



Universidad Autónoma de Querétaro
Facultad de Ciencias Naturales
Licenciatura en Microbiología

NOMBRE DEL TRABAJO:

"Análisis *in situ* de la diversidad taxonómica del microbioma y su relación con estados de salud y enfermedad en poblaciones humanas de diferente ancestría"

Tesis Individual

Que como parte de los requisitos para obtener el grado de:

Licenciado en Microbiología

Presenta:

Erick Fidel Ramírez Cruz

Dirigido por:

Dra. Lucía Guadalupe Morales Reyes

SINODALES

Dra. Lucía Guadalupe Morales Reyes
Presidente

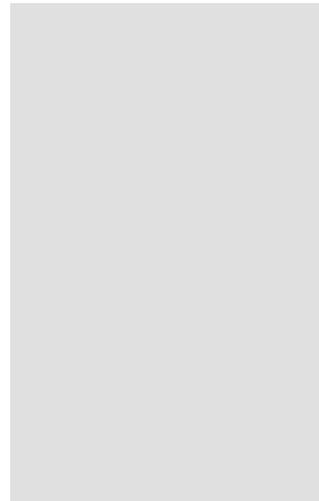
Dr. Carlos Saldaña Gutiérrez
Secretario

Dra. María del Carmen Avila Arcos
Vocal

Dr. Roberto Carlos Álvarez Martínez
Suplente

Dra. Angelina Rodríguez Torres
Suplente

Centro Universitario
Querétaro, Qro.
Marzo del 2019
México



Dirección General de Bibliotecas de la UAQ



Dirección General de Bibliotecas y Servicios Digitales
de Información



Análisis In silico de la diversidad taxonómica del
microbioma y su relación con estados de salud y
enfermedad en poblaciones humanas de diferente
ancestría

por

Erick Fidel Ramírez Cruz

se distribuye bajo una [Licencia Creative Commons
Atribución-NoComercial-SinDerivadas 4.0
Internacional](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Clave RI: CNLIN-230169

Agradecimientos

Agradecimientos técnicos y becas:

Agradezco al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) con el número de proyecto **IA207817**, titulado “Estabilización y evolución de genomas híbridos”, por otorgarme una beca durante la realización de mi tesis.

Agradezco el apoyo técnico presentado por el Laboratorio Nacional de Visualización Científica Avanzada (LAVIS) de la UNAM Campus Juriquilla, así como a todo su personal técnico, especialmente a Luis Alberto Aguilar.

Agradecimientos personales:

El presente trabajo de tesis está dedicado a mi madre, **Leticia Cruz Jiménez**, quien me ha apoyado en todo momento durante este camino, y a todas las personas que me han apoyado para terminar este ciclo tan importante en mi vida, en especial a **Luis Daniel Ramírez**.

Quiero extender mi agradecimiento con especial afecto a mi directora de tesis **Lucía Morales**, por todo su apoyo y tiempo invertido en este trabajo, al igual que a **María Ávila**.

Desde el fondo de mi corazón, **muchas gracias**.

Índice

Resumen	1
Abstract	1
1. Introducción	2
1.1. El microbioma humano	3
1.1.1. El microbioma del tracto gastrointestinal.....	4
1.2. El micobioma humano.....	5
1.3. Enfermedades humanas relacionadas a la composición del microbioma intestinal	7
1.4. Principales miembros del microbioma humano en distintas zonas del cuerpo: piel, mucosas y tracto gastrointestinal.	9
1.5. Diferencias del microbioma entre poblaciones humanas.....	13
1.6. Principales proyectos para la determinación del microbioma humano: HMP, MetaHIT-Eur y MetaHIT-Ch	17
1.6.1. El Proyecto del Microbioma Humano (HMP).....	18
1.6.2. El metagenoma del tracto intestinal humano de poblaciones Europeas, MetaHIT-Eur, asociado a enfermedad inflamatoria intestinal.....	23
1.6.3. El metagenoma del tracto intestinal humano de una población China, MetaHIT-Ch, asociado a diabetes tipo 2.....	27
1.7. Aportes de estudios metagenómicos recientes.....	30
1.8. Tecnologías para determinar la diversidad del microbioma.....	32
1.9. Biodiversidad e índices de diversidad en el análisis de datos metagenómicos	34
1.10. Herramientas de alineamiento de secuencias y bases de datos.....	36
1.10.1. Alineamiento de secuencias.....	37
1.11. Herramientas para el análisis taxonómico de comunidades microbianas. MEGAN	39
1.11.1. MEGAN	39
2. Hipótesis	45
3. Objetivos	45
3.1. Objetivo general.....	45
3.2. Objetivos específicos.....	45
4. Metodología	46
4.1. Recopilación de secuencias ensambladas derivadas de estudios de microbiomas humanos.....	46
4.2. Alineamiento de las secuencias de contigs usando la herramienta BLAST 2.3.0 contra la base de datos “no redundante de nucleótidos (nr-nt)” de NCBI.....	49

4.3.	Análisis taxonómico con MEGAN6	51
4.4.	Análisis de diversidad utilizando vegan.....	53
4.5.	Representación gráfica de los resultados utilizando ggplot2	54
5.	Resultados	55
5.1.	Proyecto del Microbioma Humano (HMP).....	55
5.1.1.	Análisis de asignación taxonómica.....	58
5.1.2.	Perfil taxonómico	60
5.2.	Metagenoma del Tracto Gastrointestinal Humano - Europeo (MetaHIT-Eur).....	63
5.2.1.	Perfil taxonómico	64
5.3.	Metagenoma del Tracto Gastrointestinal Humano - Chino (MetaHIT-Ch).....	68
5.3.1.	Perfil taxonómico	70
5.4.	Comparación del perfil taxonómico entre poblaciones China-Europea	74
6.	Discusión	76
7.	Conclusión	85
	Bibliografía	87
	ANEXOS	96

Análisis *In silico* de la diversidad taxonómica del microbioma y su relación con estados de salud y enfermedad en poblaciones humanas de diferente ancestría

Resumen

El microbioma es el conjunto de microorganismos y su contenido genético que se encuentran habitando un ambiente determinado. Los miembros del microbioma pueden ser: bacterias, virus, hongos, y protozoarios. El cuerpo humano representa un ecosistema complejo dentro del cual se distinguen diferentes hábitats. Las bacterias son los microorganismos de mayor abundancia y que presentan un mayor potencial en el mantenimiento de estados de salud o el desarrollo de cuadros de enfermedad en el hospedero. Otro grupo importante dentro del microbioma humano son los hongos, sin embargo, las comunidades fúngicas en el cuerpo humano no han sido extensamente estudiadas. En este trabajo mostramos la diversidad de hongos de 3 poblaciones humanas derivado del análisis metagenómico de datos públicos de 3 proyectos de microbiomas humanos: **(1)** Estadounidenses / HMP: 16 zonas del cuerpo; **(2)** Europeos / MetaHIT-Eur: tracto gastrointestinal; y **(3)** Chinos / MetaHIT-Ch: tracto gastrointestinal; haciendo una comparación entre grupos de casos (enfermos) y controles (sanos). Nuestros resultados muestran que en individuos estadounidenses sanos las zonas de la piel y mucosas presentan baja diversidad y se encuentran gobernadas por los géneros *Malassezia* y *Candida*, respectivamente. En las 3 poblaciones evaluadas, el tracto gastrointestinal es una zona que presenta elevada diversidad y en donde destacan dos géneros: *Enterocytozoon* y *Saccharomyces*. En individuos europeos, al comparar casos (IBD) y controles no se encuentran diferencias a nivel de composición y diversidad de hongos encontrados. Por su parte, en la población de individuos chinos, el grupo de casos (diabetes) muestra mayor abundancia del género *Saccharomyces* en comparación con el grupo control.

Abstract

The microbiome is the collection of microorganisms and their genetic content that inhabits a certain environment. The members of the microbiome can be bacteria, viruses, fungi, and protozoa. The human body represents a complex environment in which can be distinguished different habitats. Bacteria are the most abundant microorganisms, and they represent a larger potential in the maintenance of states of health and disease in the host. Another group of great importance within the human microbiome is the fungal component. However, the fungal communities inside the human body have not been extensively studied. Here, we

show the fungal diversity of 3 different human populations derived from the metagenomic analysis of public datasets, which comprise 3 human microbiome projects: (1) Americans / HMP: 16 body sites; (2) Europeans / MetaHIT-Eur: intestinal tract; and (3) Chinese / MetaHIT-Ch: intestinal tract; we made comparisons between case (disease) and control (health) groups. Our results show that in healthy individuals from the United States, skin and mucosal related-sites show a low diversity and they both are governed by the *Malassezia* and *Candida* genera, respectively. Within the 3 evaluated populations, the intestinal tract is the zone that represents the higher diversity and where the genera *Enterocytozoon* and *Saccharomyces* are highly represented. In the European individuals, by comparing the groups of cases (IBD) and controls (No IBD), we could not track differences in the composition and diversity of the identified fungi. While in the chinese population, the group of cases (diabetics) shows a higher abundance of the genus *Saccharomyces* compared to the control group.

1. Introducción

El microbioma puede ser definido como la comunidad microbiana total dentro de un ambiente específico, incluyendo tanto sus productos generados, como sus biomoléculas (Morgan, Segata and Huttenhower, 2013). Estos ambientes pueden ser el cuerpo humano o partes del mismo, como las mucosas, incluyendo a las distintas partes de la boca y la zona vaginal, o los epitelios. Los microorganismos en las distintas zonas de nuestro cuerpo representan alrededor de 3.8 trillones de células, igualando al total de células humanas estimadas en 3 trillones (Sender *et al.*, 2016), entre las que se encuentran bacterias, hongos y protozoarios, además de virus (Qin *et al.*, 2010). Aunque esta proporción y diversidad varía de acuerdo a la zona del cuerpo, el tracto gastrointestinal es la zona que presenta mayor abundancia y diversidad (Peterson *et al.*, 2009). Actualmente sabemos que muchos de estos microorganismos tienen una función importante en procesos metabólicos, nutricionales, fisiológicos e inmunológicos (Ottman *et al.*, 2012). Por ejemplo, a nivel bacteriano el filo Bacteroidetes es uno de los más abundantes en el tracto gastrointestinal; comprende 4 clases y alrededor de 7,000 especies; los integrantes de este Filo se conocen por su función como especialistas en la degradación de materia orgánica de alto peso molecular como proteínas y carbohidratos (Thomas *et al.*, 2011), lo cual apoya a los procesos fisiológicos gastrointestinales.

Una gran parte de la importancia conferida a los componentes de la microbiota y el microbioma humano ha sido su relación con estados de salud y enfermedad. Con el objetivo

de indagar sobre la relación entre estos componentes, se han realizado grandes esfuerzos para conocer la composición del microbioma en estados de salud en diversas poblaciones humanas. Uno de los estudios más destacados es el Proyecto del Microbioma Humano (HMP, *The Human Microbiome Project*) (Human Microbiome Project Consortium, 2012) . En dicho proyecto, se estudiaron 15 zonas distintas del cuerpo de individuos norteamericanos y se establecieron los microorganismos centrales presentes en cada una de estas zonas mediante un análisis taxonómico y funcional. Otro esfuerzo por conocer la abundancia y diversidad del microbioma humano, pero con un enfoque en salud y enfermedad fue el Metagenoma del Tracto Intestinal Humano (MetaHIT, *Metagenome of Human Intestinal Tract*) (Qin *et al.*, 2010). En dicho estudio, los autores describieron las poblaciones bacterianas y su función en individuos europeos sanos (dentro de sus estándares asignados) contra un grupo de individuos con Enfermedad Inflamatoria Intestinal (IBD, *Inflammatory Bowel Disease*). Un siguiente estudio por el mismo autor surgió dos años después con la publicación del metagenoma intestinal de individuos pertenecientes a una población china asociado a diabetes tipo 2 (Qin *et al.*, 2012).

1.1. El microbioma humano

El concepto del “microbioma humano” fue propuesto por primera vez por Joshua Lederberg, quien utilizó este término para referirse a la “comunidad ecológica de microorganismos comensales, simbióticos y patogénicos que comparten nuestro espacio corporal” (Lederberg & McCray 2001). Actualmente se sabe que la mayor parte de nuestro cuerpo se encuentra habitada por bacterias y hongos, y que una gran parte de estos lleva a cabo funciones importantes en nuestra nutrición y fisiología, reflejándose así, al final, en nuestro estado de salud (Ottman *et al.*, 2012). Sin embargo, la composición del microbioma humano varía de acuerdo a diversos factores, la mayor parte de ellos ambientales y del propio hospedero; involucrando así factores genéticos, alimenticios, de consumo de fármacos, edad y sistema inmune; además, las interacciones naturales entre los componentes del microbioma propias de una comunidad ecológica, son también un factor fundamental (Blekhman *et al.*, 2015). Datos de secuenciación masiva de muestras fecales del HMP han mostrado que la mayor parte del microbioma humano se compone de bacterias, seguido por hongos, protozoarios y virus (Peterson *et al.*, 2009). Algunas de las observaciones más relevantes sobre la composición bacteriana del microbioma humano son las siguientes: i) *E. coli* se encuentra presente en el intestino de la mayoría de los individuos pero a muy baja abundancia, ii) especies del género *Campylobacter*, que incluye a patógenos oportunistas,

se encuentran en la cavidad oral como microbiota normal de personas sanas, iii) el género *Bacteroides* es el más abundante a nivel intestinal en sujetos sanos, iv) el género *Streptococcus* se encuentra en mayor proporción en la cavidad oral, y v) *Staphylococcus epidermidis* coloniza la mayor parte de las zonas externas del cuerpo (Morgan, Segata & Huttenhower, 2013).

A pesar de los grandes esfuerzos por conocer de manera más profunda todos los componentes del microbioma humano, su diversidad taxonómica y funcional, sus interacciones, los componentes principales de éste y su variación en distintos estados fisiológicos del hospedero, aún falta mucho trabajo por hacer para conocer al microbioma humano en su totalidad. El estudio del microbioma es complejo, y más aún el estudio del mismo a nivel del tracto gastrointestinal, que es a donde se han dirigido actualmente la mayor parte de los estudios.

Sin embargo, algunos estudios muestran que el microbioma humano difiere incluso entre individuos sanos pertenecientes a un mismo grupo poblacional (Human Microbiome Project Consortium, 2012). Esto indica que los componentes del microbioma no son estáticos, y que se encuentran en constante cambio en respuesta a distintos factores que ya conocemos y muchos otros que aún no, como las complejas interacciones entre los diversos grupos taxonómicos que lo componen y de estos con factores propios del hospedero.

1.1.1. El microbioma del tracto gastrointestinal

El estudio del microbioma humano comenzó con el objetivo de conocer ampliamente la distribución y diversidad de microorganismos y sus productos génicos en las distintas zonas del cuerpo, tanto en mucosas como en epitelios, incluyendo la boca, lengua, piel, vagina, heces fecales, entre otros (Peter J. Turnbaugh *et al.*, 2007). Sin embargo, después del HMP en el año 2007, los estudios han centrado su atención en la investigación del microbioma intestinal debido, en gran parte, a la alta diversidad de bacterias y hongos observada en esta zona. Otra de las razones por el creciente interés en estos estudios se debe a la relación de este microbioma con distintos desordenes y enfermedades del tracto gastrointestinal, como la enfermedad de Crohn, la colitis ulcerativa (Qin *et al.*, 2010) y los adenomas colorectales (Sun & Kato, 2016). Aunque también tiene relación con otro tipo de enfermedades y desordenes metabólicos como la diabetes tipo 2 (Qin *et al.* 2012), algunos

procesos neurodegenerativos (Erny & Prinz, 2017), la obesidad (Parekh, Balart & Johnson, 2015), entre otros.

1.2. El microbioma humano

Si bien los esfuerzos por caracterizar los distintos componentes del microbioma humano han aumentado durante los últimos años—con más de 3,000 artículos científicos publicados únicamente en el año 2016, comparado con menos de 90 publicados en el año 2006 (Cani, 2017)—, la investigación en este campo se ha centrado principalmente en el componente bacteriano del microbioma, dejando a un lado a otros componentes importantes como los hongos.

El componente fúngico del microbioma humano se conoce como micobioma (Ghannoum *et al.*, 2010). El microbioma de cada individuo contiene miles de diferentes especies de microorganismos, sin embargo, sólo una pequeña cantidad de especies dominan la mayor parte de la comunidad, representando al 99.9% del microbioma. El 0.1% restante se compone de especies menos abundantes pero más diversas; a este componente de menor abundancia pero mayor diversidad se le denomina biosfera extraña (Huffnagle & Noverr, 2013). Dentro de la biosfera extraña podemos encontrar al micobioma (**Figura 1**).

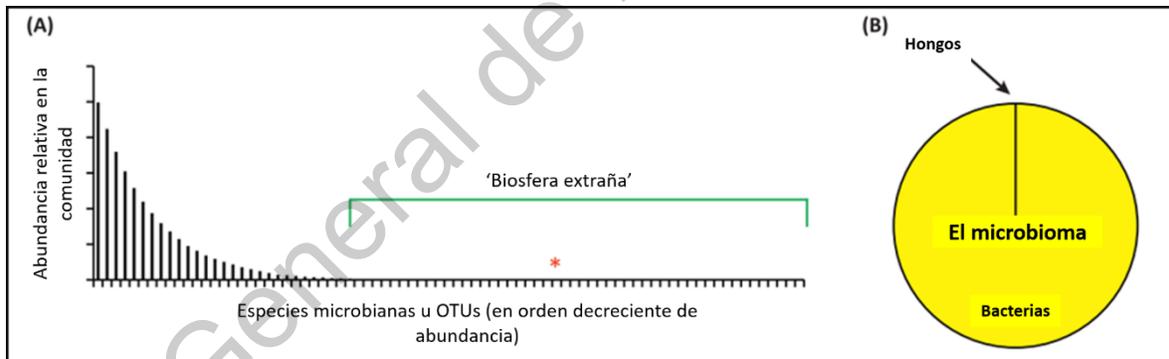


Figura 1. Abundancia de especies en el microbioma. Panel A) Muestra un gráfico de la abundancia relativa y el número de especies en una comunidad. Panel B) Muestra un gráfico de pay en donde el microbioma se encuentra conformado casi en su totalidad por bacterias, y una parte muy pequeña por hongos (Tomado y modificado de Huffnagle & Noverr 2013).

Parte de la importancia de conocer cuáles son los componentes del micobioma es su relación con estados de enfermedad (Cui, Morris & Ghedin, 2013). Este tipo de estudios se basan en un diseño experimental que involucra la participación de dos grupos de individuos:

(1) Individuos sanos (sin la condición de estudio), (2) individuos con enfermedad (con la condición en estudio), ambos grupos pertenecientes a una misma población. El objetivo principal de estos estudios es observar dos propiedades de las comunidades ecológicas: estructura taxonómica y funcional de cada grupo, y observar las diferencias que se presentan en estos. En muestras orales de individuos sanos se ha reportado la presencia de ciertos taxa como: *Candida*, *Cladosporium*, *Aspergillus*, *Fusarium*, *Glomus*, *Penicillium*, *Alternaria* y miembros del orden Saccharomycetales (Ghannoum *et al.*, 2010). Otros estudios con individuos en estados de enfermedad como el síndrome de inmunodeficiencia adquirida (SIDA), han mostrado que la diversidad de especies de hongos en la boca disminuye, siendo el género *Candida* el que se encuentra en una proporción de casi el 100% (Lúcia Bosco *et al.*, 2003). Aunque la diversidad de hongos en el cuerpo humano también difiere en función de las distintas zonas del cuerpo, el género *Malassezia* es el que mayor proporción alcanza en zonas de epitelio expuestas, como son las plantas de los pies, las manos, los brazos, entre otros (Findley *et al.*, 2013). Entre las especies de *Malassezia* descritas se encuentran *M. restricta*, *M. globosa* y *M. sympodialis*. A nivel gastrointestinal, en adultos sanos, se han descrito diferentes taxa como *Wallemia*, Trichomaceae, Saccharomycetaceae, *Rhodotorula*, y Pleosporaceae (Dollive *et al.* 2012). También, a nivel gastrointestinal, pero en individuos con enfermedad inflamatoria intestinal (Enfermedad de Crohn y colitis ulcerativa) las especies de hongos descritas se encuentran principalmente dentro de los los géneros *Rhodotorula*, *Galactomyces*, *Trametes*, *Pleospora*, *Sclerotinia*, *Penicillium*, *Bullera*, *Ustilago*, *Candida*, entre otros (Ott *et al.*, 2008a).

En mucosas como el tracto urogenital, las distintas partes de la cavidad oral, y en el tracto gastrointestinal es común que se genere una colonización por especies de *Candida*, tan sólo en individuos sanos la prevalencia de *Candida* oscila entre el 30 y el 70% (Huffnagle & Noverr, 2013). La prevalencia de *Candida* en individuos con algún estado de enfermedad que comprometa su sistema inmune también es elevada, algunas especies de importancia clínica en humanos son *Candida albicans*, *Candida glabrata*, *Candida rugosa*, *Candida parapsilosis*, *Candida tropicalis*, *Candida dubliniensis*, *Candida krusei* y *Candida lusitaniae* (Odds, 1984).

A pesar de que el estudio del micobioma ha aumentado en los últimos años, aún quedan varias limitantes por vencer para conocer la diversidad total de hongos en grupos de individuos sanos como enfermos. Para ello, es indispensable establecer metodologías cuyo fin particular sea explorar el micobioma. Actualmente una gran cantidad de protocolos de

extracción de ADN y metodologías para el manejo de muestras que se utilizan en el estudio del microbioma humano se encuentran optimizadas para realizar la extracción de ADN bacteriano, generando un sesgo en la obtención de ADN fúngico en estudios metagenómicos (Kong & Morris, 2017). Además, la manipulación de las muestras fecales para el estudio del microbioma intestinal también es un paso crítico durante la recuperación de ADN fúngico, congelar las muestras y no realizar extracción de ADN con material fresco disminuye el rendimiento de ADN obtenido asociado a hongos (Huseyin *et al.*, 2017).

1.3. Enfermedades humanas relacionadas a la composición del microbioma intestinal

Uno de los mayores intereses en el estudio del microbioma, y en especial en la zona gastrointestinal es su relación con estados de salud y enfermedad. Durante cuadros de enfermedad ciertos componentes del microbioma aumentan y otros disminuyen de manera diferencial. A pesar de que son muchos los casos en los que se han reportado estas diferencias de abundancia y composición (ver **Tabla 1**), ha sido difícil describir el impacto preciso del microbioma intestinal en la salud humana y cómo éste se encuentra involucrado en estados de enfermedad (Cani, 2017) (Carding *et al.* 2013).

Tabla 1. Enfermedades causadas por miembros propios de la microbiota humana (Tomada y modificada de Wilson 2016).

Microorganismo	Enfermedad
<i>Enterococcus faecalis</i>	Infecciones del tracto urinario, endocarditis, meningitis, infecciones en heridas
<i>Moraxella catarrhalis</i>	Broncopneumonia, sinusitis, otitis media
<i>Haemophilus influenzae</i>	Meningitis, neumonía, sinusitis, otitis media, epiglotitis
<i>Proteus mirabilis</i>	Infecciones del tracto urinario
<i>Helicobacter pylori</i>	Gastritis, úlceras, carcinoma
<i>Streptococcus mutans</i>	Caries dentales, endocarditis

<i>Porphyromonas gingivalis</i>	Periodontitis
<i>Actinomyces israelii</i> actinomycosis	Actinomicosis
<i>Staphylococcus saprophyticus</i>	Infecciones del tracto urinario
<i>Neisseria meningitidis</i>	Meningitis
<i>Malassezia</i> spp.	Dermatitis atópica, dermatitis seborreica, foliculitis
<i>Gardnerella vaginalis</i>	Vaginitis bacteriana

Como ahora sabemos, la mayor parte del cuerpo humano se encuentra colonizada por microorganismos, los cuales se han adaptado específicamente a cada uno de los distintos nichos de nuestro cuerpo. Se reconoce que debe existir un balance entre los organismos que forman parte del mismo y el hospedero humano. Este balance es importante en el mantenimiento de un estado de salud o en el desarrollo de alguna enfermedad. La microbiota intestinal, por ejemplo, tiene un profundo impacto en la fisiología humana, el metabolismo, la organización del tejido local, y en el desarrollo del sistema inmune. Una alteración desfavorable en el balance microbioma-hospedero se denomina disbiosis (Hill & Artis, 2010). Este fenómeno se caracteriza por alteraciones en las poblaciones microbianas en alguna de las zonas del cuerpo, y puede tener efectos profundos en el mantenimiento de un estado de salud (Frank *et al.*, 2011). Algunas de las enfermedades asociadas producto de una disbiosis se muestran en la **Tabla 2**.

Tabla 2. Enfermedades asociadas a disbiosis bacterianas en distintas zonas del cuerpo (Tomada y modificada de Wilson, 2016).

Enfermedad	Microbiota involucrada
Obesidad	Tracto intestinal
Rinosinusitis	Cavidad nasal
Desorden crónico obstructivo pulmonar	Pulmones
Asma	Pulmones

Autismo	Tracto intestinal
Enfermedad inflamatoria intestinal	Tracto intestinal
Esclerosis múltiple	Tracto intestinal
Artritis	Tracto intestinal
Cancer colorectal	Tracto intestinal
Periodontitis	Cavidad oral
Diabetes tipo II	Tracto intestinal

1.4. Principales miembros del microbioma humano en distintas zonas del cuerpo: piel, mucosas y tracto gastrointestinal.

La cantidad de estudios basados en la determinación del perfil taxonómico de comunidades fúngicas ha aumentado considerablemente durante la última década, aunado a estos esfuerzos, el conocimiento de miembros clave del micobioma humano también ha aumentado. Se han descrito distintos géneros de hongos en distintas partes del cuerpo humano, mostrando una presencia específica de ciertos géneros de acuerdo a la zona en estudio, entre ellos se distingue la piel con el género *Malassezia*, las mucosas con el género *Candida*, y el tracto gastrointestinal con una alta diversidad de géneros, de entre los cuales destaca *Saccharomyces* y *Candida*.

La **piel** representa la interfase y la primera línea de defensa del cuerpo humano con el ambiente externo, la diversidad de microorganismos que tienen la capacidad de colonizar la piel es determinada por una gran cantidad de factores como la predisposición genética, el estilo de vida, y de características físicas del ambiente como del propio hospedero, entre ellas la humedad, el pH, la salinidad, y el contenido sebáceo (Zeeuwen & Schalkwijk, 2016). A pesar de la elevada diversidad de microorganismos con capacidad de colonizar la piel, son pocos los géneros que gobiernan estas comunidades, entre ellos destacan miembros del género *Malassezia*.

Malassezia es un género de hongos perteneciente al filo Basidiomycota, este género alberga a un total de 17 especies con capacidad de colonizar una gran variedad de nichos. Las especies descritas que se encuentran formando parte de comunidades de la piel en

seres humanos son principalmente dos: *Malassezia restricta* y *Malassezia globosa* (Findley *et al.*, 2013). De manera normal este género se encuentra en una relación comensal con el ser humano, sin embargo, en individuos inmunocomprometidos se puede presentar como un patógeno oportunista (Grice and Dawson, 2017). La principal característica de la piel para ser colonizada por hongos es su naturaleza sebácea, alta humedad, y la ausencia de luz en algunas zonas. La relación simbiótica entre *Malassezia* y el ser humano se ha desarrollado por lo menos durante los últimos 120 años, esto con base en los primeros registros de colonización de especies de *Malassezia* en epitelios humanos. Esto ha generado que miembros del género *Malassezia* hayan perdido contenido genómico relacionado con el procesamiento de carbohidratos, amplificación de lípidos y digestión de proteínas (Grice and Dawson, 2017). Disminuyendo así su capacidad de sintetizar lípidos y por lo tanto desarrollando una elevada preferencia a colonizar zonas sebáceas, este fenómeno se conoce como evolución específica de nicho.

Las zonas del cuerpo humano en las que se encuentra una mayor abundancia de miembros de *Malassezia* son: la espalda, el occipucio, la cresta inguinal, las crestas retroauriculares, la nariz, las fosas antecubitales, las palmas de las manos, entre otras. Por otro lado, las zonas asociadas a los pies, como la planta, los dedos y el espacio interdigital también muestran presencia de este género pero en menor proporción, siendo zonas que se destacan por una mayor diversidad de hongos, entre los que se encuentran miembros de los taxa: *Chrysosporium*, *Arthrodermataceae*, *Aspergillus*, *Candida*, *Rhodotorula*, entre otros (Findley *et al.*, 2013).

Las **superficies mucosas** por sus características fisiológicas forman un nicho ecológico diferente a la piel, el cual es colonizado por miembros de hongos distintos a los encontrados en la piel, entre ellos del género *Candida*.

Las superficies mucosas llevan a cabo un rol esencial en el mantenimiento de la inmunidad y en la protección contra infecciones. Estas áreas se caracterizan por la presencia de una barrera epitelial semipermeable que es reforzada por mecanismos de inmunidad innata y adaptativa (Nagler & Feehley, 2013). Las células epiteliales han desarrollado distintos mecanismos para el control de la colonización de especies de *Candida* mediante la distinción de especies avirulentas de entre aquellas con capacidad de causar enfermedad (Verma, Gaffen & Swidergall, 2017).

Candida es un hongo que habita asintómicamente en la zona bucal de aproximadamente 50% de la población. El sobrecrecimiento de *Candida* es evitado y mediado por linfocitos T locales e interleucinas (IL-17). De esta manera, cuando la inmunidad de los individuos se encuentra comprometida, se desarrolla una infección oportunista. Los cuadros de candidiasis en su mayoría se producen por *Candida albicans*, aunque otras especies como *Candida tropicalis* y *Candida stellatoidea* también ocurren pero en menor proporción, esto en individuos sumamente inmunocomprometidos cursando algún cuadro de VIH o diabetes (Verma, Gaffen & Swidergall, 2017). Por otro lado, la prevalencia de *Candida* en mujeres también se presenta en zonas asociadas a la vagina. Existen distintas similitudes estructurales entre las superficies epiteliales de las amígdalas palatinas en la boca y del cérvix en la vagina, estas similitudes se asocian a los eventos primarios de colonización de *Candida* y su evolución a una etapa de infección en ambos sitios. La superficie de las amígdalas palatinas se encuentra cubierta por un epitelio escamoso estratificado en donde las células más externas son continuamente perdidas por acción de los movimientos bucales, y éstas son reemplazadas por células que proliferan de las capas suprabasales. Un epitelio escamoso estratificado protege la superficie de la vagina y el ectocérvix. Este mecanismo de pérdida y reemplazo celular origina que la barrera epitelial en estas zonas sea un arreglo de unicapa celular en donde la proporción de leucocitos es baja, propiciando así una zona con mayor susceptibilidad de infección por *Candida* y otros microorganismos (Southern *et al.*, 2008).

El **tracto gastrointestinal** representa un nicho ecológico atractivo para distintos microorganismos debido a su abundante flujo y concentración de nutrientes. Los nutrientes obtenidos por el hospedero suplen las necesidades energéticas de los microorganismos que ahí habitan. Sin embargo, para ser capaces de colonizar el tracto gastrointestinal, los microorganismos deben ser capaces de tolerar una serie de condiciones fisiológicas especiales como la ausencia de oxígeno, temperatura fisiológica (~37°C), contracciones peristálticas; y un pH variable, desde una elevada acidez en el estómago, hasta un pH alcalino en la mucosa intestinal (Hellen-Adams & Sur, 2017).

Distintos géneros de hongos son capaces de colonizar el tracto gastrointestinal, algunos reportes muestran miembros de los géneros *Saccharomyces*, *Candida*, y *Malassezia* gobernando las comunidades gastrointestinales de hongos (Nash *et al.*, 2017).

