococcus pyogenes A20	1828	NC_018936	Homo sapiens: fascitis necrozante	Patógeno	Zheng *et al*., 2013
Streptococcus pyogenes Alab49	1773	NC_017596	*Homo sapiens*: impétigo	Patógeno	Bessen *et al*., 2011
Streptococcus pyogenes HSC5	1744	NC_021807	N/D	Patógeno	Port *et al*., 2013
Streptococcus pyogenes M1 476	1572	NC_020540	*Homo sapiens: STSS*	Patógeno	Miyoshi-Akiyama *et al*., 2012
Streptococcus pyogenes MGAS10270	1987	NC_008022	*Homo sapiens*: faríngе	Patógeno	http://genomesonline.org/ cgi-bin/GOLD/GOLDCards.cgi? goldstamp=Gc00378
streptococcus pyogenes MGAS10394	1886	NC_006086	*Homo sapiens*: faringe	Patógeno	Banks *et al*., 2004
Streptococcus pyogenes MGAS10750	1978	NC_008024	*Homo sapiens*: faringe	Patógeno	http://genomesonline.org/ cgi-bin/GOLD/GOLDCards.cgi? goldstamp=Gc00376
Streptococcus pyogenes MGAS15252	1662	NC_017040	*Homo sapiens*: infección de tejido blando	Patógeno	Fittipaldi *et al*., 2012
Streptococcus pyogenes MGAS1882	1691	NC_017053	*Homo sapiens*	Patógeno	Fittipaldi *et al*., 2012
Streptococcus pyogenes MGAS2096	1898	NC_008023	Homo sapiens: glomerulonefrits poststreptococal	Patógeno	http://genomesonline.org/ cgi-bin/GOLD/GOLDCards.cgi? goldstamp=Gc00377
Streptococcus pyogenes MGAS315	1865	NC_004070	*Homo sapiens*: STSS	Patógeno	Beres *et al*., 2002
Streptococcus pyogenes MGAS5005	1865	NC_007297	*Homo sapiens*	Patógeno	Sumby *et al*., 2005

Tabla 7. Continuación.

Nombre	CDS	No. de acceso	Sitio de asilamiento	Estilo de vida	Referencia
Streptococcus pyogenes MGAS6180	1894	NC_007296	*Homo sapiens*	Patógeno	Green *et al*., 2005
Streptococcus pyogenes MGAS8232	1839	NC_003485	Homo sapiens: garganta: fiebre reumática	Patógeno	Beres *et al*., 2002
Streptococcus pyogenes MGAS9429	1877	NC_008021	*Homo sapiens*: faringe	Patógeno	http://genomesonline.org/cgi-bin/GOLD/GOLDCards.cgi?goldstamp=Gc00379
Streptococcus pyogenes NZ131	1700	NC_011375	*Homo sapiens*: glomerlonefritis	Patógeno	McShan *et al*., 2008
Streptococcus pyogenes SSI-1	1859	NC_004606	*Homo sapiens*: choque tóxico	Patógeno	Nakagawa *et al*., 2003
Streptococcus pyogenes Str. Manfredo	1745	NC_009332	*Homo sapiens*: fiebre reumática	Patógeno	Holden *et al*., 2007
Streptococcus salivarius 57.I	1941	NC_017594	*Homo sapiens*	Patógeno	Geng *et al*., 2011
Streptococcus salivarius CCHSS3	2027	NC_015760	*Homo sapiens*: hemocultivo	Patógeno	Delorme *et al*., 2011
Streptococcus salivarius JIM8777	1979	NC_017595	*Homo sapiens*: cavidad oral sana	No patógeno	Guédon *et al*., 2011
Streptococcus sanguinis SK36	2270	NC_009009	Homo sapiens: placa dental	Patógeno Oportunista	Xu *et al*., 2007
Streptococcus suis 05ZYH33	2186	NC_009442	*Homo sapiens*: STSS	Patógeno	Chen *et al*., 2007
Streptococcus suis 98HAH33	2185	NC_009443	*Homo sapiens*: STSS	Patógeno	Chen *et al*., 2007
Streptococcus suis A7	1974	NC_017622	*Homo sapiens*	Patógeno	Zhang *et al*., 2011
Streptococcus suis BM407 *	1947	NC_012926	*Homo sapiens*: meningitis	Patógeno	Holden *et al*., 2009
Streptococcus suis D12	2078	NC_017621	N/D	Patógeno	Zhang *et al*., 2011
Streptococcus suis D9	2074	NC_017620	N/D	Patógeno	Zhang *et al*., 2011
Streptococcus suis GZ1	1977	NC_017617	*Homo sapiens*	Patógeno	Ye *et al*., 2009
Streptococcus suis JS14	2066	NC_017618	*Homo sapiens*	Patógeno	Holden *et al*., 2009
Streptococcus suis P1/7	1824	NC_012925	*Sus scrofa y Homo sapiens*	Patógeno	Boyle *et al*., 2012
Streptococcus suis S735	1882	NC_018526	*Sus scrofa*	Patógeno	Boyle *et al*., 2012
Streptococcus suis SC070731	1933	NC_020526	N/D	N/D	N/D
Streptococcus suis SC84	1898	NC_012924	*Homo sapiens*: STSS-L	Patógeno	Holden *et al*., 2009
Streptococcus suis SS12	2079	NC_017619	N/D	N/D	Zhang *et al*., 2011
Streptococcus suis ST1	1987	NC_017950	N/D	N/D	Zhang *et al*., 2011
Streptococcus suis ST3	1952	NC_015433	*Sus scrofa*: pneumonía	Patógeno	Hu *et al*., 2011
Streptococcus suis TL13	1939	NC_021213	N/D	N/D	Wang *et al*., 2013
Streptococcus thermophilus CNRZ1066	1915	NC_006449	Yoghurt	No patógeno	Bolotin *et al*., 2004
Streptococcus thermophilus JIM 8232	2145	NC_017581	Leche	No patógeno	Delorme *et al*., 2011
Streptococcus thermophilus LMD-9 *	1715	NC_008532	N/D	N/D	Makarova *et al*., 2006
Streptococcus thermophilus LMG 18311	1888	NC_006448	Yoghurt en Francia	No patógeno	Bolotin *et al*., 2004
Streptococcus thermophilus MN-ZLW-002	1910	NC_017927	Proctos lácteos fermentados en China	No patógeno	Kang *et al*., 2012
Streptococcus thermophilus ND03	1919	NC_017563	Productos lácteos en china	No patógeno	Sun *et al*., 2011
Streptococcus uberis 0140J	1762	NC_012004	Bovino	Patógeno	Ward *et al*., 2009

Tabla 7. Continuación.

Utilizando el mismo enfoque, se buscaron los números de genes compartidos de las especies *S. gallolyticus, S. mutans, S. pneumoniae, S. salivarius, S. thermophilus, S. suis, S. agalactiae, S. dysgalactiae, S. equii* y *S. pyogenes* (ver metodología para los genomas de referencia). Los resultados correspondientes al número de proteínas que conforma a cada genoma núcleo, número de genomas disponibles y el promedio de proteínas que presenta cada especie se resume en la tabla 8. La tabla se organiza del mayor al menor número de proteínas por genoma núcleo, siendo *S. gallolyticus* la especie con el mayor número de genes dentro del genoma núcleo, seguido de *S. salivarius*, *S. mutans*, *S. dysgalactiae*, *S. equii, S. agalactiae*, *S. suis*, *S.thermophilus*, *S. pneumoniae* y *S. pyogenes* en último lugar con el menor número de genes.

# Genomas	Especie	# Proteínas del núcleo	Promedio de proteínas
3	*S. gallolyticus*	1964	2266 ± 56
3	*S. salivarius*	1646	1982 ± 43
4	*S. mutans*	1627	1914 ± 36
4	*S. dysgalactiae*	1491	2061 ± 140
4	*S. equii*	1438	1962 ± 101
9	*S. agalactiae*	1370	1973 ± 207
16	*S. suis*	1334	1999 ± 104
6	*S. thermophilus*	1314	1915 ± 137
24	*S. pneumoniae*	1310	2035 ± 165
19	*S. pyogenes*	1213	1808 ± 111

Tabla 8. Genoma núcleo por especie. Se muestra el número de genomas secuenciados por cada especie, el número de genes que conforman al genoma núcleo y el promedio de genes por genoma por especie.

El número de genes compartidos, por cada especie, fue graficado para conocer la tendencia que presenta el genoma núcleo tras realizar una regresión logarítmica a los datos de la adición secuencial del número de proteínas compartidas por cada genoma añadido (Figura 7). En el eje de las abscisas se ubica cada genoma añadido y en el eje de las ordenadas se encuentran el número de genes codificantes compartidos. En la Figura 7, se observa que a medida que se añaden genomas el número de genes compartidos disminuye, teniendo un comportamiento asintótico cerca de los 400 genes. De igual manera, con los datos resultantes de los genes compartidos en cada especie (ver Figura 8) se pueden observan las líneas de tendencia de la regresión logarítmica de los genes compartidos de cada uno de los genomas núcleo de las especies anteriormente mencionadas. La Figura 8, muestra la extrapolación de los datos, mediante la ecuación resultante de la regresión logarítmica de cada serie de datos (Tabla 9) hasta un máximo de 24 genomas, suponiendo

de esta manera la adición secuencial de más genomas en todos los casos menos para *S. pneumoniae* debido a que es la especie que cuenta con más genomas secuenciados. Las curvas resultantes de los genomas núcleo, muestran comportamientos distintos entre las especies. En las curvas de *S. pneumoniae*, *S. agalactiae* y *S. pyogenes* se aprecia que comienzan a tener un comportamiento asintótico alrededor de los 1325, 1275 y 1175 genes respectivamente, tras 21 genomas añadidos mientras que en las curvas de las demás especies no se observa un comportamiento asintótico a pesar de la extrapolación realizada.

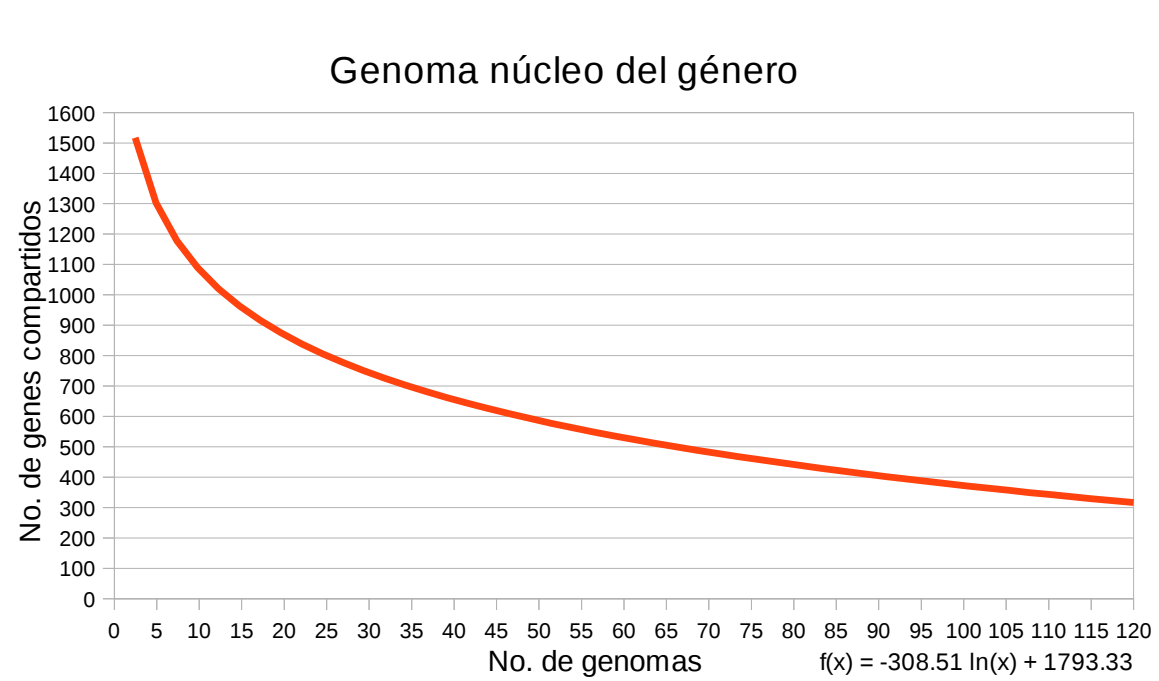

Figura 7. Genoma núcleo del género Streptococcus. En el eje X se muestran el número de genomas añadidos secuencialmente, en el eje Y se muestran el número de genes compartidos, por todas las especies del género.

En cuanto a la caracterización del pangenoma, se encontró que éste cuenta con 212,348 proteínas redundantes totales, calculados como la suma total de CDS presentes en los 108 genomas de *Streptococcus*, de los cuales se agrupan en 33,039 familias de proteínas con un nivel de identidad del 80%. La tabla S.2 que contiene un listado de las familias de proteínas encontradas está disponible en la siguiente dirección URL (http://dx.doi.org/10.6084/m9.figshare.956204). Esta tabla contiene a los representantes de cada familia (descripción y número de acceso), el conteo de presencia-ausencia en cada genoma, los identificadores de las secuencias y la suma del número de aciertos encontrados

para cada familia.

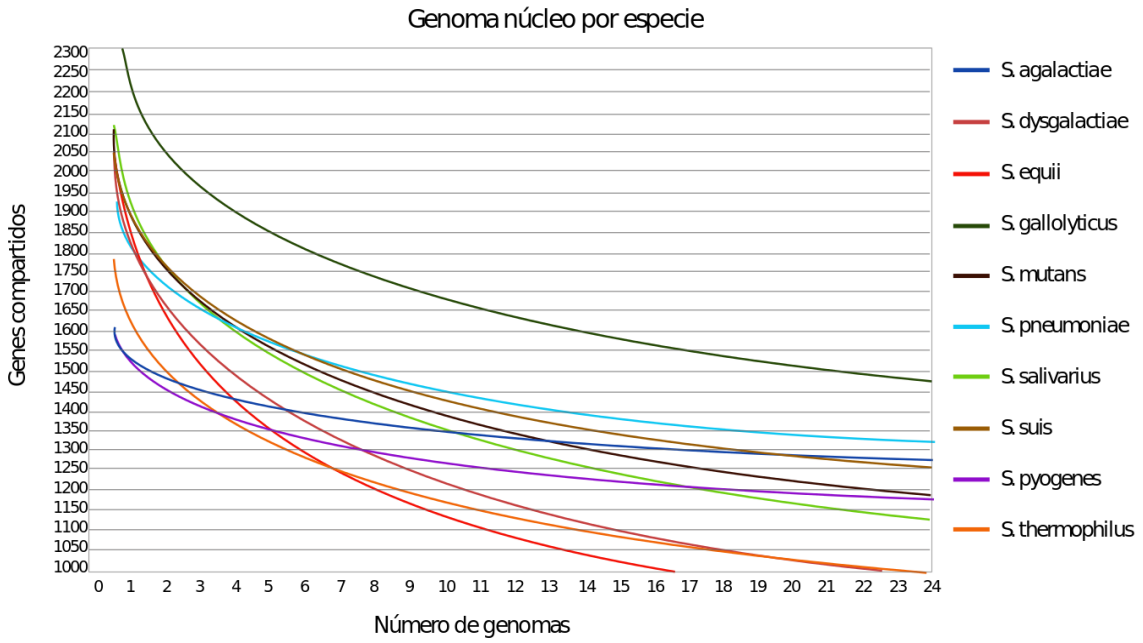

Figura 8. Genoma núcleo por especie de Streptococcus. En el eje X se muestra el número de genomas añadidos secuencialmente, en el eje Y se colocan el número de genes compartidos.