A pesar de que se han logrado detectar 267 especies de hongos en el conjunto de estudios de microbioma intestinal durante el periodo de los años 1917-2015, estas especies son

altamente variables entre individuos. Presentando una inter e intra variabilidad de los géneros de hongos detectados.

Debido a la naturaleza variable de los miembros del microbioma en el tracto gastrointestinal en función de variables ambientales como la dieta del hospedero y el estilo de vida, se han generado dos categorías descriptivas de los hongos que habitan este sitio: hongos residentes y no residentes. Uno de los requerimientos mínimos para que un hongo sea considerado como residente del tracto gastrointestinal es su capacidad de crecimiento y desarrollo a 37°C. Especies del género *Candida* cumplen con esta característica.

Por su parte, la presencia del género *Malassezia* en el tracto gastrointestinal llama la atención debido a que éste es un género especializado en la colonización de epitelios superficiales expuestos. Algunos autores argumentan que la presencia de este género en muestras fecales se debe a la contaminación cruzada de los alimentos al entrar en contacto con la piel durante su manipulación en el proceso de preparación y consumo (Nash *et al.*, 2017).

Dentro del dominio Eukarya, los hongos se encuentran en mayor proporción en el tracto gastrointestinal, sin embargo, en algunas poblaciones humanas, además de hongos, se presentan miembros del dominio protista, dentro de ellos destacan los géneros *Blastocystis* y *Enterocytozoon*. *Enterocytozoon bienersi* es una especie de parásitos unicelulares formadores de esporas pertenecientes al filo Microsporidia. Los miembros del grupo Microsporidia poseen la capacidad de infectar a un gran rango de linajes animales, desde gusanos hasta humanos (Stentiford *et al.*, 2016). Las células de microsporidios son estructuralmente simples, no cuentan con mitocondrias convencionales, tienen un aparato de Golgi atípico, son formadoras de esporas, y cuentan con moléculas de rRNA con características procariotas (Zelman & Corradi, 2011). La presencia de estas estructuras no convencionales en los microsporidios refleja su posible naturaleza eucariótica primitiva. La naturaleza primitiva y degenerada de estos microorganismos derivado de su parasitismo intracelular obligado ha dificultado la determinación de su taxonomía y su lugar a nivel evolutivo. Distintos autores posicionan al filo Microsporidia dentro de las líneas basales del reino Fungi (Hirt *et al.*, 1999; Hand & Weiss, 2017), esto mediante el análisis de secuencias de RNA polimerasas y otras proteínas, sin embargo, otros autores posicionan a este Filo dentro del reino Protista (Choi & Kim, 2017), esto mediante una construcción taxonómica derivada de genoma completo.

1.5. Diferencias del microbioma entre poblaciones humanas.

Todos los componentes de la microbiota humana se encuentran en constante interacción con factores del ambiente y del propio hospedero, como resultado de esto es que obtenemos la naturaleza dinámica y variable de la microbiota en función del tiempo y de los factores que a ésta afectan. Sabemos que la microbiota humana varía incluso entre individuos sanos pertenecientes a una misma población (Human Microbiome Project Consortium, 2012). El estilo de vida, la alimentación, el clima, la altitud, son factores importantes en la determinación de una estructura taxonómica específica en la microbiota al llegar a la edad adulta (Yatsunenکو *et al.*, 2012). A pesar de que una gran cantidad de factores externos determinen la composición general de la microbiota humana, existe un grupo de microorganismos centrales -core- que se comparten y se mantienen estables entre los distintos individuos dentro o fuera de una población. A nivel de bacterias se distingue un grupo central de 14 géneros compartidos entre el 95% de individuos (Falony *et al.*, 2016) de distintas nacionalidades que han participado en los mayores estudios metagenómicos como el HMP (n=300), el MetaHIT-Chino (n=145), y el MetaHIT-Europeo (n=124); además del Proyecto de la Flora Intestinal Belga-Flamenca (*Belgian Flemish Gut Flora Project*, n=1106) (Falony *et al.*, 2016), y el estudio Belga LifeLines-DEEP (*Dutch LifeLines-DEEP study*, n=1135) (Zhernakova *et al.*, 2016). Otros estudios metagenómicos de menor tamaño pero que también se han analizado para determinar este core bacteriano corresponden a lugares como Papúa Nueva Guinea (Martínez *et al.*, 2015), Perú (Obregon-Tito *et al.*, 2015), y Tanzania (Schnorr *et al.*, 2014). Los taxa bacterianos compartidos entre todas estas poblaciones estudiadas son: Lachnospiraceae, Ruminococcaceae, *Bacteroides*, *Faecalibacterium*, *Blautia*, *Roseburia*, Erysipelotrichaceae, *Coprococcus*, *Dorea*, Clostridiaceae, Hyphomicrobiaceae, Clostridiales, Veillonellaceae y *Clostridium* (Falony *et al.*, 2016).

Las comunidades microbianas del tracto gastrointestinal representan una fuente de diversidad genética y metabólica en los seres humanos (Yatsunenکو *et al.*, 2012). El ambiente tiene una influencia directa en el estilo de vida de las personas, y éste en la naturaleza y diversidad del microbioma humano. Debido a lo anterior, distintos estudios han basado su interés en la determinación del microbioma humano en relación a su biogeografía. El estilo de vida occidental ha sido adoptado por una gran cantidad de poblaciones humanas. A este estilo de vida se le atribuyen prácticas específicas como la cesárea, el uso de antibióticos, la alimentación de infantes con fórmulas sintéticas, y la

sanitización del ambiente (Tasnim *et al.*, 2017). Además de estas prácticas, otro tipo de comportamientos se asocian a dicho estilo de vida, como una menor actividad física, ocupaciones con mayor generación de estrés, sedentarismo, y un mayor consumo de tabaco y bebidas alcohólicas. Este tipo de prácticas, junto con el tipo de alimentación, se asocian a un decremento en la diversidad de la microbiota intestinal humana. El tipo de alimentos que se consumen, la cantidad, y la calidad de los mismos son características que han cambiado dentro del estilo de vida occidental. Actualmente existe un alto consumo de alimentos procesados, carnes rojas, harinas, productos altos en grasas, huevo, y granos refinados (Rodríguez-Castaño, Caro-Quintero & Reyes, 2017).

La adopción de un estilo de vida occidental se asocia al decremento en la diversidad de las comunidades microbianas en el tracto gastrointestinal humano. Asociado a esta baja diversidad de la microbiota se encuentra el aumento en la prevalencia de enfermedades inflamatorias crónicas, como la enfermedad inflamatoria intestinal, diabetes, obesidad, alergias y asma (Broussard & Devkota, 2016). El estilo de vida y dieta occidental no se restringe solamente a sociedades de Occidente, la globalización y la urbanización han incrementado la exposición mundial a este estilo de vida. Actualmente podemos observar cómo otros países han comenzado a adoptar un tipo de dieta occidental, tal es el caso de los japoneses, algunos países del centro y sur de América, además de Sudáfrica (Rodríguez-Castaño, Caro-Quintero & Reyes, 2017).

El nivel de industrialización del ambiente en el que vivimos y nos desarrollamos es uno de los factores con mayor influencia en la determinación de la estructura taxonómica y funcional de la microbiota humana (Valle *et al.*, 2016). Debido a lo anterior, instituciones y consorcios de distintas regiones y países alrededor del mundo han llevado a cabo proyectos para determinar la estructura del microbioma en diferentes poblaciones de interés. Es así que han surgido proyectos sobre microbiomas humanos en Estados Unidos (HMP); alrededor de Europa: Dinamarca y España (MetaHIT-Eur), Francia, Alemania, Italia, y Suecia (Mueller *et al.*, 2006); China (MetaHIT-Ch), Bélgica (LifeLines-DEEP), Rusia (Tyakht *et al.*, 2013), entre otros.

La cantidad de datos metagenómicos derivados de estudios de microbiomas humanos actualmente es muy grande. Esto permite realizar comparaciones de la estructura del microbioma entre distintas poblaciones. Un ejemplo de esto es un estudio liderado por el EMBL (*European Molecular Biology Laboratory*) en Heidelberg Alemania (Costea *et al.*, 2017). En este estudio se evalúan las principales diferencias en la estructura taxonómica

de bacterias y arqueas en 9 sets de datos correspondientes a poblaciones provenientes de: América, Austria, China, Dinamarca, Francia, Alemania, Kazajistán, España y Suecia. De las 71 especies de procariontes más abundantes, las que gobiernan las comunidades del tracto gastrointestinal en todas las poblaciones son: *Bacteroides vulgatus/dorei*, *Prevotella copri*, y *Eubacterium rectale*. Sin embargo, existen diferencias en la composición del microbioma entre las 69 especies restantes, estas diferencias se explican por el enriquecimiento geográfico y la generación de subespecies. En concordancia con lo anteriormente señalado - sobre la influencia del estilo de vida y alimentación Occidental - las poblaciones de América, Austria, Dinamarca, Francia, España y Alemania muestran una distribución similar entre la diversidad de especies que se presentan. Sin embargo, la población China y de Kazajistán muestran un enriquecimiento geográfico distintivo de miembros de los géneros *Eubacterium*, *Faecalibacterium*, *Ruminococcus*, y *Coprococcus*, este enriquecimiento podría estar explicado por las diferencias en estilos de vida, nivel de industrialización y hábitos alimenticios de estas dos poblaciones. Además, en cada una de las diferentes poblaciones evaluadas se observa una especialización en la estructura taxonómica mediante el surgimiento de distintas subespecies (Costea *et al.*, 2017).

Además del ambiente y el estilo de vida, los factores genéticos del propio hospedero se relacionan fuertemente con la estructura de la microbiota que posee cada individuo. La variación genética en los seres humanos explica una parte de la diversidad de los miembros del microbioma humano.

La genética del hospedero define las características físicas y químicas de las zonas del cuerpo que serán habitadas por los distintos microorganismos. Estas características incluyen principalmente el metabolismo y disponibilidad de nutrientes, así como la modulación de la respuesta inmune. A nivel del sistema inmune la regulación de la microbiota intestinal por parte del hospedero se lleva a cabo mediante mecanismos de secreción de inmunoglobulinas A (IgA), péptidos antimicrobianos y microRNAs. Como consecuencia de esto, las comunidades microbianas intestinales presentarán una estructura especializada con microorganismos capaces de metabolizar moléculas específicas del hospedero y evadir mecanismos del sistema inmune (Brantley Hall, Tolonen & Xavier, 2017). La respuesta y modulación de procesos inflamatorios del hospedero juegan un papel clave en el establecimiento de la diversidad en las comunidades bacterianas. Un proceso inflamatorio ocurre cuando el sistema inmune envía moléculas señalizadoras y células leucocitarias a un sitio de daño para controlar a los posibles invasores microbianos

y facilitar la reparación de tejido. En individuos que poseen un polimorfismo de nucleótido sencillo (SNP) del gen *MEFV* que regula la formación de pirina, una molécula involucrada en el control de procesos inflamatorios, la estructura de las comunidades de bacterias intestinales es distintiva en comparación a aquellos individuos que no portan este polimorfismo (Khachatryan *et al.*, 2008).

Los miembros de la microbiota humana se encuentran en constante interacción con las células del hospedero. En las distintas partes del cuerpo el control del crecimiento de las comunidades de microorganismos es esencial para evitar el desarrollo de cuadros de enfermedad, algunos ejemplos son las bacterias del género *Bacillus* en la nariz, los hongos del género *Candida* en la vagina, o la bacteria *Helicobacter pylori* en el estómago, estos microorganismos forman parte de la flora microbiana normal, sin embargo, cuando existe un desbalance de estas comunidades se producen cuadros de enfermedad, generando el desarrollo de sinusitis (Zhang *et al.*, 2015) , candidiasis vaginal (Southern *et al.*, 2008) y úlceras gástricas (Diaconu *et al.*, 2017), respectivamente. Para mantener a las comunidades microbianas en homeostasis con el hospedero el umbral de expresión de componentes moleculares de diferentes vías de señalización es fundamental. Los niveles de expresión de estas moléculas están relacionados con la variación genética del hospedero. Algunos ejemplos de lo anterior son el polimorfismo con pérdida de función del gen *FUT2*, el cual se sabe es un factor de riesgo en el desarrollo de la enfermedad de Chron. Este polimorfismo se asocia a la modulación energética del metabolismo del microbioma intestinal, lo cual puede inducir el desarrollo de una estructura específica de la microbiota en sujetos que lo presenten. Por otro lado, en individuos con IBD se ha demostrado que el alelo de riesgo *NOD2* está relacionado con un incremento en la abundancia relativa de miembros de la familia Enterobacteriaceae (Blekhman *et al.*, 2015).

La variación genética de los individuos de una misma especie dentro de una población se explica por múltiples factores. Dentro de estos se encuentran la tasa de mutación, deriva genética, endogamia, distribución geográfica, ascendencia genética (ancestría), entre otros (Klug *et al.*, 2013). En las poblaciones humanas, el mayor porcentaje de variación genética (85%) entre individuos ocurre dentro de las poblaciones, mientras que solo el 15% de variación existe entre poblaciones (NIH, 2017). Debido a estos factores es común encontrar individuos dentro de una población que comparten ciertos polimorfismos genéticos en mayor proporción. Como resultado de estas observaciones se han llevado a cabo estudios de asociación tipo GWAS en poblaciones humanas que presentan una ascendencia

genética particular con el objetivo de asociar miembros de la microbiota con su fondo genético. Uno de los principales descubrimientos de estos estudios es la identificación de una propiedad de heredabilidad de ciertos miembros dentro de la microbiota humana en función de las características fisiológicas del hospedero dada su ascendencia genética, de estos se destacan los géneros *Methanobrevibacter*, *Bifidobacterium*, y *Turicibacter*; las familias Clostridiaceae y Christensellaceae; y miembros de la clase Actinobacteria (Goodrich *et al.*, 2016). Estos hallazgos indican que estos taxa bacterianos son responsivos a las características genóticas del hospedero dentro de las poblaciones. De esta manera se evidencia la importancia de la genética del hospedero y de la heredabilidad de estas características en el mantenimiento de una estructura particular de la microbiota en individuos dentro de poblaciones humanas.

El estudio de las comunidades microbianas humanas alrededor del mundo nos ha permitido obtener un panorama general sobre la composición y estructura de estas comunidades. Ahora sabemos qué bacterias se comparten entre las distintas poblaciones, cuáles son más abundantes, y cuáles muestran niveles de especialización en función a la zona geográfica, estilo de vida, y ascendencia genética de la población evaluada. Sin embargo, aún queda mucho trabajo por hacer. A pesar de los grandes esfuerzos para determinar las comunidades microbianas en seres humanos, sigue existiendo un sesgo muy grande en el estudio exclusivo de bacterias, mientras que otros miembros importantes como los hongos son poco estudiados. Para obtener una visión más amplia sobre el funcionamiento de las comunidades ecológicas en el tracto gastrointestinal humano se debe realizar un esfuerzo por incluir en este tipo de estudios al mayor número de integrantes del microbioma: bacterias, arqueas, hongos, y virus, lo que permitirá generar un análisis que permita la obtención de resultados más robustos.

1.6. Principales proyectos para la determinación del microbioma humano: HMP, MetaHIT-Eur y MetaHIT-Ch

El actual entendimiento entre el microbioma humano y su asociación con estados de enfermedad y desórdenes metabólicos como en la obesidad, enfermedad inflamatoria intestinal, artritis y autismo, se ha extendido rápidamente. Las mejoras en el rendimiento y precisión de las técnicas de secuenciación de ADN de genomas provenientes de comunidades microbianas asociadas a muestras humanas, complementado con análisis de transcriptomas, proteomas, metabolomas, inmunomas y otro tipo de estudios mecanísticos

en organismos modelo, han mejorado nuestra habilidad de entender la estructura y función del microbioma en estados de salud y enfermedad (Caporaso *et al.*, 2018).

Los estudios de microbioma humano revelan que incluso entre individuos sanos, la diversidad de microorganismos que habitan los distintos ambientes como el tracto gastrointestinal, la vagina y la piel, difieren fuertemente (Human Microbiome Project Consortium, 2012). La mayor parte de esta diversidad permanece sin explicación aparente, sin embargo, la dieta, el ambiente, la genética del hospedero, y la exposición temprana a distintos microorganismos, juegan un papel importante en el establecimiento de una microbiota específica en los diferentes individuos dentro de una población. Por ello, caracterizar la ecología de las comunidades microbianas asociadas a individuos humanos se ha vuelto un tema de gran avance e investigación dentro de la comunidad científica durante los últimos años.

1.6.1. El Proyecto del Microbioma Humano (HMP)

El HMP es un proyecto establecido en el 2008 como parte de una iniciativa de los Institutos Nacionales de Salud de Estados Unidos (*National Institutes of Health*, NIH), con la misión de generar fuentes que permitan la caracterización comprensible del microbioma humano y el análisis de su rol con estados de salud y enfermedad. Dentro de este esfuerzo interdisciplinario, los centros de investigación que destacan son el *Broad Institute*, *Baylor College of Medicine*, *Washington University School of Medicine*, *J. Craig Venter Institute* y el *Data and Analysis Coordination Center*.

En la primera etapa del HMP se realizó el análisis de 300 individuos adultos **sanos, 151 mujeres y 149 hombres**, entre las edades de 18 y 40 años de edad de dos regiones geográficas de Estados Unidos: Houston, Texas y St. Louis, Missouri. De cada individuo se obtuvieron de 15 a 18 muestras de sitios específicos en función al sexo, y abarcando 5 regiones del cuerpo: **(1) la cavidad oral**: mucosa bucal, placa supragingival, placa subgingival, encía (gingiva) queratinizada, paladar duro, amígdalas palatinas, saliva, garganta y dorso lingual; **(2) la cavidad nasal**: nariz anterior; **(3) la piel**: cresta retroauricular izquierda y derecha, fosa antecubital izquierda y derecha; **(4) el tracto gastrointestinal**: heces fecales; y **(5) el tracto urogenital**: fórnix posterior, vagina media, y el introito vaginal. El estudio se realizó de manera longitudinal. Se tomaron 3 muestreos de cada individuo en 3 diferentes tiempos, el primer muestreo fue realizado en noviembre

del 2008, y la muestra final en octubre del 2010. Resultando así en un total de ~11,000 muestras primarias derivadas de los 300 individuos en estudio.

El análisis de las comunidades microbianas se llevó a cabo mediante técnicas de 16S y de secuenciación metagenómica tipo *shotgun* (MSS, *Metagenomic Shotgun Sequencing*). Para la generación de datos metagenómicos se partió de un total de 749 muestras que representan las 15 zonas del cuerpo de los 300 individuos en el estudio. Derivado del proceso de filtración y análisis de calidad, se obtuvo un conjunto de 681 muestras (**Tabla 3**) de alta calidad incluyendo a 242 individuos, 129 hombres y 113 mujeres.

Cada una de las 681 muestras contiene un total de 13 Gb (± 4.3) de datos de secuencias, generando un set total de 8.8Tb. Posteriormente, mediante una limpieza de *reads* provenientes de contaminación humana se filtró un 49% del total de secuencias. Las muestras provenientes de tejidos suaves fueron las que presentaron la mayor contaminación por ADN humano: vagina media (96%), nariz anterior (82%), y garganta (75%). La saliva también presentó una alta proporción de ADN humano (80%), mientras que las heces fecales son la zona con la menor cantidad de secuencias humanas (~1%). Después de aplicar un protocolo de control de calidad, el cual involucra la eliminación de lecturas y bases de baja calidad al igual que de secuencias humanas, se obtuvo un total de 3.8Tb de datos provenientes de las 681 muestras. **Este set de datos es uno de los utilizados en el presente trabajo de tesis.**

Tabla 3. Muestras del HMP examinadas por MSS.

Región del cuerpo	Sitio del cuerpo	Número total de muestras
Tracto Gastrointestinal	Heces fecales	136
Cavidad Oral	Mucosa bucal	107
	Paladar duro	1
	Encía queratinizada	6
	Amígdalas palatinas	6
	Saliva	5
	Placa Subgingival	7

	Placa Supragingival	115
	Garganta	7
	Dorso lingual	122
Cavidad Nasal	Nariz anterior	84
Piel	Fosa antecubital izquierda	0
	Cresta retroauricular izquierda	9
	Fosa antecubital derecha	0
	Cresta retroauricular derecha	15
Vagina	Vagina Media	2
	Fórnix posterior	53
	Introito vaginal	3
TOTAL		681

Este primer estudio clínico del **HMP**, además de caracterizar la estructura de las comunidades microbianas humanas en estados de salud y enfermedad, se tenía trazados otros 5 objetivos que incluyen:

- 1) El desarrollo de un set de referencia de 3,000 secuencias genómicas de aislados microbianos.
- 2) Un estudio metagenómico inicial basado en 16s y MSS para generar un estimado de la complejidad de las comunidades microbianas en cada zona del cuerpo.
- 3) Proyectos de demostración para determinar la relación entre enfermedad y cambios en el microbioma humano.
- 4) El desarrollo de nuevas herramientas y tecnologías para el análisis computacional, y el establecimiento de un Centro para la Coordinación y Análisis de Datos (DACC).
- 5) La examinación de las implicaciones éticas, legales y sociales, a ser consideradas en el estudio y aplicación de análisis metagenómicos de la microbiota humana.

Los primeros resultados del **HMP** se dieron a conocer en el 2012 con un artículo en la revista Nature que describe la estructura, función y diversidad del microbioma humano en individuos humanos saludables (Human Microbiome Project Consortium, 2012).

La diversidad microbiana se asocia fuertemente a condiciones de salud humana. Esta diversidad de microorganismos puede ser definida como la distribución del número y abundancia de los distintos microorganismos presentes en un hábitat determinado. Una menor diversidad de microorganismos en el tracto gastrointestinal se asocia a cuadros de obesidad (Peter J Turnbaugh *et al.*, 2007) y enfermedad inflamatoria intestinal (Qin *et al.*, 2010). Por otro lado, una alta diversidad en la vagina se asocia a una vaginosis bacteriana (Fredricks, Fiedler & Marrazzo, 2005).

Como parte de los resultados del HMP, se observa que la variación de microorganismos entre las muestras se agrupa por área del cuerpo (**Figura 2**), con la cavidad oral, la zona gastrointestinal, la piel y el hábitat urogenital separados; mientras que la cavidad nasal y la piel comparten una gran cantidad de componentes de la microbiota, seguidos de la cavidad oral. Con respecto a la diversidad, las comunidades microbianas de la cavidad oral y las heces fecales son especialmente diversas en términos de miembros de la comunidad. Mientras que las áreas vaginales analizadas muestran las comunidades más simples. Este estudio muestra que los patrones de diversidad alfa (dentro de las muestras) difieren marcadamente entre comparaciones de muestras del mismo hábitat entre distintos individuos (diversidad beta). Por ejemplo, la saliva muestra la mayor diversidad alfa de especies, pero uno de los valores más bajos de diversidad beta. Esto nos indica que la saliva es una de las zonas más ricas y diversas en cuanto a especies de microorganismos en todos los individuos, y que, además, la mayor parte de estos miembros de la comunidad se comparten entre los distintos individuos analizados. Por su parte la nariz anterior muestra el mayor nivel de beta diversidad, con un nivel de alfa diversidad intermedio. La vagina presenta el menor nivel de diversidad alfa, con un nivel relativamente bajo de diversidad beta a nivel de género. Por otro lado, la variación dentro de un mismo individuo durante los 3 muestreos a través del tiempo es consistentemente menor que la variación entre los diferentes individuos analizados, tanto a nivel de composición como de función. Por último, se observa que la composición microbiana única de cada individuo parece ser estable a través del tiempo, lo cual podría ser una característica específica de un microbioma humano asociado con un 'correcto' estado de salud.

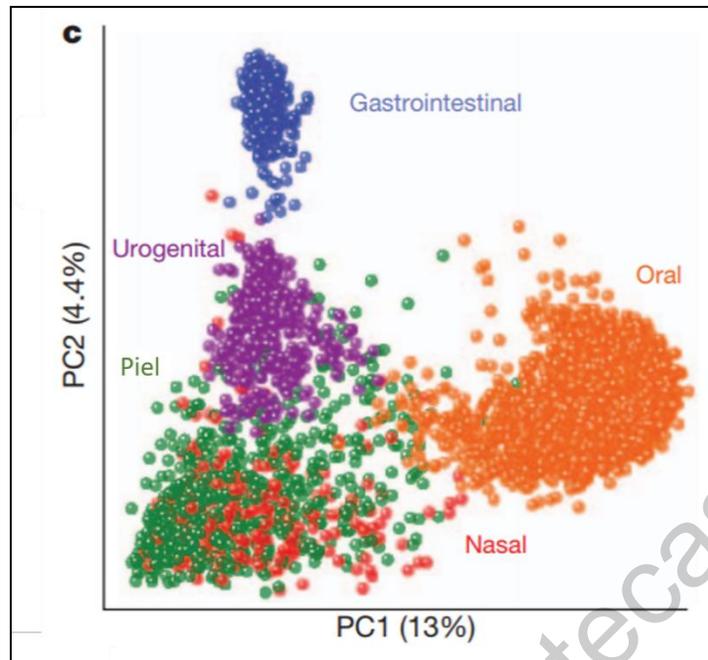


Figura 2. Gráfico de Componentes Principales entre las distintas zonas del cuerpo del HMP. Muestra la variación entre las muestras demostrando que se agrupan por sitio del cuerpo, además, que existen hábitats con miembros distantes, mientras que otros se sobrelapan (Tomado y modificado de *HMP Consortium, 2012*).

En cuanto a la composición a nivel de Filo (**Figura 3**) de las distintas regiones representativas del cuerpo, se observa que 5 de los 10 fila bacterianos mayormente representados gobiernan la estructura microbiana de estas comunidades. Como se observaba anteriormente en la gráfica de componentes principales una gran parte de la variabilidad entre las muestras se explica por la propia zona del cuerpo. En congruencia con estos resultados, observamos que las heces fecales muestran una composición distintiva, con una riqueza de miembros de los fila Firmicutes y Bacteroidetes. Mientras que las regiones asociadas a la piel, mucosa nasal y oral muestran una interrelación entre los componentes a nivel de Filo entre los cuales se observan: Firmicutes, Actinobacteria, Bacteroidetes, Proteobacteria, Fusobacteria, entre otros. Por su parte la zona vaginal, como ya se había observado en los análisis de alfa diversidad, muestra una composición simple, con miembros del filo Firmicutes gobernando la comunidad.

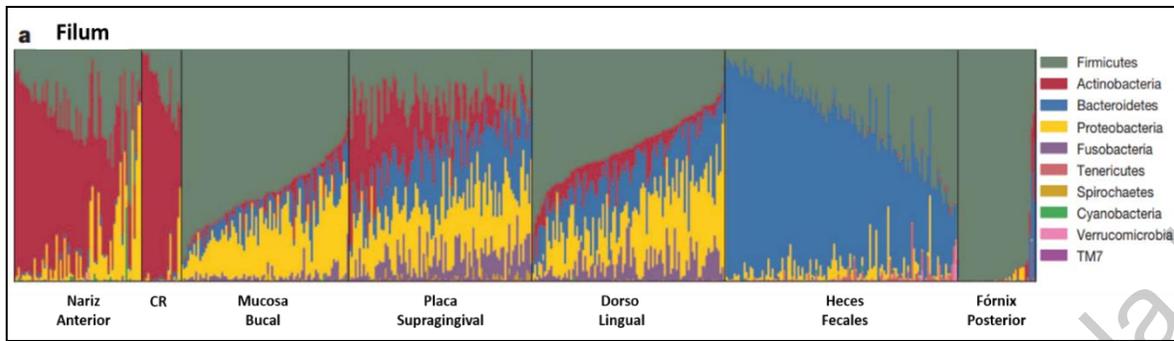


Figura 3. Análisis taxonómico a nivel de Filo en sitios del cuerpo que reflejan las regiones representativas del HMP. CR = Cresta Retroauricular (Tomado y modificado de *HMP Consortium, 2012*).

Actualmente 42 Instituciones de Salud e Investigación de Estados Unidos conforman el Consorcio de trabajo del HMP. La investigación del HMP ya no se basa únicamente en conocer la estructura y función de las comunidades microbianas en individuos sanos, sino también en estados de enfermedad; y otro gran grupo de investigaciones basadas en metabolismo (Cantarel, Lombard & Henrissat, 2012) y ecología (Li *et al.*, 2012) de estas comunidades.

1.6.2. El metagenoma del tracto intestinal humano de poblaciones Europeas, MetaHIT-Eur, asociado a enfermedad inflamatoria intestinal.

Después de los primeros resultados del HMP y de observaciones previas sobre la gran complejidad y diversidad de la microbiota del tracto intestinal humano y su particular importancia en el mantenimiento de un correcto estado de salud, surge el proyecto **MetaHIT**.

El MetaHIT-Eur es un proyecto financiado por la Comisión Europea, fue desarrollado durante un periodo que abarcó desde el 1ro de enero del 2008, hasta el 30 de junio del 2012. El principal objetivo de este proyecto se basa en el establecimiento de asociaciones entre los genes de la microbiota intestinal humana relacionados a estados de salud y enfermedad. Para ello se estudiaron dos desórdenes de salud en aumento durante los últimos años en poblaciones Europeas: Enfermedad Inflamatoria Intestinal (IBD, *Inflammatory Bowel Disease*), incluyendo Colitis Ulcerativa y Enfermedad de Crohn; y Obesidad. Durante el desarrollo del estudio participaron 124 individuos, 55 hombres y 69 mujeres, de origen Danés del Norte de Copenhague, y de origen Español. Del total de

individuos, 99 no presentan IBD, y 25 presentan IBD (21 Colitis Ulcerativa, 4 Enfermedad de Crohn) como se muestra en la **Figura 4**.

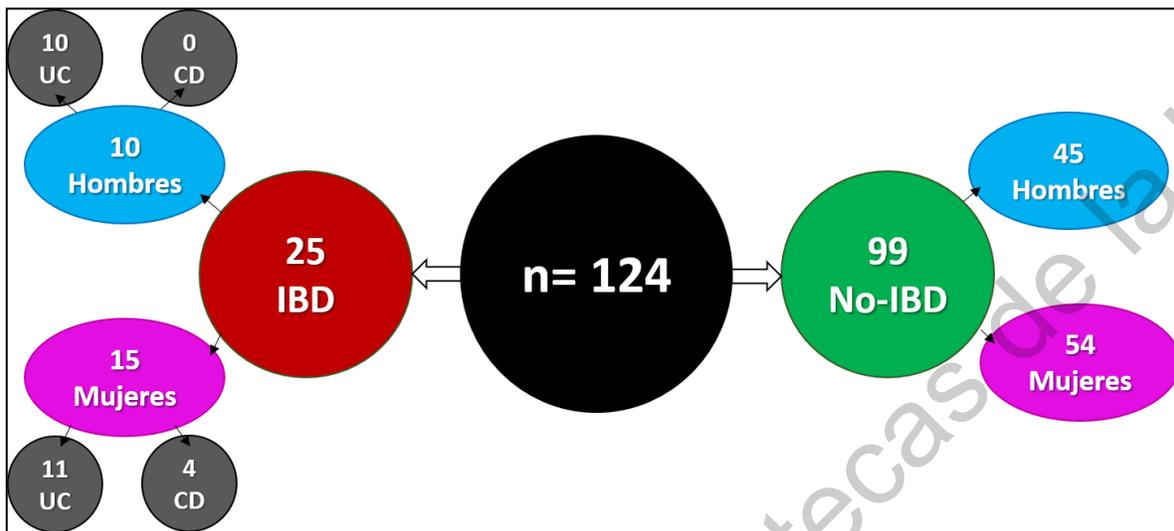


Figura 4. Número de individuos participantes en el MetaHIT-Eur y su clasificación de acuerdo a su estatus de presencia o ausencia de IBD. UC= Colitis Ulcerativa, CD= Enfermedad de Crohn.

Para alcanzar el objetivo central del proyecto y lograr establecer exitosamente una relación entre los genes de la microbiota intestinal, y su asociación con IBD y obesidad, se trazaron 4 pasos esenciales: (1) El establecimiento de un catálogo de referencia extensivo de genes microbianos presentes en el intestino humano; (2) el desarrollo de herramientas bioinformáticas para el almacenamiento, organización, e interpretación de los datos; (3) el desarrollo de herramientas que determinen qué genes del catálogo de referencia se encuentran presentes en los diferentes individuos y a qué frecuencia; y (4) reunir cohortes de individuos con estados de salud y enfermedad y determinar qué genes se encuentran presentes y con mayor representación en cada cohorte (MetaHITConsortium, 2010).

A nivel metodológico se realizó la colecta de heces fecales de los 124 individuos que participaron en el estudio. Se distinguieron 4 diferentes grupos de individuos: (1) Saludables, (2) con sobrepeso, (3) con obesidad, y (4) con IBD. El ADN total extraído de cada una de las muestras fecales generó en promedio 4.5 Gb (con un rango de 2 a 7.3 Gb) de secuencia por cada muestra. Obteniendo un set total de 576.7Gb de secuencia.

Siguiendo un proceso de ensamblaje para la generación de secuencias más largas tipo cóntigos. (**Figura 5**). Los cóntigos (que de aquí en adelante en este escrito serán llamados

“contigs”) son secuencias ensambladas continuas derivadas de lecturas de procesos de secuenciación, estas se forman por traslapes entre las secuencias. Los súper contigs (de aquí en adelante en este escrito serán llamados “scaffolds”) consisten en contigs sobrelapados con huecos de un tamaño conocido entre las secuencias (Koren & Phillippy, 2015), (Sangwan, Xia & Gilbert, 2016). Para ensamblar las secuencias en el proyecto del **MetaHIT-Eur** se utilizó SOAPdenovo (Lam *et al.*, 2013), ésta es una herramienta diseñada para el ensamblaje de *reads* muy cortos basada en Grafos de De Bruijn. Debido a que se espera una alta diversidad entre individuos, se ensambló cada muestra de manera independiente. Los *reads* que no pudieron ser ensamblados de manera independiente en cada una de las muestras se agruparon en un *pool* de *reads* en común, los cuales fueron ensamblados en conjunto como se representa en la **Figura 5**. Al menos un 42.7% de los *reads* derivados de Illumina GA fue ensamblado en un total de 6.58 millones de contigs con un tamaño mayor a 500pb, dando un total de 10.3Gb de secuencia de contigs con un tamaño N50 de 2.2kb y un rango de 12.3 a 237.6Mb. **Este set de datos de secuencias de contigs es uno de los utilizados en el presente trabajo de tesis.**

Aproximadamente el 35% de los *reads* de cualquier muestra pudo ser mapeado a los contigs de otras muestras, indicando la existencia de un *core* de secuencias comunes (Qin *et al.*, 2010).

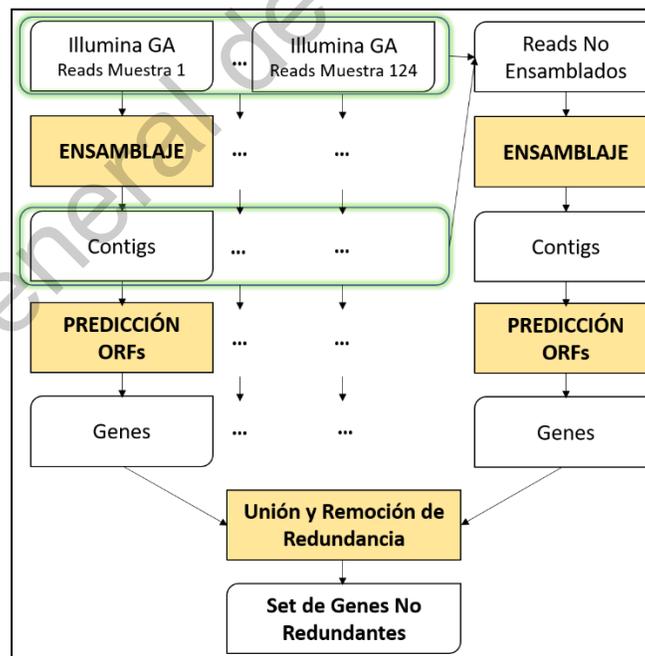


Figura 5. Flujo de trabajo en el MetaHIT-Eur para el ensamblaje, predicción de ORFs y generación de un set de genes no redundantes para cada una de las 124 muestras en estudio.

1.6.2.1. Diversidad de la microbiota intestinal humana, MetaHIT-Eur.

Los seres humanos vivimos en constante interacción con los diferentes componentes del ambiente que nos rodea, entre ellos, los microorganismos. Naturalmente, los microorganismos como las bacterias tienen una alta capacidad de adaptación a distintos tipos de ambientes. El ser humano por su parte provee una gran cantidad de microambientes distintos que han sido aprovechados por distintas comunidades microbianas, estableciendo así lo que hoy conocemos como el microbioma humano. La microbiota humana es compleja y dinámica, en particular la microbiota intestinal (Peter J Turnbaugh *et al.*, 2007), la cual tiene una profunda influencia en la fisiología y nutrición humana. Los miembros de la microbiota intestinal realizan funciones de digestión de alimentos, y de síntesis de vitaminas y aminoácidos esenciales para el cuerpo humano (Rowland *et al.*, 2018). Los cambios en la estructura normal de los componentes de la microbiota (disbiosis) dentro de un individuo generan afecciones del sistema digestivo, es por ello que es importante entender la estructura de estos cambios y cómo pueden estar aparentemente asociados con desórdenes crónicos como en el síndrome de IBD y la obesidad.

Para entender el impacto entre la microbiota intestinal y la salud humana se vuelve crucial evaluar el potencial genético de ésta. Como parte del proyecto **MetaHIT-Eur** se realizó la secuenciación metagenómica, ensamblaje y caracterización de 3.3 millones de genes microbianos no redundantes, derivados de 576.7 gigabases de secuencias, provenientes de 124 muestras de individuos europeos. El set de genes microbianos es ~150 veces más grande que el set completo de genes humano. Cerca del 99% de los genes identificados son bacterianos. En la cohorte de 124 individuos se lograron caracterizar entre 1,000 y 1,150 especies de bacterias prevalentes. Cada individuo porta un aproximado de 160 de estas especies, las cuales son altamente compartidas entre individuos (Qin *et al.*, 2010).

En los primeros resultados del HMP se observó que los miembros pertenecientes a dos Fila bacterianos: Bacteroidetes y Firmicutes, constituyen el 90% del total de la comunidad bacteriana en el tracto gastrointestinal (**Figura 3**). En este estudio (**MetaHIT-Eur**) se logró identificar un *core* de 57 especies bacterianas compartidas entre los 124 individuos analizados, las cuales en su mayoría pertenecen a los dos fila ya mencionados. Además, un análisis de asignación filogenética a partir de la base de datos NCBI-nr (*non redundant*) de secuencias proteicas determinó que el 99.96% del total de secuencias asignadas del **MetaHIT-Eur** pertenece a los Dominios Bacteria y Archaea, mostrando la predominancia

de estos en el tracto intestinal humano. Observaciones de estudios anteriores indican que la diversidad bacteriana en individuos que sufren de IBD es menor en comparación a individuos saludables (Manichanh *et al.*, 2006). Como parte de los resultados del **MetaHIT-Eur** esta observación se relaciona fuertemente al hecho de que los individuos con IBD de este estudio presentan 25% menos genes bacterianos que aquellos que no sufren de IBD (Qin *et al.*, 2010). Por otra parte, un análisis de componentes principales muestra la agrupación de los grupos de individuos con base en las diferencias de la abundancia de especies bacterianas en individuos con IBD (Colitis Ulcerativa y Enfermedad de Crohn) e individuos sanos (**Figura 6**).

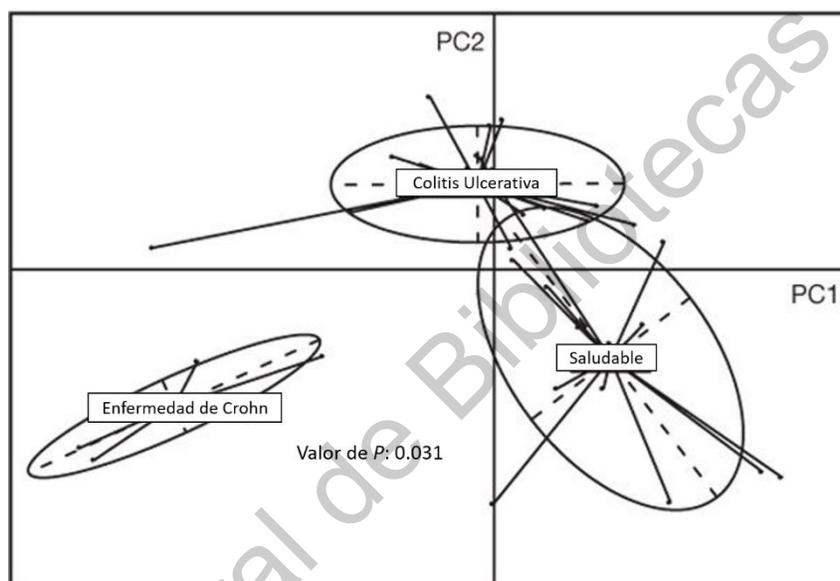


Figura 6. Análisis de componentes principales sobre la abundancia diferencial de especies bacterianas en individuos con IBD e individuos sanos (Tomada y modificada de Qin *et al.*, 2010).

1.6.3. El metagenoma del tracto intestinal humano de una población China, **MetaHIT-Ch**, asociado a diabetes tipo 2

El proyecto **MetaHIT-Eur** y **MetaHIT-Ch** forman parte del Consorcio Internacional del Microbioma Humano (*International Human Microbiome Consortium*, IHMC). Este consorcio tiene el objetivo de trabajar bajo un conjunto de políticas y principios comunes en el estudio y entendimiento de la función del microbioma humano y su relación con el mantenimiento de estados de salud y causas de enfermedad, utilizando este conocimiento para mejorar nuestra habilidad en la prevención y tratamiento de enfermedades (IHMC, 2008). Por ello, como parte de una iniciativa en común con el **MetaHIT-Eur**, se llevó a cabo un estudio

semejante (**MetaHIT-Ch**) en un grupo poblacional de individuos chinos. Este último estudio con características distintivas se realizó con el objetivo de conocer las comunidades microbianas intestinales asociadas a desórdenes de diabetes tipo 2 (T2D).

El **MetaHIT-Ch** es un estudio de casos y controles que involucra la participación de 368 individuos provenientes de la Ciudad de Shenzhen, al sur de China. Incluye 345 muestras fecales utilizadas para un estudio de asociación de metagenomas completos (MGWAS, *Metagenome Wide Association Study*) y 23 muestras adicionales utilizadas para la validación de los resultados en individuos con T2D. Como parte de este estudio, un grupo de muestras de 145 individuos (71 casos y 74 controles) (**Figura 7**) fue utilizado para una secuenciación metagenómica tipo *shotgun*, obteniendo un promedio de 2.61Gb (15.8 millones) de *reads* pareados por muestra. Siguiendo un proceso de filtrado, análisis de calidad y ensamblaje similar al del MetaHIT-Eur (**Figura 5**), se obtuvo un conjunto de 4.01 millones de contigs ensamblados mayores a 500 pb de longitud. **Este conjunto de datos de contigs es uno de los utilizados en el presente trabajo de tesis.** Mediante un análisis de asignación taxonómica cerca del 61% de los *reads* de las 145 muestras pudieron ser asignados a un taxón exitosamente.

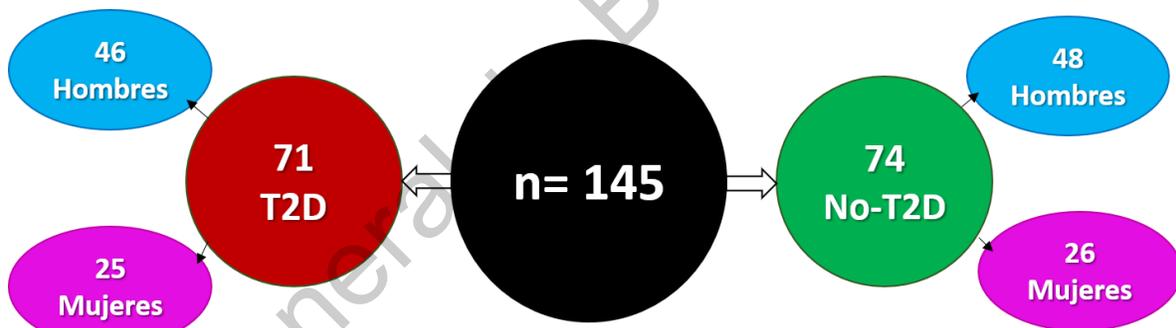


Figura 7. Número de individuos participantes en el MetaHIT-Ch y su clasificación de acuerdo a la presencia o no de T2D.

1.6.3.1. El microbioma intestinal humano y su relación con T2D.

La diabetes tipo 2 es un trastorno metabólico caracterizado por la desregulación del metabolismo de carbohidratos, lípidos y proteínas, lo cual resulta en una secreción de insulina alterada, resistencia a la insulina, o una combinación de ambas. La T2D es un problema de salud pública global en expansión con una fuerte relación a la actual epidemia de obesidad. Los individuos con T2D se encuentran en un mayor riesgo de desarrollar

complicaciones microvasculares como la retinopatía, nefropatía y neuropatía; además de desórdenes cardiovasculares. Los factores ambientales (obesidad, dieta poco saludable, e inactividad física) y genéticos contribuyen a las múltiples alteraciones fisiopatológicas responsables de la homeostasis de la glucosa en individuos con T2D. Cerca del 90% de los casos de diabetes registrados corresponden a T2D, mientras que el 10% restante engloba a la diabetes mellitus tipo 1 y a la diabetes gestacional (DeFronzo *et al.*, 2015).

La reciente investigación metagenómica de comunidades microbianas intestinales ha cambiado nuestro entendimiento sobre enfermedades humanas y su potencial impacto médico. Desde una perspectiva taxonómica y funcional, la microbiota humana podría estar ligada a la contribución etiológica de diferentes enfermedades complejas, entre ellas, la diabetes tipo 2 (Qin *et al.* 2012). Los metabolitos generados por la microbiota intestinal contribuyen a la integridad de la barrera intestinal. Por su parte, una barrera intestinal comprometida resultado de una alteración en los componentes de la microbiota intestinal conduce a la filtración de mediadores inflamatorios dentro de la circulación sistémica, resultando en un aumento en la resistencia a la insulina (Upadhyaya & Banerjee, 2015).

La obesidad predispone el desarrollo de T2D y enfermedades cardiovasculares. Estas dos patologías son parte de lo que se conoce como el síndrome metabólico. La perturbación de la composición y actividad de la microbiota se encuentra involucrada en la emergencia del síndrome metabólico.

Una de las actividades con el mayor potencial de cambio en las comunidades microbianas intestinales es una buena alimentación acompañada del consumo regulado de probióticos y prebióticos. Los probióticos son microorganismos vivos que administrados en cantidades adecuadas confieren un beneficio a la salud del hospedero (Hill *et al.*, 2014). Mientras que los prebióticos son un sustrato que es selectivamente utilizado por los microorganismos del hospedero que confiere un beneficio a su salud (Gibson *et al.*, 2017). En este sentido, una gran parte de proyectos de casos y controles asociados a la identificación de los componentes estructurales de la microbiota intestinal humana se realizan con este objetivo en particular: la identificación de microorganismos asociados a un buen estado de salud o a una mejora en el estado de enfermedad que puedan ser utilizados como probióticos.

Durante los últimos años, y como resultado de la gran cantidad de proyectos de microbiomas intestinales humanos, se han identificado microorganismos que cumplen con esta descripción.

Los microorganismos benéficos de siguiente generación son aquellos que han sido identificados mediante el uso de tecnologías de secuenciación de nueva generación y no por métodos de cultivo convencionales (*Next-Generation Sequencing*, NGS).

Akkermansia muciniphila es una de las especies bacterianas más abundantes de la microbiota intestinal (0.5 – 5% del total de bacterias). Es una bacteria degradadora de mucina. La mucina es una glicoproteína producida y degradada específicamente en el colon y que forma parte de la mucosa intestinal. Interesantemente, *A. muciniphila* se encuentra en menor abundancia en la microbiota intestinal de ratones obesos y diabéticos (Cani & De Vos, 2017). En humanos, la abundancia de *A. muciniphila* disminuye en distintas situaciones patológicas como en la obesidad, diabetes tipo 2, enfermedad inflamatoria intestinal, hipertensión y enfermedades del hígado. Por otra parte, se encuentra el caso de la bacteria *Christensenella minuta*, una bacteria asociada a individuos delgados identificada como parte de los esfuerzos del HMP. Estudios en ratones con sobrepeso alimentados con *C. minuta* muestran una disminución de peso (Rosa *et al.*, 2017).

1.7. Aportes de estudios metagenómicos recientes

A pesar de los grandes esfuerzos de estos consorcios y de otros grandes proyectos, aún se desconoce una gran parte del repertorio de microorganismos asociados al cuerpo humano, principalmente a la zona gastrointestinal. Esto se sabe debido a que en proyectos como el **HMP** y en ambos estudios del **MetaHIT** una elevada proporción (entre el 30% al 50%) de lecturas y secuencias ensambladas generadas por secuenciación metagenómica no han podido ser mapeadas a las diferentes bases de datos de microorganismos, evidenciando así que existen miembros de la microbiota humana de los cuales aún no se cuenta con genomas de referencia para su identificación. Elucidar la totalidad de microorganismos que forman parte del microbioma es de vital importancia debido a que la presencia de diferentes géneros y especies de bacterias, virus y hongos se asocian con el desarrollo de fenotipos de importancia médica en el hospedero. Para abordar esta problemática recientemente han sido publicados algunos trabajos en los cuales se ensamblan y recuperan genomas microbianos, principalmente bacterias, provenientes del total de conjuntos de datos de metagenomas públicos disponibles. Uno de ellos es el trabajo de Donovan H. Parks y colaboradores (Parks *et al.*, 2017) en donde se reporta la reconstrucción de 7,903 genomas de bacterias y arqueas derivado del análisis de 1,550 metagenomas. A pesar de que este estudio se basó principalmente en el ensamblaje de

genomas microbianos provenientes de metagenomas no asociados a humanos, los resultados nos permiten contextualizar la cantidad de bacterias que permanecen aún sin ser identificadas. Otro estudio con una metodología similar pero asociado a metagenomas humanos es el de Alexandre Almeida y colaboradores (Almeida *et al.*, 2019), en donde logran ensamblar 92,143 genomas microbianos derivado del análisis de 11,850 microbiomas del tracto gastrointestinal humano, incrementando así en ~281% la diversidad filogenética asociada a la microbiota intestinal. De estos genomas ensamblados 1,952 corresponden a nuevas especies candidatas de bacterias. Además, logran aumentar en 200% el número de genomas microbianos de poblaciones poco representadas como son el caso de África y Sudamérica, al aplicar técnicas de ensamblaje *de novo* de los sets de metagenomas disponibles derivados de dichas poblaciones. De acuerdo a distintos autores, estos dos estudios han permitido aumentar significativamente el número de taxa bacterianos existentes en las diferentes bases de datos, lo que permitirá expandir substancialmente el árbol de la vida, principalmente a nivel de microorganismos.

La mayoría de los estudios sobre microbiomas humanos se han realizado en poblaciones europeas y americanas. Esto se debe a que los consorcios más grandes de investigación sobre este tema provienen de Estados Unidos y Europa. Generando así una elevada representación de metagenomas provenientes de estas poblaciones en las diferentes bases de datos. Gracias al estudio de Almeida y colaboradores citado en el párrafo anterior se incrementó el número de genomas microbianos de poblaciones poco representadas como África y Sudamérica. De igual forma, gracias al trabajo de Almeida y colaboradores se logró disminuir la proporción de *reads* o contigs que no recibían una asignación taxonómica en estudios de microbiomas humanos. En su conjunto esto permitirá observar nuevos taxa de microorganismos que se encuentran en asociación con seres humanos, principalmente a nivel del tracto gastrointestinal.

Los componentes del micobioma muestran una naturaleza variable. En distintos estudios longitudinales se ha observado que existe una intra-variabilidad elevada entre los distintos géneros de hongos que se pueden detectar en un individuo en un momento determinado, y que esta composición varía fuertemente en el mismo individuo al realizarse un segundo muestreo en una temporalidad diferente (Nash *et al.*, 2017). Otros estudios además muestran una alta inter-variabilidad mediante comparaciones del perfil taxonómico fúngico entre individuos dentro de un grupo de estudio, mostrando que el número de individuos que explica la mayor parte de la diversidad de hongos es muy bajo en comparación al total de

individuos participantes. Un ejemplo de esto es el estudio de Ghannoum y colaboradores (Ghannoum *et al.*, 2010), quienes se enfocaron en estudiar la diversidad del microbioma oral en un grupo reducido de 20 individuos, detectando 85 géneros de hongos, de los cuales, 39 géneros se explican por la diversidad de hongos que posee 1 solo individuo. Estas observaciones nos muestran la complejidad y la variabilidad de las comunidades microbianas que habitan en el cuerpo humano. Si bien, actualmente los grandes consorcios dedicados a la investigación de microbiomas humanos han establecido protocolos de investigación más detallados, es indispensable el control de la mayor cantidad de variables en este tipo de estudios, y aún más en aquellos que se enfocan en generar diferencias en la composición en grupos de casos y controles, de esta manera se disminuye el error provocado por variables no controladas y se aumenta la certeza de que dichos cambios observables se deben con mayor probabilidad a la presencia o ausencia de la característica en estudio.

1.8. Tecnologías para determinar la diversidad del microbioma.

El estudio del microbioma humano ha avanzado drásticamente durante los últimos años gracias al surgimiento y mejora de las nuevas tecnologías de secuenciación. Estas tecnologías actualmente nos permiten obtener un mejor panorama de las comunidades microscópicas que viven dentro y sobre el cuerpo humano, tanto a nivel de prevalencia como de diversidad (Clemente *et al.*, 2012). Anteriormente, las aproximaciones metodológicas para la detección e identificación de microorganismos se basaban en técnicas de cultivo, pruebas bioquímicas, microscopía y cultivo en medios selectivos (Huffnagle & Noverr, 2013), sin embargo, la mayor parte de las bacterias y hongos son incultivables en medios de cultivo convencionales, generando una pérdida de información en estudios de este tipo.

Una de las principales herramientas que ha permitido estudiar las secuencias genómicas de comunidades microbianas completas es la secuenciación de última generación (NGS, *next generation sequencing*). Estas herramientas permiten la secuenciación de ADN de alto rendimiento. De esta manera millones o billones de hebras de ADN pueden ser secuenciadas en paralelo, permitiendo una mayor calidad y minimizando la necesidad de incluir métodos de clonación de segmentos de ADN como era el caso en la secuenciación tipo Sanger. Debido al progreso en estas tecnologías de secuenciación se ha podido disminuir el costo por megabase secuenciada e incrementar el número y diversidad de genomas secuenciados (Goodwin, McPherson & McCombie, 2016), como en el caso de los

estudios de metagenomas. Además de las características anteriores, las tecnologías de NGS presentan una diversidad en la estructura y tamaño de secuencias que se pueden obtener, por un lado, Illumina domina las tecnologías de secuenciación de lecturas cortas, mientras que otras tecnologías como Oxford NanoPore permiten la generación de secuencias largas (~200kb).

Actualmente existen dos principales aproximaciones metodológicas para estudiar el perfil microbiano de una comunidad: los estudios metagenómicos, y la secuenciación de marcadores como 16S, 18S o secuencias espaciadoras transcritas internas (ITS, *Internal Transcribed Spacer*). (Tang *et al.*, 2015). En este sentido, la metagenómica puede ser definida como el análisis directo de genomas contenidos en una muestra ambiental sin la necesidad de generar cultivos clonales (Sharpton, 2014). Una de las mayores ventajas que nos ofrece la metagenómica es el conocimiento de la abundancia de los componentes de la comunidad, y de su estructura mediante el análisis de índices de diversidad: alfa (α), beta (β) y gama (γ). Los índices de diversidad anteriores son variables descriptivas fundamentales a nivel ecológico (Jost, 2005). De manera general lo que cada índice nos indica es lo siguiente: la diversidad alfa es la diversidad a nivel local en una comunidad cerrada; la diversidad beta es la medida de qué tan diferentes son dos muestras tomadas de una misma área; y la diversidad gama es la diversidad total de una región a larga escala que contiene diversas comunidades (Krebs, 1999). La escala de los índices de diversidad puede ser ajustada de acuerdo a las comunidades de estudio, desde comunidades de organismos superiores hasta de microorganismos. Por otra parte, conocer la composición de una comunidad en estudio es importante para la determinación de un análisis funcional adecuado que nos permita reconstruir vías metabólicas, de esta manera es posible evaluar el posible impacto que estas comunidades de microorganismos tienen en la fisiología del hospedero.

La secuenciación metagenómica tipo *shotgun* es una herramienta completa que permite obtener datos de comunidades microbianas enteras. Este método es costoso y necesita de recursos computacionales potentes para su análisis, sin embargo, el volumen de datos producidos y la calidad en el análisis de estos es mucho mayor, permitiendo así dar respuesta a una gran cantidad de preguntas biológicas bajo estudio. Una alternativa a esta última metodología es la amplificación mediante reacciones en cadena de la polimerasa (PCR, *Polymerase Chain Reaction*) de marcadores blanco específicos como las regiones ribosomales variables 16S y 18S, así como las regiones espaciadoras ITS. Estas regiones

son amplificadas y secuenciadas con el fin de identificar las poblaciones de bacterias u hongos (de acuerdo al marcador utilizado) a través de una comparación con bases de datos de referencia como SILVA (Quast *et al.*, 2013) en el caso de 16S y 18S, y UNITE (Nilsson *et al.*, 2018) en el caso de ITS (Tang *et al.*, 2015).

Actualmente existen muchas bases de datos con información de secuencias metagenómicas como las mencionadas anteriormente: HMP, MetaHIT-Europeo y MetaHIT-Chino. Estos proyectos cuentan con información derivada de procesos de ensamblado tipo contigs o scaffolds.

1.9. Biodiversidad e índices de diversidad en el análisis de datos metagenómicos

La diversidad biológica es la variabilidad entre los organismos vivos de todas las fuentes, incluidos, entre otros, los ecosistemas terrestres, marinos y otros ecosistemas acuáticos y los complejos ecológicos de los que forman parte; esto incluye la diversidad dentro de las especies, entre las especies y de los ecosistemas (CBD, 1992). Se distinguen tres niveles de biodiversidad: (1) genética, (2) de especies y (3) de ecosistemas. El interés de los estudios de biodiversidad en sistemas naturales se basa en representar la variabilidad de todos o alguno de estos tres componentes en la comunidad en estudio.

La biodiversidad puede verse como una propiedad multidimensional de los sistemas naturales (Morris *et al.*, 2014). El cuerpo humano contiene una gran cantidad de microorganismos, los cuales varían de individuo a individuo de acuerdo a diversos factores del ambiente o del propio hospedero. Conocer cuál es la naturaleza y la dinámica de cambio del microbioma humano es parte fundamental en estudios que evalúan grupos de casos y controles.

Con el objetivo de describir las propiedades generales de las comunidades ecológicas que nos permitan comparar distintas regiones, taxa, y niveles tróficos se han creado diferentes medidas de biodiversidad. Los índices de diversidad nos permiten obtener un estimado cuantitativo de la variabilidad biológica que puede ser usado para comparar entidades biológicas que difieren entre ellas en espacio o tiempo.

La diversidad usualmente implica la medida de dos propiedades fundamentales en un contexto ecológico: (1) el número de especies (riqueza) y (2) la equidad entre las especies.

La riqueza de especies es una medida del número total de especies en una comunidad, mientras que la equidad de especies expresa qué tan igualitaria es la distribución de cada una de ellas dentro de la población.

Estas dos medidas permiten obtener una primera visión sobre la estructura de la comunidad en estudio (riqueza y equidad), sin embargo, para conocer la identidad de los miembros de la comunidad es necesario hacer uso de índices taxonómicos. Este tipo de índices clasifican a los miembros de la comunidad de acuerdo a su relación taxonómica, reflejando así la distancia taxonómica promedio entre dos organismos tomados de manera aleatoria de la muestra en estudio. Es importante mencionar que los índices de riqueza y equidad (diversidad α) pueden ser utilizados sin distinción en el estudio comparativo espacial de diferentes comunidades. Dentro de los índices de diversidad alfa más utilizados en ecología encontramos el índice de entropía de Shannon y el índice de dominancia de Simpson. Por su parte, dentro de los índices de diversidad beta, el más utilizado es el índice de Bray-Curtis (McHugh *et al.*, 2018).

El índice de entropía de Shannon cuantifica la falta de certeza en la predicción de la identidad de un individuo que es tomado de manera aleatoria de un set de datos (Grabchak *et al.*, 2017). Así, mientras mayor sea la diferencia entre el tipo de individuos dentro de una población, y más proporcionales – equitativas - sean las abundancias de estos, mayor será el grado de incertidumbre al intentar predecir la identidad del individuo tomado de manera aleatoria. La fórmula para estimar la entropía de Shannon (Oksanen, 2018) es la siguiente:

$$H = - \sum_{i=1}^S p_i \log_b p_i$$

En donde p_i es la proporción de individuos pertenecientes a la especie i , mientras que S el número de especies en la población de interés.

El índice de dominancia de Simpson es una medida que iguala la probabilidad de que dos entidades tomadas de manera aleatoria de un set de datos de interés representen el mismo tipo. Sin embargo, han surgido variaciones a este índice como el inverso de Simpson, que por su parte representa la probabilidad de que dos entidades tomadas de manera aleatoria de un set de datos pertenezcan a diferentes especies. Esta variación es la más utilizada en estudios ecológicos (Morris *et al.*, 2014). La fórmula para estimar el índice de Simpson inverso (Oksanen, 2018) es la siguiente:

$$D_1 = 1 - \sum_{i=1}^S p_i^2$$

En donde S es la riqueza o el número total de especies distintas dentro de la muestra. Esta ecuación también es igual a la media aritmética ponderada de las abundancias proporcionales (p_i) de las diferentes especies.

El índice de Bray-Curtis es utilizado para cuantificar las diferencias en la estructura entre dos poblaciones distintas cuando se cuenta con datos de abundancia de sus integrantes. En el caso de datos de presencia/ausencia se tendría que utilizar una variante de Bray Curtis denominada Sørensen. Los valores del índice de Bray Curtis se encuentran restringidos a un rango de 0 a 1. Cuando la distancia es cero, dos sitios son completamente distintos para cada especie (Kindt, 2008). Cuando la distancia iguala a 1 son completamente similares. En el caso del índice de disimilitud el razonamiento es el inverso: “Disimilitud = 1 – Similitud”. La fórmula de disimilitud es la siguiente:

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

En donde “ ij ” son dos sitios, “ S_i ” es el número total de especímenes contados en el sitio i , “ S_j ” es el número total de especímenes contados en el sitio j , “ C_{ij} ” es la suma que representa solo a los recuentos menores para cada especie encontrada en ambos sitios. De esta manera si una especie se encuentra presente en ambos sitios, solo se tomará en cuenta para la sumatoria “ C_{ij} ” el número de aquel sitio en donde se encuentra en menor abundancia.

1.10. Herramientas de alineamiento de secuencias y bases de datos

Uno de los primeros pasos para el entendimiento de datos generados por métodos de secuenciación masiva, es el alineamiento. A partir de este, se realizan estudios posteriores que involucran la generación de análisis filogenéticos, de función, de estructura y predicción proteica, entre otros.