Especie	Ecuación
S. agalactiae	f(x) = -73.6452ln(x) + 1531.0882
S. dysgalactiae	f(x) = -272.0606ln(x) + 1849.1558
S. equi	f(x) = -293.4686ln(x) + 1825.4148
S. gallolyticus	f(x) = -230.0161ln(x) + 2205.3778
S. mutans	f(x) = -232.3956ln(x) + 1928.8914
S. pneumoniae	f(x) = -156.3664ln(x) + 1814.9372
S. pyogenes	f(x) = -111.0286ln(x) + 1529.6765
S. salivarius	f(x) = -252.6959ln(x) + 1936.5901
S. suis	f(x) = -201.2154ln(x) + 1869.5467
S. thermophilus	f(x) = -199.3594ln(x) + 1632.6060

Tabla 9. Ecuaciones de las lineas de tendencia de la figura 8. Se presenta la ecuación resultante de la regresión logarítmica de cada serie de datos.

La Figura 9 muestra una gráfica de la acumulación de genes por cada genoma secuenciado del género *Streptococcus* que se añade. Esta gráfica es una representación del comportamiento del pangenoma, en la que se observa que a pesar de incluir 108 genomas, éste continua incrementando.

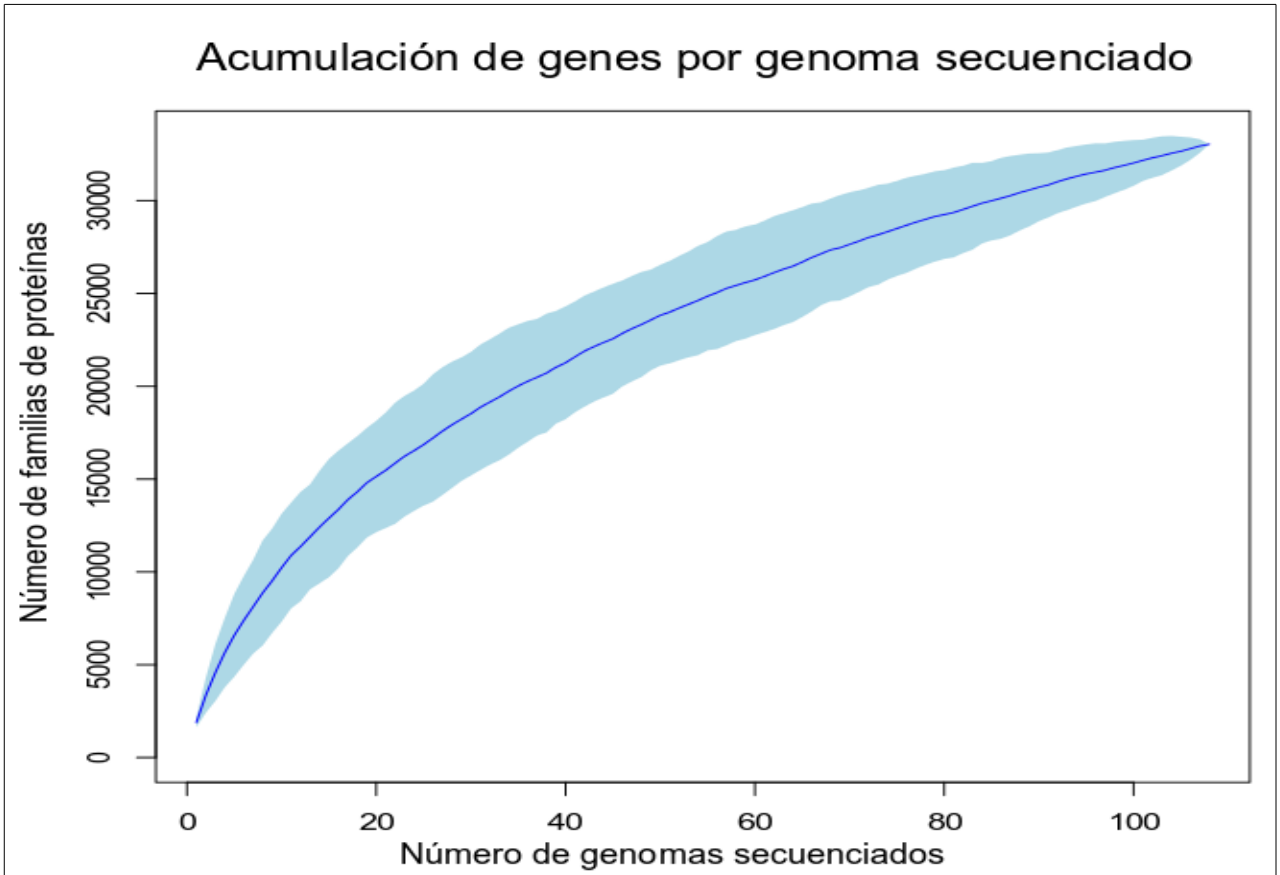

Figura 9. Gráfica de acumulación de genes por genoma añadido. La gráfica muestra el comportamiento del pangenoma del género Streptococcus por cada genoma añadido. En el eje X se encuentra el número de genomas añadidos y el eje Y presenta el número de familias de proteínas.

Composición funcional de los genomas núcleo y pangenoma

Para comprender el rol funcional de los genes presentes en el pangenoma, genoma núcleo del género y genomas núcleo de especies selectas, se llevó a cabo la anotación de las secuencias que conforman a éstos, utilizando una clasificación jerárquica por subsistemas y una clasificación funcional mediante los COGs (Tatusov et al., 2003) (Figuras 10-A y 10-B). La primera es una clasificación en la cual se agrupan secuencias que cumplen roles funcionales similares y que juntas representan procesos biológicos específicos (e.g. rutas metabólicas), denominados subsistemas (Aziz et al., 2008) y la segunda es una clasificación basada en las relaciones ortólogas entre genes (Tatusov *et al*., 2003). Las jerarquías y funciones que resultan tras la anotación se resumen en las gráficas (*heatmaps*)

que se presentan en las figura 10, donde cada rectángulo representa la abundancia de genes encada uno de las categorías de los subsistemas o COG (renglones), normalizada contra el total de secuencias en cada muestra (columnas), resultando entonces en abundancias relativas. La escala en la parte superior derecha de los gráficos tiene un código de color en gradiente el cual representa estas abundancias relativas, representadas como valores Z (normalizados); en amarillo las menos representadas, en azul claro las medianamente representadas y en azul obscuro las funciones con mayor representación. Los gráficos agrupan a los subsistemas o categorías funcionales, así como a los distintos genomas núcleo, mediante dendrogramas en los ejes horizontal superior y vertical izquierdo

En la Figura 10-A se comparan los genomas núcleo del género (primera columna) y las especies: *S. thermophilus, S. salivarius, S. mutans, S. gallolyticus, S. pyogenes, S. suis, S. pneumoniae, S. agalactiae, S. equii* y *S. dysgalactiae* en la clasificación por subsistemas. En el eje horizontal se observa la formación de tres grupos; en el primero se coloca solo el genoma núcleo del género, en el segundo, se agrupan las espeices *S. thermophilus* y *S. salivarius* y un tercer grupo mas grande se forma conteniendo a *S. mutans, S. gallolyticus, S. pyogenes, S. suis, S. pneumoniae, S. agalactiae, S. equii* y *S. dysgalactiae* que a su vez se divide en dos grupos, uno con dos especies (*S. mutans* y *S. gallolyticus*) y un segundo con 6 especies (*S. pyogenes, S. suis, S. pneumoniae, S. agalactiae, S. equii* y *S. dysgalactiae).* En el eje vertical el dendrograma resultante agrupa a las funciones de los Subsistemas de acuerdo a su abundancia, mostrándose en la parte inferior las más representadas (Z-*score* ≥*1.5*) , en la parte media las menos representadas (valor Z ≤ 0.15) y en la parte superior las medianamente representadas (z-score = 0.15 ≥ valor Z ≤ 1.5). De estas funciones se observa que las que cuentan con una mayor representación son la categoría 5 (subsistemas basados en clustering), siendo la más abundante de todas las categorías en todos los genomas núcleo (promedio = 3.4); la categoría 2 (carbohidratos), que en el caso del genoma núcleo del género, *S. thermophilus* y *S. salivarius* tienen una representación media (Z-score ≤ 1.5) en relación a los demás genomas núcleo, lo cual concuerda con la distancia entre estos grupos; la categoría 21 (metabolismo de proteínas), en la que se observa una mayor representación en el genoma núcleo del género (Z-score = 2.02) y *S. thermophilus (Z-score = 1.70)* al compararse con las demás especies y la categoría 13 (misceláneos; promedio =

1.70) se encuentra medianamente representada en todos los casos.

Las sección media baja del gráfico, resulta poco informativa debido a que las categorías 20 (metabolismo de potasio; promedio = -0.76), 15 (metabolismo de nitrógeno; promedio = -0.77), 12 (metabolismo de compuestos aromáticos; promedio = -0.75), 27 (metabolismo de azufre; -0.74), 14 (movilidad y quimiotaxis; promedio = -0.79), 8 (dormancia y esporulación; promedio = -0.80), 25 (metabolismo secundario; promedio = -0.81) y 19 (fotosíntesis; promedio = -0.82) presentan la menor abundancia en todos los casos, formando un gran grupo. En la sección media superior se observa que las categorías 18 (metabolismo de fósforo; promedio = -0.65), 10 (adquisición y metabolismo de hierro; promedio = -0.62), 24 (respiración; promedio = -0.60), 17 (plásmidos; promedio = -0.55), 23 (regulación y señalización celular; promedio = -0.33) y 3 (división celular y ciclo celular; promedio = -0.36) cuentan con baja abundancia, destacando que en esta última categoría hay una mayor representación en el genoma núcleo (valor Z = -0.20) del género en comparación con las especies (promedio = -0.38). Es interesante notar la diferencia de abundancia en la categoría 28 (virulencia, enfermedad y defensa) entre las especies *S. salivarius, S. gallolyticus, S. pyogenes, S.pneumoniae, S. agalactiae, S. equi* y *S. dysgalactiae* (promedio = -0.05) *y* los genomas núcleo del género y las especies *S. thermophilus, S. mutans* y *S. suis* (promedio = -0.30); donde los últimos se encuentran menos representados. Las categorías 9 (ácidos grasos, lípidos e isoprenoides) y 26 (respuesta al estrés) muestran una mayor abundancia en los genomas núcleo del género, *S. dysgalactiae* y *S. suis*.

En la sección superior del gráfico se observan funciones medianamente representadas. En el caso de la categoría 22 (metabolismo de RNA; promedio = 0.48) el genoma núcleo del género (valor Z= 0.79) muestra una mayor abundancia en comparación con los de las especies, situación que resulta inversa en las categorías 1 (aminoácido y derivados; promedio = 0.52; género = -0.5) y 6 (cofactores, vitaminas, grupos prostéticos y pigmentos; promedio = 0.41; género = 0.22). La categoría 4 (pared celular y cápsula; promedio = 0.24) presenta una abundancia relativamente homogénea en todos los casos. Para las categorías 11 (transporte de membrana; promedio = 0.12; *S. mutans*= 0.29), 16 (nucleótidos y nucleósidos; promedio = 0.09) y 7 (metabolismo de DNA; promedio = 0.15) hay una

abundancia baja en todos los casos.

La Figura 10-B muestra la misma comparación de la Figura 10-A, pero en este caso la clasificación de las funciones se ha hecho con base en COGs. De nuevo el genoma núcleo del género forma un grupo separado a los genomas núcleo de las especies en la primera columna; en la segunda y tercera columna se agrupan *S. gallolyticus* y *S. mutans*; en la cuarta y quinta de nuevo se agrupan las especies *S. thermophilus* y *S. salivarius* y de la sexta a la doceava columna se forma un grupo grande que contiene a las demás especies, que a su vez se divide en dos más pequeños conteniendo de derecha a izquierda a *S. pyogenes, S. pneumoniae* y *S. equi* en un grupo y a *S. dysgalactiae, S. suis* y *S. agalactiae* en otro.

El dendrograma en el eje vertical agrupa a las categorías de la clasificación por COGs, mostrándose las más representadas en la parte inferior (valor Z ≥ 1.0) , las menos representadas (valor Z ≤ -0.2) en un grupo grande en la parte media y las medianamente representadas (-0.2 ≤ valor Z ≤ 1.0) en un tercer grupo en la parte superior. En cuanto a la abundancia de funciones mediante esta clasificación, se encontró que la categoría J (traducción y estructura ribosomal; promedio especies = 1.51; género = 3.08) se encuentra altamente representada en los genomas núcleo de todas las especies, pero se puede observar que en el genoma núcleo del género se tiene la mayor abundancia de esta categoría. Entre las categorías que se encuentran altamente representadas están la categoría R (función desconocida; promedio = 1.94), S (funciones generales; 2.40) y E (transporte y metabolismo de aminoácidos; promedio especies = 0.95; género = -0.16).

Entre las categorías que se encuentran medianamente representadas, encontramos a la categoría K (transcripción; promedio = 0.41), la cual se observa con una abundancia uniforme en todos los casos menos en *S. thermophilus* (valor Z = 0.02); la categoría G (transporte y metabolismo de carbohidratos; promedio = 0.80), en la cual el genoma núcleo del género (valor Z = 0.09), junto con el de *S. thermophilus* (valor Z= -0.16) presentan la menor abundancia; P (transporte y metabolismo de iones inorgánicos; promedio = 0.05), la cual se observa pobremente representada en el caso de *S. suis* (valor Z = -0.20); M

(biogénesis de membrana, envoltura y pared celular; promedio = 0.03), la cual es más abundante en el genoma núcleo del género (valor Z = 0.22); F (metabolismo y transporte de nucleótidos; promedio = -0.16), que resulta con baja abundancia para *S. mutans* (valor Z = -0.31), *S. gallolyticus* (valor Z = -0.28), *S. equi* (valor Z = -0.23) y *S. dysgalactiae* (valor Z = -0.25) y L (replicación, recombinación y reparación; promedio = -0.05; género = 0.42), la que presenta mayor abundancia en el núcleo del género.

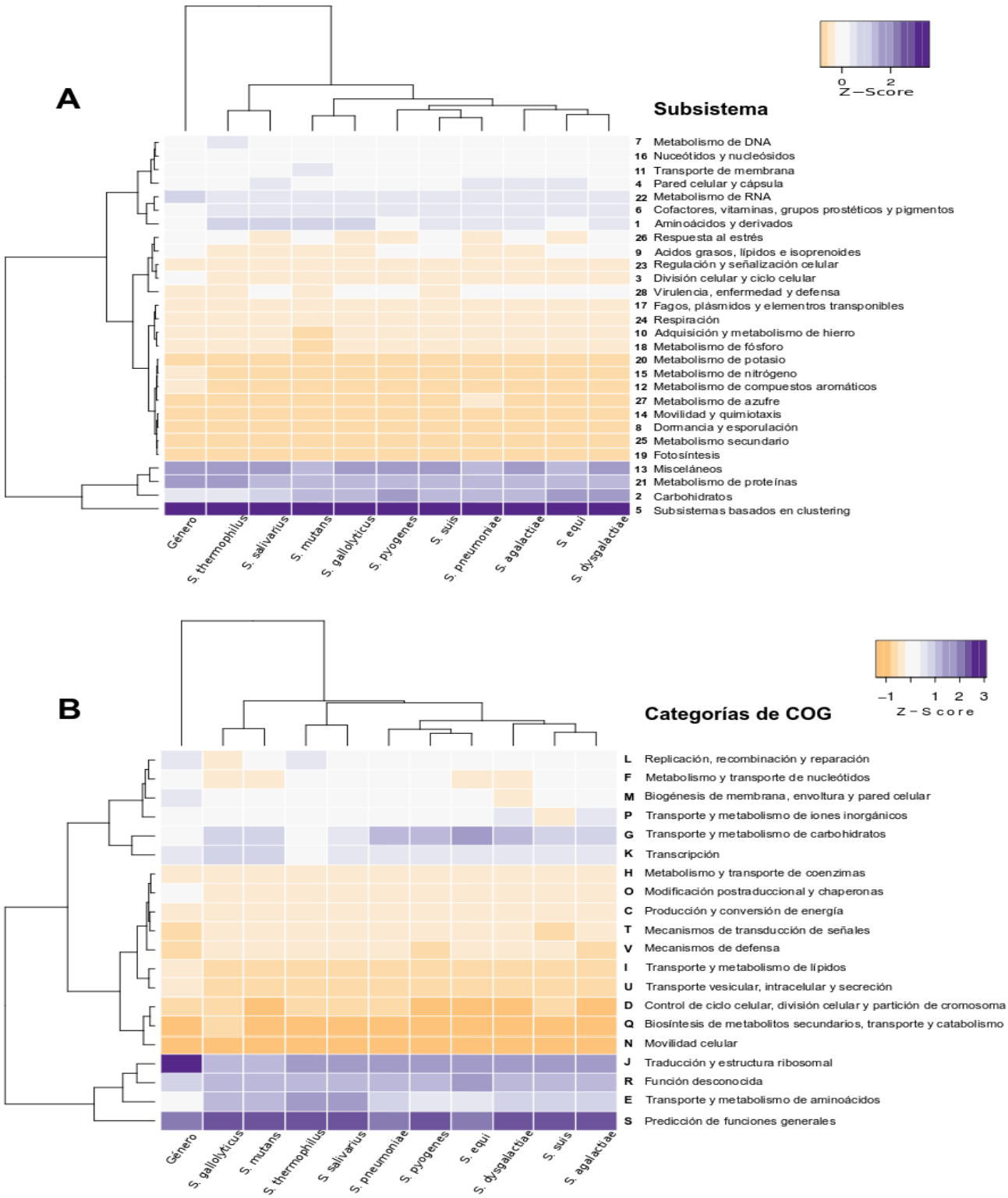

Figura 10. Heatmaps de la comparación de funciones entre los genomas núcleo de las especies del género Streptococcus. A) Clasificación jerárquica por subsistmas. B) Clasificación jerárquica por COGs.

Para conocer las diferencias funcionales que pueden existir entre el número total de

genes presentes en los streptococci y el conjunto de genes compartidos por todos, cuyo número es solo una pequeña fracción del primero, se realizó la comparación de la representación de las funciones entre el genoma núcleo del género y el pangenoma mediante la clasificación por subsistemas (Figura 11-A). En este gráfico se agrupan las funciones más representadas (valor Z ≥ 1.7) en la parte inferior, las medianamente representadas (0 ≤ valor Z ≤ 1.7) en la sección media y las poco representadas (valor Z ≤ 0) en la parte superior como se puede apreciar en el dendrograma del eje vertical. Entre las más abundantes se encuentran la categoría 5, que se observa igualmente representada para ambos casos; la categoría 21, más abundante en el genoma núcleo, de igual manera que la categoría 13 y la categoría 2, que se observa con mayor abundancia en el pangenoma. Entre las categorías parcialmente representadas encontramos que las categorías 1, 6, 4, 17 y 28 son más abundantes en el pangenoma; la categoría 22 (metabolismo de RNA) es más abundante en el genoma núcleo y las categorías 11 (transporte de membrana) y 16 (nucleótidos y nucleósidos) cuentan con la misma abundancia en ambos casos.

Las categorías con baja abundancia (valor Z ≤ 0) son las categorías 19, 25, 8, 14, 15, 27, 20, 12, 18, 24, 10, 3, 23, 9 y 26. De éstas se aprecian diferencias, donde el genoma núcleo presenta mayor abundancia, en la categoría 3 (división y ciclo celular), 9 (ácidos grasos, lípidos e isoprenoides) y 26 (respuesta al estrés).

La misma comparación entre pangenoma y genoma núcleo del género fue realizada utilizando la clasificación por COGs (figura 11-B). En este caso las categorías más representadas (valor Z ≥ 0.65) se encuentran agrupadas en la parte superior del gráfico, las parcialmente representadas en la parte media (-0.32 ≤ valor Z ≤ 0.65) y las pobremente representadas (valor Z ≤ -0.32) en la parte inferior. Entre las más representadas se observa una gran diferencia en la categoría J (biogenésis, estructura ribosomal y traducción), la cual se encuentra medianamente representada en el pangenoma (valor Z = 0.88) y altamente representada en el genoma núcleo (valor Z = 3.08); las categorías E (transporte y metabolismo de aminoácidos; núcleo = -0.16; pangenoma = 0.81) y G (transporte y metabolismo de carbohidratos; núcleo = 0.09; pangenoma = 1.01) se muestran medianamente representadas en el genoma núcleo, mientras que en el pangenoma se

encuentra en mayor abundancia. En las categorías L, K y R se aprecia una representación media en ambos casos sin presentar mayor diferencia entre el pangenoma y el genoma núcleo; la categoría S se encuentra ligeramente más representada en el pangenoma.

Las categorías parcialmente representadas (-0.32 ≤ valor Z ≤ 0.65) de COGs son las categorías C, T, V, H, O, F, M y P; de las cuales T (transducción de señales; valor Z= -0.68) y V (mecanismos de defensa; valor Z= -0.87) están pobremente representadas en el genoma núcleo y la categoría M (biogénesis de membrana, envoltura y pared celular; valor Z núcleo= 0.22; valor Z pangenoma= 0.10) se encuentra más representada en este mismo. En las categorías H, O F, y P no se observan diferencias.

De las categorías menos representadas en la parte inferior del gráfico podemos observar que las categorías I (transporte y metabolismo de lípidos; valor Z núcleo= -0.22; valor Z pangenoma= -0.81) y U (transporte vesicular, intracelular y secreción; valor Z núcleo=-0.55; valor Z pangenoma= -0.85) presentan una abundancia baja en ambos casos pero más representada en el genoma núcleo. Las demás categorías en esta sección son la D, Q y N; de éstas la categoría D (control de ciclo celular, división celular y partición de cromosoma; valor Z núcleo= -0.87; valor Z pangenoma=-1.06) se encuentra más representada en el genoma núcleo, Q (metabolitos secundarios) y N (movilidad celular) son las menos abundantes para ambos casos sin que se aprecien diferencias entre pangenoma y genoma núcleo (valores Z < 1.10).

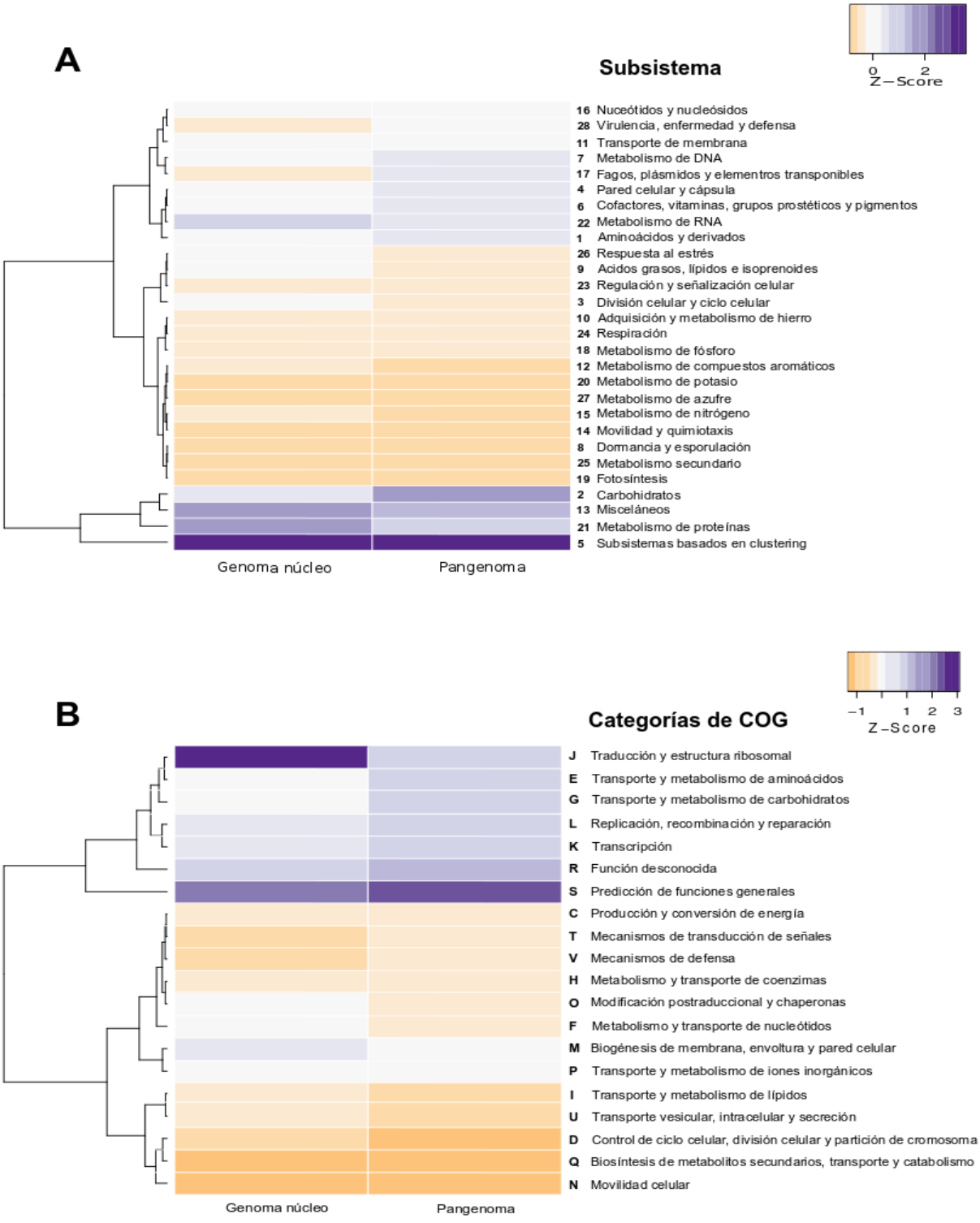

Figura 11. Heat maps de la comparación de funciones entre el pangenoma y genoma núcleo del género Streptococcus. A. Clasificación jerárquica por subsistmas. **B.** Clasificación jerárquica por COGs.

Reconstrucción filogenética

Para conocer la estructura filogenética de los microorganismos utilizados en este estudio se alinearon las secuencias de RNA ribosomal 16S para generar un árbol por el método de Neighbor-Joining (figura 12), el cual fue enraizado utilizando como grupos hermanos a *B. subtilis* y *B. licheniformis*. El gen de rRNA 16S se encuentra, típicamente, en múltiples copias dentro de los organismos, variando desde 1 hasta 15 (Kembel et al., 2012), debido a esto se utilizó a una secuencia por genoma, utilizando a las que tuvieran longitudes similares para generar el alineamiento múltiple utilizado para graficar el árbol filogenético.

Se formaron cuatro grupos principales tomando los nodos internos, en azul se agrupan las especies *S. pyogenes, S. dysgalactiae, S. agalactiae, S. equii, S. uberis, S. parauberis* y *S. iniae* (bootstrap=82), en café se agrupa solamente *S. suis* (bootstrap=99), que de la misma manera no ha sido clasificado en ningún grupo en el estudio anteriormente citado; en verde claro se agrupan *S. thermophilus* y *S. salivarius (bootstrap*=99), en verde obscuro encontramos a *S. mutans, S. infantarius, S. lutetiensis, S. macedonicus y S. gallolyticus* (bootstrap=38). En el caso de esta agrupación, el soporte de *bootstrap* es muy bajo y la distancia entre S. mutans y las especies del sub árbol verde obscuro es grande, por lo que se puede considerar a S. mutans como un grupo separado. Finalmente en rojo, se agrupan *S. pneumoniae, S. mitis, S. oralis, S. pasteurianus, S. parasanguinis, S. sanguinis, S. gordonii, S. oligofermentans y S. intermedius* (bootstrap=44).

Al observar los nodos más internos en la filogenia por rRNA 16S, se obtienen valores de soporte de bootsrap más altos (>90), agrupando solamente a las especies en los casos de: *S. pyogenes, S. dysgalactiae, S. agalactiae y S. equii, S. mutans y S. pneumoniae.* En el caso de *S. thermophilus y S. salivarius,* ambas especies forman un solo grupo, al igual que *S. infantarius, S. lutetiensis, S. macedonicus y S. gallolyticus.* Las especies *S. mitis, S. oralis, S. pasteurianus, S. parasanguinis, S. sanguinis, S. gordonii, S. oligofermentans y S. intermedius* consolidan un grupo en el cual cada especie resulta distante y quedan agrupados junto con *S. pneumoniae* en el nodo más interno pero observándose en el nodo más externo una clara separación de esta última especie.

Posteriormente se generó una matriz con la puntuación de GSS entre los ortólogos compartidos por pares para medir la distancia entre las especies. Esta matriz se graficó como un árbol por el método de Neighbor-Joining (Figura 13) para realizar la evaluación de la distancia genómica entre todas las especies; esta estrategia puede utilizarse como una herramienta complementaria para aclarar las relaciones entre organismos utilizando secuencias completas de ortólogos compartidos por pares (Alcaraz et al., 2010). En este árbol se aprecia la formación de los siguientes grupos: Tomando los nodos internos se sigue observando la agrupación de *S. pyogenes, S. dysgalactiae* y *S. equi* (azul obscuro), en este caso *S. agalactiae* forma un grupo más alejado (azul claro); *S. thermophilus* y *S. salivarius* continúan agrupándose juntos (verde claro); *S. mutans* y *S. gallolyticus* también agrupados junto con *S. infantarius, S. macedonicus, S. lutetiensis* y *S. macedonicus* (verde obscuro); *S. suis* sigue mostrándose como un solo grupo (amarillo) y *S. pneumoniae* se agrupa de nuevo con *S. mitis, S. oralis, S. pasteurianus, S. parasanguinis, S. sanguinis, S. gordonii, S. oligofermentans y S. intermedius.* En este árbol se observa que *S. infantarius* presenta una distancia mucho mayor con respecto a la que se tiene en el árbol de rRNA 16S.

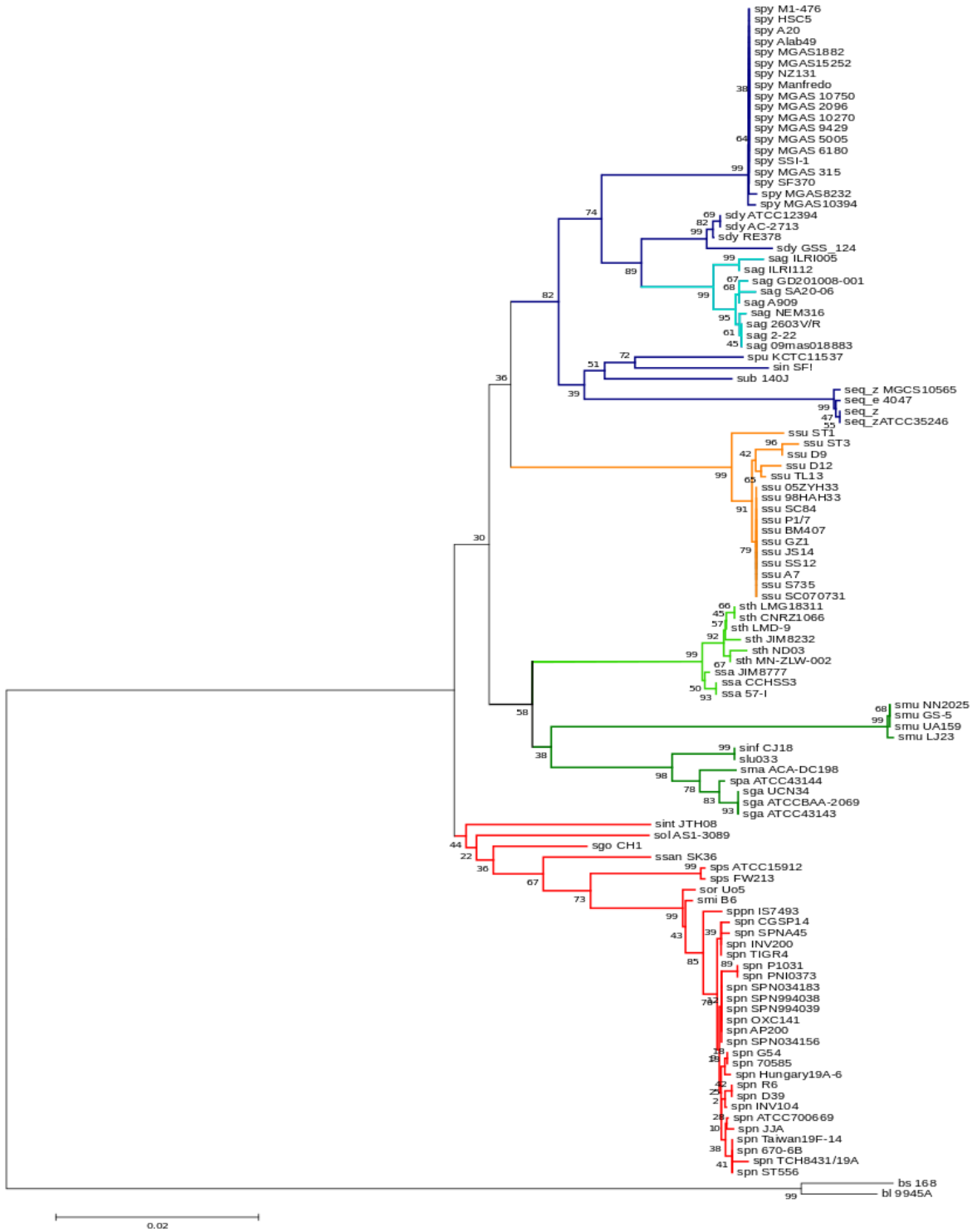

Figura 12. Reconstrucción filogenética de Streptococcus. Árbol filogenético generado por el método de Neighbor-Joining utilizando secuencias de rRNA 16S en MEGA5. *(**spy**=S. pyogenes, **sdy**=S.dysgalactiae, **sag**=S. agalactiae, **spu**=S. parauberis, **sin**=S. iniae, **sub**=S.uberis, **seq_z**=S.equi subsp. zooepidemicus, **seq_z**=S.equi subsp. equi, **ssu**=S. suis, **sth**=S.thermophilus, **ssa**=S.salivarius, **smu**=S. mutans, **sint**=S. intermedius, **sol**=S. oligofermentans, **ssan**=S. sanguinis, **sgo**=S. gordonii, **sps**=S. parasanguinis, **spas**=S. pasteurianus, **sor**=S. oralis, **spn**=S. pneumoniae, **sppn**=S. pseudopneumoniae, **smi**=S. mitis, **sga**=S. gallolyticus, **sma**=S. macedonicus, **slu**=S. lutetiensis, **sinf**=S. infantarius, **bs**=B. subtilis, **bl**=B. licheniformis.)*