En paralelo con el crecimiento de las nuevas tecnologías de secuenciación masiva, una gran cantidad de algoritmos que alinean los datos generados por éstas han sido desarrollados. Una de las cosas que no ha cambiado durante los últimos años, es la salida de datos crudos generada por los equipos de secuenciación, la cual es por sí misma poco informativa. Con el fin de darle un sentido biológico a los datos obtenidos, estos deberán pasar por una tubería de análisis informáticos *-pipeline-*. La cual nos permite obtener resultados informativos que nos acercarán cada vez más a resolver las preguntas biológicas que indujeron la generación de los datos.

1.10.1. Alineamiento de secuencias

El alineamiento es el proceso por el cual se determina la localización más probable de una lectura dentro de una referencia de secuencias genómicas. Dicha referencia es seleccionada en función al conocimiento previo de la fuente original de la lectura (Flicek & Birney, 2009). Las herramientas actuales realizan el alineamiento de secuencias de ADN de especies de interés a genomas de referencia ensamblados de dichas especies.

Uno de los programas más utilizados para la búsqueda de similitud de secuencias es la herramienta BLAST (*Basic Local Alignment Search Tool*) (Altschup *et al.*, 1990). La búsqueda de similitud de secuencias puede llevar a la identificación de proteínas o genes homólogos, esto con base en análisis estadísticos que nos permitan inferir una ascendencia común entre las secuencias comparadas (Dalal & Atri, 2014). Debido a lo anterior, evaluar la similitud en secuencias de ADN se ha vuelto un entendimiento crucial para la biología evolutiva.

BLAST es una herramienta de búsqueda de alineamiento local. El alineamiento de secuencias es una forma de comparación de secuencias de ADN, ARN o aminoácidos, que nos permite identificar regiones de similitud que pueden ser una consecuencia de relaciones funcionales, estructurales, o evolutivas entre las secuencias. La búsqueda de similitud de secuencias generalmente puede ser clasificada como global o local de acuerdo al algoritmo utilizado para su búsqueda. Las herramientas de búsqueda de similitud global se basan en el algoritmo Needleman-Wunsch, el cual optimiza el alineamiento completo entre dos secuencias, incluyendo así el alineamiento de secuencias largas con similitud baja. Este tipo de alineamiento es más útil cuando las secuencias a comparar comparten mayor similitud y son del mismo tamaño. Por su parte, las herramientas de alineamiento local como BLAST, hacen uso del algoritmo Smith-Watermann, optimizando la búsqueda de

subsecuencias relativamente conservadas; alineamientos más cortos con mayor similitud (Altschup *et al.*, 1990) (**Figura 8**). Las medidas de similitud local se prefieren para su búsqueda en bases de datos, en donde las secuencias pueden ser comparadas contra secuencias parciales de genes. Además de la generación de alineamientos, BLAST provee información estadística acerca del alineamiento, tales como el valor esperado (*expected value*).

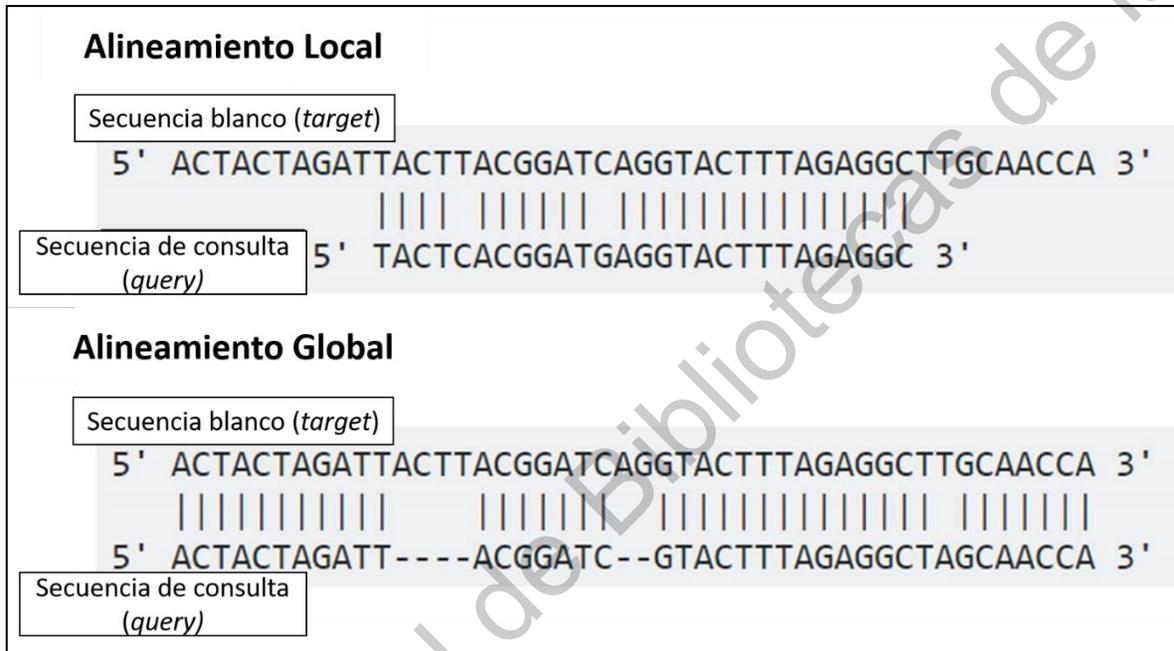


Figura 8. Representación de un alineamiento local y un alineamiento global (Tomada y modificada de Altschup *et al.*, 1990).

Los programas de alineación normalmente siguen un procedimiento que involucra múltiples pasos para realizar el mapeo con precisión de una secuencia. El primer paso es el uso de técnicas heurísticas en donde el objetivo es identificar rápidamente un pequeño set de zonas en la secuencia de referencia en donde es más probable que se localice el mejor mapeo. Una vez que se ha definido el *subset* de zonas con un posible mapeo, se realiza un alineamiento más lento y más preciso con el uso del algoritmo Smith-Waterman. Correr estos algoritmos de alta precisión en todas las posibles regiones en donde podría existir un mapeo es computacionalmente inviable. Además, todos los programas de alineamiento implementan una “política de mapeo” que rigen los aspectos clave de rendimiento de dicha herramienta. Independientemente del algoritmo que se utilice, una regla general es que existe un equilibrio entre velocidad y sensibilidad (Flicek & Birney, 2009).

1.11. Herramientas para el análisis taxonómico de comunidades microbianas. MEGAN.

Una gran cantidad de estudios han documentado las diferencias en composición de comunidades microbianas asociadas a sus hospederos durante estados de salud y enfermedad. Para un creciente número de enfermedades, un microbioma alterado no es solamente un marcador de enfermedad, sino también un contribuyente activo a la patología (Chassaing *et al.*, 2012). Debido a la importancia en la determinación de los componentes taxonómicos de las comunidades microbianas y su relación con distintas patologías, han surgido un gran número de herramientas computacionales con el fin de evaluar con mayor detalle y precisión las comunidades microbianas derivadas de datos metagenómicos (Goodrich *et al.*, 2014).

La determinación de la estructura dentro de las comunidades microbianas asociadas a humanos es crucial para el entendimiento ecológico de éstas cuando se presentan cuadros de enfermedad. Para este fin se han creado distintas herramientas de asignación taxonómica, que con el surgimiento de más y mejores tecnologías de secuenciación, han sido mejoradas a través del tiempo. Las características que se buscan en una herramienta de asignación taxonómica es rapidez y sensibilidad.

Dentro de las herramientas de asignación taxonómica podemos distinguir dos grupos: (1) aquellas que basan su clasificación en *sets* de genes blanco (ej. QIIME, mOTU, MetaPhyler, MetaPhlAn), y (2) aquellas que utilizan todas las secuencias disponibles – *reads* o secuencias ensambladas – en el *set* de datos para realizar la determinación (ej. MG-RAST, Kraken, MEGAN) (Lindgreen, Adair & Gardner, 2016).

1.11.1. MEGAN.

La metagenómica es el estudio del contenido genómico de una muestra de organismos obtenidos de un hábitat común utilizando secuenciación dirigida (marcadores) o aleatoria (*shotgun*). El principal objetivo de este campo de estudio se basa en entender el contenido y función de la diversidad microbiana en un ambiente determinado. MEGAN (*MEtaGenome Analyzer*) es una herramienta que permite el análisis de datos metagenómicos, creada por Daniel Huson en el año 2007 (Huson, Auch, Qi, & Schuster, 2007). El funcionamiento de MEGAN se basa en el (1) pre-procesamiento, (2) filtración, (3) asignación taxonómica y (4) representación visual de los datos metagenómicos (**Figura 9**). En un paso de pre-

procesamiento, el *set* de lecturas de ADN o contigs es comparado contra una base de datos de referencia de interés que contiene secuencias ensambladas y anotadas de distintos organismos, usando para ello herramientas de comparación basadas en alineamientos locales como BLAST. Después de este análisis de comparación, las herramientas de MEGAN son utilizadas para estimar y explorar el contenido taxonómico del set de datos, usando como base de referencia el árbol taxonómico de NCBI para generar una representación visual de un árbol taxonómico de los datos metagenómicos evaluados.

El funcionamiento de MEGAN está basado en un algoritmo de “Ancestro común más bajo” (*Lowest Common Ancestor algorithm, LCA*). Este tipo de algoritmos nacen junto con la teoría de grafos, haciendo uso de nodos y uniones (aristas). En principio, el ancestro común más bajo entre dos nodos x y y , se define como el nodo más bajo que tiene a x y y como descendientes (**Figura 10**). Este tipo de organización jerárquica en los sistemas biológicos siguen un principio central basado en la historia evolutiva de distintos grupos de organismos (Federhen, 2012).

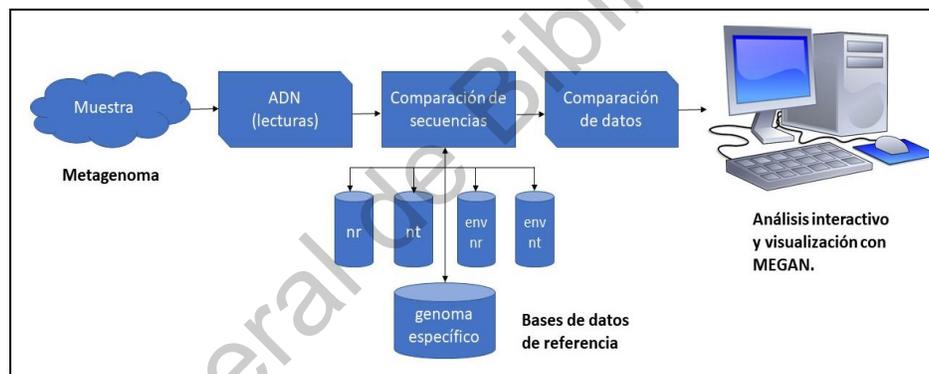


Figura 9. Diagrama de flujo de MEGAN. Preprocesamiento y asignación taxonómica (Tomada y modificada de *Huson et al., 2007*).

El algoritmo LCA asigna las lecturas derivadas del pre-procesamiento y las asigna a los taxa correspondientes, de manera que el nivel taxonómico del taxón asignado refleja el nivel de conservación de la secuencia.

En sus inicios, MEGAN fue utilizado para el análisis de un *subset* de datos del Mar de los Sargazos de Craig Venter; y de una muestra de un hueso de Mamut con una composición aproximada de 300,000 *reads*.

Junto con el crecimiento de los proyectos metagenómicos, tanto en número de muestras consideradas como de *reads* colectados, y el interés en el uso de secuenciación tipo *shotgun*, surge una versión de MEGAN que provee las herramientas necesarias para el análisis y comparación de datos de este tipo: MEGAN-CE (*Community Edition*) (Huson *et al.*, 2016). Esta versión de MEGAN contiene 3 algoritmos LCA: (1) “naive-LCA”, (2) “weighted-LCA” y (3) “LR-LCA” (*long reads*). Debido al interés actual en la generación de lecturas largas de alta calidad (10kb – 1Mb) como resultado de las nuevas tecnologías de secuenciación surge el algoritmo LR-LCA de MEGAN (Huson *et al.*, 2018). Este algoritmo está explícitamente diseñado para el análisis de lecturas largas (generalmente mayores a 1000 pb, y que se encuentran entre los rangos de 10 a 15 kb) o contigs ensamblados, y es una adaptación de los algoritmos anteriores para el análisis de lecturas cortas (~100 a 600 pb, o hasta 1000 pb).

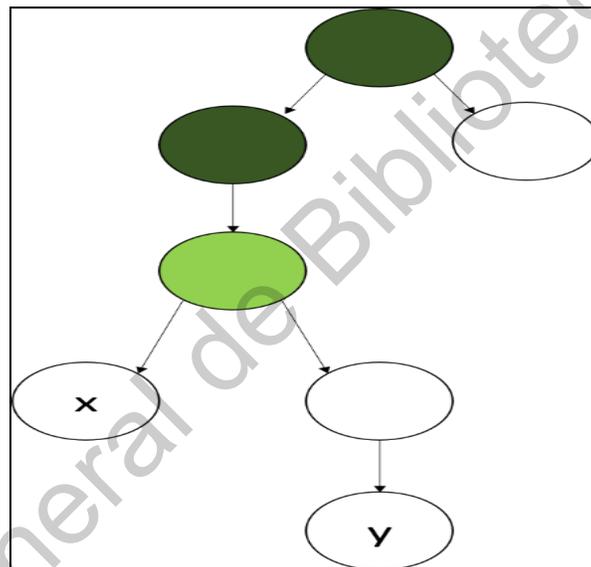


Figura 10. Árbol de ancestro común más bajo. En este árbol el ancestro común más bajo de los nodos **x** y **y** está marcado con un círculo color verde claro. Otros ancestros en común más lejanos de **x** y **y** se muestran en color verde oscuro.

Trabajar con lecturas cortas en el análisis de datos metagenómicos genera limitantes en la resolución de ciertas preguntas biológicas, dificultando la respuesta de éstas. Un ejemplo muy claro es si dos genes que son detectados en la misma muestra de microbioma también ocurren juntos en el mismo genoma de un organismo. Abordar esta pregunta haciendo uso de secuencias cortas puede ser posible con un procedimiento de ensamblaje adecuado en algunos casos, sin embargo, se vuelve una limitante extra a resolver, problemática que

puede ser evitada haciendo uso de lecturas largas, en donde ambos genes pueden o no ocurrir en un mismo genoma de un organismo en particular.

Algoritmos LCA en MEGAN:

1. **Naive-LCA:** Es el primer algoritmo desarrollado para su uso en MEGAN, además, es el algoritmo que MEGAN utiliza por default en un análisis taxonómico. El objetivo de este algoritmo es realizar una asignación taxonómica. En este enfoque, cada *read* es asignado al nodo del “ancestro común más bajo” de la taxonomía. De esta manera la asignación se realizará al taxón/nodo que se encuentra por encima de todas las especies en las cuales dicho *read* tuvo un alineamiento significativo (**Figura 11**). Así, los *reads* que alinean a genes altamente conservados son asignados a niveles taxonómicos más altos (ej. Filo, Clase), mientras que los *reads* que alinean a un gen que es específico para un cierto tipo de organismos pueda ser asignado a un nivel taxonómico más bajo (Género o Especie). Como consecuencia de lo anterior, los *reads* son asignados a través de todos los niveles taxonómicos. Este algoritmo provee un método de asignación taxonómica rápido y directo, corriendo a una tasa de ~100 millones de *reads* (cortos) y 2 billones de alineamientos por hora. Sin embargo, es menos adecuado para propósitos de perfiles taxonómicos, en donde el objetivo es obtener una estimación precisa del contenido taxonómico de una muestra. Una de las razones del bajo rendimiento de este algoritmo como una herramienta de perfilamiento es el hecho de que procesa cada *read* por separado, independientemente de todos los otros *reads*.

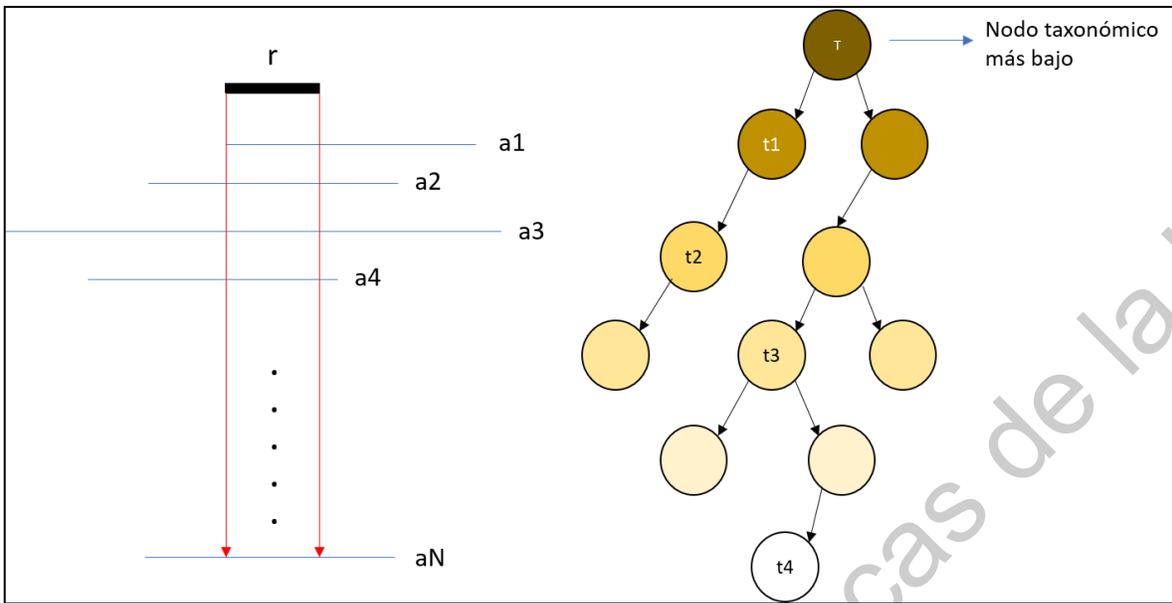


Figura 11. Algoritmo “naive-LCA” MEGAN-CE. Considerando un *read* r que presenta un alineamiento significativo a las secuencias de referencia $a1$, $a2$, $a3$, $a4$... aN , que se encuentran asociadas a los taxa $t1$, $t2$, $t3$, $t4$. El algoritmo **naive-LCA** asigna a r al nodo taxonómico más bajo (T) que representa al nodo que se encuentra por encima de todas las especies en donde se presentó un alineamiento significativo de r .

2. **Weighted-LCA:** Este algoritmo es diseñado e implementado en MEGAN con el objetivo de abordar las limitantes de perfilación taxonómica presentadas por el algoritmo naive-LCA. Este algoritmo funciona de la siguiente manera. En una primera fase, a cada secuencia de referencia (S) le es asignado un peso (weight). Este valor es el número de *reads* (R) que únicamente alinearon a S . Entonces, en una segunda fase, cada *read* R es colocado en el nodo más bajo de la taxonomía que se encuentra arriba del 75% o más del peso total de todas las referencias para las cuales R tuvo un alineamiento significativo. De esta manera la asignación taxonómica no se realizará únicamente en el nodo taxonómico más bajo, sino en todos los nodos taxonómicos arriba del 75% del valor del peso de todas las referencias. Esto mejora la especificidad de la asignación taxonómica, pero requiere más tiempo y trabajo de computo.
3. **LR-LCA:** Este algoritmo es utilizado para realizar una asignación taxonómica de un conjunto de *reads* largos o contigs. Hace uso del algoritmo naive-LCA. Una de las mayores asunciones al trabajar con *reads* cortos en procesos de alineamiento es que cualquier *read* alineará solamente a un solo gen y por lo tanto

a todos los taxones asociados a ese mismo gen. Cuando se trabaja con secuencias largas esta asunción no se cumple, por el contrario, se espera que un *read* largo se traslape con uno o más genes.

Para hacer el algoritmo *naive*-LCA aplicable al alineamiento de *reads* largos o contigs, primero se parte de la idea general de detectar los diferentes genes a lo largo del *read* en donde se presenten alineamientos significativos, para después aplicar el algoritmo *naive*-LCA a cada una de las regiones detectadas de manera individual. Arrojando así una asignación taxonómica basada en secuencias largas.

Todos los seres vivos nos encontramos en constante interacción con el ambiente que nos rodea. Esta interacción nos permite generar asociaciones de convivencia ecológica que nos brindan beneficios para un mejor aseguramiento de nuestra eficacia biológica. Los seres humanos y los microorganismos no son la excepción a este fenómeno. El cuerpo humano posee al menos un número semejante de células humanas como de células microbianas. Se estima que al menos entre 500-1,000 especies de bacterias viven sobre y dentro del cuerpo humano en un momento determinado, y que la naturaleza de éstas varía de acuerdo a distintos factores ambientales y del propio hospedero. Cada bacteria posee un genoma que a su vez posee cientos o miles de genes; la diversidad genética que ofrecen los miembros del microbioma al ser humano es al menos 10 veces mayor que la propia diversidad genética humana. Los componentes del microbioma humano son cambiantes, cada individuo humano posee una estructura taxonómica distintiva, sin embargo, ciertos componentes microbianos se comparten entre individuos dentro y fuera de una población. Muchos miembros del microbioma son necesarios para el mantenimiento de un buen estado de salud en los seres humanos. El cambio drástico en la estructura y dinámica de la microbiota ocasionada por factores externos o del propio hospedero generan un desbalance en la comunidad ecológica, lo que resulta en un desequilibrio homeostático y puede llevar a la aparición de cuadros de enfermedad o en un aumento de la severidad de estos. Los recientes avances tecnológicos a nivel de herramientas de secuenciación nos han permitido obtener un mejor y más amplio panorama acerca de estas importantes comunidades. Hoy en día conocemos que distintas especies y géneros bacterianos se encuentran relacionadas con una mejor prognosis y diagnóstico de enfermedades. Las bacterias conforman la mayor parte de la microbiota humana. Sin embargo, otros integrantes de la microbiota también llevan a cabo procesos esenciales para el mantenimiento de la

comunidad ecológica, este es el caso de los hongos. Los hongos, a diferencia de las bacterias de la microbiota humana, son menos abundantes, pero son muy diversos. Distintas especies de hongos han sido descritas formando parte normal de la microbiota humana como miembros de los géneros *Candida* y *Malassezia*. Conocer la estructura completa de las comunidades microbianas en los seres humanos nos permitirá establecer relaciones entre sus miembros y entender mejor cómo el desbalance de alguno de ellos puede llevar a estados de enfermedad.

En el presente trabajo se realizará un análisis exhaustivo de las comunidades poco estudiadas de hongos derivados de datos metagenómicos, y se harán relaciones sobre los miembros de las comunidades fúngicas y su posible relación con estados de salud y enfermedad en individuos humanos.

2. Hipótesis

Existe una diferencia en el componente fúngico del microbioma humano relacionado a estados de salud y enfermedad, además, éste varía en función de la ancestría de la población en la que se observa.

3. Objetivos

3.1. Objetivo general

Determinar la diversidad taxonómica del microbioma en distintas poblaciones humanas de acuerdo a un estatus de salud o enfermedad.

3.2. Objetivos específicos

- 1) Recopilar información de datos derivados de secuenciación metagenómica (contigs) de 3 de los principales estudios de microbiomas en poblaciones humanas: HMP, MetaHIT-Eur y MetaHIT-Ch.
- 2) Alinear las secuencias de contigs obtenidas contra una base de datos de referencia.
- 3) Filtrar los alineamientos obtenidos de acuerdo a la calidad de estos.
- 4) Asignar los alineamientos de las secuencias a un perfil taxonómico con un énfasis en el reino 'Fungi'.
- 5) Describir la diversidad taxonómica encontrada en los distintos grupos (casos y controles).

- 6) Representar visual y gráficamente los datos generados para una mejor interpretación de los mismos.

4. Metodología

La metodología, en concordancia con los objetivos planteados, se divide en las siguientes seis etapas: **1)** Selección y recopilación de datos de secuencias y ensamblajes metagenómicos tipo contigs derivados de 3 de los principales proyectos de microbiomas en poblaciones humanas: **HMP, MetaHIT-Eur, y MetaHIT-Ch**; **2)** Generación de alineamientos de los datos de secuencias tipo contigs obtenidos utilizando la herramienta **BLAST 2.3.0** contra la base de datos de referencia *nr-nt* de NCBI; **3)** Filtración y **4)** análisis taxonómico de los alineamientos generados mediante el uso de la herramienta **MEGAN6-CE**; **5)** Derivado del perfil taxonómico se utilizará la herramienta **vegan 2.5.2** (RStudio) para realizar una descripción de la diversidad encontrada mediante el uso de índices de biodiversidad tipo alfa: usando los índices no inversos de Shannon y Simpson; **6)** Finalmente, se realizará una representación visual gráfica de los resultados obtenidos mediante el uso de la herramienta **ggplot2** (RStudio).

La metodología propuesta involucra el análisis de datos metagenómicos derivados de microbiomas de distintas partes del cuerpo humano (HMP), sin embargo, se hará un énfasis en el componente fúngico de aquellos obtenidos del tracto gastrointestinal (MetaHIT) que se encuentran relacionados a estados de salud y enfermedad.

4.1. Recopilación de secuencias ensambladas derivadas de estudios de microbiomas humanos.

Se trabajó con archivos de secuencias ensambladas tipo contigs generadas originalmente por cada uno de los dos grandes consorcios: *The Human Microbiome Project*, y *The Metagenome of the Human Intestinal Tract*.

1) Proyecto del Microbioma Humano, HMP (1er set de datos)

Se obtuvieron las secuencias de contigs derivados de un estudio de MSS de 300 individuos estadounidenses en 15 zonas del cuerpo. El set de datos disponible es una representación de 681 muestras de 242 individuos debido a un filtro de calidad del HMP. Todos los individuos se encuentran catalogados como individuos sanos.

Los datos con los que se trabajó se obtuvieron durante el mes de Mayo del 2017 del sitio web **NIH Human Microbiome Project** siguiendo la vía: >home >data browser >hmbsa >hmbsa:healthy como se muestra en la siguiente liga: <https://www.hmpdacc.org/hmp/HMBSA/healthy/index.php>

Los archivos se encuentran organizados como un conjunto de contigs de todos los individuos por sitio del cuerpo, sumando un total de 15 archivos. Se realizó la descarga de los documentos vía FTP en línea de comandos *bash* en Ubuntu/Linux con el comando *wget* para cada uno de los archivos como se ejemplifica a continuación:

```
wget ftp://public-ftp.ihmpdacc.org/HMBSA/Anterior_nares.tar.bz2
```

Tabla 4. Vista general de la composición de los archivos obtenidos del **HMP**.

Zona del cuerpo	No. de muestras	Número de contigs	Tamaño promedio de contigs	Tamaño del archivo
Nariz anterior	84	67, 612	635 bp	25.0 MB
Mucosa bucal	107	358, 115	494 bp	103.3 MB
Gingiva queratinizada	6	54, 234	475 bp	14.0 MB
Cresta retroauricular izquierda	9	61, 316	563 bp	16.9 MB
Vagina media	2	353	695 bp	87.2 KB
Amígdalas palatinas	6	106, 953	592 bp	29.3 MB
Fórnix posterior	53	17, 820	540 bp	5.8 MB
Cresta retroauricular derecha	15	118, 851	438 bp	47.2 MB
Saliva	5	21, 693	408 bp	5.3 MB
Heces fecales	136	2, 225, 881	471 bp	769.1 MB
Placa subgingival	7	275, 970	413 bp	79.7 MB

Placa supragingival	115	784, 017	498 bp	247.3 MB
Garganta	7	105, 394	450 bp	27.7 MB
Dorso de la lengua	122	264, 916	615 bp	95.0 MB
Introito vaginal	3	2, 280	437 bp	595.9 KB
TOTAL	681	4, 465, 405		

2) Metagenoma del Tracto Intestinal Humano, MetaHIT-Eur (2do set de datos)

El proyecto MetaHIT-Eur contiene una base de datos de **6, 577, 036** contigs con un tamaño mayor a 500 pares de bases generados de la secuenciación de **muestras fecales** de 124 individuos europeos (39 casos y 85 controles). El set de datos completos se muestra en la **Tabla A1** de la sección de **Anexos** al final de este documento.

Los archivos de contigs se encuentran disponibles en el sitio web: http://www.bork.embl.de/~arumugam/Qin_et_al_2010/ organizados por individuo, sumando un total de 124 archivos. Estos fueron descargados durante el mes de Mayo del 2017. Se utilizó el comando `wget` más la dirección FTP de cada archivo como se ejemplifica a continuación:

```
wget http://www.bork.embl.de/~arumugam/Qin_et_al_2010/MH0001.seq.fa.gz
```

3) Metagenoma del Tracto Intestinal Humano, MetaHIT-Ch (3er set de datos)

El proyecto MetaHIT-Ch contiene una base de datos de **4, 019, 997** contigs con un tamaño mayor a 500 pares de bases. Estas secuencias fueron generadas de la secuenciación y ensamblaje de **muestras fecales** pertenecientes a 145 individuos chinos (71 casos y 74 controles). El set de datos completo se muestra en la **Tabla A2** de la sección de **Anexos** al final de este documento.

Los archivos de contigs se encuentran disponibles en el sitio FTP del repositorio GigaDB: ftp://parrot.genomics.cn/gigadb/pub/10.5524/100001_101000/100036/AssemblyContigs/ . Cada archivo representa a un individuo del estudio, sumando un total de 145 archivos. Se realizó la descarga de todos los archivos dentro de la carpeta *AssemblyContigs* mediante el siguiente comando:

```
wget -r
ftp://parrot.genomics.cn/gigadb/pub/10.5524/100001_101000/100036/Assembly
Contigs/
```

4.2. Alineamiento de las secuencias de contigs usando la herramienta BLAST 2.3.0 contra la base de datos “no redundante de nucleótidos (nr-nt)” de NCBI.

Se realizaron alineamientos de las secuencias de contigs de los 3 sets de datos de microbiomas. Los alineamientos fueron realizados con la herramienta BLAST 2.3.0 (Altschul *et al.* 1990) en su versión de línea de comandos. Blast es una herramienta de búsqueda de alineamiento local (*Basic Local Alignment Search Tool*), genera la alineación de dos secuencias de ADN de acuerdo a algoritmos de homología y similitud mediante la comparación de combinaciones de nucleótidos en bases de datos establecidas. Si una secuencia desconocida genera un alineamiento en una secuencia blanco que cumple con los parámetros establecidos, entonces se le asigna un identificador a dicha secuencia (Ye, McGinnis & Madden, 2006).

Descarga y almacenamiento de la base de datos *nr-nt*

Los alineamientos se realizaron contra la base de datos de secuencias de nucleótidos “no redundantes” (*non redundant*) de NCBI “*nr-nt*”. Se llevó a cabo la descarga de la base de datos del sitio de FTPs de NCBI: <ftp://ftp.ncbi.nlm.nih.gov/blast/db/> el día **20 de Abril del 2017**. La base de datos fue almacenada dentro del clúster “DNA HPCC” del Laboratorio de Visualización Científica Avanzada (LAVIS) de la UNAM Campus Juriquilla.

La base de datos no redundantes de NCBI se construye con secuencias idénticas que son fusionadas en una misma entrada de las distintas bases de datos. Para que dos secuencias sean fusionadas éstas deben tener tamaños idénticos, y cada residuo en cada posición debe ser el mismo (NCBI, 2017).

La base de datos *nr-nt* corresponde a todas las secuencias no redundantes reportadas en bases de datos como GenBank, EMBL y DDBJ (NCBI, 2017). Tan solo en la actualización de Abril del 2003 esta base de datos contaba con secuencias que representan a 4,653 familias, 26,427 géneros, 130,207 especies, y un total de 176,890 taxas (Federhen, 2013).

Alineamiento de secuencias

Para realizar el alineamiento de secuencias el primer paso es darle un formato específico a la base de datos que se utilizará para que blast pueda hacer uso correcto de ella.

Posteriormente se realiza un alineamiento de los archivos de contigs utilizando los parámetros por default de la herramienta BLAST2.3.0 contra la base de datos *nr-nt*.

Comando:

```
makeblastdb -in mydb.fsa -dbtype nucl -parse_seqids
```

Después de haber dado formato a la base de datos. Se llevó a cabo el alineamiento de las secuencias. En este caso, al realizar alineamientos de archivos grandes se paralelizó el *script* y se aumentó el número de núcleos por trabajo (job) mandado al clúster, de la siguiente manera:

Comando:

```
blastn -db path/to/database -query path/to/contigs -num_threads  
4 -out path/to/out/directory -outfmt 6
```

Los parámetros anteriores son los que aseguran un rendimiento óptimo de BLAST: 4 núcleos con 4 Gb de RAM cada uno, para un total de 16Gb de RAM de trabajo continuo. Asegurando así no hacer uso de núcleos ni memoria excesiva que BLAST no utilizaría correctamente. Debido a la cantidad de secuencias y la naturaleza de los datos (metagenomas microbianos), el tiempo de alineamiento fue de 25 (+/- 5) días naturales para cada uno de los sets de datos.

Los archivos esperados como resultado del alineamiento de BLAST son un formato tabular que contienen información de las secuencias alineadas y la calidad del alineamiento, el archivo que se espera como resultado es llamado archivo "output".

Tabla 5. Valores generados como resultado de un alineamiento en BLAST con la salida de texto por default.

Concepto/Valor	Significado
qseqid	Nombre del identificador de la secuencia de consulta
sseqid	Nombre del identificador de referencia

pident	Porcentaje de identidad
length	Tamaño
mismatch	Número de incompatibilidades
gapopen	Número de aberturas tipo gap
qstart	Inicio del alineamiento en la secuencia de consulta
qend	Fin del alineamiento en la secuencia de consulta
sstart	Inicio del alineamiento en la referencia
send	Fin del alineamiento en la referencia
evaluate	Valor de e
bitscore	Valor de bitscore

4.3. Análisis taxonómico con MEGAN6.

El análisis taxonómico se llevó a cabo con la herramienta MEGAN6 (Huson *et al.*, 2007). La cual hace uso preferente del formato tabular (outfmt 6) generado por BLAST. MEGAN (*MetaGenome Analyzer*) es una herramienta de análisis de datos metagenómicos, la cual computa y explora el contenido taxonómico y funcional de datos de este tipo de muestras. Esta herramienta trabaja con un algoritmo de asignación taxonómica denominado LCA (Lowest Common Ancestor).

MEGAN6-CE es una versión que incluye una interfaz gráfica de trabajo (**Figura 12**). Haciendo uso de esta interfaz se realizó la asignación taxonómica de todas las secuencias de contigs. Estas secuencias se asignaron en todos los reinos, sin embargo, para los fines de este trabajo se realizó un énfasis en aquellas obtenidas en el reino Fungi a nivel de Filo y género.

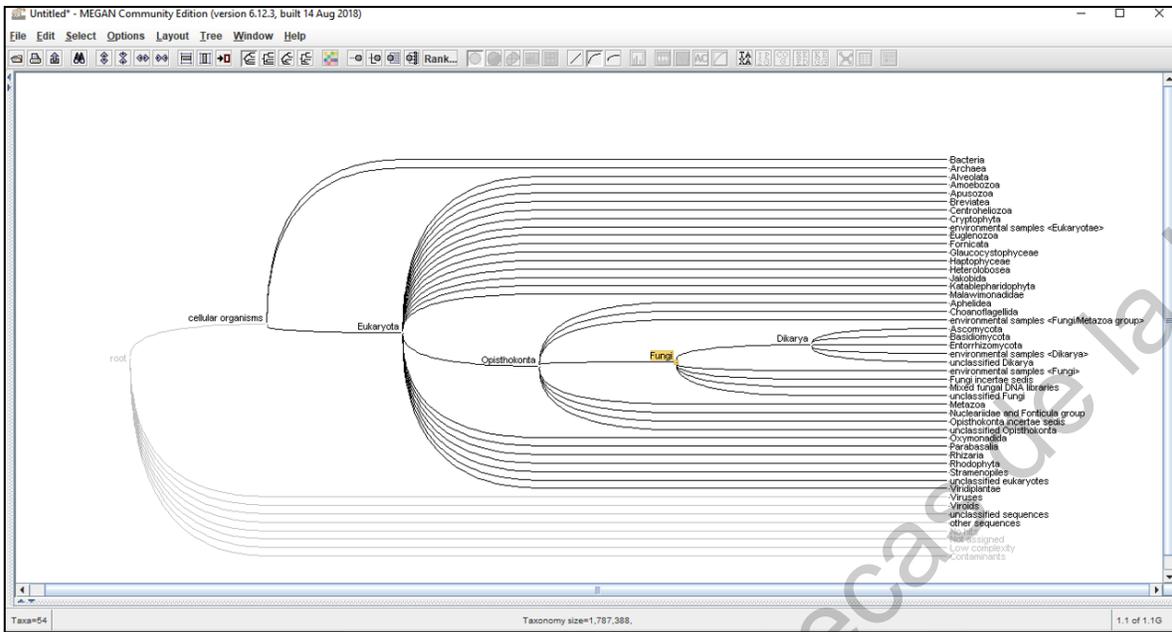


Figura 12. Interfaz gráfica de MEGAN 6-CE.

Los parámetros que se utilizaron en MEGAN6 para realizar una asignación taxonómica de los alineamientos generados por BLAST son los siguientes:

Tabla 6. Parámetros a usar para análisis taxonómicos en MEGAN 6-CE. Valores por default y valor utilizado.

Parámetro de asignación taxonómica	Valor por Default	Valor Utilizado
Puntuación mínima (min score).	50.0	80.0
Valor esperado (e-value).	0.01	0.001
Porcentaje mínimo de identidad.	0	75
Porcentaje top.	10.0	50.0
Porcentaje mínimo de soporte.	0.0	0.0(off)
Soporte mínimo	1	5
Complejidad mínima	0.0	0.3
Algoritmo LCA	<i>naive</i>	<i>LongReads</i>

Porcentaje de cobertura	80	80
Modo de asignación	<i>readCount</i>	<i>alignedBases</i>
Porcentaje mínimo del read a ser cubierto	0	75

Los valores utilizados diferentes a los valores por *default* son los que nos permiten obtener un perfil taxonómico con un mayor nivel de confianza. Al aumentar parámetros como el *e-value*, el porcentaje mínimo del *read* a ser cubierto, y el porcentaje mínimo de cobertura, generamos un análisis más estricto en el cual la presencia de falsos positivos se ve disminuida. En el caso del modo de asignación, el parámetro '*alignedBases*' nos permite obtener un perfil taxonómico más certero en niveles menores de organización, como en el caso de género. El algoritmo LCA utilizado es '*LongReads*' debido a que se trabaja con secuencias de contigs. Estos parámetros fueron evaluados en *tests* de prueba y error, y son los que mejor se ajustan al tipo de datos con los que se trabajó.

MEGAN6-CE genera diversos archivos *output* con información ordenada de manera tabular en formato *csv* (*comma-separated values*) ó *tsv* (*tabular-separated values*) acerca de la composición taxonómica de las secuencias analizadas. Estos archivos de salida de texto son los utilizados para posteriores análisis de diversidad y representación gráfica.

4.4. Análisis de diversidad utilizando vegan. 2.5-2

Los análisis estadísticos se realizaron con el paquete R 3.3.2 mediante la interfaz de RStudio.

Para calcular los índices de diversidad alfa y beta se utilizó el paquete '**vegan 2.5-2**'. Vegan es un paquete especializado en la ejecución de análisis multivariantes (principalmente de ordenación), y métodos de análisis de diversidad de comunidades ecológicas: diversidad taxonómica y funcional, modelos de abundancia de especies, acumulación de especies y diversidad beta, modelos de rarefacción, entre otros.

Vegan utiliza como archivo *input* matrices de abundancia, por lo cual los archivos de texto tipo *dataframe* generados por MEGAN fueron modificados para su utilización con el fin de generar este tipo de matrices, utilizando la función *transpose*:

```
t(dataframe)
```

Los comandos utilizados para calcular el índice de Shannon y Simpson, respectivamente, son los siguientes:

```
diversity(DataSet, index= "shannon")  
diversity(DataSet, index= "simpson")
```

4.5. Representación gráfica de los resultados utilizando ggplot2.

ggplot2 es un paquete de R que permite la visualización de datos por medio de la generación de salidas gráficas. ggplot cuenta con una gran cantidad de variables y funciones para generar distintos tipos de gráficos. Para la generación de un gráfico se requiere que se den instrucciones basadas en una sintaxis establecida que permite el reconocimiento, características y elementos básicos de los componentes de los datos que darán lugar a los gráficos.

Algunos ejemplos de la sintaxis y comandos utilizados para la generación de los distintos gráficos utilizados en el presente trabajo son los siguientes:

Función: Composición de hongos a nivel de Género en el set de datos MetaHIT-Ch

```
ggplot(Comp_ChineseMetLongGenus, aes(x=variable, y=value,  
fill=Genus)) + geom_col() + scale_fill_brewer(palette = "Set3") +  
labs(x="Group", y="Relative Abundance (%)") + ggtitle("Fungal  
Genus Composition. ChineseMet Dataset.") + theme(plot.title =  
element_text(hjust = 0.5))
```

Función: Número de asignaciones de hongo por individuo en los sets de datos MetaHIT-Eur y MetaHIT-Ch

```
ggplot(MetaHIT_Complete_ReadName_TaxonName, aes(Individual)) +  
geom_bar(aes(fill=Status)) +  
theme(axis.text.x=element_text(angle=90,hjust=0.4, vjust=0.5,  
size=6)) + facet_grid(. ~ Status, scales="free") +  
labs(x="Individual", y="Number of Assignments") + ggtitle("Number  
of Assignments per Individual. MetaHIT Dataset.") +  
theme(plot.title=element_text(hjust=0.5))
```

Función: Diferencias en la composición fúngica a nivel de Filo entre casos y controles en el set de datos MetaHIT-Eur

```
ggplot(LongMetaHITComp, aes(x = variable, y=value, fill=Phyllum))
+ geom_col() + labs(x="Group", y= "Relative Abundance (%)") +
ggtitle("Differences in Fungi at Phyllum. MetaHIT Dataset.") +
theme(plot.title = element_text(hjust = 0.5))
```

Función: Comparación del tamaño de contigs en función de su cobertura

```
ggplot() + geom_point(data=HashOut_AnteriorNares_AllGenus_C,
aes(CONTIG_LENGTH2, P_CONTIG_COVERAGE2), color="black", shape=16,
size=0.5) + geom_point(data=HashOut_AnteriorNares_FungalGenus_C,
aes(CONTIG_LENGTH, P_CONTIG_COVERAGE), shape=17, color="red",
size=1.2) + xlim(0,5000) + ggtitle("Anterior Nares ") +
labs(x="Contig Length", y="% Contig Coverage")
```

5. Resultados

5.1. Proyecto del Microbioma Humano (HMP)

Derivado de los análisis de alineamiento local con la herramienta BLAST, y la determinación del perfil taxonómico con la herramienta MEGAN6 en los datos del **HMP (Tabla 3)** se obtuvieron los siguientes resultados.

Se observa que el número total de contigs por sitio del cuerpo varía (**Figura 13**). De los 15 sitios del cuerpo evaluados, 8 de ellos que corresponden en su mayoría a la zona oral no presentan alguna abundancia de hongos detectables por la metodología realizada. Sin embargo, lo anterior contrasta con la mucosa bucal y la placa supragingival, sitios correspondientes también a la zona oral, pero en los cuáles sí se detectó una abundancia de hongos. En la **Figura 13** también se muestran 7 sitios del cuerpo en donde se logró observar que existe abundancia de hongos. De estos sitios destaca el tracto gastrointestinal, representado por muestras de heces fecales, siendo el sitio con el mayor número de contigs asociados a hongos, seguido de la placa supragingival, mucosa bucal, y las crestas retroauriculares. De las 3 zonas vaginales analizadas, únicamente el fórnix posterior que corresponde a la zona más interna de la vagina, tuvo una asignación de contigs asociados a hongos, mientras que el introito vaginal (vagina externa) y la vagina

media no presentan esta asignación. Otros de los sitios en donde se detecta abundancia de hongos son la nariz anterior y las crestas retroauriculares.

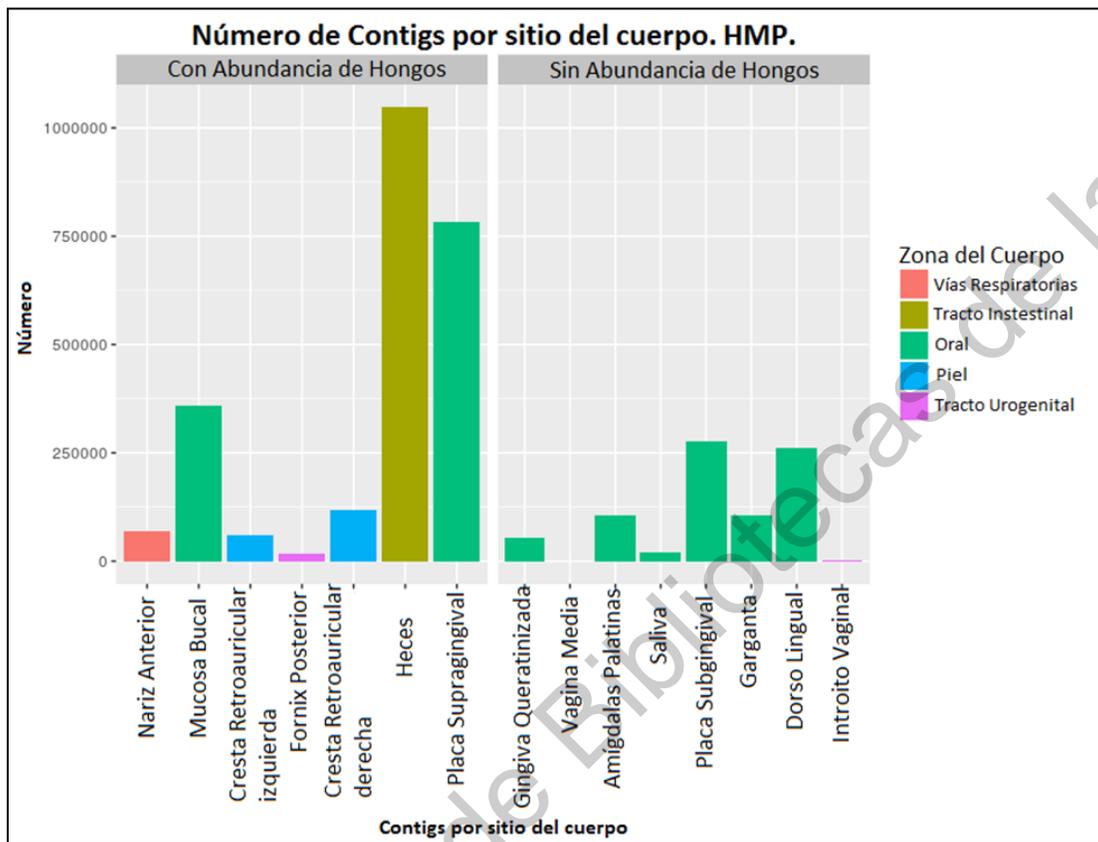


Figura 13. Número de contigs por zona del cuerpo en el HMP. Se muestran los 15 sitios del cuerpo y el número de contigs de cada uno de ellos. El gráfico se encuentra dividido en dos grupos: aquellos en los que se detecta una asignación taxonómica a hongos, y aquellos en los que no se detecta esta asignación.

Observando las variaciones que existen entre el número de contigs en cada sitio del cuerpo, se evaluó el tamaño promedio de estos contigs (**Figura 14**).

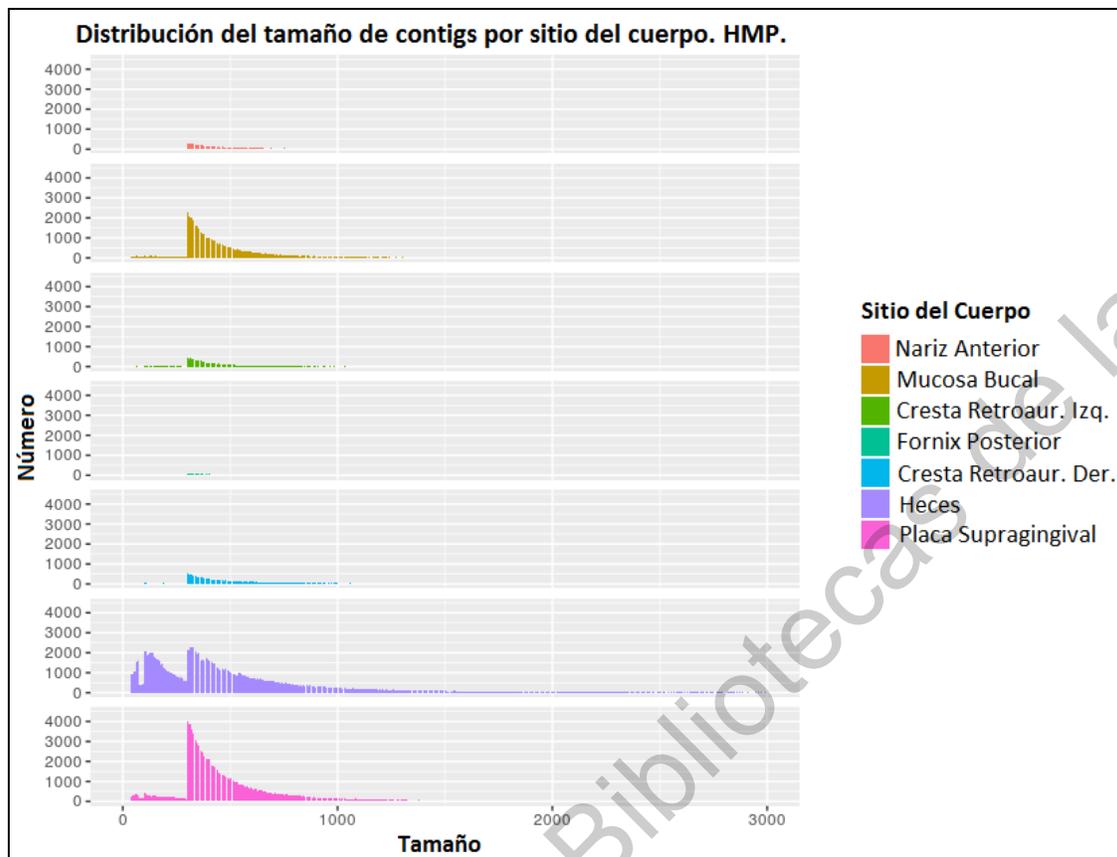


Figura 14. Distribución del tamaño de contigs por sitio del cuerpo. Se muestra el tamaño de los contigs en pares de bases y el número de contigs asociados dicho tamaño de los 7 sitios del cuerpo que presentan abundancia de contigs asociados a hongos.

Los contigs en este set de datos (**HMP**) tienen un tamaño promedio que va de los 475 pares de bases (pb) en el caso de la cresta retroauricular izquierda y llega a los 695 pb en la cresta retroauricular derecha como se muestra en la **Tabla 4**, sin embargo, algunos contigs tienen tamaños menores a 100 pb, lo anterior es más notorio en las heces fecales (**Figura 14**). Para los análisis subsecuentes se realizará un filtro para únicamente evaluar los contigs mayores o iguales a 100 pb.

Tabla 7. Número de contigs por cada zona del cuerpo y tamaño promedio. En amarillo se muestran las zonas en las que se detecta abundancia de hongos y en rojo las que no presentan esta abundancia.

Sitio del Cuerpo	Número de Muestras	Número de Contigs	Tamaño Promedio
Nariz Anterior	84	67, 612	635 bp
Mucosa Bucal	107	358, 115	494 bp
Cresta Retroauricular Izquierda	9	61, 316	475 bp
Fornix Posterior	53	17, 820	563 bp
Cresta Retroauricular Derecha	15	118, 851	695 bp
Heces Fecales	136	2, 225, 881	592 bp
Placa Supragingival	115	784, 017	540 bp
Encía Queratinizada	6	54, 234	438 bp
Vagina Media	2	353	408 bp
Tonsils Palatinos	6	106, 953	471 bp
Saliva	5	21, 693	413 bp
Placa Subgingival	7	275, 970	498 bp
Garganta	7	105, 394	450 bp
Dorso Lingual	122	264, 916	615 bp
Introito Vaginal	3	2, 280	437 bp

5.1.1. Análisis de asignación taxonómica.

Derivado del análisis de asignación taxonómica con MEGAN6 se realizó un análisis de proporción de asignaciones totales (**Tabla 8**). Del total de contigs en este set de datos una muy alta proporción no recibió una asignación taxonómica (78.6%), mientras que el 21.32% de los datos sí recibió una asignación taxonómica. Estos resultados son comparables con datos reportados del HMP, en donde siguiendo una metodología de tres niveles logran mapear aproximadamente el 57% de las secuencias a genomas de referencia (HMP Consortium, 2012). Del total de asignaciones generadas, el 4.75% corresponde a una asignación a Hongo. El sitio que presenta la mayor proporción de contigs asignados es el

Fórnix Posterior seguido de la Nariz Anterior. Por otro lado, los sitios con la menor proporción de asignación taxonómica son las Heces Fecales y la Mucosa Bucal, lo que contrasta con el hecho de que son de los sitios que presentan mayor número de contigs. En cuanto a la asignación taxonómica en Hongo se destacan fuertemente las zonas asociadas al epitelio externo. Por su parte la Cresta Retroauricular Izquierda presenta una proporción de asignación a Hongo del 44.32% mientras la Cresta Retroauricular Derecha presenta una proporción del 27.24% en esta misma asignación. Por otro lado, los sitios con la menor proporción de contigs asignados en Hongo son la Nariz Anterior (0.04%) y las Heces Fecales (0.04%).

Tabla 8. Proporción de contigs asignados en cada sitio del cuerpo.

Sitio del cuerpo	Número de muestras	Número de contigs	Contigs Asignados	Contigs asignados a Hongo	Proporción de contigs asignados	Proporción de contigs asignados a Hongo	Proporción de los contigs asignados que mapean a Hongos
Nariz Anterior	84	67,762	29,622	2,348	43.71	0.06	7.93
Mucosa Bucal	107	358,155	107,138	45	29.91	0.008	0.04
Cresta Retroauricular Izquierda	9	61,316	19,917	8,828	32.48	0.05	44.32
Fornix Posterior	53	17,820	11,860	75	66.55	0.4	0.63
Cresta Retroauricular Derecha	15	118,851	37,315	10,166	31.40	0.03	27.24
Heces fecales	136	2,225,881	387,696	156	17.42	0.0008	0.04
Placa Supragingival	115	784,017	181,346	15,221	23.13	0.003	8.39

TOTAL		3,633,802	774,894	36,839	21.32	0.001	4.75
--------------	--	-----------	---------	--------	--------------	--------------	-------------

Debido a que se observa una notoria diferencia entre el número de contigs asociados al número de muestras de cada sitio del cuerpo en este set de datos, se procedió a evaluar si existe una relación lineal entre estos factores con relación a la detección de secuencias asociadas a hongos.

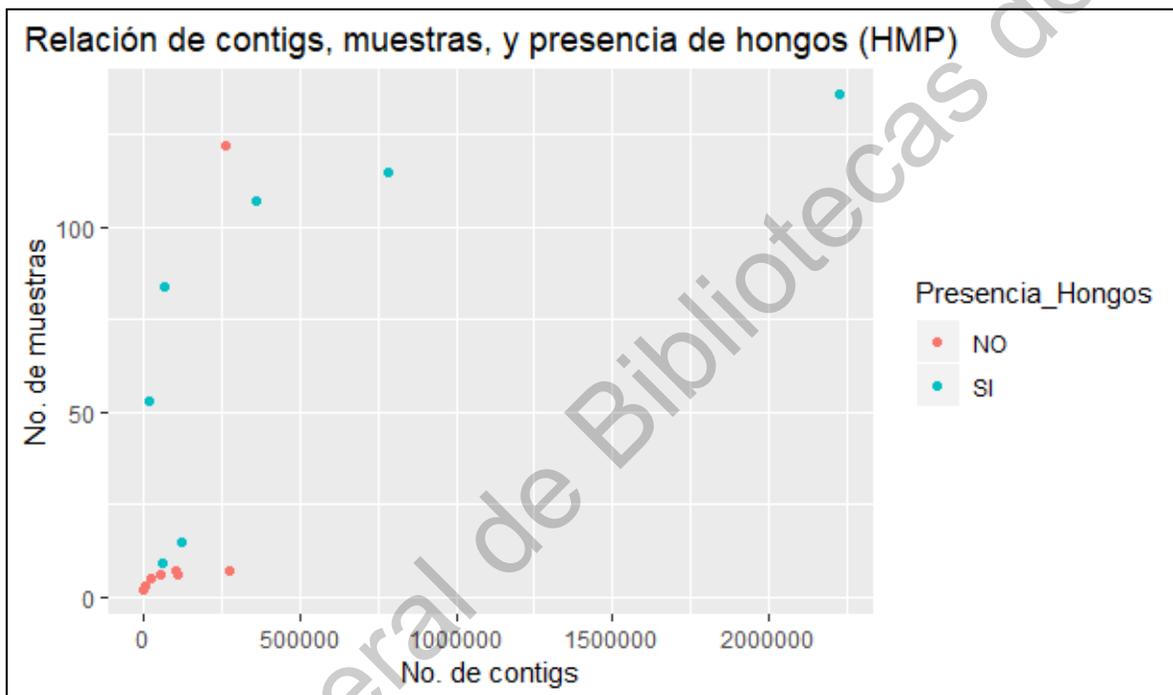


Figura 15. Se observa la nula relación lineal entre el número de muestras y el número de contigs de cada una de los 15 sitios del cuerpo del HMP. Cada punto representa un sitio del cuerpo evaluado.

En la **Figura 15** se observa que no existe una relación lineal entre el número de muestras asociadas a cada sitio del cuerpo en función al número de contigs, y que, además, estos factores no se asocian a la detección de secuencias de contigs que mapean a hongos en este set de datos.

5.1.2. Perfil taxonómico

Se generó un perfil taxonómico de todas las comunidades microbianas en los 7 sitios del cuerpo donde se observó que existe una abundancia de hongos. Sin embargo, a pesar de

que la asignación se realizó a todos los niveles taxonómicos (bacterias, virus, hongos, protozoarios, etc) dada la importancia de los hongos en este estudio, solamente se evaluó a detalle la asignación taxonómica en este reino, tanto a nivel de Filo como de género.

En la asignación a nivel de Filo (**Figura 16**) se observó que, en los sitios asociados a mucosas, como la Mucosa Bucal, el Fórnix Posterior y la Placa Supragingival, la mayoría o todos los Hongos asignados pertenecen al filo Ascomycota, mientras que los sitios asociados a epitelios externos como las crestas retroauriculares y la nariz anterior, presentan una asignación mayor al filo Basidiomycota. Las heces fecales por su parte son un sitio característico, que se distingue de los demás sitios por presentar una asignación de los fila Basidiomycota, Ascomycota, Mucoromycota y Microsporidia, estos dos últimos fila encontrándose exclusivamente en heces fecales.

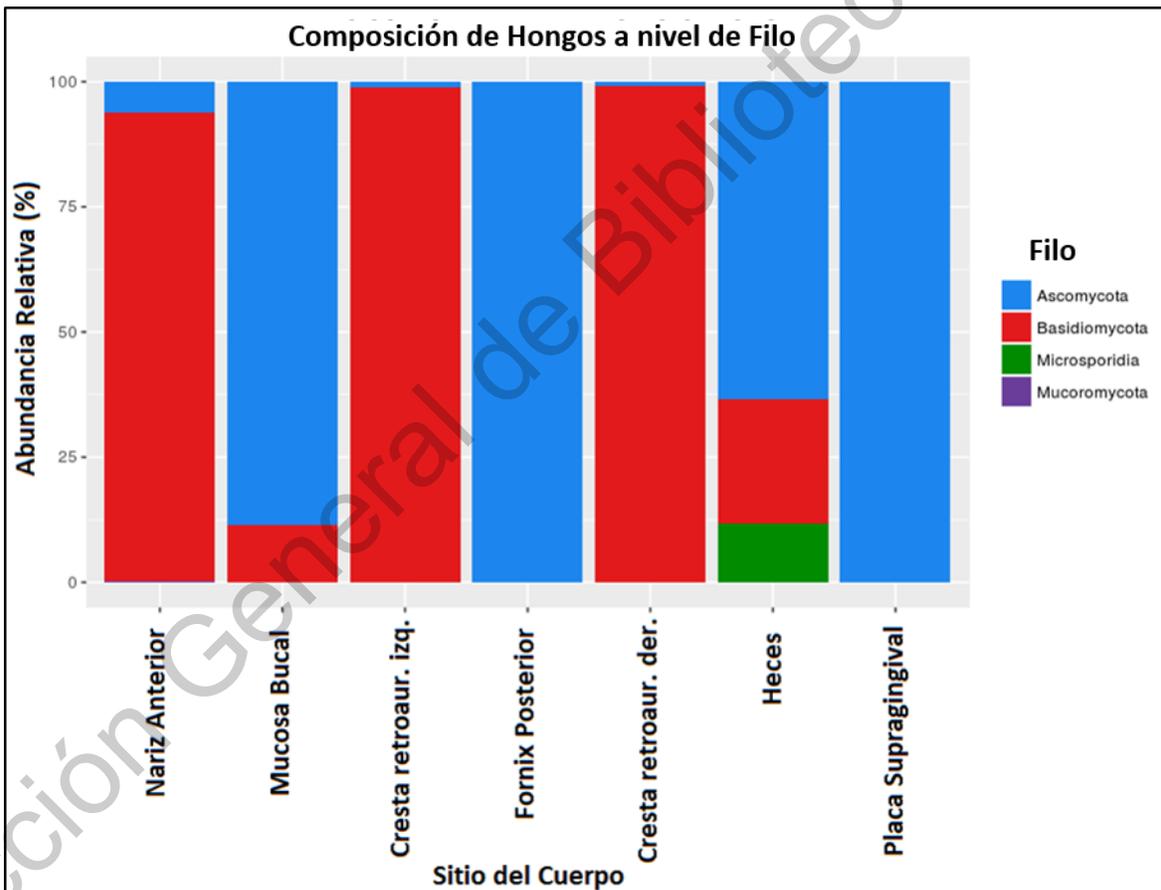


Figura 16. Perfil taxonómico a nivel de Filo en zonas del cuerpo del HMP.

La asignación taxonómica realizada a nivel de género se muestra en la **Figura 17**. Ahí podemos observar que existe una relación directa entre la asignación a nivel de Filo y a

nivel de género, debido a que el Filo asignado se refleja directamente en un género asignado de manera mayoritaria. Por ejemplo, las zonas asociadas a mucosas presentan una elevada proporción de asignación del género *Candida* (Ascomycota), por su parte las zonas asociadas a epitelios son ricas en abundancia del género *Malassezia* (Basidiomycota). Al igual que a nivel de Filo, las heces fecales destacan en su asignación taxonómica a nivel de género. La principal característica de esta asignación en heces fecales es su diversidad comparada con los otros sitios evaluados, se observan géneros como *Aspergillus*, *Auricularia*, *Saccharomyces* y *Enterocytozoon*, siendo este último el que se encuentra en mayor proporción. A nivel de diversidad, la nariz anterior es la segunda zona más diversa, a pesar de que está dominada por el género *Malassezia*, se logran detectar otros géneros como *Aspergillus*, *Trichosporon*, *Pseudozyma*, *Sporisorium*, entre otros.