Figura 13. Matriz de puntuación de similitud genómica (GSS) graficada como un árbol por el método de Neighbor-Joining en MEGA 5. *(**spy**=S. pyogenes, **sdy**=S.dysgalactiae, **sag**=S. agalactiae, **spu**=S. parauberis, **sin**=S. iniae, **sub**=S.uberis, **seq_z**=S.equi subsp. zooepidemicus, **seq_z**=S.equi subsp. equi, **ssu**=S. suis, **sth**=S.thermophilus, **ssa**=S.salivarius, **smu**=S. mutans, **sint**=S. intermedius, **sol**=S. oligofermentans, **ssan**=S. sanguinis, **sgo**=S. gordonii, **sps**=S. parasanguinis, **spas**=S. pasteurianus, **sor**=S. oralis, **spn**=S. pneumoniae, **sppn**=S. pseudopneumoniae, **smi**=S. mitis, **sga**=S. gallolyticus, **sma**=S. macedonicus, **slu**=S. lutetiensis, **sinf**=S. infantarius, **bs**=B. subtilis, **bl**=B. licheniformis.)*

Análisis en metagenomas selectos.

Se analizaron los metagenomas orales humanos disponibles en el servidor MG-RAST para determinar las especies del género *Streptococcus* que se encuentran presentes en cada uno de éstos, así como la abundancia global que presenta el género. La gráfica de la figura 14 muestra en las columnas cada uno de los metagenomas, comenzando con los de los individuos con caries y en las últimas dos columnas los individuos sanos; en el eje vertical se muestra el porcentaje de la abundancia. Se encontró que la abundancia de los streptococci con respecto al total de OTUs es mayor en los metagenomas de los individuos sanos (NOCA_01P y NOCA_03P) con una abundancia del 20.1% y 20.8%, mientras que en los individuos enfermos con caries (Figura 14) va desde el 4.2% hasta el 10.6%.

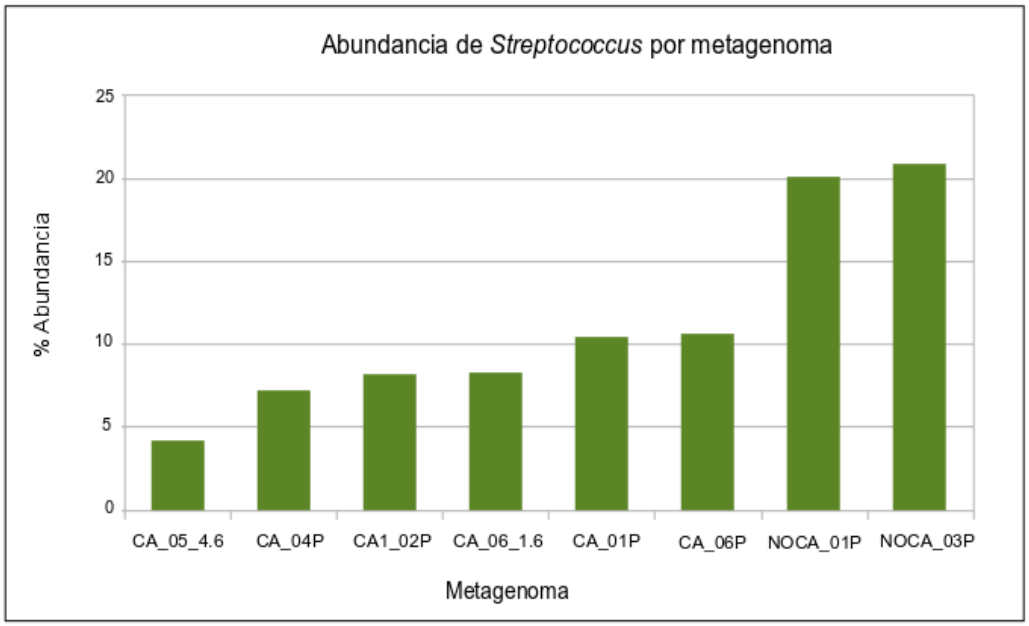

Figura 14. Abundancia de Streptococcus en metagenomas orales humanos. La gráfica presenta el porcentaje de los streptococci con respecto al total de organismos presentes en cada metagenoma.

Posteriormente se analizó la composición de las especies de *Streptococcus* por cada uno de estos metagenomas, encontrándose que ésta es relativamente constante. En los ocho metagenomas se encontró que la especie predominante de *Streptococcus* es *S. pneumoniae* (30.5% - 18.3%) y en orden decreciente, siguiéndole, las especies *S. mitis*, *S. gordonii*, *S.*

sanguinis y *S.oralis* (Figura 15). La abundancia de *S. mutans,* considerado el principal agente etiológico de las caries dentro del género representa, en los metagenomas con esta afección, entre un 1.2% a 3.0% y en los metagenomas sin caries menos del 1.1%, y no se encuentra entre las especies más representadas.

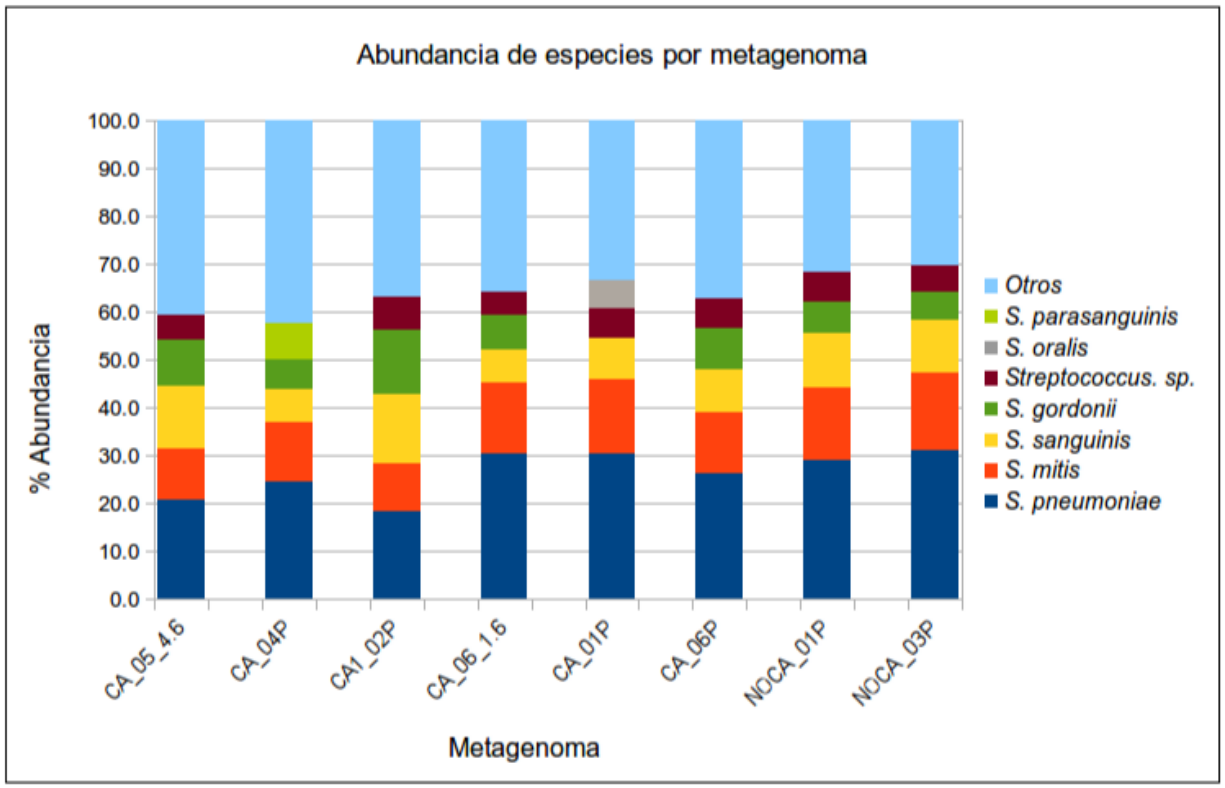

Figura 15. Abundancia de especies de Streptococcus por metagenoma. Se presenta un gráfico de barras donde se grafican por cada metagenoma las 5 especies más abundantes.

Funciones de los streptococci presentes en los metagenomas orales

Para tener una visión de las funciones que llevan a cabo los streptococci dentro de los metagenomas orales humanos, se hizo la comparación con un gráfico *heat map* de éstas en cada metagenoma mediante la clasificación jerárquica por subsistemas (Figura 16-A) y COG (Figura 16-B) donde cada columna representa a un metagenoma y cada renglón es una categoría, ya sea de subsistemas o de COG. Se presentan dendrogramas en cada *heatmap* los cuales indican la distancia que existe entre cada metagenoma (eje horizontal) y entre las categorías de funciones (eje vertical). De los metagenomas analizados se cuentan con los

siguientes metadatos. Los individuos NOCA_03P y NOCA_01P nunca han tenido caries, los individuos CA1_01P y CA1_02P han presentado caries y han sido tratados; los individuos CA_04P y CA_06P tienen caries activas y por último los individuos CA_05_4.6 y CA_06_1.6 son lo que cuentan con el peor estado de salud dental y presentan cavidades. (Belda-Ferre *et al.*, 2012).

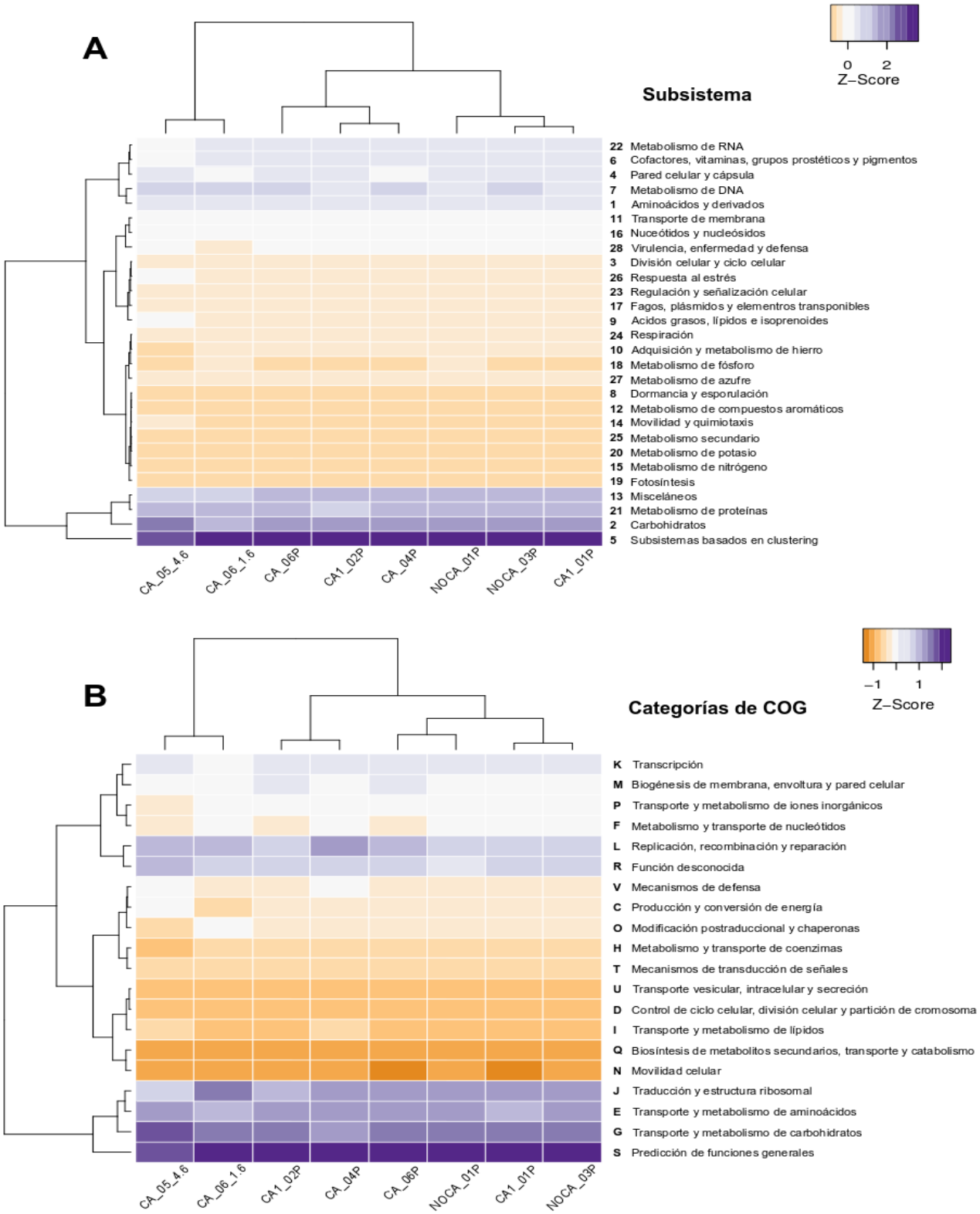

Figura 16. Heatmaps de la comparación de funciones de los streptococci dentro de los metagenomas orales humanos. A) Clasificación jerárquica por subsistemas. B) Clasificación jerárquica por COGs.

Dentro de la clasificación por subsistemas (Figura 16-A) se observan agrupadas a las funciones más representadas (valor Z >1.1) en la parte inferior, las medianamente representadas (0.20 < valor Z < 1.1) en la parte superior y las menos representadas (valor Z < 0.20) en la sección media.

Los metagenomas CA_05_4.6 y CA_06_1.6 forman un grupo más lejano a los demás metagenomas y los metagenomas de los individuos que nunca han tenido caries forman un grupo junto con un metagenoma de un individuo que ha sido tratado. En particular el metagenoma CA_05_4.6 muestra la mayor cantidad de diferencias en relación a los demás metagenomas. Estas diferencias radican en las categorías 2 (carbohidratos; valor Z = 2.40; promedio = 1.93) y 26 (respuesta a estrés; valor Z = -0.19; promedio = -0.36), las cuales están más representadas, y las categorías 22 (metabolismo de RNA; valor Z = 0.00; promedio = 0.36) y 6 (cofactores y vitaminas; valor Z = 0.23; promedio = 0.31) que muestran menos abundancia con respecto a los demás metagenomas. El metagenoma CA_06_1.6 presenta diferencias en la categoría 4 (pared celular y cápsula; valor Z = 0.16; promedio = 0.32) que se observa menos representada.

Con la clasificación por COG (Figura 16-B) se observan agrupadas a las funciones más representadas (valor Z >0.72) en la parte inferior, las medianamente representadas (-0.20 < valor Z < 0.72) en la parte superior y las menos representadas (valor Z < -0.20) en la sección media.

Los metagenomas CA_05_4.6 y CA_06_1.6 de nuevo forman un grupo lejano a los demás metagenomas y los de individuos que nunca han tenido caries se encuentran en dos grupos distintos pero cercanos. El metagenoma CA_05_4.6 presenta menos representadas las categoría J (traducción y estructura ribosomal; valor Z = 0.74; promedio = 1.20), P (transporte de iones inorgánicos; valor Z = -0.31; promedio = -0.04) y F (metabolismo y transporte de nucleótidos; valor Z = -0.36; promedio = -0.15); y sobre representadas las categorías V (mecanismos de defensa; valor Z=-0.07; promedio = -00.27) y C (producción y conversión de energía; valor Z = 0.05; promedio = -0.31). En el metagenoma CA_06_1.6 se aprecia a la categoría O (modificación postraduccional y chaperonas; valor Z = -0.12; promedio = -0.34)

más representada en comparación con los demás metagenomas pero sin estar en la categoría de media representación.

Discusión de resultados

Análisis del genoma núcleo

Tras los análisis de genómica comparativa, realizados en el genoma núcleo del género Streptococcus encontramos que el genoma núcleo se compone de 405 genes, lo cual se contrapone con lo anteriormente descrito tanto por Lefébure & Stanhope (2007), como por Van der Bogert, *et al.* (2013) quienes estipulan que el tamaño del genoma núcleo de este género es de 611 genes y 547 genes respectivamente. La diferencia en el número de genes, es debida a que en el estudio del 2007, los autores utilizaron 26 genomas en el análisis; y en el estudio de 2013 utilizaron 64 genomas, mientras que en el presente estudio se analizaron 108 genomas. El genoma núcleo disminuye, su número de genes, a medida que se incorporan genomas en el análisis, ya que la cantidad de genes compartidos entre todas los genomas analizados disminuye; ya que éste es un pequeño subconjunto del pangenoma en el cual se encuentran los genes más conservados, cómo los involucrados en replicación, transcripción y traducción, que a su vez se encuentran bajo una gran presión selectiva para prevenir cambios drásticos y perder funciones (Lapierre & Gogarten, 2009). El genoma núcleo, representa también casi una quinta parte del promedio de genes calculado para el género, lo cual refleja el número tan pequeño de genes que se comparten dentro de un género bacteriano, pensando que en el caso de las diferencias en regiones codificantes entre chimpancés y humanos no excede el 1.23% (The Chimpanzee Analysis Consortium, 2005).

También existe un punto de comparación en el trabajo de Lefébure *et al.,* (2007) para los genomas núcleo de las especies *S. agalactiae* (1472 genes), *S. pyogenes* (1376 genes) y *S. thermophilus* (1487 genes). En este estudio se encontró que estas especies cuentan con 1370, 1213 y 1314 genes en sus genomas núcleo, respectivamente. Nuevamente, esta discrepancia se origina debido al tamaño muestral, donde Lefébure y colaboradores utilizan

solamente 3 genomas para la construcción del genoma núcleo de cada especie y en este estudio se utilizó un número mayor de genomas en cada caso (ver Tabla 8). En las especies restantes analizadas encontramos que *S. gallolyticus* cuenta con el mayor número de proteínas

Dadas las comparaciones mencionadas, podemos observar que las líneas de tendencia de los genomas núcleo de las especies *S. pneumoniae*, *S. agalactiae* y *S. pyogenes* (ver Figura 8) son asintóticas. Estas curvas fueron extrapoladas hasta la adición de 24 genomas y este comportamiento indica que al continuar añadiendo genomas, el número de genes compartidos se mantendrá relativamente constante, aunque siempre existe la posibilidad de que se aumente el tamaño muestral y este resultado pueda volver a ser refutado como esta sucediendo con el autor del 2007.

En cuanto al pangenoma podemos observar en la Figura 9 que éste, a pesar de la adición de más genomas sigue comportándose como un pangenoma abierto, lo cual es esperado ya que este género tiene es propenso a la transferencia de material genético y esta propiedad hace que el número de genes que se encuentran dentro de éste incremente constantemente (Medini et al., 2005).

Análisis filogenético

Al comparar los árboles generados con las secuencias de rRNA 16S y por medio del GSS (Figura 17) podemos notar que con la estrategia del GSS (Figura 17-A), que utiliza la distancia genómica entre los organismos, se observa una mayor distancia y discriminación entre las cepas en los grupos de cada especie, que en el árbol de rRNA 16S (Figura 17-B), no se obtiene la mínima distancia genética, debido al alto grado de conservación que este gen, teniendo entre un 96% y 99% de identidad entre las especies de los streptococci (Kawamura et al., 1995). Se ha estipulado que el umbral de corte parra determinar a una especie es del 97% de identidad en la secuencia de rRNA 16S y valores mas bajos hablan de distintas especies (Alcaraz, 2013). Aún así, el agrupamiento de los taxones utilizados en este estudio ha sido más o menos similar en ambos árboles (ver Figuras 12 y 13), formándose en

ambos casos los grupos descritos con anterioridad por Kawamura *et al.* (1995), los cuales se describen a continuación:

El grupo *pyogenes* (subárbol en azul obscuro y claro), en el cual Kawamura y colaboradores en 1995, han ubicado las especies *S. canis*, *S. hyointestinalis* y *S. porcinus*, las cuales no han sido incluidas en este estudio, por la falta de genomas secuenciados. Dentro de este grupo, en ambos árboles, se separa con una mayor distancia a *S. agalactiae*; en el caso del árbol de GSS se aprecia una distancia mucho mayor con respecto a los demás integrantes del grupo pyogenes, indicando que *S. agalactiae* podría ser considerado, a nivel genómico, como un grupo hermano al pyogenes.

El grupo *mitis* (subárbol rojo), en donde se observan a todos los integrantes propuestos con anterioridad por Kawamura *et al*. y a *S. intermedius*, que pertenece al grupo “anginosus” y se observa en el árbol de rRNA 16S como la especie más lejana dentro del subárbol rojo, lo cual es esperado debido a que la distancia entre los grupos mitis y anginosus es la más cercana y *S. intermedius* fue el único integrante de este grupo analizado en este estudio. Cabe mencionar que este subárbol cuenta con un valor de *bootstrap* muy bajo (44), lo que indica que la probabilidad de que la topología sea de la manera en la que se reporta no sea certera. Al tomar el enfoque de la distancia genómica (árbol de GSS), el acomodo de esta especie difiere y se observa que es más cercana a las especies *S. gordonii*, *S. oligofermentans* y *S. sanguinis* lo que sugiere que hay un mayor grado de parentesco con el grupo “mitis” a nivel genómico que a nivel de un solo gen.

El grupo *salivarius* (subárbol verde claro) se mantiene conteniendo a las especies no patógenas *S. thermophilus*, el cual ha tenido un amplio uso industrial en la producción de derivados lácteos (Bolotin et al., 2004) y a *S. salivarius* del cual se ha propuesto su uso como probiótico, ya que es productor de bacteriocinas lantibióticas y ha demostrado inhibir el crecimiento de *S. pyogenes in vitro (Burton et al., 2006)*.

El grupo *bovis*, donde se agrupan a las especies *S. bovis*, *S. equinus* y *S. alactolyticus*. De este grupo, la primera especie mencionada ha sido dividida y renombrada como las

especies: *S. gallolyticus* (antes *S. bovis* I), *S. pasteurianus* (antes *S. bovis* II/2) y *S. infantarius* (antes *S. bovis* II/1) (Corredoira et al., 2008). En este estudio, dichas especies forman un solo grupo junto con *S. lutetiensis* y *S. macedonicus* (subárbol verde obscuro), guardando coherencia con lo anteriormente señalado por Kawamura *et al.* (1995). Dentro de este grupo hay un resultado muy relevante que es que; al comparar los árboles producidos por GSS y rRNA 16S podemos observar una gran diferencia entre la distancia que presenta *S. infantarius* dentro del primero, siendo ésta mucho mayor que en el segundo. Esta observación nos permite afirmar que esta métrica implementada puede aportar nuevos horizontes dentro de la clasificación taxonómica y corrobora las descripción hecha por Alcaraz *et al.,* (2010) *a*cerca del uso de esta herramienta como un complemento para aclarar las relaciones entre organismos, ya que las descripciones taxonómicas de esta especie habían sido hechas con base en secuencias de rRNA 16S y del gen *sodA* (Poyart *et al.,* 2002) y no a un nivel genómico como el realizado en este trabajo. Otro resultado que es interesante notar, es que la cepa específica utilizada de *S. macedonicus* se agrupe dentro de este grupo de streptococci patógenos, ya que ésta ha sido aislado de productos lácteos (Papadimitriou *et al*., 2012) y probablemente no represente un riesgo a la salud, como ha sido probado en *S. thermophilus*, especie que ha evolucionado mediante pérdida de funciones debido al ambiente en el cual se encuentra (Bolotin *et al*., 2004). Dentro de este subárbol se coloca a *S. mutans* como un grupo más alejado, el cual se ha descrito como un solo grupo (grupo mutans) por Kawamura *et al.* (1995) junto con otras especies no utilizadas en este trabajo y esto se corrobora debido al bajo soporte de *bootstrap* (38) que tiene el nodo donde convergen las especies de este subárbol.

Finalmente, al comparar los dendrogramas que agrupan a las especies en la Figura 10-A y 10-B (eje horizontal superior), y cotejarlos con los grupos formados en los árboles de las Figuras 12 y 13 podemos apreciar que en estos dendrogramas, los cuales están agrupando a las especies con base en la abundancia y similitud de funciones, el grupo salivarius se forma de nuevo y se coloca de manera distante a los demás grupos; lo que nos indica un mayor soporte de la estructura filogenética de este grupo. A nivel funcional, los integrantes del grupo pyogenes se agrupan dentro del dendrograma de la clasificación por COG (Figura 10-B) en subgrupos más cercanos que en el de la clasificación por subsistemas (Figura

10-A). El hecho de que el genoma núcleo del género presente la mayor distancia a los genomas núcleo de las especies en los dendrogrogramas horizontales de la Figura 5 se debe a que la abundancia de aciertos del genoma núcleo se relaciona con su tamaño (405 proteínas predichas), el cual es, en el mejor de los casos, tres veces más pequeño que los genomas núcleo de las especies (ver tabla 8) y la distancia se calcula con base en estas abundancias.

Al comparar los árboles de GSS y rRNA 16S podemos observar que el agrupamiento de las especies no difiere en gran medida, pero la resolución (distancias) que otorga el árbol de GSS es mucho mayor a la de rRNA 16S. En el árbol de GSS se resuelven las ramas terminales; por lo tanto es posible realizar la diferenciación entre cada una de las cepas en cada grupo de especies, situación que resulta imposible en el árbol de rRNA 16S. Se podría esperar que la comparación genómica resultara en una topología distinta, ya que el uso de las distancias genómicas por GSS incluye al número máximo de genes compartidos comparables por pares (Alcaraz et al., 2010) en contraste con la reconstrucción filogenética clásica que solo utiliza a un gen en este caso.

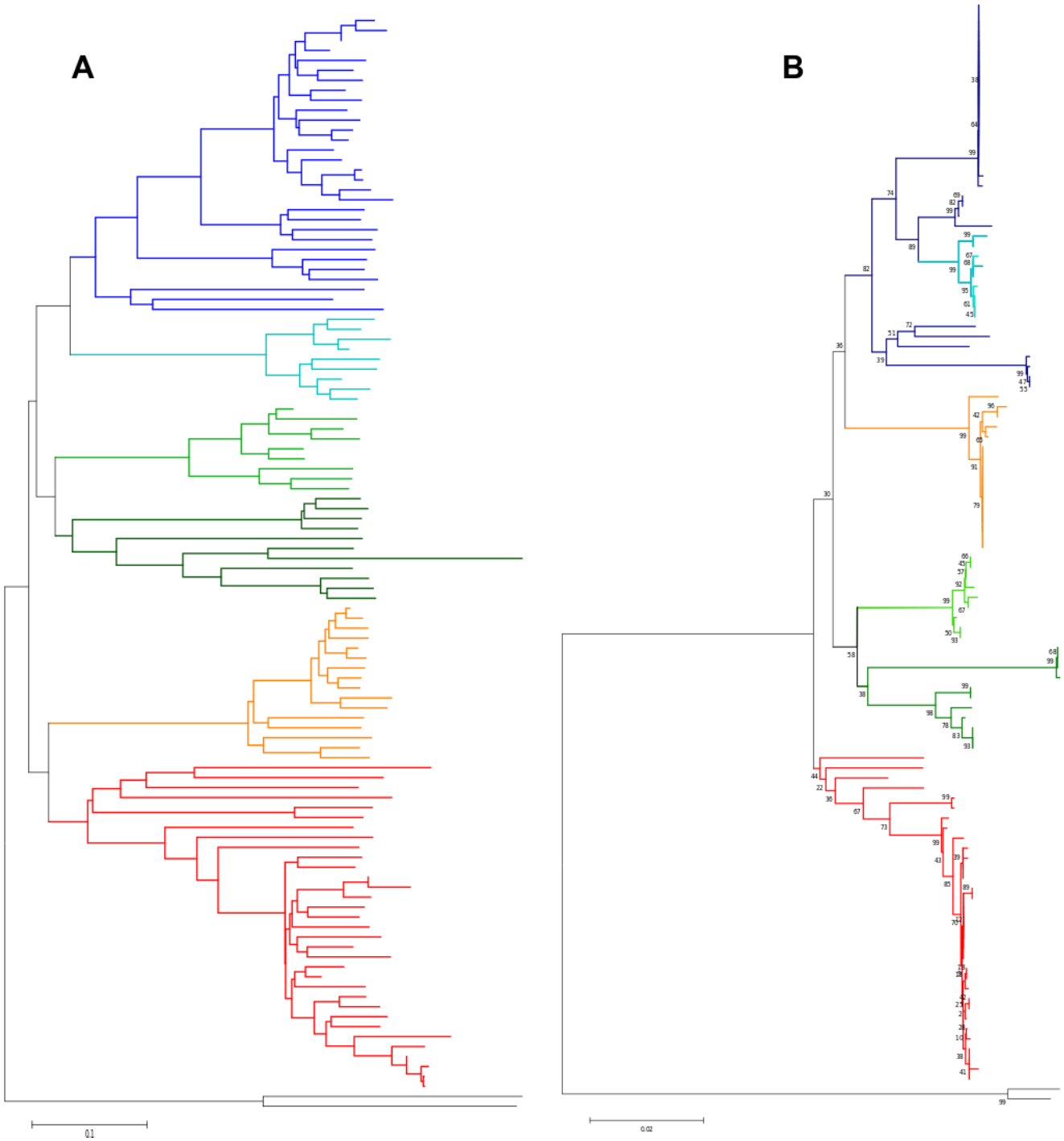

Figura 17. Topologías de las reconstrucciones filogenéticas. A. Matriz de distancias GSS graficada como un árbol por el método Neighbor-Joining. **B.** Filogenia por el método Neighbor-Joining de las secuencias de rRNA 16S.