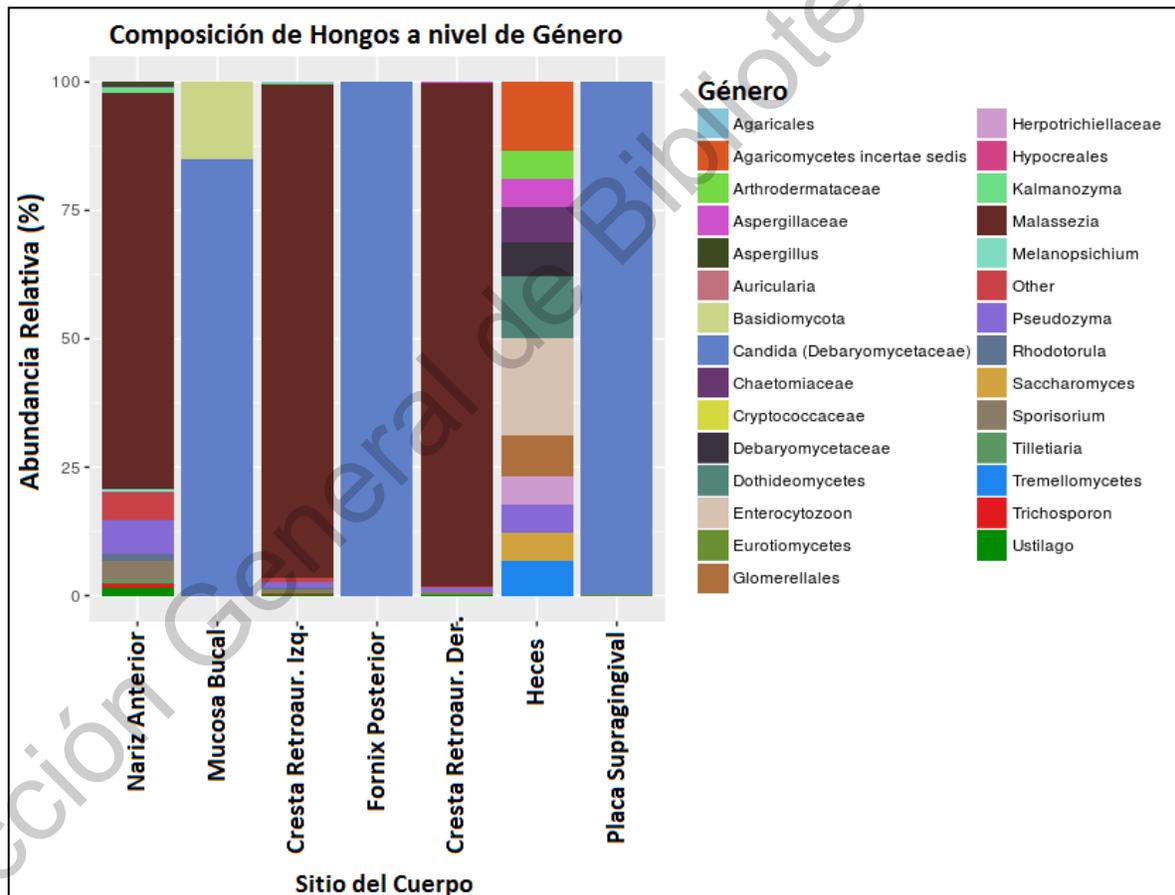


Figura 17. Perfil taxonómico a nivel de Género en las distintas zonas del HMP.

5.2. Metagenoma del Tracto Gastrointestinal Humano - Europeo (MetaHIT-Eur).

Siguiendo una metodología similar a la realizada con el set de datos del HMP, se realizó un perfil taxonómico de los datos del **MetaHIT-Eur**. Este set de datos incluye a un total de 124 individuos, 55 hombres y 69 mujeres, divididos en dos grupos: Individuos con IBD (n=25), e Individuos sin IBD (n=99). De cada uno de los individuos se cuenta con un archivo de contigs ensamblados.

Se analizó la distribución del tamaño de contigs por individuo (**Figura 18**), observando que la distribución es relativamente homogénea en la mayoría de los participantes, encontrándose en un rango promedio de 40 mil a 70 mil contigs por cada individuo.

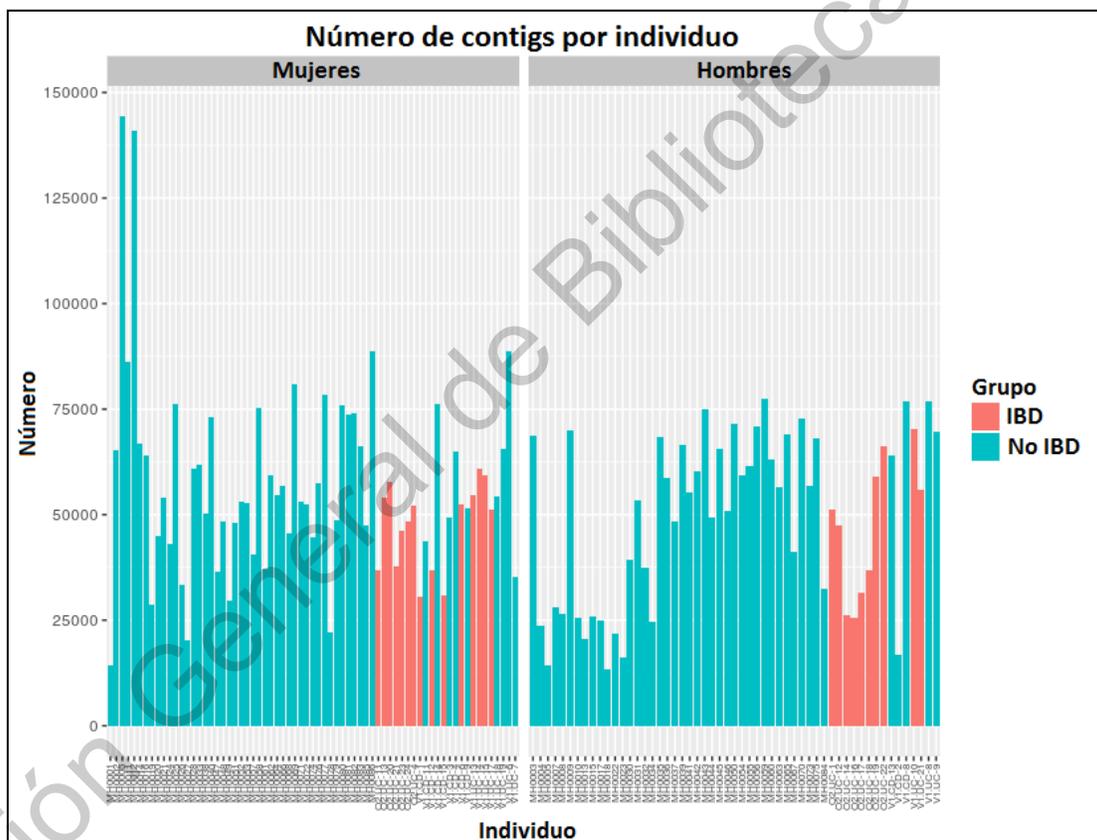


Figura 18. Número de contigs por individuo en el MetaHIT-Eur.

El set de datos completo incluye un total de 6, 577, 036 contigs con un tamaño promedio de 1, 571 pb (**Figura 19**), un tamaño mínimo de contigs de 500 pb y un tamaño máximo de 192, 092 pb.

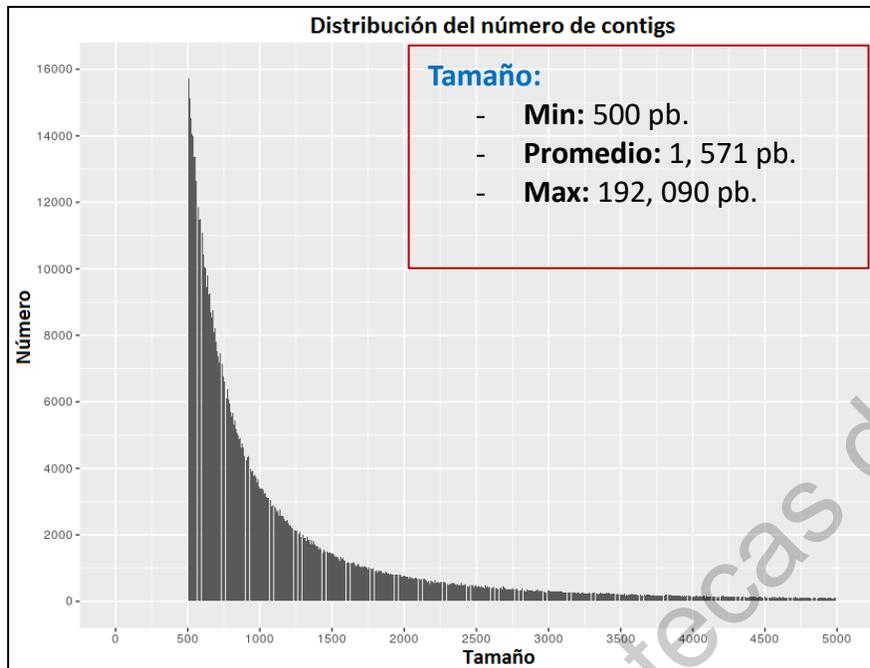


Figura 19. Distribución del tamaño de contigs en el set de datos MetaHIT-Eur.

5.2.1. Perfil taxonómico

Debido a la variabilidad en el número de contigs por muestra de cada individuo, se procedió a establecer si existe una relación entre el número de hongos que pueden ser detectados en cada muestra con respecto al número de contigs de la misma. Como se muestra en la **Figura 20** existe una fuerte relación entre el número de hongos que pueden ser detectados en cada individuo en función al número de contigs de su muestra.

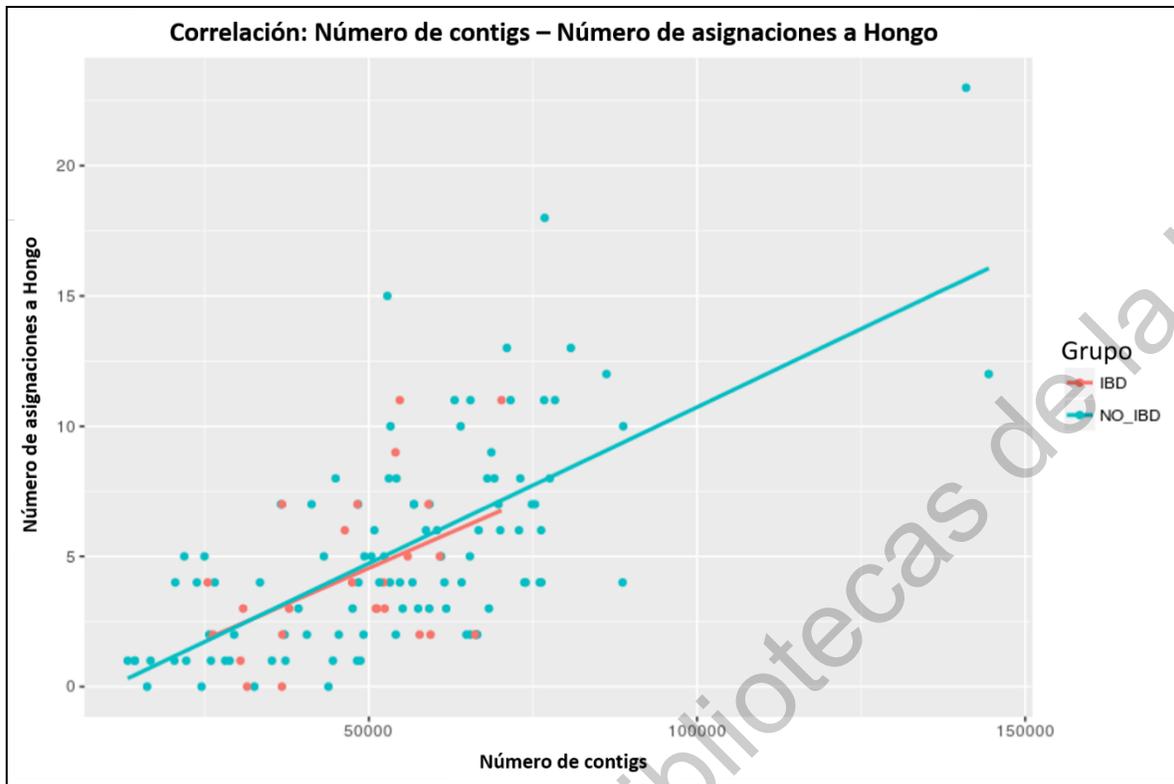


Figura 20. Relación del número de asignaciones a hongos en función al número de contigs en las muestras de cada individuo. Cada punto representa a un individuo distinto del grupo MetaHIT-Eur.

Mediante un análisis complementario utilizando la herramienta MEGAN6-CE con el modo de asignación *readCount* se determinó el número de géneros de hongos que pueden ser detectados por individuo en este *set* de datos, debido a que como se observa en la **Figura 20**, este número se encuentra fuertemente asociado al número de contigs que posee el archivo asociado a cada individuo. Como se observa en la **Figura 21**, existe una alta variación del número de géneros de hongos asignados por individuo. Esta variación se explica principalmente por el número de contigs representado en cada individuo.

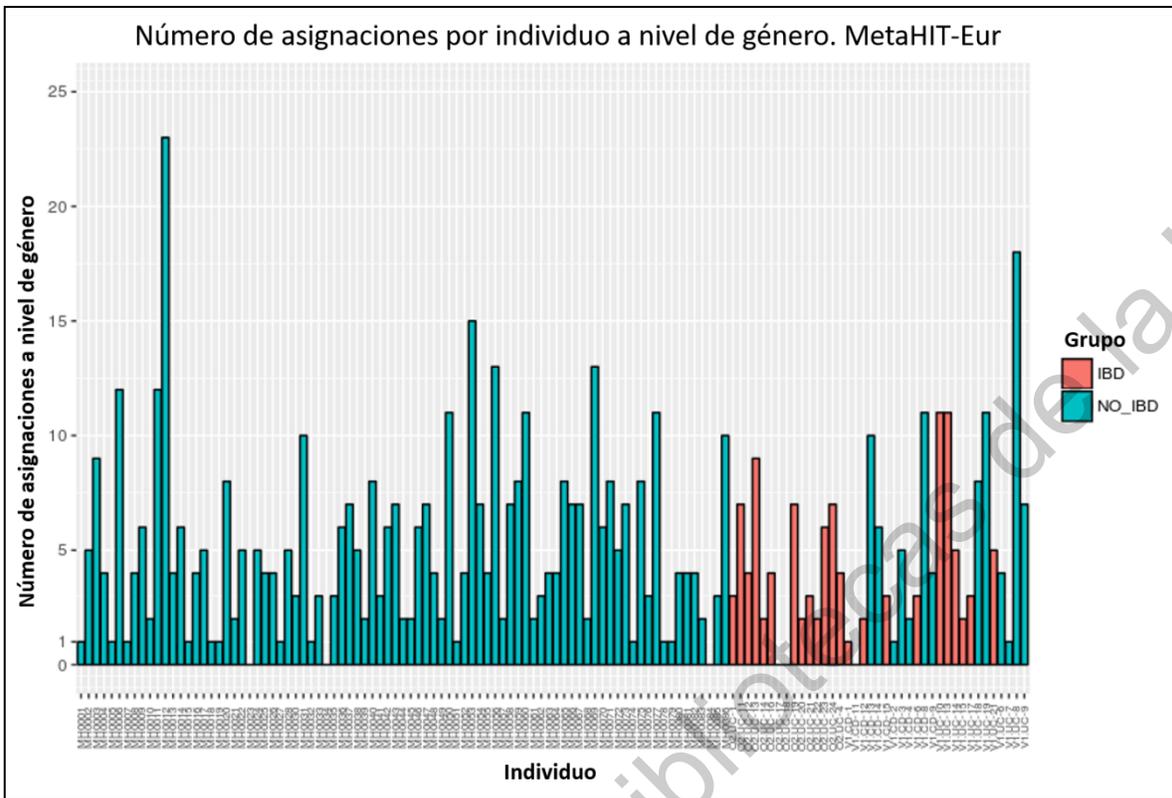


Figura 21. Número de géneros de hongos asignados por individuo en el set de datos MetaHIT-Eur.

Se llevó a cabo una asignación taxonómica a nivel poblacional, dividiendo los datos en dos grupos: **Personas con IBD**, y **Personas Sin IBD**.

La asignación taxonómica se realizó primeramente a nivel de Filo (**Figura 22**). En los dos grupos evaluados (IBD, No IBD) se observan asignaciones en los fila Basidiomycota, Ascomycota, Chitridiomycota y Microsporidia. La abundancia de estos fila se comparte entre ambos grupos. En el grupo sin IBD se observa además el filo Mucoromycota, el cual no fue detectado en personas con IBD.

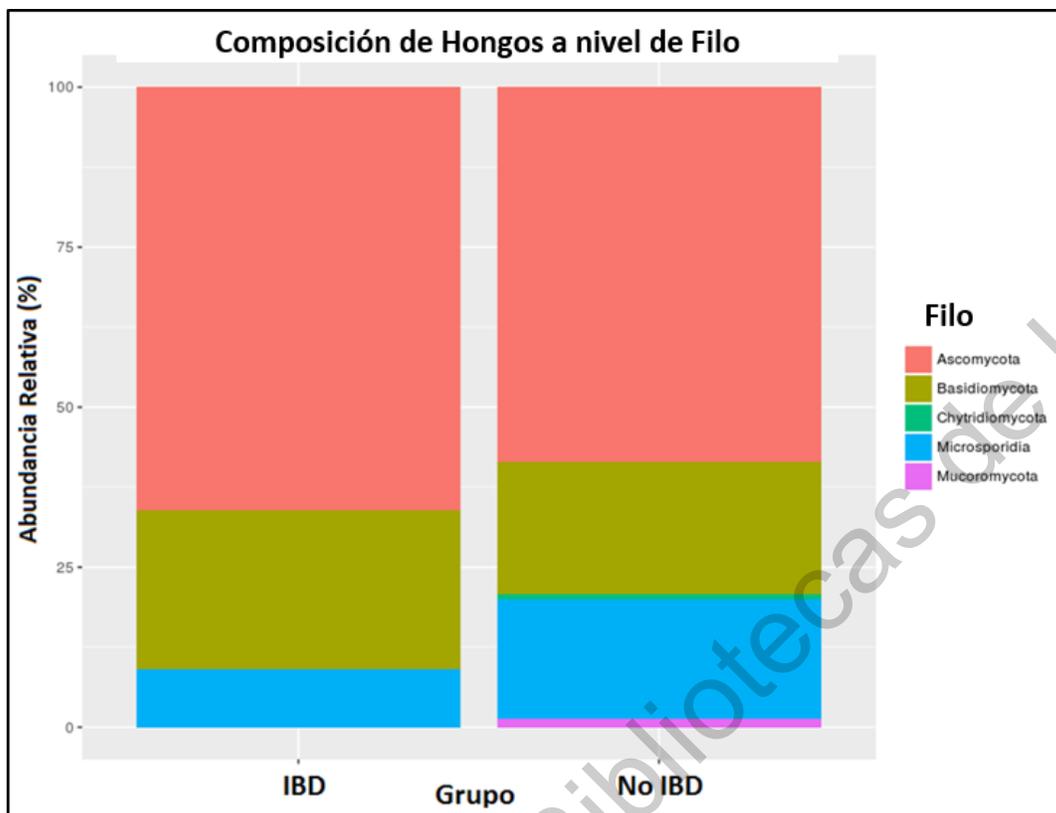


Figura 22. Perfil taxonómico a nivel de Filo en individuos con IBD y sin IBD. Grupos del MetaHIT-Eur.

A nivel de género (**Figura 23**), evaluando nuevamente ambos grupos, se observa que la diversidad en el grupo de individuos que no presenta IBD es mayor que la de aquellos que sí presentan IBD, aunque esto se relaciona directamente con el número total de individuos de cada grupo. En el grupo con IBD se lograron detectar 61 géneros de hongos, en comparación con 103 géneros detectados en el grupo sin IBD. Sin embargo, gracias al índice de Shannon podemos observar que la diversidad de ambos grupos en función a su número de individuos es muy similar, obteniendo un índice de diversidad de Shannon de 2.31 en individuos con IBD, y de 2.51 en individuos sin IBD. En la **Figura 23** se puede observar que en ambos grupos existen géneros que gobiernan la mayor parte de la comunidad, este es el caso de *Saccharomyces* y *Enterocytozoon*, estas observaciones se refuerzan con los valores obtenidos en el índice de dominancia de Simpson, el cual nos indica que en ambas comunidades existen miembros que gobiernan la comunidad en un nivel muy semejante.

Tabla 9. Índices de diversidad en ambos grupos del MetaHIT-Eur.

	IBD	No IBD
Núm. individuos	25	99
Núm. géneros detectados	61	103
Bases alineadas	43 984	176 707
Shannon (diversidad α)	2.31	2.51
Simpson (diversidad α)	0.73	0.77

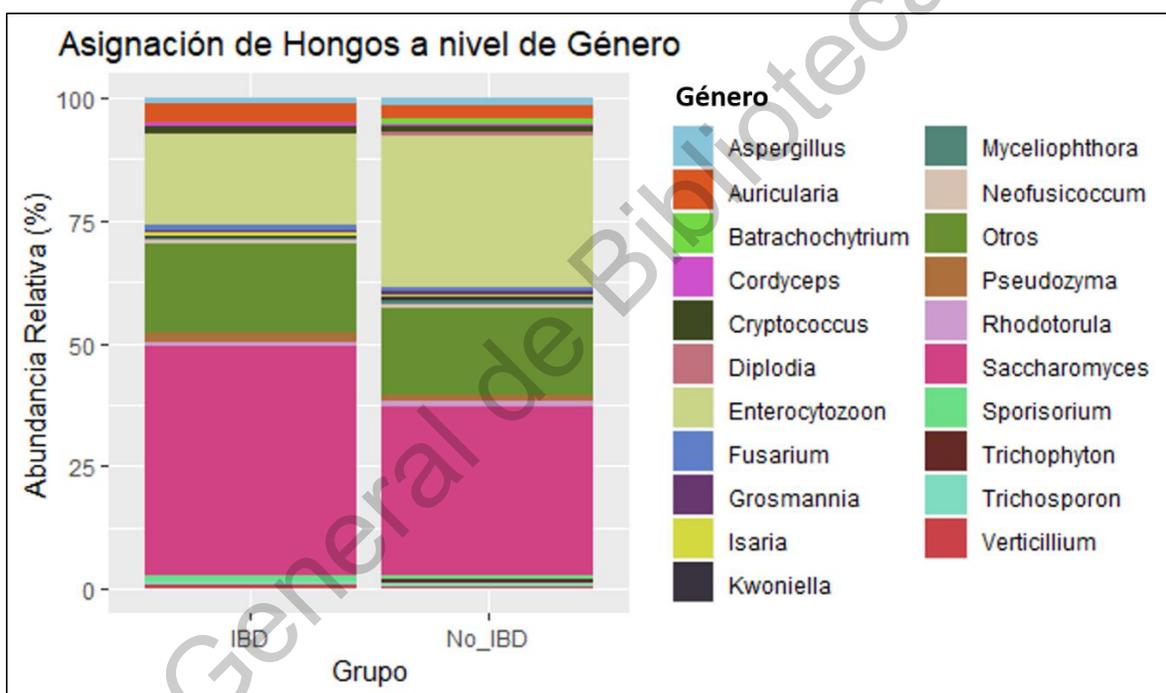


Figura 23. Asignación taxonómica a nivel de Género en grupos del MetaHIT-Eur.

5.3. Metagenoma del Tracto Gastrointestinal Humano - Chino (MetaHIT-Ch)

Siguiendo la metodología utilizada para el análisis de los dos sets de datos anteriores, se realizó el perfil taxonómico del **MetaHIT-Ch**.

El **MetaHIT-Ch** es un set de datos que incluye 145 individuos, 94 mujeres y 51 hombres, se encuentra dividido en dos grupos: Individuos con Diabetes tipo 2 (n=71) (T2D) y sin

Diabetes Tipo 2 (n=74). De cada uno de los individuos de este set de datos se cuenta con un archivo de contigs ensamblados derivado de un estudio de secuenciación metagenómica tipo *shotgun* del microbioma intestinal (heces fecales).

Se evaluó la distribución del número de contigs por individuo (**Figura 24**). Se observa una distribución relativamente homogénea, el rango promedio del número de contigs en este set de datos va de 20,000 a 30,000 contigs por individuo. Además, el número de individuos en el grupo de estudio (con T2D) y en el grupo control (sin T2D) son similares.

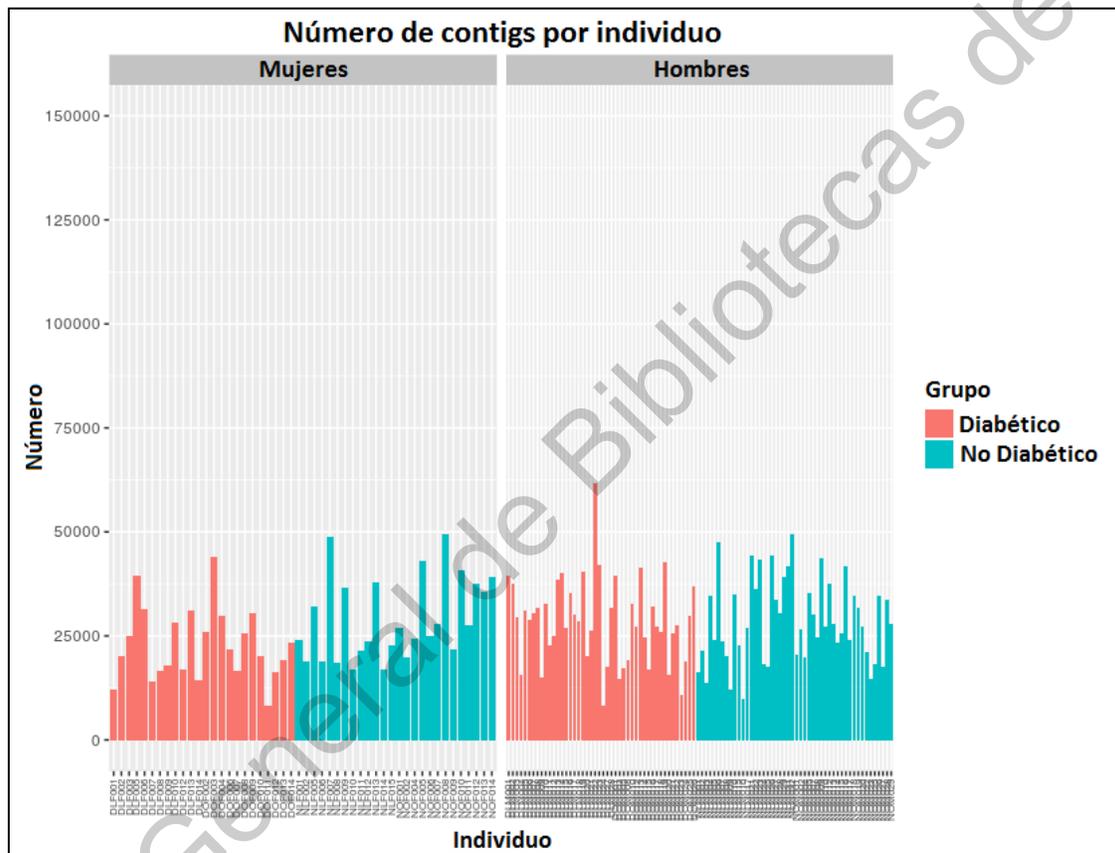


Figura 24. Distribución del número de contigs por individuo. MetaHIT-Ch.

El set de datos completo posee un total de 4,019,997 contigs. El tamaño promedio de los contigs (**Figura 25**) es de 2,132 pb, con un tamaño mínimo de 500 pb y máximo de 554,332pb.

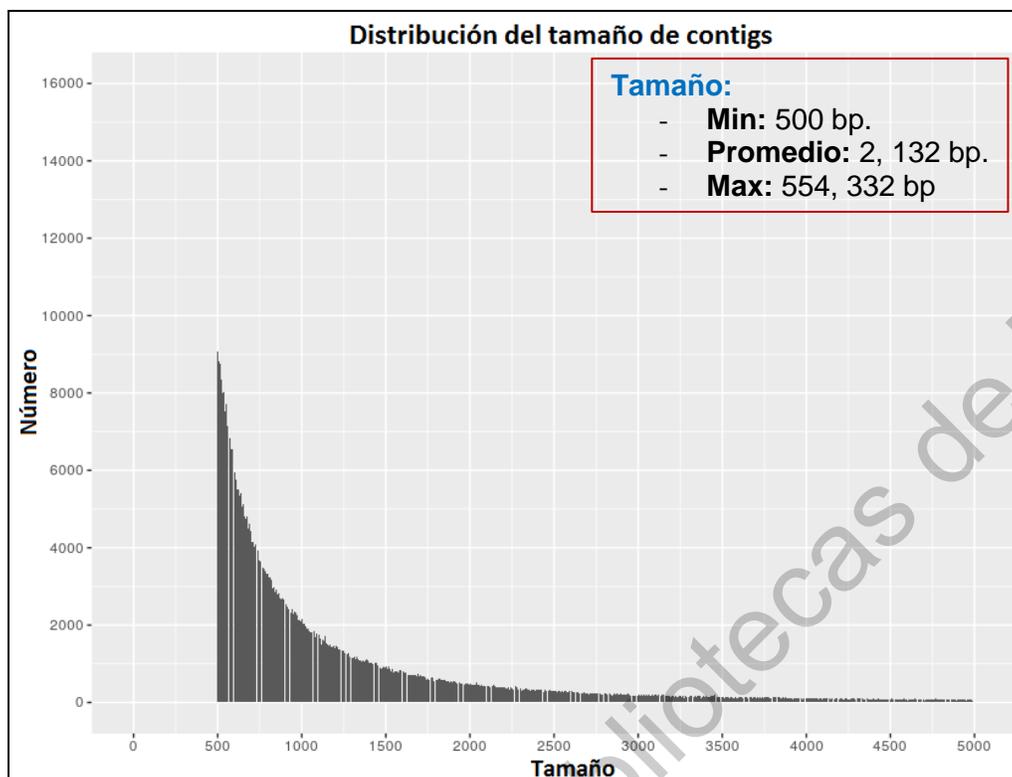


Figura 25. Distribución del tamaño de contigs. MetaHIT-Ch.

5.3.1. Perfil taxonómico

Debido a que se observan diferencias en el número de contigs de cada individuo se procedió a realizar una correlación sobre el total de asignaciones a hongos en función del número de contigs de cada individuo como se observa en la **Figura 26**. En dicha figura se muestra una tendencia positiva en la detección de hongos mientras mayor número de contigs posea la muestra de cada individuo.