Funciones de los genomas núcleo y pangenoma

Las comparaciones que se hacen mediante la anotación con los COG son más burdas debido a que esta clasificación tiene un número menor de categorías y genes anotados (21 categorías, 4873 COGs y 138,458 genes) , pero da una idea más general sobre las funciones que se llevan a cabo (Tatusov et al., 2003).

Como ha sido estipulado, el genoma núcleo, es el conjunto de genes que se comparten entre todas las cepas (Tettelin et al., 2005) y por esto se piensa que estos genes se encontrarán más conservados de entre todas las muestras que se utilicen, no importando el nivel taxonómico al que se esté trabajando.

Al realizar la clasificación funcional del genoma núcleo del género tanto por subsistemas como por COG (ver figura 10-A y 10-B), encontramos una serie de características esperadas. Categorías de los subsistemas como la 22 (metabolismo de RNA) o de COG como L (replicación, recombinación y reparación) se encuentran en una mayor abundancia en el genoma núcleo del género en comparación con los de las especies, lo que es normal debido a que este tipo de funciones son requeridas por todos los organismos para los procesos básicos de la vida.

Encontramos que las categorías de los subsistemas relacionadas a fotosíntesis, metabolismo secundario, esporulación y movilidad son las de menor abundancia en todos los casos, lo que es un resultado congruente debido a que los streptococci no son bacterias formadoras de espora, no son móviles, no llevan a cabo la producción de metabolitos secundarios de manera extensiva como lo hacen los bacilli (Chaabouni *et al.,* 2012, pp 347-366) y tampoco son organismos que se caractericen por utilizar la fotosíntesis como fuente de energía (Kayser *et al.,* 2005, pp234).

El hecho de que la categoría 3 de subsistemas y D de COG (división celular y ciclo celular), en la cual se encuentran genes correspondientes al operón *Fts*, los cuales están

involucrados en la división celular y se reportan 10 genes esenciales de este operón para la forma y división celular (Kobayashi et al., 2003) , se encuentre poco representada en todos los genomas núcleo y con mayor abundancia en el del género, es señal de que no se cuentan con muchas copias de los genes asociados a estos procesos y dentro de los núcleos de cada especie, al estar reportando la abundancia relativa de cada función, se observan menos representados. Resulta esperado que en el genoma núcleo del género haya una mayor representación de esta categoría, ya que al igual que el metabolismo de RNA, ésta es una función que debe estar compartida y conservada entre todos los streptococci por ser fundamental para la vida y en relación al pequeño conjunto de genes del genoma núcleo la abundancia de ésta categoría se observa más elevada.

Las funciones de transporte de membrana, se aprecian con mayor abundancia en *S. mutans*, lo que puede estar relacionado al estilo de vida que tiene esta especie; al ser un patógeno oral y llevar a cabo la formación de biopelículas, tiene sistemas de secreción eficiente para secretar tanto polisacáridos extracelulares como factores de virulencia, como la hemolisina (Mitchell, 2003).

La categoría 6 (vitaminas, cofactores) muestra una menor representación en el genoma núcleo del género y esto se debe a que existen funciones dentro de esta categoría como la biosíntesis de ubiquinona o menaquinona las cuales están presentes dentro de algunas de las especies, como *S. gallolyticus* o *S. thermophilus,* pero no se encuentran en genoma núcleo del género lo cual hace sentido, ya que el género es considerado como anaerobio facultativo (Kayser *et al,* 2005, pp. 234), hecho que puede estar dado por la transferencia horizontal de genes que se reporta en varias especies del género como *S. thermophilus (Delorme et al., 2007)*, *S. pneumoniae* o *S. mitis (Donati et al., 2010)* o por casos particulares como el de *S. agalactiae*, la cual puede obtener grupos hemo y menaquinona exógena para activar la cadena respiratoria en ambientes aerobios (Yamamoto et al., 2006).

De las comparaciones hechas mediante COGs también surgen resultados esperados y que ayudan a corroborar los resultados de la clasificación por subsistemas. Dentro del genoma núcleo del género, se encuentra la categoría COG J (traducción y estructura

ribosomal) como la más abundante de entre todos los genomas núcleo (ver Figura 10-B) y esto es normal, debido a que este tipo de funciones deben de estar compartidas por todas las especies, tanto así que las secuencias de rRNA 16S son utilizadas para definir a las especies bacterianas hoy en día, aunque en ocasiones nos lleva a definiciones engañosas (Georgiades & Raoult, 2010) . La categoría E (transporte y metabolismo de aminoácidos) y G (metabolismo de carbohidratos) son menos abundantes en el núcleo del género y variables dentro de las especies, situación que acontece también en la clasificación por subsistemas (Figura 10-A; categoría 2); este es un resultado esperado, ya que que cada una de las especies varía en cuanto a sus características bioquímicas y nutricionales, adquiriendo y procesando sus fuentes de carbono y aminoácidos en diversas maneras. De igual manera hay una conservación de funciones como es en las categorías L (replicación, recombinación y reparación) y M (membrana, pared y envoltura celular) dentro del genoma núcleo del género, situación que hemos discutido con anterioridad en la clasificación por subsistemas, y el hecho de que se encuentren en baja abundancia puede estar indicando que el número de copias de los genes involucrados en estos procesos (topoisomerasas, helicasas, polimerasas) no es muy grande en comparación con el número total de genes en cada genoma.

La comparación de las funciones presentes en el pangenoma y genoma núcleo del género, tanto por subsistemas como por COG (ver Figura 11), nos demuestra el conjunto de genes que se conservan en todos los streptococci y los que resultan accesorios y se encuentran solo en algunas especies o cepas. Algunas de las funciones necesarias para el desarrollo de la vida bacteriana, como son las asociadas al procesamiento de información, envoltura y forma de la célula, división celular y obtención de energía (Kobayashi et al., 2003), contenidas en las categorías J y M de COG; y 22 y 3 de subsistemas (metabolismo de RNA, membrana y pared celular, división celular), se encuentran más representadas en el genoma núcleo, corroborando lo anteriormente discutido; mientras que la abundancia de funciones pobremente caracterizadas (categorías R y S del COG) es mayor dentro del pangenoma, lo cual es relevante debido a que dentro de esta categoría se encuentran una gran cantidad de genes hipotéticos, que a su vez tienen un mayor potencial de aportar nuevas funciones que confieran la especificidad de habitar cierto ambiente, estando

relacionados a mecanismos de defensa o patogenicidad que a estar vinculados a funciones relacionadas al dogma central de la biología y que aún no han sido caracterizados, por lo que debería existir un mayor interés en la caracterización y descripción de las secuencias clasificadas dentro de esta categoría.

Al observar que los elementos transponibles y plásmidos (Figura 11-B; categoría 17) son más abundantes en el pangenoma y en el genoma núcleo son prácticamente inexistentes podemos pensar en la versatilidad de funciones que pueden acarrear los elementos móviles y por esto se encuentran en el pangenoma como parte de los genes accesorios y que están confiriendo una serie de funciones específicas a cada especie o cepa.

Entre otras funciones interesantes que se encuentran más representadas en el genoma núcleo, en comparación con el pangenoma, observamos a las de respuesta a estrés (categoría 26 de subsistemas), donde se encuentran proteínas como DnaK, GrpE o SmpB, que es una proteína muy conservada dentro del reino bacteriano (Karzai et al., 1999), las cuales son necesarias para lidiar con problemas de mensajeros defectuosos y proteínas desnaturalizadas.

La categoría 7 (metabolismo de DNA) contiene a las funciones replicación, reparación, recombinación y competencia tanto en el genoma núcleo como en el pangenoma pero está más representada en el pangenoma debido a que dentro de éste, podemos encontrar a las secuencias CRISPR, que confieren resistencia a los procariontes contra los virus (Sorek et al., 2008) y que han sido encontradas en diversas cepas de *S. mutans* (Maruyama et al., 2009) y otros streptococcci. El hecho de que este tipo de secuencias se localicen en el pangenoma hace que la categoría 7 de subsistemas se encuentre más representada y al mismo tiempo indica que no todas las especies o cepas han adquirido estas secuencias, ya que no todas las cepas son infectadas por los mismos fagos y las especificidad de esto puede estar mediada por esta clase de elementos.

De las funciones relacionadas a la categoría 4 (pared, membrana y cápsula) se espera que se encuentren conservadas debido a que son estructuras producidas por todos o la

mayoría de los streptococci, como es el caso de los genes relacionados a la biosíntesis de peptidoglicano, propios de las bacterias Gram + y presentes en el genoma núcleo del género. En el caso de la cápsula, tenemos dos clases de cápsula producida por los streptococci; una de ácido hialurónico, producida comúnmente por los integrantes del grupo A de Lancefield (Crater & van de Rijin, 1995) y la otra de polisacáridos, típica en *S. pneumoniae*. (Wessels, 1997). En ambos casos se ha reportado que éstas se encuentran involucradas en el proceso de patogénesis y presentan una gran variabilidad, identificándose en *S. pneumoniae* hasta 90 distintos serotipos capsulares (Henrichsen, 1995). Dentro del pangenoma se presenta esta categoría con mayor abundancia relativa, ya que en éste hay una gran cantidad de genes que no se encuentran en el genoma núcleo y tienen funciones asociadas a síntesis de exopolisacáridos o biosíntesis de polisacáridos capsulares, las cuales son funciones particulares de cepas patógenas como algunos de los *S. suis* o *S. pneumoniae* del cual es característica la formación de una cápsula de polisacáridos (Wessels, 1997)

En cuanto a los mecanismos de virulencia y enfermedad (categoría 28 de subsistemas), encontramos en esta categoría a toda la batería de elementos génicos asociados a resistencia a antibióticos y compuestos tóxicos; bacteriocinas y péptidos antibacterianos, los cuales en el genoma núcleo del género no están presentes, ya que cada especie y cepa cuenta con distintas estrategias para evadir al sistema inmune de su huésped o competir en el medio circundante, el cual varía dependiendo de cada especie. Dentro de esta categoría se encuentra la función de adhesión, que contiene la enzima sortasa, responsable del anclaje de proteínas con el motivo LPXTG a la pared celular mediante el corte entre la treonina y la glicina, uniendo el extremo amino de la treonina a la pentaglicina de la pared celular y que se encuentra en todas las bacterias Gram positivas (Paterson & Mitchell, 2004),esta asociada a la patogénesis de especies como *S. pneumoniae* (Paterson & Mitchell, 2006) y está presente en el genoma núcleo del género, indicando que se encuentra conservada dentro del género.

Funciones dentro de los metagenomas

En cuanto a las funciones que están asociadas a las comunidades de los streptococci

dentro de los metagenomas orales humanos, encontramos que en el metagenoma CA_05_1.4 los streptococci aportan una mayor cantidad de genes relacionados al metabolismo de carbohidratos puede estar relacionado al estado de salud del individuo del cual proviene esta muestra (cavidades), ya que al tener un conjunto mayor de genes involucrados con la degradación de carbohidratos, implica la producción de una mayor cantidad de ácido debido que os streptococci son bacterias ácido lácticas y el producto final de la degradación de los carbohidratos será ácido láctico (Kilian, 2007). Esto sucede también en el metagenoma CA_06_1.6, que es un individuo en el mismo estado de salud que el anterior, aunque se observa la abundancia relativa de las funciones asociadas a metabolismo de carbohidratos menos representada y esto es debido a la variación en la abundancia taxonómica de las comunidades de este género bacteriano mostradas en la Figura 14.

Aunque la proporción de especies de S*treptococcus* en relación al total de especies presentes en cada metagenoma es variable , la composición de las mismas es similar dentro de cada metagenoma (ver Figura 15) y por ello no es posible distinguir funciones específicas que puedan ser asociadas a un organismo en particular. Es interesante notar que los individuos más lejanos en el dendrograma son individuos que presentan cavidades (ver Figura 16), pero al mismo tiempo tienen la menor abundancia de streptococci; por lo tanto podríamos asociar una abundancia baja de streptococci con estados de salud oral pobre y viceversa, a pesar de que *S. mutans* ha sido identificado como uno de los principales agentes etiológicos de la enfermedad (Loesche, 1986).

Conclusiones

El presente trabajo nos demuestra la importancia de realizar trabajos de genómica comparativa en bacterias. Mediante la implementación de estrategias bioinformáticas es posible analizar la enorme cantidad de información que resulta de la secuenciación de genomas completos y que se encuentra disponible en las bases de datos públicas.

La caracterización del pangenoma y genoma núcleo de grupos bacterianos complejos, diversos y con problemas en su taxonomía, como sucede con las especies del género Streptococcus, nos permite conocer la diversidad y variabilidad genética que éstas pueden tener. Hemos encontrado que el total de genes compartidos por los streptococci analizados es de 405, siendo ésta una pequeña fracción del total de genes, 212,348 genes redundantes, conformando 33,039 familias dentro del pangenoma. Aunado a esto, la clasificación jerárquica de funciones y el análisis de la abundancia del genéro brinda información que resulta útil para el entendimiento de la biología y el estilo de vida de estos microorganismos y los roles que pueden estar cumpliendo dentro de ambientes como la cavidad oral humana, en donde se tiene a un grupo bacteriano (los streptococci) del cual la mayor parte de sus integrantes son patógenos pero a pesar de esto, es interesante notar que los estados de salud oral se relacionan a una mayor abundancia de los integrantes de este género.

Se ha demostrado la utilidad del de *Genomic Similarity Score* (GSS) como una herramienta eficaz para aclarar las relaciones entre especies filogenéticamente cercanas en comparación con los métodos tradicionales de reconstrucción filogenética, los cuales utilizan comparaciones a nivel de un solo gen o un pequeño conjunto de genes. En cambio, esta novedosa estrategia permite diferenciar a especies muy relacionadas, ya que utiliza las distancias genómicas proporcionadas por los genes ortólogos.

Perspectivas

El análisis del pangenoma y genoma núcleo de un grupo bacteriano nos permite llevar a cabo descubrimientos en cuanto a la evolución de dicho grupo y la definición de las especies bacterianas, la cual aún se encuentra en discusión, mediante la búsqueda de marcadores moleculares que proporcionen mejores señales filogenéticas; la búsqueda de blancos apropiados para vacunas dentro del pangenoma y posteriormente la construcción de genomas núcleo de las especies relacionadas a cierta enfermedad para generar vacunas específicas contra estos grupos; trabajos sobre epidemiología pueden ser realizados para encontrar patrones en cuanto a los mecanismos de virulencia conservados entre grupos bacterianos y poder proponer estrategias adecuadas para evitar estos problemas de salud pública.

Llevar a cabo el análisis detallado del pangenoma de *Streptococcus* en cuanto a las categorías relacionadas al metabolismo de DNA, debido a que en ésta se han encontrado secuencias relacionadas a las regiones CRISPR, que a su vez se asocian a la infección por bacteriófagos y pueden hablar sobre el ecosistema en el que habitan los streptococci, lo cual podría dar pistas sobre la clase de fagos a los que podrían ser vulnerables las especies y cepas patógenas más importantes y llevar a cabo terapia con fagos para solucionar problemas de salud.

Bibliografía

Ajdić D., McShan W. M., McLaughlin R. E., Savić G., Chang J., Carson M. B., … Ferretti J. J. (2002). Genome sequence of Streptococcus mutans UA159, a cariogenic dental pathogen. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22), 14434–9. doi:10.1073/pnas.172501299

Alcaraz, L. D. (2013). Pan-genomics: Unmasking the gene diversity hidden in the bacteria species., 1–7. doi:10.7287/peerj.preprints.113v1

Alcaraz L., Moreno-Hagelsieb G., Eguiarte L. E., Souza V., Herrera-Estrella L., & Olmedo G. (2010). Understanding the evolutionary relationships and major traits of Bacillus through comparative genomics. *BMC Genomics*, *11*(1), 332. doi:10.1186/1471-2164-11-332

Altschul S. F., Gish W., Miller W., Myers E. W. & Lipman D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–10. doi:10.1016/S0022-2836(05)80360-2

Aziz R. K., Bartels D., Best A., DeJongh M., Disz T., Edwards R., … Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, *9*, 75. doi:10.1186/1471-2164-9-75

Banks, D., Porcella, S., Barbian, K., Beres, S., Philips, L., Voyich, J., … Martin, J. (2004). Progress toward Characterization of the Group A streptococcus Metagenome: Complete Genome Sequence of a Macrolide-Resistnat Serotype M6 Strain. *The Journal of Infectious Disease*, 190, 727–782.

Belda-Ferre P., Alcaraz L. D., Cabrera-Rubio R., Romero H., Simón-Soro A., Pignatelli M., & Mira A. (2012). The oral metagenome in health and disease. *The ISME Journal*, *6*(1), 46–56. doi:10.1038/ismej.2011.85

Berendsen R., Pieterse C. & Bakker P. (2012). The rhizosphere microbiome and plant health. *Trends in Plant Science*, *17*(8), 478–86. doi:10.1016/j.tplants.2012.04.001

Beres, S. B., Sylva, G. L., Barbian, K. D., Lei, B., Hoff, J. S., Mammarella, N. D., … Musser, J. M. (2002). Genome sequence of a serotype M3 strain of group A Streptococcus: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proceedings of the National Academy of Sciences of the United States of America*, 99(15), 10078–83. doi:10.1073/pnas.152298499

Beres, S. B., Sesso, R., Pinto, S. W. L., Hoe, N. P., Porcella, S. F., Deleo, F. R., & Musser, J. M. (2008). Genome sequence of a Lancefield group C Streptococcus zooepidemicus strain causing epidemic nephritis: new information about an old disease. *PloS one*, 3(8), e3026. doi:10.1371/journal.pone.0003026

Bessen, D. E., Kumar, N., Hall, G. S., Riley, D. R., Luo, F., Lizano, S., … Tettelin, H. (2011). Whole-genome association study on tissue tropism phenotypes in group A Streptococcus. *Journal of bacteriology*, 193(23), 6651–63. doi:10.1128/JB.05263-11

Biswas S. & Biswas I. (2012). Complete genome sequence of Streptococcus mutans GS-5, a serotype c strain. *Journal of bacteriology*, 194(17), 4787–8. doi:10.1128/JB.01106-12

Bolotin A., Quinquis B., Renault P., Sorokin A., Ehrlich S. D., Kulakauskas S., … Hols, P. (2004). Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. *Nature Biotechnology*, *22*(12), 1554–8. doi:10.1038/nbt1034

Boyle, B., Vaillancourt, K., Bonifait, L., Charette, S. J., Gottschalk, M., & Grenier, D. (2012). Genome sequence of the swine pathogen Streptococcus suis serotype 2 strain S735. *Journal of bacteriology*, 194(22), 6343–4. doi:10.1128/JB.01559-12

Burton J. P., Wescombe P. A., Moore C. J., Chilcott C. N. & Tagg J. R. (2006). Safety Assessment of the Oral Cavity Probiotic. *Applied and Environmental Microbiology*, *72*(4), 3050–3053. doi:10.1128/AEM.72.4.3050

Camilli R., Del Grosso M., Iannelli F. & Pantosti A. (2008). New genetic element carrying the erythromycin resistance determinant erm(TR) in Streptococcus pneumoniae. *Antimicrobial agents and chemotherapy*, 52(2), 619–25. doi:10.1128/AAC.01081-07

Caporaso J., Kuczynski J. & Stombaugh, J. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature …*, *7*(5), 335–336. doi:10.1038/nmeth0510-335

Castilla, L. M. (2007). Reconstrucción de la historia de cambio de los caracteres. In L. Eguiarte, V. Souza, & X. Aguirre (Eds.), *Ecología Molecular.* México D.F.: Conabio.

Chaabouni I., Guesmi A., Cherif A. (2012) Secondary metabolites of Bacillus: Potentials in Biotechnology. In: Sansinenea E. (Ed.), Bacillus thuringensis biotechnology. Springer Science+Business Media: 347-366.

Chen C., Tang J., Dong W., Wang C., Feng Y., Wang J., … Yu J. (2007). A glimpse of streptococcal toxic shock syndrome from comparative genomics of S. suis 2 Chinese isolates. *PloS one*, 2(3), e315. doi:10.1371/journal.pone.0000315

Ciccarelli F. D., Doerks T., von Mering C., Creevey C. J., Snel B. & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*, *311*(5765), 1283–7. doi:10.1126/science.1123061

Corredoira J., Alonso M. P., Coira A. & Varela, J. (2008). Association between Streptococcus infantarius (formerly S. bovis II/1) bacteremia and noncolonic cancer. *Journal of Clinical Microbiology*, *46*(4), 1570. doi:10.1128/JCM.00129-08

Crater D. & van de Rijin I. (1995). Hyaluronic Acid Synthesis Operon (has) Expression in Group A Streptococci. *The Journal of Biological Chemistry*, *270*(31), 18452–18458.

Croucher N. J., Walker D., Romero P., Lennard N., Paterson G. K., Bason N. C., … Mitchell T. J. (2009a). Role of Conjugative Elements in the Evolution of the Multidrug-Resistant Pandemic Clone Streptococcus pneumoniae ST81. *Journal of bacteriology*, 191(5), 1480–1489. doi:10.1128/JB.01343-08

Delorme C., Poyart C., Ehrlich S. D. & Renault, P. (2007). Extent of horizontal gene transfer in evolution of Streptococci of the salivarius group. *Journal of Bacteriology*, *189*(4), 1330–41. doi:10.1128/JB.01058-06

Delorme, C., Guédon, E., Pons, N., Cruaud, C., Couloux, A., Loux, V., … Renault, P. (2011). Complete genome sequence of the clinical Streptococcus salivarius strain CCHSS3. *Journal of bacteriology*, 193(18), 5041–2. doi:10.1128/JB.05416-11

Delorme C., Bartholini C., Luraschi M., Pons N., Loux V., Almeida M., … Renault P. (2011).

Complete genome sequence of the pigmented Streptococcus thermophilus strain JIM8232. *Journal of bacteriology*, 193(19), 5581–2. doi:10.1128/JB.05404-11

Denapaite D., Brückner R., Nuhn M., Reichmann P., Henrich B., Maurer P., … Hakenbeck R. (2010). The genome of Streptococcus mitis B6--what is a commensal? *PloS one*, 5(2), e9426. doi:10.1371/journal.pone.0009426

Ding F., Tang P., Hsu M. H., Cui P., Hu S., Yu J., & Chiu C. H. (2009). Genome evolution driven by host adaptations results in a more virulent and antimicrobial-resistant Streptococcus pneumoniae serotype 14. *BMC genomics*, 10, 158. doi:10.1186/1471-2164-10-158

Donati C., Hiller N. L., Tettelin H., Muzzi A., Croucher N. J., Angiuoli S. V, … Masignani V. (2010). Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biology*, *11*(10), R107. doi:10.1186/gb-2010-11-10-r107

Facklam, R. (2002). What happened to the streptococci: overview of taxonomic and nomenclature changes. Clinical microbiology reviews, 15(4). doi:10.1128/CMR.15.4.613

Fitch W. M. (2000). Homlogy a personal view on some of the problems. *Trends in Genetics : TIG*, *16*(5), 227–231.

Fox G., Wisotzkey J. & Jurtshuk, P. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *International Journal of Systematic Bacteriology*, *42*(1), 166–70.

Geng J., Chiu C. H., Tang P., Chen Y., Shieh H. R., Hu S. & Chen Y. Y. M. (2012). Complete genome and transcriptomes of Streptococcus parasanguinis FW213: phylogenic relations and potential virulence mechanisms. *PloS one*, 7(4), e34769. doi:10.1371/journal.pone.0034769

Georgiades K. & Raoult, D. (2010). Defining pathogenic bacterial species in the genomic era. *Frontiers in Microbiology*, *1*(January), 151. doi:10.3389/fmicb.2010.00151

Glaser P., Rusniok C., Buchrieser C., Chevalier F., Frangeul L., Msadek T., … Kunst F. (2002). Genome sequence of Streptococcus agalactiae, a pathogen causing invasive neonatal disease. *Molecular …*, *45*, 1499–1513. Retrieved from http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2958.2002.03126.x/full

Goris J., Konstantinidis K. T., Klappenbach J. a, Coenye T., Vandamme P. & TiedjeJ. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, *57*(Pt 1), 81–91. doi:10.1099/ijs.0.64483-0

Guédon E., Delorme C., Pons N., Cruaud C., Loux V., Couloux A., … Renault P. (2011). Complete genome sequence of the commensal Streptococcus salivarius strain JIM8777. *Journal of bacteriology*, 193(18), 5024–5. doi:10.1128/JB.05390-11

Henrichsen J. (1995). Six Newly Recognized Types of Streptococcus Pneumoniae. *Journal of Clinical Microbiology*, *33*(10), 2759–2762.

Hinse D., Vollmer T., Rückert C., Blom J., Kalinowski J., Knabbe C. & Dreier J. (2011). Complete genome and comparative analysis of Streptococcus gallolyticus subsp. gallolyticus, an emerging pathogen of infective endocarditis. *BMC genomics*, 12(1), 400.

doi:10.1186/1471-2164-12-400

Holden M. T. G., Scott A., Cherevach I., Chillingworth T., Churcher C., Cronin A., … Parkhill J. (2007). Complete genome of acute rheumatic fever-associated serotype M5 Streptococcus pyogenes strain manfredo. *Journal of bacteriology*, 189(4), 1473–7. doi:10.1128/JB.01227-06

Holden M. T. G., Hauser H., Sanders M., Ngo T. H., Cherevach I., Cronin A., … Parkhill J. (2009). Rapid evolution of virulence and drug resistance in the emerging zoonotic pathogen Streptococcus suis. *PloS one*, 4(7), e6072. doi:10.1371/journal.pone.0006072

Hoskins J., Alborn W., Arnold J., Blaszack L., Burgett S., DeHoff B., … Al., E. (2001). Genome of the Bacterium Streptococcus pneumoniae Strain R6. *Journal of bacteriology*, 183(19), 5709–5717. doi:10.1128/JB.183.19.5709

Huang Y., Niu B., Gao Y., Fu L. & Li W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* (Oxford, England), 26(5), 680–2. doi:10.1093/bioinformatics/btq003

Hu P., Yang M., Zhang A., Wu J., Chen B., Hua Y., … Jin M. (2011). Complete genome sequence of Streptococcus suis serotype 14 strain JS14. *Journal of Bacteriology*, 193(9), 2375–6. doi:10.1128/JB.00083-11

Hu P., Yang M., Zhang A., Wu J., Chen B., Hua Y., … Jin M. (2011). Complete genome sequence of Streptococcus suis serotype 3 strain ST3. *Journal of Bacteriology*, 193(13), 3428–9. doi:10.1128/JB.05018-11

Jans C., Follador R., Lacroix C., Meile L., & Stevens M. J. A. (2012). Complete genome sequence of the African dairy isolate Streptococcus infantarius subsp. infantarius strain CJ18. *Journal of Bacteriology*, 194(8), 2105–6. doi:10.1128/JB.00160-12

Jin D., Chen C., Li L., Lu S., Li Z., Zhou Z., … Xu J. (2013). Dynamics of fecal microbial communities in children with diarrhea of unknown etiology and genomic analysis of associated Streptococcus lutetiensis. *BMC Microbiology*, 13(1), 141. doi:10.1186/1471-2180-13-141

Kang X., Ling N., Sun G., Zhou Q., Zhang L., & Sheng Q. (2012). Complete genome sequence of Streptococcus thermophilus strain MN-ZLW-002. *Journal of Bacteriology*, 194(16), 4428–9. doi:10.1128/JB.00740-12

Karzai W., Susskind M. M. & Sauer R. T. (1999). SmpB, a unique RNA-binding protein essential for the peptide-tagging activity of SsrA (tmRNA). *The EMBO Journal*, 18(13), 3793–9. doi:10.1093/emboj/18.13.3793

Kawamura Y., Sultana F. & Miura H. (1995). Determination of 1 6 s rRNA Sequences of Streptococcus mitis and Streptococcus gordonii and Phylogenetic relationships among Members of the Genus Streptococcus. *International Journal of Systematic Bacteriology*, 45(2), 406–408.

Kayser F. H., Bienz K. A., Eckert J., Zinkernagel R. (2005). *Medical Microbiology*. Thieme. Stuttgart. Nueva York

Kembel S. W., Wu M., Eisen J. & Green J. L. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS*

Computational Biology, 8(10), e1002743. doi:10.1371/journal.pcbi.1002743

Kilian M. (2007). Streptococcus and enterococcus. Medical Microbiology. En: Greenwood D., Slack R., Peutherer J., Barer M. (Eds.), *Medical Microbiology. A Guide to Microbial Infections: Pathogenesis, Immunity, Laboratory Diagnosis and Control*. 17th ed. Churchill Livingstone Elsevier, London: 178–193.

Kobayashi K., Ehrlich S., Albertini A., Amati G., Andersen K. & Arnaud M. (2003). Essential Bacillus subtilis genes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(8), 4678–83. doi:10.1073/pnas.0730515100

Koonin, E. V. Genome Sequences: Genome sequence of a model prokaryote. *Current Biology*. 1997 Oct 1;7(10):R656-9

Lancefield R. C. (1932). A serological differentiation of human and other groups of hemolytic streptococci, 1919(1), 571–595.

Lanie J. A, Ng W.-L., Kazmierczak K. M., Andrzejewski T. M., Davidsen T. M., Wayne K. J., … Winkler M. E. (2007). Genome sequence of Avery's virulent serotype 2 strain D39 of Streptococcus pneumoniae and comparison with that of unencapsulated laboratory strain R6. *Journal of Bacteriology*, 189(1), 38–51. doi:10.1128/JB.01148-06

Lapierre P. & Gogarten J. P. (2009). Estimating the size of the bacterial pan-genome. *Trends in Genetics*, 25(3), 107–10. doi:10.1016/j.tig.2008.12.004

Lefébure T. & Stanhope M. J. (2007). Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. *Genome Biology*, 8(5), R71. doi:10.1186/gb-2007-8-5-r71

Li G., Hu F. Z., Yang X., Cui Y., Yang J., Qu F., … Zhang J. R. (2012). Complete genome sequence of Streptococcus pneumoniae strain ST556, a multidrug-resistant isolate from an otitis media patient. *Journal of Bacteriology*, 194(12), 3294–5. doi:10.1128/JB.00363-12

Lin I. H., Liu T. T., Teng Y. T., Wu H. L., Liu Y. M., Wu K. M., … Hsu M. T. (2011). Sequencing and comparative genome analysis of two pathogenic Streptococcus gallolyticus subspecies: genome plasticity, adaptation and virulence. *PloS One*, 6(5), e20519. doi:10.1371/journal.pone.0020519

Liu G., Zhang W. & Lu C. (2012). Complete genome sequence of Streptococcus agalactiae GD201008-001, isolated in China from tilapia with meningoencephalitis. *Journal of Bacteriology*, 194(23), 6653. doi:10.1128/JB.01788-12

Loesche W. (1986). Role of Streptococcus mutans in human dental decay. *Microbiological Reviews*, 50(4), 353–380.