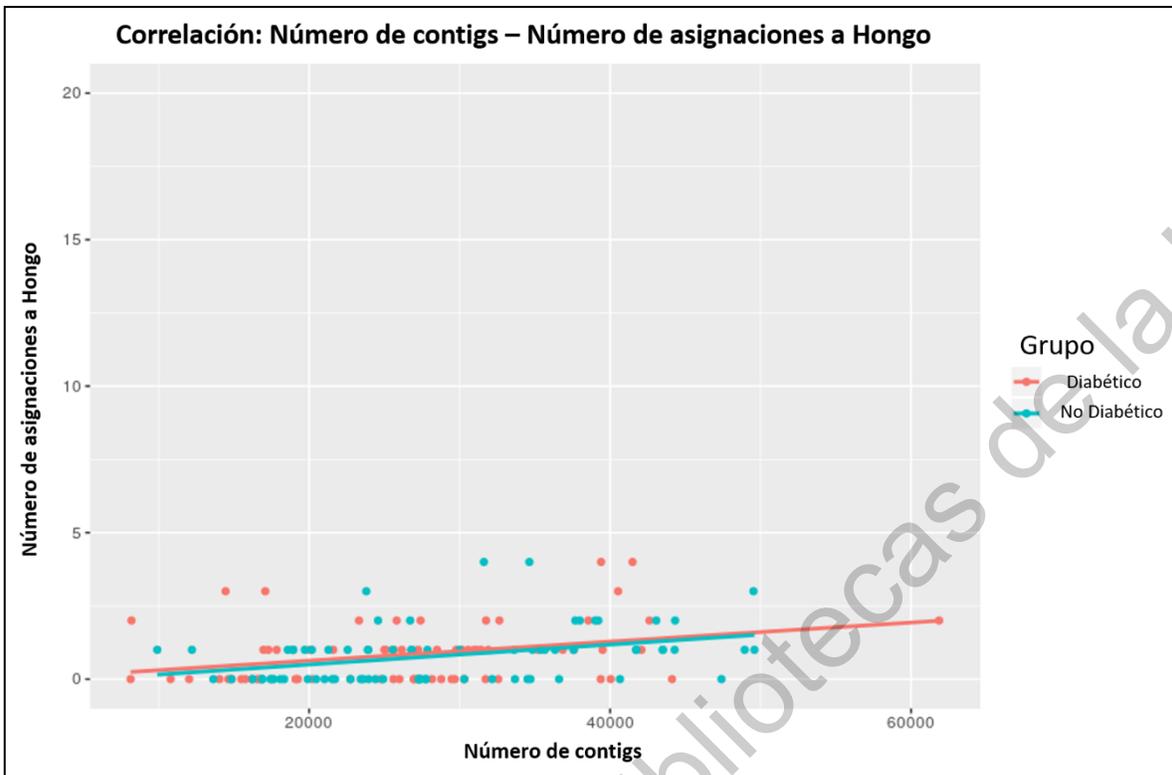


Figura 26. Se muestra la relación del número de asignaciones a hongos en función del número de contigs de cada individuo. Cada punto representa a un individuo distinto del grupo MetaHIT-Ch.

Debido a lo anterior, se procedió a realizar un análisis complementario con la herramienta MEGAN6-CE mediante el modo de asignación *readCount* para observar el número de géneros detectados por individuo debido a que éste se encuentra fuertemente relacionado al número de contigs. Como resultado de este análisis podemos observar en la **Figura 27** que el número de asignaciones de hongo a nivel de género varía entre individuo, además, el número de asignaciones a hongo se encuentra en muy baja proporción en este set de datos. También es posible observar que varios individuos no obtuvieron asignación alguna de hongos a nivel de género.

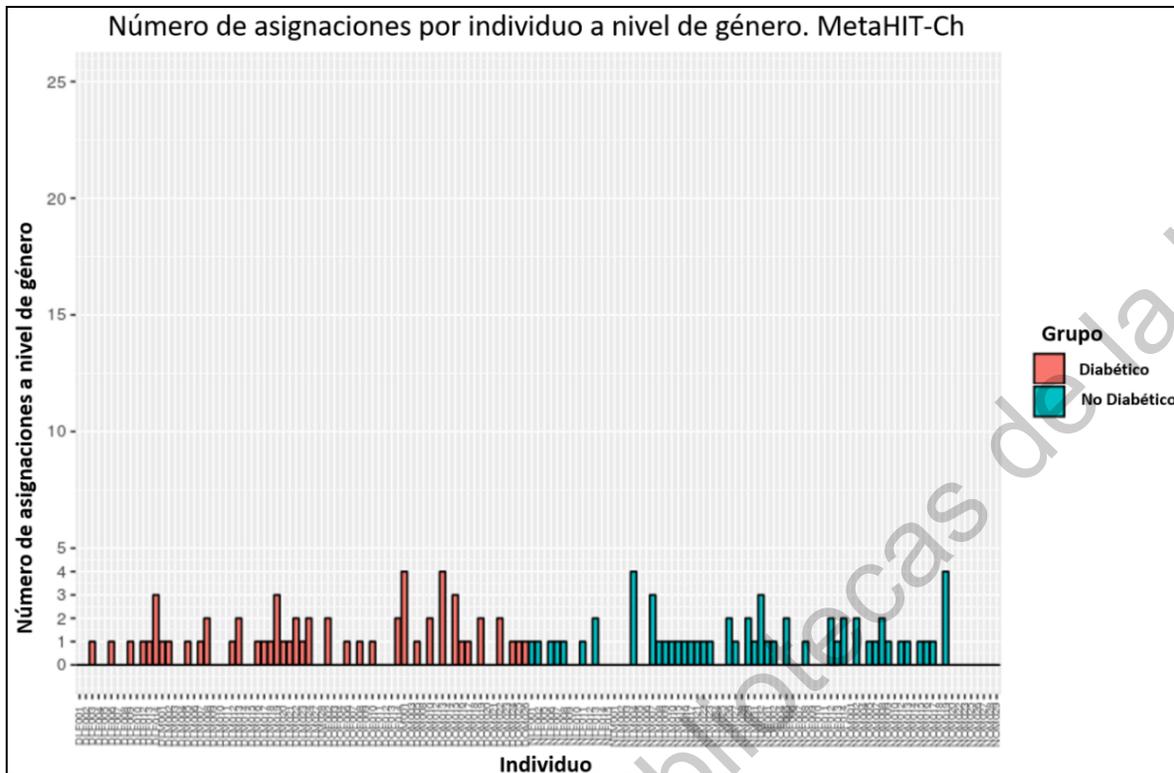


Figura 27. Número de géneros de hongos asignados por individuo en el set de datos MetaHIT-Ch.

Se realizó la construcción de un perfil taxonómico de ambos grupos, con el fin de observar si existen cambios asociados a un estado de salud y enfermedad en Diabetes Tipo 2.

A nivel de Filo (**Figura 28**), la abundancia de los distintos fila asignados es prácticamente similar en ambos grupos, estos son: Ascomycota, Basidiomycota y Microsporidia.

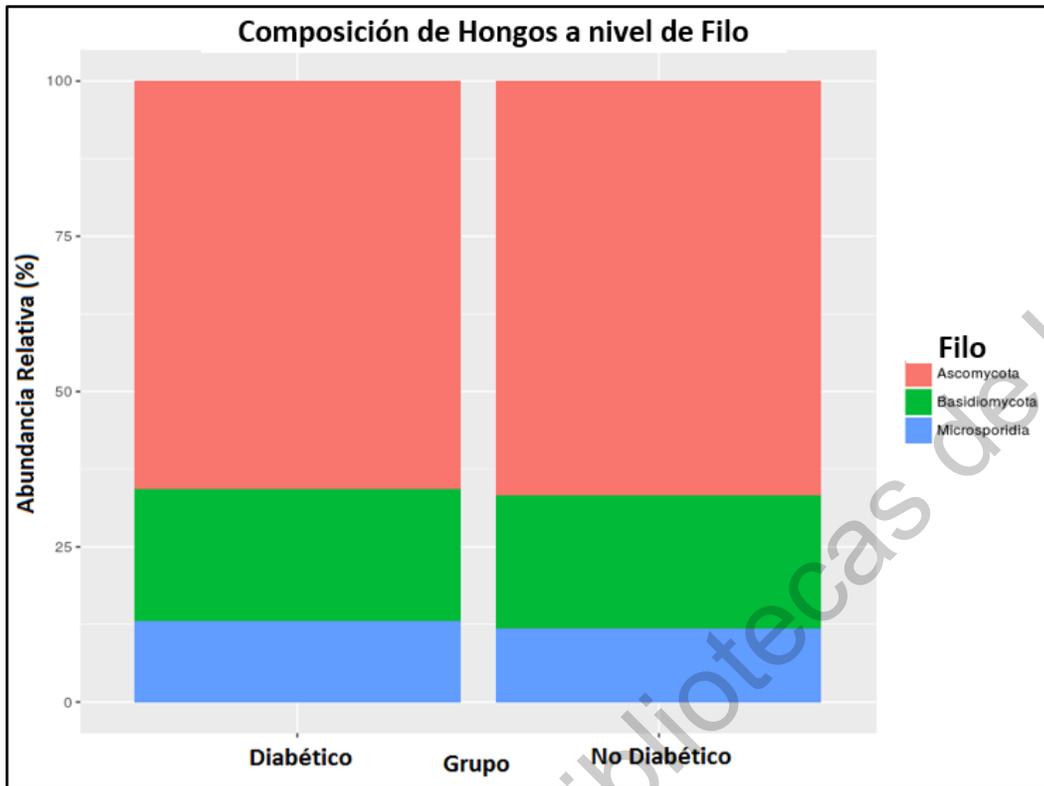


Figura 28. Asignación taxonómica a nivel de Filo en grupos del MetaHIT-Ch.

A nivel de Género (**Figura 29**) se realizó la asignación taxonómica en ambos grupos, observándose diferencias en cuanto a composición y abundancia. En el grupo de individuos diabéticos se lograron detectar 59 géneros de hongos distintos, mientras que en el de no diabéticos se detectaron 53. A pesar de que se logró la detección de mayor número de géneros en las personas diabéticas, el índice de Shannon nos indica que el grupo de individuos no diabéticos muestra mayor diversidad de hongos, esto debido a que con un número menor de bases alineadas se pudo detectar un mayor número de hongos en comparación con el grupo de personas diabéticas. En la **Figura 29** podemos observar que en ambos grupos el género que presenta mayor abundancia es *Enterocytozoon*, además, en el grupo de individuos diabéticos el género *Saccharomyces* se encuentra más representado en comparación al grupo que no presenta diabetes. En ambos grupos no destaca un género que gobierne fuertemente la comunidad, la mayor parte de los miembros de la comunidad se encuentran en equilibrio de abundancias, lo cual se refleja en los índices de Simpson obtenidos.

Tabla 10. Índices de diversidad en ambos grupos del MetaHIT-Ch.

	Diabético	No Diabético
Núm. individuos	71	74
Núm. géneros detectados	59	53
Bases alineadas	30 177	18 411
Shannon (diversidad α)	2.78	3.09
Simpson (diversidad α)	0.84	0.87

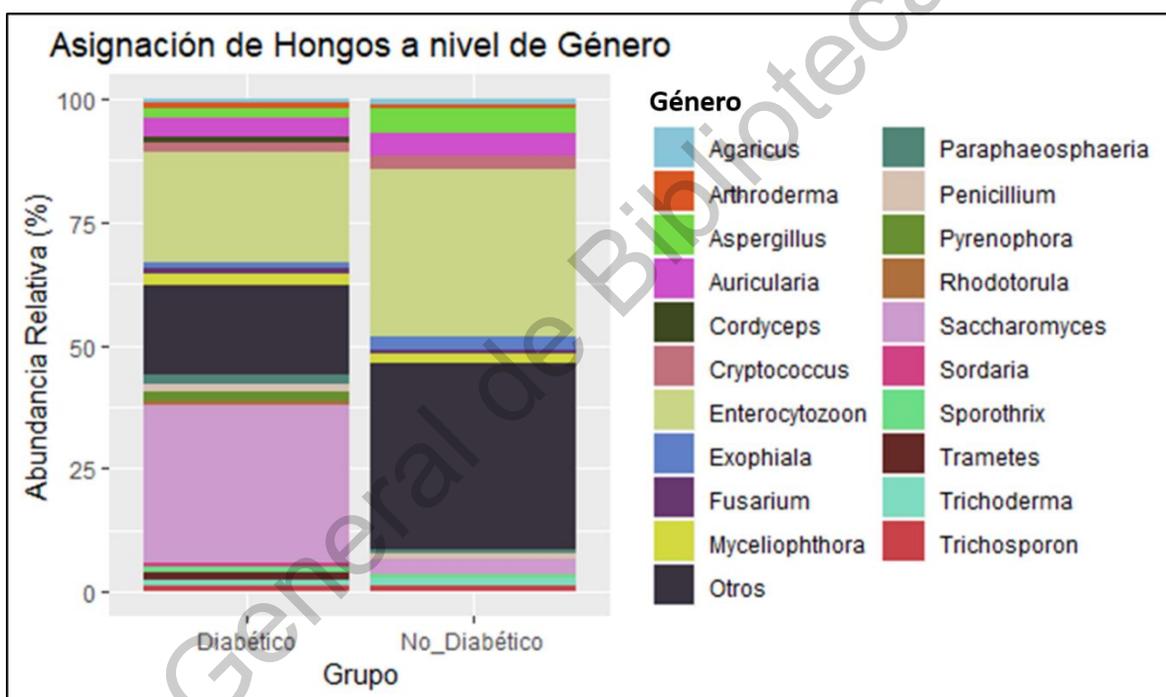


Figura 29. Asignación taxonómica a nivel de género en grupos del MetaHIT-Ch.

5.4. Comparación del perfil taxonómico entre poblaciones China-Europea

Con el objetivo de realizar una comparación entre el perfil taxonómico de ambas poblaciones pertenecientes al **MetaHIT**, se realizó una asignación taxonómica utilizando MEGAN6-CE de todos los individuos sanos de la población China (n=74) y de la población Europea (n=99).

A nivel de Filo (**Figura 30**) se lograron detectar 5 fila de hongos distintos: Ascomycota, Basidiomycota, Chytridiomycota, Microsporidia y Mucoromycota. La proporción obtenida de cada Filo entre poblaciones se comparte fuertemente. Siendo el filo Ascomycota el que presenta la mayor proporción (42.9% Población China, 56.47% Población Europea), seguido de los fila Microsporidia y Basidiomycota.

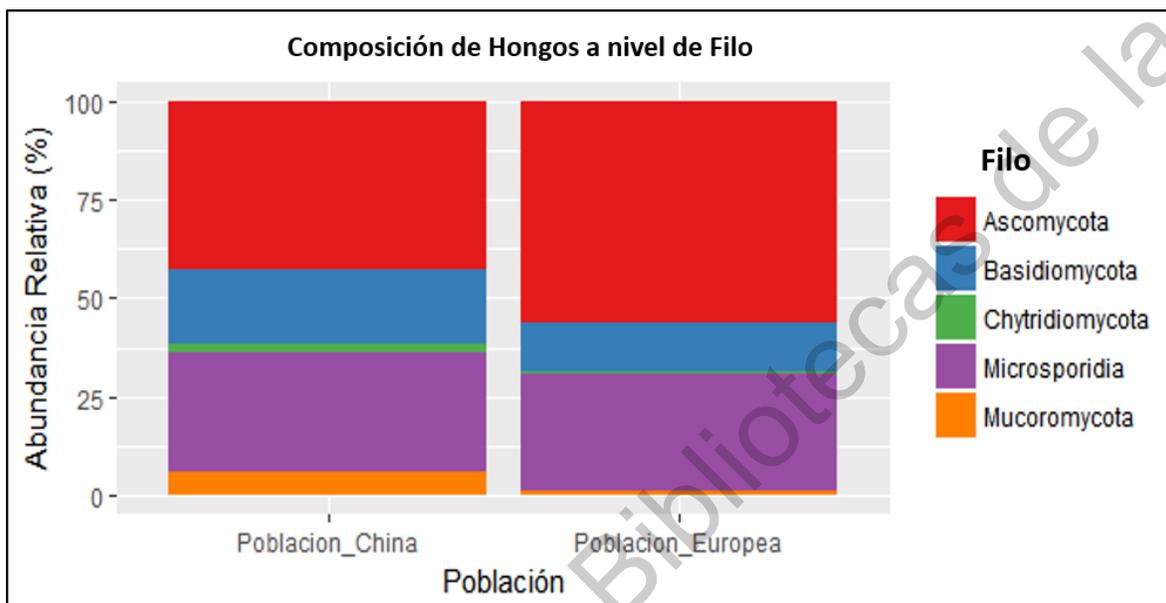


Figura 30. Perfil taxonómico a nivel de Filo entre las dos poblaciones del MetaHIT, Europeos y Chinos.

A nivel de género (**Figura 31**), se observa la presencia de *Aspergillus*, *Auricularia*, *Candida*, *Enterocytozoon*, *Saccharomyces*, entre otros. *Enterocytozoon* es el género con mayor abundancia que se comparte en ambas poblaciones, por su parte, *Saccharomyces* en la población europea representa el 34.6% del total de géneros asignados, mientras en la población china representa solamente el 3.1% del total. De acuerdo al índice de Shannon, la población de individuos chinos presenta mayor diversidad, esto debido a que, a pesar de que el número de bases alineadas sea menor, se logró una detección elevada de géneros. En ambas poblaciones el índice de Simpson es similar, lo que nos indica que no hay un género específico gobernando la comunidad a pesar de las diferencias en las proporciones obtenidas.

Tabla 11. Índices de diversidad en ambas poblaciones: Chinos, Europeos.

	Europeos	Chinos
Núm. individuos	99	74
Núm. géneros detectados	103	53
Bases alineadas	176 707	18 411
Shannon (diversidad α)	2.51	3.09
Simpson (diversidad α)	0.77	0.87

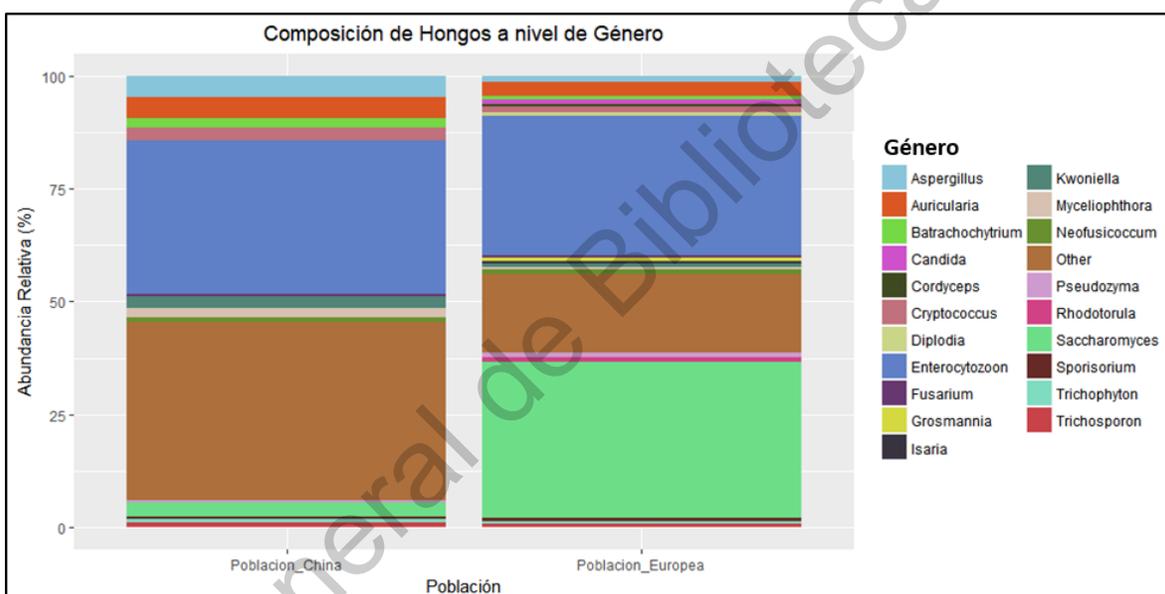


Figura 31. Perfil taxonómico a nivel de género entre las dos poblaciones del MetaHIT, Europeos y Chinos.

6. Discusión

Las distintas zonas del cuerpo humano como la piel, mucosas y el tracto gastrointestinal representan hábitats potenciales para la colonización de distintos microorganismos. A las distintas comunidades de microorganismos que se encuentran habitando las diferentes partes del cuerpo humano se les conoce como microbiota. Los miembros de la microbiota pueden ser procariontes como bacterias y arqueas, así como eucariontes como los

protozoarios y hongos. Otro grupo de microorganismos que también se distinguen dentro de estas comunidades son los virus. Los miembros de la microbiota humana se encuentran en una relación de comensalismo con el hospedero. Sin embargo, en ocasiones, cuando las condiciones del ambiente lo permiten, algunos de los miembros de la microbiota pueden reproducirse de manera exponencial generando el desarrollo de cuadros de enfermedad, como en el caso de algunos géneros de hongos. Por ello, el estudio de la diversidad de los miembros de la microbiota humana es esencial para comprender de qué manera estas comunidades se encuentran asociadas al mantenimiento de estados de salud o el desarrollo de cuadros de enfermedad en el hospedero humano.

A pesar de los esfuerzos por parte de consorcios como el **HMP** y el **MetaHIT** por conocer la diversidad de microorganismos en el cuerpo humano, aún se desconoce gran parte de su estructura. Los mayores avances que se han realizado durante los últimos años se basan principalmente en la descripción de las comunidades bacterianas. Sin embargo, otros miembros de la microbiota humana como los hongos permanecen poco estudiados.

Los hongos, a pesar de su baja abundancia llevan a cabo funciones esenciales que permiten el mantenimiento de las comunidades microbianas en el ser humano. Debido a lo anterior, es esencial conocer cómo se encuentran organizadas las comunidades fúngicas dentro de la microbiota humana.

Con el objetivo de aportar información a este ramo de estudio, en el presente trabajo de tesis se realizó la determinación del perfil taxonómico de hongos, derivado de secuencias metagenómicas provenientes de tres proyectos de investigación asociados a diferentes poblaciones humanas. Los datos analizados fueron archivos de secuencias de contigs de tres estudios de los dos mayores consorcios de investigación sobre microbiomas humanos durante la última década: **HMP** y **MetaHIT**. La importancia del análisis de secuencias asociadas a hongos de estos conjuntos de datos radica en que, en el caso del **HMP**, las secuencias provienen de un estudio metagenómico tipo *shotgun* que involucra a un total de 300 individuos sanos en donde se muestrearon 16 zonas del cuerpo. Los resultados de este set de datos nos permitieron observar la diversidad y abundancia de hongos presentes en las distintas zonas del cuerpo de individuos sanos, lo cual, junto con datos de la literatura, nos permitió obtener un mayor panorama de las comunidades de hongos asociadas a los seres humanos. Por otra parte, en el caso del consorcio **MetaHIT**, en este estudio se identificaron las secuencias asociadas a hongos en individuos sanos y en individuos que presentan una enfermedad crónica, ya sea diabetes en el caso de la población china, o

enfermedad inflamatoria intestinal en el caso de la población europea. Esto nos permitió: (1) obtener un panorama general de las comunidades de hongos en individuos sanos de diferentes poblaciones, y (2) observar si existía una diferencia general de la estructura de las comunidades de hongos a nivel de diversidad y abundancia entre los grupos de casos y controles.

La metodología empleada se basa principalmente en el uso de una herramienta de alineación de secuencias de ADN (blast 2.3.0), y una herramienta que asigna dichas secuencias a un árbol taxonómico para determinar su perfil (MEGAN6-CE). Esta metodología cuenta con dos propiedades principales: es robusta y reproducible. La robustez de la metodología está dada por el uso de herramientas ampliamente distribuidas y utilizadas en los diferentes sistemas operativos, esto permite, además de otras cosas, que el flujo de análisis sea resiliente a errores y *bugs* por parte de las herramientas utilizadas. Aunado a esto, el uso de dichas herramientas también permite que los resultados obtenidos sean reproducibles, debido a que los datos metagenómicos con los que se trabajó se encuentran disponibles de manera pública en los repositorios mencionados en la sección de Metodología, además de que se detallan las versiones de los programas utilizados, así como los parámetros con los que se trabajó en cada fase de la metodología previamente descrita.

A nivel general, los resultados obtenidos en este trabajo concuerdan con datos reportados en la literatura sobre la composición de hongos en la microbiota humana mediante distintas metodologías, entre las que se encuentran micobiotas intestinales determinados por secuencias ITS en individuos sanos del HMP (Nash *et al.*, 2017), micobiotas obtenidos de individuos con IBD mediante la amplificación de la región 18S ribosomal (Ott *et al.*, 2008), micobiotas orales mediante ITS (Diaz *et al.*, 2017), y el micobioma de la piel mediante 18S e ITS (Findley *et al.*, 2013).

En este trabajo observamos una gran diferencia en la composición de los datos de secuencias entre los diferentes grupos de cada uno de los *sets* de datos evaluados. Por una parte, en los datos del **HMP** se evaluaron 15 sitios del cuerpo que pertenecen a 5 zonas generales: vías respiratorias (1 sitio), tracto gastrointestinal (1 sitio), zona oral (8 sitios), piel (2 sitios), y tracto urogenital (3 sitios). En la **Tabla 7** de la sección de **Resultados** se observa el número de muestras y el número de contigs asociado a cada sitio del cuerpo. El número de muestras por sitio es muy variable, presentándose en el rango de 2 muestras en el caso de la vagina media, y hasta 136 muestras en el caso de las heces fecales. El número de

muestras está directamente relacionado con el número de contigs asociado a cada sitio, y también se correlaciona con la presencia o no de hongos detectados por la metodología propuesta. En la **Tabla 7** se observa que en aquellos sitios en donde el número de muestras evaluadas fue mayor, se logran detectar secuencias asociadas a hongos. Esta observación aplica principalmente para la nariz anterior, mucosa bucal, fórnix posterior, heces fecales y la placa supragingival. Sin embargo, al evaluar el total de muestras de cada sitio en función del número de contigs y de la presencia y/o ausencia de hongos, no se logra observar una relación lineal entre estas variables como se ilustra en la **Figura 15**. Esto nos indica que, si bien el número de muestras y el número de contigs asociados a dicho esfuerzo de muestreo son un factor importante en la determinación de las comunidades de hongos en estudios metagenómicos, estos factores no son absolutos. Existen otro tipo de factores, en su mayoría biológicos que explican estas diferencias. Un ejemplo de esto lo podemos observar en la zona del dorso lingual en la **Tabla 7**. Esta zona tiene una representación de 122 muestras con un total de 264,916 contigs, siendo la segunda zona con el mayor número de muestras en el set de datos. A pesar de estas características, en esta zona no se logró realizar la detección de hongos, esto sugiere que, el dorso lingual por sus propiedades biológicas no es una zona en la cual se genere el establecimiento, desarrollo y colonización de hongos.

En el caso de los proyectos del **MetaHIT**, la principal diferencia observable en los conjuntos de datos radica en el número de individuos dentro de los grupos de casos y controles. Esto es más notorio en el **MetaHIT-Eur**. En este set de datos el grupo de casos con IBD involucra a 25 individuos con un total de 1,179,609 contigs, representando así al 17.93% del total de secuencias. Mientras que el grupo control se compone por 99 individuos con un total de 5,397,427 contigs, lo que representa el 82.06% del total de secuencias. Por otro lado, en el **MetaHIT-Ch** se observan diferencias menores entre el total de individuos participantes en cada uno de los grupos evaluados. El grupo de casos con T2D agrupa a un total de 71 individuos con un total de 1,885,547 contigs, representando así el 46.97% del total de secuencias. Mientras que el grupo control agrupa a 74 individuos con un total de 2,131,450 contigs, lo que representa el 53.02% del total de secuencias.

Como se logra observar, en cada uno de los tres conjuntos de datos analizados existen diferencias en la estructura y cantidad de secuencias disponibles para su análisis. Conscientes de estas diferencias y de los posibles sesgos que estas pueden generar en el perfil taxonómico obtenido en cada uno de los conjuntos de datos, continuamos

cautelosamente con este trabajo de tesis, y obtuvimos las observaciones que a continuación se describen.

- **El Proyecto del Microbioma Humano (HMP)**

De los 15 sitios del cuerpo evaluados en este set de datos, se logró detectar la abundancia de hongos en 7 de ellos: nariz anterior, mucosa bucal, cresta retroauricular izquierda y derecha, fórnix posterior, heces fecales y placa supragingival. Los sitios asociados a la zona oral, a pesar de que representan la mayor parte de los sitios evaluados (8 sitios), fueron en los que en su mayoría no se logró realizar la detección de secuencias asociadas a hongos, como se muestra en la **Figura 13**. De acuerdo a nuestros resultados, podemos sugerir que la placa supragingival es el hábitat bucal que presenta las características ideales para la colonización y establecimiento de comunidades fúngicas, en especial de miembros del género *Candida*, cuya presencia en esa zona se encuentra gobernando la comunidad. Esta observación además concuerda con información de la literatura en donde se hace mención de que los tejidos de las placas de las amígdalas (gingiva) palatinas presentan un recambio constante de células, por lo cual su organización tisular muestra una conformación de unicapa, lo que hace a esta zona más susceptible a la colonización de distintos microorganismos como *Candida*, debido a los bajos niveles de respuesta inmune que se presentan en dicha zona (Southern *et al.*, 2008).

Las zonas en las que se logró asignar la mayor cantidad de contigs a cualquiera de los taxa representados en el árbol taxonómico de NCBI son la nariz anterior y el fórnix posterior. Mientras que la zona en donde se observa la menor proporción de esta asignación de contigs son las heces fecales (ver **Tabla 8**). Esto puede indicar, por un lado, que tanto en la nariz anterior como en el fórnix posterior existe una sobrerrepresentación (riqueza) de un número limitado de taxa microbianos que permiten que la cantidad de contigs asignados sea elevada; o que los distintos miembros de la microbiota en esas zonas se encuentran bien representados en la base de datos no redundante de nucleótidos, y que la proporción de especies que faltan por describir en esas zonas son relativamente pocas. En las heces fecales la proporción de contigs asignados es del 17.42%, esto representa alrededor de 388 mil contigs de los 2,225,881 que se analizaron. Esto sugiere que existe una elevada cantidad de especies de microorganismos del tracto gastrointestinal que aún no se encuentran representadas en la base no redundante de NCBI. Sin embargo, gracias a algunos trabajos recientes en donde se describen nuevos taxa de microorganismos derivado del análisis de metagenomas del tracto gastrointestinal humano (como el trabajo

de Almeida y colaboradores detallado la sección 1.7 de este documento), puede ser posible que la proporción de contigs no asignados a algún taxa en muestras de heces fecales, como lo sucedido en este trabajo de tesis, disminuyan en trabajos futuros.

Las zonas asociadas a epitelios expuestos son las que presentan la mayor proporción de contigs asignados que mapean a Hongo. En este set de datos estos sitios son las crestas retroauriculares izquierda y derecha, con un 44.23% y un 27.24% del total de contigs asignados que mapean a Hongo, respectivamente. Estos números relativamente elevados se deben a la elevada presencia de miembros del género *Malassezia* asociadas a la piel. Por su parte las heces fecales y la mucosa bucal son los sitios en los cuales esta proporción de mapeo a hongos fue menor, encontrándose en ambos casos en 0.04%.

Las heces fecales son el sitio que presenta la mayor diversidad de hongos asignados. Una característica importante de esta zona en comparación con los demás sitios, se basa en que la mayor parte de los miembros del microbioma se encuentran distribuidos de manera relativamente homogénea, y no se encuentra un género gobernando fuertemente la comunidad como en el caso de la piel o mucosas bucal y vaginal.

De manera general las principales observaciones que podemos hacer en este set de datos son que:

- 1) *Malassezia* se encuentra en mayor abundancia en zonas asociadas a la piel
- 2) *Candida* se encuentra en mayor abundancia en zonas asociadas a mucosas
- 3) El tracto intestinal presenta al filo Microsporidia, representado por el género *Enterocytozoon* abundante en heces fecales
- 4) Las heces fecales son la zona con la mayor diversidad de hongos

- **El metagenoma del tracto intestinal humano (MetaHIT)**

En ambos estudios del **MetaHIT** se obtuvo un archivo de contigs que pertenece a cada individuo. Debido a este tipo de estructura y organización de los datos fue posible realizar un análisis más puntual por cada individuo, a diferencia de los resultados mostrados en el **HMP**.