Lundberg D. S., Lebeis S. L., Paredes S. H., Yourstone S., Gehring J., Malfatti S., … Dangl J. L. (2012). Defining the core Arabidopsis thaliana root microbiome. *Nature*, 488(7409), 86–90. doi:10.1038/nature11237

Makarova, K., Slesarev, a, Wolf, Y., Sorokin, a, Mirkin, B., Koonin, E., … Mills, D. (2006). Comparative genomics of the lactic acid bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 103(42), 15611–6. doi:10.1073/pnas.0607117103

Maruyama F., Kobata M., Kurokawa K., Nishida K., Sakurai A., Nakano K., … Nakagawa I. (2009). Comparative genomic analyses of Streptococcus mutans provide insights into chromosomal shuffling and species-specific content. *BMC Genomics*, 10, 358. doi:10.1186/1471-2164-10-358

Ma Z., Geng J., Zhang H., Yu H., Yi L., Lei M., … Hu S. (2011). Complete genome sequence of Streptococcus equi subsp. zooepidemicus strain ATCC 35246. *Journal of Bacteriology*, 193(19), 5583–4. doi:10.1128/JB.05700-11

McDonald D., Clemente J. C., Kuczynski J., Rideout J. R., Stombaugh J., Wendel D., … Caporaso J. G. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience*, 1(1), 7. doi:10.1186/2047-217X-1-7

McShan W. M., Ferretti J. J., Karasawa T., Suvorov A. N., Lin S., Qin B., … Savic D. J. (2008). Genome sequence of a nephritogenic and highly transformable M49 strain of Streptococcus pyogenes. *Journal of Bacteriology*, 190(23), 7773–85. doi:10.1128/JB.00672-08

Medini D., Donati C., Tettelin H., Masignani V. & Rappuoli R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, 15(6), 589–94. doi:10.1016/j.gde.2005.09.006

Metzker M. L. (2009). Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1), 31–46. doi:10.1038/nrg2626

Meyer F., Paarmann D., D'Souza M., Olson R., Glass E. M., Kubal M., … Edwards R. A. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9, 386. doi:10.1186/1471-2105-9-386

Mitchell T. J. (2003). The pathogenesis of streptococcal infections: from tooth decay to meningitis. *Nature Reviews. Microbiology*, 1(3), 219–30. doi:10.1038/nrmicro771

Miyoshi-Akiyama T., Watanabe S., & Kirikae T. (2012). Complete genome sequence of Streptococcus pyogenes M1 476, isolated from a patient with streptococcal toxic shock syndrome. *Journal of Bacteriology*, 194(19), 5466. doi:10.1128/JB.01265-12

Moreno-Hagelsieb G. & Janga S. (2008). Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. *Proteins: Structure, Function, Bioinformatics*, 344–352. doi:10.1002/prot

Moreno-Hagelsieb G. & Latimer K. (2008). Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* (Oxford, England), 24(3), 319–24. doi:10.1093/bioinformatics/btm585

Muzzi A., Masignani V. & Rappuoli R. (2007). The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. *Drug Discovery Today*, 12(11-12), 429–39. doi:10.1016/j.drudis.2007.04.008

Nakagawa I., Kurokawa K., Yamashita A., Nakata M., Tomiyasu Y., Okahashi N., … Hamada S. (2003). Genome Sequence of an M3 Strain of Streptococcus pyogenes Reveals a Large-Scale Genomic Rearrangement in Invasive Strains and New Insights into Phage

Evolution. *Genome Research*, 13(6), 1042–1055. doi:10.1101/gr.1096703.1

Nawrocki E. P., Structural RNA Homology Search and Alignment using Covariance Models , Ph.D. thesis, Washington University in Saint Louis, School of Medicine, (2009).

Nho S. W., Hikima J., Cha I. S., Park S. Bin, Jang H. Bin, del Castillo C. S., … Jung T. S. (2011). Complete genome sequence and immunoproteomic analyses of the bacterial fish pathogen Streptococcus parauberis. *Journal of Bacteriology*, 193(13), 3356–66. doi:10.1128/JB.00182-11

Oksanen J, Blanchet G., Kindt R., Legendre P., Minchin P. R., O'Hara R. B., Simpson G.L., Solymos P., Stevens M.H.H. and Wagner H (2013). vegan: Community Ecology Package. R package version 2.0-10.

Okumura K., Shimomura Y., Murayama S. Y., Yagi J., Ubukata K., Kirikae T. & Miyoshi-Akiyama T. (2012). Evolutionary paths of streptococcal and staphylococcal superantigens. *BMC Genomics*, 13(1), 404. doi:10.1186/1471-2164-13-404

Overbeek R., Begley T., Butler R. M., Choudhuri J. V, Chuang H.Y., Cohoon M., … Vonstein V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17), 5691–702. doi:10.1093/nar/gki866

Pagani I., Liolios K., Jansson J., Chen I. M., Smirnova T., Nosrat B., … Kyrpides N. C. (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 40(Database issue), D571–9. doi:10.1093/nar/gkr1100

Papadimitriou K., Ferreira S., Papandreou N. C., Mavrogonatou E., Supply P., Pot B. & Tsakalidou E. (2012). Complete genome sequence of the dairy isolate Streptococcus macedonicus ACA-DC 198. *Journal of Bacteriology*, 194(7), 1838–9. doi:10.1128/JB.06804-11

Papadimitriou K., Ferreira S., Papandreou N. C., Mavrogonatou E., Supply P., Pot B. & Tsakalidou E. (2012). Complete genome sequence of the dairy isolate Streptococcus macedonicus ACA-DC 198. *Journal of Bacteriology*, 194(7), 1838–9. doi:10.1128/JB.06804-11

Paterson G. K. & Mitchell T. J. (2004). The biology of Gram-positive sortase enzymes. *Trends in Microbiology*, 12(2), 89–95. doi:10.1016/j.tim.2003.12.007

Paterson G. K. & Mitchell T. J. (2006). The role of Streptococcus pneumoniae sortase A in colonisation and pathogenesis. *Microbes and Infection*, 8(1), 145–153. doi:10.1016/j.micinf.2005.06.009

Pereira U. D. P., Rodrigues Dos Santos A., Hassan S. S., Aburjaile F. F., Soares S. D. C., Ramos R. T. J., … Figueiredo H. C. P. (2013). Complete genome sequence of Streptococcus agalactiae strain SA20-06, a fish pathogen associated to meningoencephalitis outbreaks. *Standards in Genomic Sciences*, 8(2), 188–97. doi:10.4056/sigs.3687314

Port G., Paluscio E., & Caparon M. (2013). Complete Genome Sequence of emm Type 14 Streptococcus pyogenes Strain H5. *Genome Announcements*, 1(4), 2–3.

doi:10.1128/genomeA.00612-13.Copyright

Poyart C., Quesne G. & Trieu-cuot P. (2002). Taxonomic dissection of the, (November 2001), 1247–1255. doi:10.1099/ijs.0.02044-0.

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., … Bork, P. (2014). eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Research, 42(Database issue), D231–9. doi:10.1093/nar/gkt1253

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reichmann P., Nuhn M., Denapaite D., Brückner R., Henrich B., Maurer P., … Hakenbeck R. (2011). Genome of Streptococcus oralis strain Uo5. *Journal of Bacteriology*, 193(11), 2888–9. doi:10.1128/JB.00321-11

Rice P., Longden I., & Bleasby A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* : TIG, 16(6), 2–3.

RosinskiChupinI., Sauvage E., Mairey B., Mangenot S., Ma L., Da Cunha V., … Glaser P. (2013). Reductive evolution in Streptococcus agalactiae and the emergence of a host adapted lineage. *BMC Genomics*, 14, 252. doi:10.1186/1471-2164-14-252

Rusniok C., Couvé E., Da Cunha V., El Gana R., Zidane N., Bouchier C., … Glaser P. (2010). Genome sequence of Streptococcus gallolyticus: insights into its adaptation to the bovine rumen and its ability to cause endocarditis. *Journal of Bacteriology*, 192(8), 2266–76. doi:10.1128/JB.01659-09

Shendure J. & Ji H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10), 1135–1145.

Shahinas D., Tamber G. S., Arya G., Wong A., Lau R., Jamieson F., … Pillai D. R. (2011). Whole-genome sequence of Streptococcus pseudopneumoniae isolate IS7493. *Journal of Bacteriology*, 193(21), 6102–3. doi:10.1128/JB.06075-11

Shimomura Y., Okumura K., Murayama S. Y., Yagi J., Ubukata K., Kirikae T. & Miyoshi-Akiyama T. (2011). Complete genome sequencing and analysis of a Lancefield group G Streptococcus dysgalactiae subsp. equisimilis strain causing streptococcal toxic shock syndrome (STSS). *BMC Genomics*, 12(1), 17. doi:10.1186/1471-2164-12-17

Smith J. M., Smith N. H., O'Rourke M. & Spratt B. G. (1993). How clonal are bacteria? Proceedings of the National Academy of Sciences of the United States of America, 90(10), 4384–8. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=46515&tool=pmcentrez&rendertype=abstract

Sorek R., Kunin V. & Hugenholtz P. (2008). CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. Nature Reviews. Microbiology, 6(3), 181–6. doi:10.1038/nrmicro1793

Stackebrandt E. & Goebel B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic Bacteriology*, 44(4), 846–849.

Sumby P., Porcella S., Madrigal A., Barbian K., Virtaneva K., Ricklefs S., … Graham M. (2005). Evolutionary Origin and Emergence of a Highly Successful Clone of Serotype M1

Group A. *The Journal of Infectious Disease*, 192(5), 771–782.

Sun Z., Chen X., Wang J., Zhao W., Shao Y., Wu L., … Chen W. (2011). Complete genome sequence of Streptococcus thermophilus strain ND03. *Journal of Bacteriology*, 193(3), 793–4. doi:10.1128/JB.01374-10

Tagg J.R., Wescombe P. A. & Burton J. P. (2011). Streptococcus: A Brief Update on the Current Taxonomic Status of the Genus. In: Sampo Lahtinen, Arthur C. Ouwehand, Seppo Salminen, Atte von Wright (Eds.), Lactic Acid Bacteria: Microbiological and Functional Aspects, 4th ed. CRC Press, US: 123-146.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731-2739.

Tatusov R. L., Fedorova N. D., Jackson J. D., Jacobs A. R., Kiryutin B., Koonin E. V, … Natale D. A. (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 4, 41. doi:10.1186/1471-2105-4-41

Tatusov R. L., Galperin M. Y., Natale D. & Koonin E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research, 28(1), 33–6. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102395&tool=pmcentrez&rendertype=abstract

Tettelin H., Masignani V., Cieslewicz M. J., Donati C., Medini D., Ward N. L., … Fraser C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome.” Proc Natl Acad Sci U S A, 102(39), 13950–13955.

Tettelin H., Masignani V., Cieslwicz M. J., Eisen J. A., Peterson S., Wessels M. & Al. E. (2002). Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V Streptococcus agalactiae. Proceedings of the National Academy of Sciences of the United States of America, 99(19), 12391–6. doi:10.1073/pnas.182380799

Tettelin, H., Masignani, V., Cieslwicz, M. J., Eisen, J. A., Peterson, S., Wessels, M., & Al., E. (2002). Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V Streptococcus agalactiae. Proceedings of the National Academy of Sciences of the United States of America, 99(19), 12391–6. doi:10.1073/pnas.182380799

Tettelin H., Riley D., Cattuto C., & Medini D. (2008). Comparative genomics: the bacterial pan-genome. Current Opinion in Microbiology, 11(5), 472–477. doi:10.1016/j.mib.2008.09.006

The Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055), 69–87. doi:10.1038/nature04072

Tomoyasu T., Tabata A., Hiroshima R., Imaki H., Masuda S., Whiley R. A, … Nagamune H. (2010). Role of catabolite control protein A in the regulation of intermedilysin production by Streptococcus intermedius. *Infection and Immunity*, 78(9), 4012–21.

doi:10.1128/IAI.00113-10

Tong H., Shang N., Liu L., Wang X., Cai J. & Dong X. (2013). Complete Genome Sequence of an Oral Commensal, Streptococcus oligofermentans Strain AS 1.3089. *Genome Announcements*, 1(3), 5–6. doi:10.1128/genomeA.00353-13.Copyright

Van den Bogert B., Boekhorst J., Herrmann R., Smid E. J., Zoetendal E. G., & Kleerebezem M. (2013). Comparative genomics analysis of Streptococcus isolates from the human small intestine reveals their adaptation to a highly dynamic ecosystem. *PloS one*, 8(12), e83418. doi:10.1371/journal.pone.0083418

Vartoukian S. R., Palmer R. M. & Wade W. G. (2010). Strategies for culture of “unculturable” bacteria. *FEMS Microbiology Letters*, 309(1), 1–7. doi:10.1111/j.1574-6968.2010.02000.x

Vela A, Perez M., Zamora L., Palacios L., Domínguez L., & Fernández-Garayzábal J. F (2010). Streptococcus porci sp. nov., isolated from swine sources. International journal of systematic and evolutionary microbiology, 60(Pt 1), 104–8. doi:10.1099/ijs.0.011171-0

Vickerman M. M., Lobst S., Jesionowski M., & Gill S. R. (2007). Genome-wide transcriptional changes in Streptococcus gordonii in response to competence signaling peptide. *Journal of Bacteriology*, 189(21), 7799–807. doi:10.1128/JB.01023-07

Wang K., Yao H., Lu C. & Chen J. (2013). Complete Genome Sequence of Streptococcus suis Serotype 16 Strain TL13. *Genome Announcements*, 1(3), 2009–2010. doi:10.1128/genomeA.00394-13.

Ward P. N., Holden M. T. G., Leigh J. A, Lennard N., Bignell A., Barron A., … Parkhill J. (2009). Evidence for niche adaptation in the genome of the bovine pathogen Streptococcus uberis. *BMC Genomics*, 10, 54. doi:10.1186/1471-2164-10-54

Warnes G.R., Bolker B., Bonebakker L., Gentleman R., Huber W., Liaw A., Lumley T., Maechler M., Magnusson A., Moeller S., Schwartz M & Venables B. (2013). gplots: Various R programming tools for plotting data. R package version 2.12.1

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., … Trüper, H. G. (1987). Report of the Ad Hoc Comitee on Reconciliation of Approaches to Bacterial Systematics. *International journal of systematic bacteriology*, 37(4), 463–463.

Wessels M. R. (1997). Biology of streptococcal capsular polysaccharides. Society for Applied Bacteriology Symposium Series, 26, 20S–31S. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9436314

Wooley J. C. J., Godzik A. & Friedberg I. (2010). A primer on metagenomics. PLoS Computational Biology, 6(2), e1000667. doi:10.1371/journal.pcbi.1000667

Xu P., Alves J. M., Kitten T., Brown A., Chen Z., Ozaki L. S., … Buck G. A. (2007). Genome of the opportunistic pathogen Streptococcus sanguinis. *Journal of Bacteriology*, 189(8), 3166–75. doi:10.1128/JB.01808-06

Yamamoto Y., Poyart C., Trieu-Cuot P., Lamberet G., Gruss A. & Gaudu P. (2006). Roles of environmental heme, and menaquinone, in streptococcus agalactiae. Biometals : An International Journal on the Role of Metal Ions in Biology, Biochemistry, and Medicine, 19(2), 205–10. doi:10.1007/s10534-005-5419-6

Ye C., Zheng H., Zhang J., Jing H., Wang L., Xiong Y., … Xu J. (2009). Clinical, experimental,

and genomic differences between intermediately pathogenic, highly pathogenic, and epidemic Streptococcus suis. *The Journal of Infectious Diseases*, 199(1), 97–107. doi:10.1086/594370

Zhang A., Yang M., Hu P., Wu J., Chen B., Hua Y., … Jin M. (2011). Comparative genomic analysis of Streptococcus suis reveals significant genomic diversity among different serotypes. *BMC genomics*, 12(1), 523. doi:10.1186/1471-2164-12-523

Zhang L. (2011). Problem Solving Handbook in Computational Biology and Bioinformatics. In L. S. Heath & N. Ramakrishnan (Eds.), . Boston, MA: Springer US. doi:10.1007/978-0-387-09760-2

Zheng P., Chung K,, Chiang-ni C., Wang S., Tsai P., Chuang W., … Wu J. (2013). Complete Genome Sequence of emm1 Streptococcus pyogenes A20, a Strain with an Intact Two-component system, CovRS, Isolated from a Patient with Necrotizing Fascitis. Genome Announcements, 1(1), 2010–2011. doi:10.1128/genomeA.00149-12.Copyright

Zubair S., Villiers E. P., De Fuxelius H. H., Andersson G., & Bishop R. P. (2013). Genome Sequence of streptococcsu agalactiae Strain 09mas018883, Isolated from a Swedish Cow. *Genome Announcements*, 1(4), 1–2. doi:10.1186/1471-2164-9-75.Carver

Zubair S., Villiers E. P., De Younan M., Tettelin H., Riley D. R., Jores J., … Bishop R. P. (2013). Genome Sequences of Two Pathogenic Streptococcus agalcatiae Isolates from the One-Humped Camel Camelus Dromedarius. *Genome Announcements,* 1(4), 13–14. doi:10.1128/genomeA.00456-13.5.