En el **MetaHIT-Eur** observamos que la cantidad de contigs asociado al archivo de cada individuo es variable, encontrándose en el rango de 13,247 a 144,440 contigs por individuo. Esta observación es importante debido a que como se muestra en la **Figura 20**, existe una

relación lineal entre el número de contigs asociados a cada individuo y el número de géneros de hongos que logran ser detectados. En este set de datos el número de individuos pertenecientes a los grupos de casos y controles son muy variables. Algunos estudios como el de Nash y colaboradores (Nash *et al.*, 2017) muestran que existe una elevada inter e intravariabilidad de los componentes del microbioma en distintos individuos pertenecientes a una misma población, además de esto, también muestran que son pocos los individuos que explican la mayor parte de la diversidad de hongos en estudios de microbiomas humanos. En este set de datos se observa un fenómeno similar. Mediante un análisis complementario en MEGAN-CE utilizando el modo de asignación *readCount* (**Figura 21**) se observa que una alta proporción de individuos (88.8%) presenta 10 o menos asignaciones taxonómicas de hongos a nivel de género, mientras que el 12.2% del total de individuos presenta más de 10 asignaciones taxonómicas. Además, 6 individuos no presentan ninguna asignación taxonómica a hongo en este análisis complementario.

Los hongos que se encuentran en mayor proporción en el tracto gastrointestinal de esta población son los que pertenecen al filo Ascomycota, seguido del filo Basidiomycota y Microsporidia. Otros filo detectados en muy baja proporción son Mucoromycota (2%) y Chytridiomycota (0.82%). En el perfil taxonómico a nivel de género, a pesar de la diferencia del número de individuos en cada grupo, se observa un notorio patrón de elevada abundancia de los géneros *Saccharomyces*, y *Enterocytozoon* en ambos grupos. En ambos grupos la suma de las abundancias de estos dos géneros es de aproximadamente el 65% del total de los integrantes de la comunidad de hongos detectados. Los géneros restantes, cada uno representa en promedio el 0.8% del total. Las diferencias que se observan entre el número de géneros detectados entre los grupos de casos y controles se explican principalmente por la diferencia de individuos de cada grupo. A pesar de la diferencia en la estructura de cada set de datos, no se observa una elevada diferencia entre el índice de Shannon del grupo de casos (2.31) y del grupo control (2.51). Sin embargo, estas observaciones son únicamente cualitativas. Con los resultados presentados en este trabajo de tesis no es posible indicar si estas diferencias observadas a nivel de diversidad y de abundancia de cada género entre ambos grupos son significativas.

Las principales observaciones en este set de datos son:

- 1) La abundancia a nivel de Filo se comparte en ambos grupos

- 2) *Saccharomyces* y *Enterocytozoon* son los dos géneros más abundantes en tanto en el grupo de casos como en los controles
- 3) El nivel de diversidad observado por el índice de Shannon en ambos grupos es muy similar

En el **MetaHIT-Ch** también se observa una elevada variación en la estructura de los archivos de contigs por cada individuo. El número de contigs de cada archivo se encuentra entre un rango de 8,138 a 61,865 contigs. Esto es importante debido a que como se muestra en la **Figura 26**, existe una correlación entre el número de contigs asociados a cada individuo y el número de asignaciones taxonómicas a Hongo. El número de individuos pertenecientes a cada grupo en este set de datos es muy similar (71 casos y 74 controles). Además, como ya se mencionó en la primera parte de la discusión, ambos grupos poseen una proporción similar de contigs. El número de contigs es una de las variables que explica el número de asignaciones de hongos entre los individuos. Esto también es cierto a nivel del set de datos. Mientras que en el set de datos **MetaHIT-Eur** los individuos poseen en promedio ~50,000 contigs, en el **MetaHIT-Ch** el promedio de número de contigs por individuo es de ~15,000. Es por ello que en un análisis complementario con MEGAN-CE utilizando un modo de asignación por *readCount* (**Figura 27**) en este set de datos encontramos que en 65 individuos (44% del total) no se logra realizar una asignación a Hongo. Mientras que 54 (37%) sólo presentan 1 asignación, 17 individuos (11%) 2 asignaciones, 5 individuos (3.44%) 3 asignaciones, y finalmente en 4 individuos participantes (2%) se logró la detección de 4 géneros de hongos en cada uno de ellos.

A nivel de Filo ambos grupos (casos y controles) presentan un nivel de abundancia muy similar entre la estructura de los fila detectados, siendo el filo Ascomycota el más abundante, en ambos grupos representando más del 50% del total de los fila detectados. Le siguen en abundancia el filo Basidiomycota y Microsporidia. A nivel de género, la principal diferencia que se observa entre los dos grupos es que el género *Saccharomyces* presenta una mayor abundancia relativa en individuos con T2D (31.8%) en comparación con el grupo control (3.1%). En ambos grupos el género *Enterocytozoon* es uno de los más abundantes, en el grupo con diabetes representa el 22.5% del total de géneros detectados, mientras que en el grupo no diabético representa el 33.88%. Los géneros restantes detectados se encuentran presentes con un nivel de abundancia más homogéneo encontrándose en un rango de 0.5 a 3% por cada género.

Las principales observaciones en este set de datos son:

- 1) La composición y abundancia a nivel de Filo se comparte en ambos grupos
- 2) *Saccharomyces* es un género mayormente representado en el grupo con diabetes tipo 2
- 3) *Enterocytozoon* es el género que presenta la mayor abundancia en ambos grupos
- 4) Ambos grupos presentan un nivel de diversidad semejante de acuerdo al índice de Shannon

- **Comparación entre poblaciones.**

Las diferencias que se observan en la estructura de la microbiota intestinal humana se explican en gran proporción por factores ambientales como el estilo de vida, la alimentación y la industrialización del ambiente. A nivel del propio hospedero un factor importante en este sentido es la ascendencia genética, la cual difiere en las diferentes poblaciones humanas.

Realizando la comparación de la estructura de hongos entre las dos poblaciones pertenecientes al **MetaHIT** (europeos y chinos) observamos que, a nivel de Filo, ambas poblaciones comparten una proporción similar de los fila detectados. Por su parte el filo Ascomycota es el más abundante en ambas poblaciones, seguido de Microsporidia y Basidiomycota. Otro Filo que se logra detectar en menor proporción es Mucoromycota. A nivel de género la principal diferencia es la alta abundancia de *Saccharomyces* en el grupo europeo (34%), mientras que en la población china solamente representa un 3% de abundancia relativa en la población. En ambos casos *Enterocytozoon* es un género con alta abundancia (~30%). El número de géneros detectados en la población europea (n=99) es de 103, mientras que en la población China (n=74) es de 53. De acuerdo al índice de Shannon la población china presenta una mayor diversidad en comparación con la europea. Esto puede deberse al número de bases alineadas con el que se determinó el perfil taxonómico. El número de bases alineadas de la población china representa solamente el 10% del total que se alineó en el grupo de europeos. Por lo cual se espera que, al aumentar el número de bases alineadas en el grupo de individuos chinos, este sobrepase el número de géneros detectados en la población europea.

7. Conclusión

Como resultado de este trabajo podemos concluir que, el análisis de secuencias derivadas de proyectos metagenómicos para la determinación de poblaciones de hongos muestra algunas limitantes. La principal de ellas es la proporción de secuencias asociadas a hongos que se obtuvieron del total de los datos, la cual se encuentra en ~0.1% del total de secuencias. También fue posible concluir que, en el caso del análisis de secuencias de hongos por individuo, existe una elevada intra variabilidad entre aquellos que pertenecen a una misma población. Esta variabilidad podría ser explicada por el estilo de vida y alimentación del propio individuo; y por la variación genética entre los individuos dentro de una población determinada.

Como parte de este trabajo se realizó un perfil taxonómico de hongos a nivel poblacional derivado de datos de secuencias metagenómicas. El análisis de los datos del **HMP** permitió concluir que las zonas asociadas a mucosas en el cuerpo humano presentan una baja diversidad y se encuentran gobernadas principalmente por géneros de *Candida*, mientras que las zonas sebáceas de la piel son gobernadas por miembros del género *Malassezia* y también presentan una baja diversidad de hongos. Por su parte la zona gastrointestinal, en ambos proyectos (**HMP** y **MetaHIT**), es la zona que presenta una mayor diversidad de géneros de hongos, entre los cuales destacan los géneros *Saccharomyces* y *Auricularia*, al igual que el protista intracelular *Enterocytozoon* clasificado como hongo basal por el programa de clasificación taxonómica MEGAN. También se encontraron otros géneros en una proporción homogénea dentro de la comunidad de hongos pertenecientes a la zona gastrointestinal. En conclusión, observamos que la zona gastrointestinal no se encuentra gobernada por un género en particular, a diferencia de las mucosas y los epitelios.

Obtuvimos los índices de Shannon y de Simpson tanto de casos como controles de ambos estudios del **MetaHIT**. Los resultados sugieren que no hay diferencias entre el grupo que presenta diabetes y el grupo control en el caso del **MetaHIT-Ch**; y entre el grupo con IBD y su grupo control en el **MetaHIT-Eur**. Sin embargo, como perspectivas a futuro de este trabajo el contraste de diversidad entre las muestras se deberá realizar formalmente mediante el uso de pruebas estadísticas. Además, como se hizo mención anteriormente, las mayores diferencias entre ambos grupos están explicadas por la variabilidad de géneros asociados a algunos individuos que explican la mayor parte de la diversidad de hongos. Y como se mencionó en la discusión, esto se encuentra completamente relacionado con el número de contigs asociados a cada individuo. Por lo cual, un análisis más detallado es

necesario para concluir si existe o no una diferencia en la estructura y composición de hongos entre ambos grupos como resultado de la presencia o ausencia de la enfermedad en estudio.

A nivel poblacional observamos que la estructura de hongos a nivel de género entre el grupo de individuos chinos y europeos es similar. Sin embargo, el género *Saccharomyces* se encontró con mayor abundancia en la población europea. Con respecto a la población china, creemos que este grupo presenta mayor diversidad de hongos detectados ya que el índice de Shannon obtenido de estos datos es el mayor entre las dos poblaciones.

En este trabajo se implementó una tubería de análisis para la generación del perfil taxonómico de comunidades fúngicas de datos de metagenomas públicos. Como resultado de estos análisis logramos identificar secuencias de contigs asociadas a hongos en distintos sitios del cuerpo, con un énfasis en la zona gastrointestinal. La proporción de secuencias asociadas a los diferentes taxa fúngicos es baja en cada uno de los sets de datos que se evaluaron. A pesar de esta limitante, logramos realizar un perfil taxonómico de hongos a nivel poblacional para cada set de datos, tanto a nivel de Filo como de género. También se realizó el cálculo de índices asociados al nivel de diversidad de las poblaciones de hongos observadas. Esto nos muestra y confirma la problemática existente sobre la necesidad de mejorar y estandarizar las técnicas en el estudio de comunidades de hongos, tanto en procesos de extracción y manipulación de la muestra, como de herramientas bioinformáticas para el mejoramiento en el estudio y comprensión de la estructura de hongos dentro de estas comunidades.

Bibliografía

Almeida, A. *et al.* (2019) 'A new genomic blueprint of the human gut microbiota', *Nature*, p. 1. doi: 10.1038/s41586-019-0965-1.

Altschup, S. F. *et al.* (1990) 'Basic Local Alignment Search Tool', *J. Mol. Biol.*, 215, pp. 403–410. Available at: http://ac.els-cdn.com/S0022283605803602/1-s2.0-S0022283605803602-main.pdf?_tid=b02bbf5a-4f90-11e7-a151-00000aab0f6c&acdnat=1497287030_bbc896b9cfe669c4cbc71ee5f17e3c94 (Accessed: 12 June 2017).

Blekhman, R. *et al.* (2015) 'Host genetic variation impacts microbiome composition across human body sites', *Genome Biology*. *Genome Biology*, 16(1), p. 191. doi: 10.1186/s13059-015-0759-1.

Brantley Hall, A., Tolonen, A. C. and Xavier, R. J. (2017) 'Human genetic variation and the gut microbiome in disease', *Nature Publishing Group*, 18. doi: 10.1038/nrg.2017.63.

Broussard, J. L. and Devkota, S. (2016) 'The changing microbial landscape of Western society : Diet , dwellings and discordance', *Molecular Metabolism*. Elsevier GmbH, 5(9), pp. 737–742. doi: 10.1016/j.molmet.2016.07.007.

Cani, P. D. (2017) 'Gut microbiota — at the intersection of everything?', *Nature Reviews Gastroenterology & Hepatology*. Nature Publishing Group, 14(6), pp. 321–322. doi: 10.1038/nrgastro.2017.54.

Cani, P. D. and de Vos, W. M. (2017) 'Next-generation beneficial microbes: The case of *Akkermansia muciniphila*', *Frontiers in Microbiology*, 8(SEP), pp. 1–8. doi: 10.3389/fmicb.2017.01765.

Cantarel, B. L., Lombard, V. and Henrissat, B. (2012) 'Complex carbohydrate utilization by the healthy human microbiome', *PLoS ONE*, 7(6), pp. 1–10. doi: 10.1371/journal.pone.0028742.

Caporaso, J. G. *et al.* (2018) 'Current understanding of the human microbiome', *Nature Medicine*, 24(4), pp. 392–400. doi: 10.1038/nm.4517.

Carding, S. *et al.* (no date) 'Dysbiosis of the gut microbiota in disease'. doi: 10.3402/mehd.v26.26191.

CBD (1992) 'CONVENTION ON BIOLOGICAL DIVERSITY UNITED NATIONS 1992'. Available at: <https://www.cbd.int/doc/legal/cbd-en.pdf> (Accessed: 8 February 2019).

Chassaing, B. *et al.* (2012) 'Gut Microbiota Drives Metabolic Disease in Immunologically Altered Mice', *Advances in Immunology*. Academic Press, 116, pp. 93–112. doi: 10.1016/B978-0-12-394300-2.00003-X.

Clemente, J. C. *et al.* (2012) 'The Impact of the Gut Microbiota on Human Health: An Integrative View'. doi: 10.1016/j.cell.2012.01.035.

Costea, P. I. *et al.* (2017) 'Subspecies in the global human gut microbiome', *Molecular Systems Biology*, (13), pp. 1–11. doi: 10.15252/msb.20177589.

Cui, L., Morris, A. and Ghedin, E. (2013) 'The human mycobiome in health and disease', *Genome Medicine*, 5(63), pp. 1–12. doi: 10.1186/gm467.

Dalal, A. and Atri, A. (2014) 'An Introduction to Sequence and Series', *International Journal of Research*, 1(10), pp. 1286–1292. doi: 10.1002/0471250953.bi0301s42.An.

DeFronzo, R. A. *et al.* (2015) 'Type 2 diabetes mellitus', *Nature Reviews Disease Primers*. Macmillan Publishers Limited, 1(July), pp. 1–23. doi: 10.1038/nrdp.2015.19.

Diaconu, S. *et al.* (2017) 'Helicobacter pylori infection: old and new.', *Journal of medicine and life*, 10(2), pp. 112–117. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/28616085> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5467250>.

Diaz, P. I. *et al.* (2017) 'Mining the oral mycobiome: Methods, components, and meaning', *Virulence*, 8. doi: 10.1080/21505594.2016.1252015.

Dollive, S. *et al.* (no date) 'A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples'. doi: 10.1186/gb-2012-13-7-r60.

Erny, D. and Prinz, M. (2017) 'Microbiology: Gut microbes augment neurodegeneration', *Nature*. doi: 10.1038/nature21910.

Falony, G. *et al.* (2016) 'Population-level analysis of gut microbiome variation', *Science*, 352(6285), pp. 560–564. doi: 10.1126/science.aad3503.

Federhen, S. (2012) 'The NCBI Taxonomy database', *Nucleic Acids Research*. Oxford University Press, 40(D1), pp. D136–D143. doi: 10.1093/nar/gkr1178.

Federhen, S. (2013) 'Chapter 4: The Taxonomy Project', in *The NCBI Handbook*, pp. 1–31. Available at: <https://pdfs.semanticscholar.org/1a81/aadadb0e16ddee0f701e69b3b8aeb64ef886.pdf> (Accessed: 14 March 2019).

Findley, K. *et al.* (2013) 'Topographic diversity of fungal and bacterial communities in human skin.', *Nature*. Nature Publishing Group, 498(7454), pp. 367–70. doi: 10.1038/nature12171.

Flicek, P. and Birney, E. (2009) 'Sense from sequence reads: Methods for alignment and assembly', *Nature Methods*, 6(11S), p. S6. doi: 10.1038/nmeth.1376.

Frank, D. N. *et al.* (2011) 'Investigating the biological and clinical significance of human dysbioses', *Trends in Microbiology*, 19(9), pp. 427–434. doi: 10.1016/j.tim.2011.06.005.

Fredricks, D. N., Fiedler, T. L. and Marrazzo, J. M. (2005) 'Molecular Identification of Bacteria Associated with Bacterial Vaginosis', *New England Journal of Medicine*, 353(18), pp. 1899–1911. doi: 10.1056/NEJMoa043802.

Ghannoum, M. A. *et al.* (2010) 'Characterization of the Oral Fungal Microbiome (Mycobiome) in Healthy Individuals', *PLoS Pathog*, 6(1). doi: 10.1371/journal.ppat.1000713.

Gibson, G. R. *et al.* (2017) 'Expert consensus document: The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of prebiotics', *Nature Reviews Gastroenterology and Hepatology*. Nature Publishing Group, 14(8), pp. 491–502. doi: 10.1038/nrgastro.2017.75.

Goodrich, J. K. *et al.* (2014) 'Conducting a microbiome study', *Cell*, 158(2), pp. 250–262. doi: 10.1016/j.cell.2014.06.037.

Goodrich, J. K. *et al.* (2016) 'Cross-species comparisons of host genetic associations with the microbiome', *Science*, 352(6285), pp. 532–535. doi: 10.1126/science.aad9379.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) 'Coming of age: ten years of next-generation sequencing technologies', *Nat Rev Genet*. Nature Publishing Group, 17(6), pp. 333–351. doi: 10.1038/nrg.2016.49.

Grabchak, M. *et al.* (2017) 'The generalized Simpson's entropy is a measure of biodiversity', *PLoS*

ONE, 12(3), pp. 1–11. doi: 10.1371/journal.pone.0173305.

Grice, E. A. and Dawson, T. L. (2017) 'Host–microbe interactions: *Malassezia* and human skin', *Current Opinion in Microbiology*, 40, pp. 81–87. doi: 10.1016/j.mib.2017.10.024.

Hill, C. *et al.* (2014) 'Expert consensus document: The international scientific association for probiotics and prebiotics consensus statement on the scope and appropriate use of the term probiotic', *Nature Reviews Gastroenterology and Hepatology*, 11(8), pp. 506–514. doi: 10.1038/nrgastro.2014.66.

Hill, D. A. and Artis, D. (2010) 'Intestinal Bacteria and the Regulation of Immune Cell Homeostasis', *Annual Review of Immunology*, 28(1), pp. 623–667. doi: 10.1146/annurev-immunol-030409-101330.

Huffnagle, G. B. and Noverr, M. C. (2013) 'The emerging world of the fungal microbiome', *Trends in Microbiology*. Elsevier Ltd, 21(7), pp. 334–341. doi: 10.1016/j.tim.2013.04.002.

Human Microbiome Project Consortium, T. (2012) 'Structure, function and diversity of the healthy human microbiome', *Nature*, 486. doi: 10.1038/nature11234.

Huseyin, C. E. *et al.* (2017) 'Forgotten fungi—the gut mycobiome in human health and disease', *FEMS Microbiology Reviews*. doi: 10.1093/femsre/fuw047.

Huson, D. *et al.* (2007) 'MEGAN analysis of metagenome data', *Genome Research*, 17, pp. 377–386. doi: 10.1101/gr.5969107.

Huson, D. H. *et al.* (2016) 'MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data', *PLoS Computational Biology*, 12(6), pp. 1–12. doi: 10.1371/journal.pcbi.1004957.

Huson, D. H. *et al.* (2018) 'MEGAN-LR: New algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs', *Biology Direct*, 13(1), pp. 1–17. doi: 10.1186/s13062-018-0208-7.

Jost, L. (2005) 'In press , Ecology , “ “ Concepts and Synthesis ” ” section components', *Ecology*, pp. 1–25. Available at:

<https://pdfs.semanticscholar.org/88c4/67e65b42eff088d19354632f41df6ca5804d.pdf> (Accessed: 10 June 2017).

Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, Yangqing Peng, Dongya Zhang, Zhuye Jie, Wenxian Wu, Youwen Qin, Wenbin Xue, Junhua Li, Lingchuan Han, Donghui Lu, Peixian W, S. Z. (2012) 'A metagenome-wide association study of gut microbiota in type 2 diabetes', *Nature*. Nature Publishing Group, 490(7418), pp. 55–60. doi: 10.1038/nature11450.

Khachatryan, Z. A. *et al.* (2008) 'Predominant role of host genetics in controlling the composition of gut microbiota', *PLoS ONE*, 3(8). doi: 10.1371/journal.pone.0003064.

Kindt, R. (2008) *TREE DIVERSITY ANALYSIS A manual and software for common statistical methods for ecological and biodiversity studies Using the BiodiversityR software within the R 2.6.1 environment*. Available at:

<http://www.worldagroforestry.org/downloads/Publications/PDFS/MN08242.pdf> (Accessed: 8 February 2019).

Kong, H. H. and Morris, A. (2017) 'The emerging importance and challenges of the human mycobiome', *Virulence*, 3 April, pp. 310–312. doi: 10.1080/21505594.2017.1279780.

Koren, S. and Phillippy, A. M. (2015) 'One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly', *Current Opinion in Microbiology*. Elsevier Ltd, 23, pp. 110–120. doi: 10.1016/j.mib.2014.11.014.

Krebs, C. J. (1999) 'Chapter 4. Estimating Density: Quadrats Counts', in *Ecological Methodology*, pp. 136–204. Available at:

http://www.zoology.ubc.ca/~krebs/downloads/krebs_chapter_13_2014.pdf (Accessed: 10 June 2017).

Lam, T.-W. *et al.* (2013) 'F1000Prime recommendation of: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.', *GigaScience*, pp. 1–6.

Lederberg, J. and Mccray, A. T. (no date) 'COMMENTARY 'Ome Sweet 'Omics-- A Genealogical Treasury of Words'. Available at: <https://lhncbc.nlm.nih.gov/files/archive/pub2001047.pdf> (Accessed: 7 June 2017).

Li, K. *et al.* (2012) 'Analyses of the microbial diversity across the human microbiome', *PLoS ONE*, 7(6), p. 32118. doi: 10.1371/journal.pone.0032118.

Lindgreen, S., Adair, K. L. and Gardner, P. P. (2016) 'An evaluation of the accuracy and speed of

metagenome analysis tools', *Scientific Reports*. Nature Publishing Group, 6(1), p. 19233. doi: 10.1038/srep19233.

Lúcia Bosco, V. *et al.* (2003) 'Yeasts from the oral cavity of children with AIDS: exoenzyme production and antifungal resistance Leveduras da cavidade oral de crianças com AIDS: produção de exoenzimas e resistência antifúngica', *Pesqui Odontol Bras*, 17(3), pp. 217–22. Available at: <http://www.scielo.br/pdf/pob/v17n3/a04v17n3.pdf> (Accessed: 7 June 2017).

Manichanh, C. *et al.* (2006) 'Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach', *Gut*, 55, pp. 205–211. doi: 10.1136/gut.2005.073817.

Martínez, I. *et al.* (2015) 'The Gut Microbiota of Rural Papua New Guineans: Composition, Diversity Patterns, and Ecological Processes', *Cell Reports*, 11(4), pp. 527–538. doi: 10.1016/j.celrep.2015.03.049.

McHugh, A. J. *et al.* (2018) 'Mesophilic Sporeformers Identified in Whey Powder by Using Shotgun Metagenomic Sequencing Downloaded from', *aem.asm.org 1 Applied and Environmental Microbiology on February*, 84, pp. 1305–1323. doi: 10.1128/AEM.01305-18.

Morgan, X. C., Segata, N. and Huttenhower, C. (2013) 'Biodiversity and functional genomics in the human microbiome', *Trends in Genetics*. Elsevier Ltd, 29(1), pp. 51–58. doi: 10.1016/j.tig.2012.09.005.

Morris, E. K. *et al.* (2014) 'Choosing and using diversity indices: Insights for ecological applications from the German Biodiversity Exploratories', *Ecology and Evolution*, 4(18), pp. 3514–3524. doi: 10.1002/ece3.1155.

Mueller, S. *et al.* (2006) 'Differences in Fecal Microbiota in Different European Study Populations in Relation to Age, Gender, and Country: a Cross-Sectional Study', 72(2), pp. 1027–1033. doi: 10.1128/AEM.72.2.1027.

Nagler, C. and Feehley, T. (2013) 'Chapter 6 – Mucosal Immunity', *Cancer Immunotherapy: Immune Suppression and Tumor Growth*, pp. 71–81. doi: 10.1016/B978-0-12-394296-8.00006-3.

Nash, A. K. *et al.* (2017) 'The gut mycobiome of the Human Microbiome Project healthy cohort', *Microbiome*. Microbiome, 5(1), p. 153. doi: 10.1186/s40168-017-0373-4.

NCBI (2017) *RefSeq non-redundant proteins*. Available at:

<https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/> (Accessed: 8 February 2019).

Obregon-Tito, A. J. *et al.* (2015) 'Subsistence strategies in traditional societies distinguish gut microbiomes', *Nature Communications*. Nature Publishing Group, 6, pp. 1–9. doi: 10.1038/ncomms7505.

Odds, F. C. (1984) *Ecology and Epidemiology of Candida Species*, *Zbl. Bakt. Hyg. A*. doi: 10.1016/S0174-3031(84)80073-6.

Oksanen, J. (2018) 'Vegan : ecological diversity'. doi: 10.1029/2006JF000545.

Ott, S. J. *et al.* (2008a) 'Fungi and inflammatory bowel diseases: Alterations of composition and diversity', *Scandinavian Journal of Gastroenterology*, 43(7), pp. 831–841. doi: 10.1080/00365520801935434.

Ott, S. J. *et al.* (2008b) 'Fungi and inflammatory bowel diseases: Alterations of composition and diversity', *Scandinavian Journal of Gastroenterology*, 43(7), pp. 831–841. doi: 10.1080/00365520801935434.

Ottman, N. *et al.* (2012) 'The function of our microbiota: who is out there and what do they do?', *Front Cell Infect Microbiol*, 2(August), p. 104. doi: 10.3389/fcimb.2012.00104.

Parekh, P. J., Balart, L. A. and Johnson, D. A. (2015) 'The Influence of the Gut Microbiome on Obesity, Metabolic Syndrome and Gastrointestinal Disease', *Clinical and Translational Gastroenterology*, 6. doi: 10.1038/ctg.2015.16.

Parks, D. H. *et al.* (2017) 'Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life'. doi: 10.1038/s41564-017-0012-7.

Peterson, J. *et al.* (2009) 'The NIH Human Microbiome Project', *Genome Research*, 19(12), pp. 2317–2323. doi: 10.1101/gr.096651.109.

Qin, J. *et al.* (2010) 'A human gut microbial gene catalog established by metagenomic sequencing', *Nature*, 464(7285), pp. 59–65. doi: 10.1038/nature08821.A.

Rodriguez-castaño, G. P., Caro-quintero, A. and Reyes, A. (2017) 'Advances in Gut Microbiome Research , Opening New Strategies to Cope with a Western Lifestyle', 7(January), pp. 1–15. doi: 10.3389/fgene.2016.00224.

- Rosa, B. *et al.* (2017) 'Genome Sequence of *Christensenella minuta* DSM 22607T', *ASM, Genome Announcements*, pp. 4–5.
- Rowland, I. *et al.* (2018) 'Gut microbiota functions: metabolism of nutrients and other food components', *European Journal of Nutrition*. Springer Berlin Heidelberg, 57(1), pp. 1–24. doi: 10.1007/s00394-017-1445-8.
- Sangwan, N., Xia, F. and Gilbert, J. A. (2016) 'Recovering complete and draft population genomes from metagenome datasets', *Microbiome*. *Microbiome*, 4(1), p. 8. doi: 10.1186/s40168-016-0154-5.
- Schnorr, S. L. *et al.* (2014) 'Gut microbiome of the Hadza hunter-gatherers', *Nature Communications*. Nature Publishing Group, 5, pp. 1–12. doi: 10.1038/ncomms4654.
- Sharpton, T. J. (2014) 'An introduction to the analysis of shotgun metagenomic data.', *Frontiers in Plant Science*, 5(June), p. 209. doi: 10.3389/fpls.2014.00209.
- Southern, P. *et al.* (2008) 'C. albicans colonization of human mucosal surfaces', *PLoS ONE*, 3(4). doi: 10.1371/journal.pone.0002067.
- Sun, J. and Kato, I. (2016) 'Gut microbiota, inflammation and colorectal cancer', *Genes & Diseases*, 3, pp. 130–143. doi: 10.1016/j.gendis.2016.03.004.
- Tang, J. *et al.* (2015) 'Mycobiome: Approaches to analysis of intestinal fungi', *Journal of Immunological Methods*. Elsevier B.V., 421(August), pp. 112–121. doi: 10.1016/j.jim.2015.04.004.
- Tasnim, N. *et al.* (2017) 'Linking the gut microbial ecosystem with the environment: Does gut health depend on where we live?', *Frontiers in Microbiology*, 8(OCT), pp. 1–8. doi: 10.3389/fmicb.2017.01935.
- Thomas, F. *et al.* (2011) 'Environmental and gut Bacteroidetes: The food connection', *Frontiers in Microbiology*, 2(MAY), pp. 1–16. doi: 10.3389/fmicb.2011.00093.
- Turnbaugh, P. J. *et al.* (2007) 'The human microbiome project: exploring the microbial part of ourselves in a changing world', *Nature*, 449(7164), pp. 804–810. doi: 10.1038/nature06244.
- Turnbaugh, P. J. *et al.* (2007) 'The Human Microbiome Project', *Nature*. Nature Publishing Group, pp. 804–810. doi: 10.1109/SAINT.2010.93.

- Tyakht, A. V *et al.* (2013) 'Human gut microbiota community structures in urban and rural populations in Russia', *Nature Communications*. Nature Publishing Group, 4, pp. 1–9. doi: 10.1038/ncomms3469.
- Upadhyaya, S. and Banerjee, G. (2015) 'Type 2 diabetes and gut microbiome: At the intersection of known and unknown', *Gut Microbes*, 6(2), pp. 85–92. doi: 10.1080/19490976.2015.1024918.
- Valle, G. *et al.* (2016) 'Impact Of Human Aging And Modern Lifestyle On Microbiota', *Critical Reviews in Food Science and Nutrition*, pp. 1–34.
- Verma, A., Gaffen, S. and Swidergall, M. (2017) 'Innate Immunity to Mucosal Candida Infections', *Journal of Fungi*, 3(4), p. 60. doi: 10.3390/jof3040060.
- Wilson, M. (2016) 'The human microbiota: An historical perspective', *The Human Microbiota and Chronic Disease: Dysbiosis as a Cause of Human Pathology*, pp. 1–36. doi: 10.1002/9781118982907.ch1.
- Yatsunenکو, T. *et al.* (2012) 'Human gut microbiome viewed across age and geography', *Nature*, 486(7402), pp. 222–227. doi: 10.1038/nature11053.
- Ye, J., McGinnis, S. and Madden, T. L. (2006) 'BLAST: Improvements for better sequence analysis', *Nucleic Acids Research*, 34(WEB. SERV. ISS.). doi: 10.1093/nar/gkl164.
- Zeeuwen, P. L. J. M. and Schalkwijk, J. (2016) 'The skin microbiota', *The Human Microbiota and Chronic Disease: Dysbiosis as a Cause of Human Pathology*, pp. 81–93. doi: 10.1002/9781118982907.ch5.
- Zhang, J. *et al.* (2015) 'Distinguishing the dominant species of pathogen in maxillary sinusitis by sequencing DNA dataset analysis', *Gene*. Elsevier B.V., 561(2), pp. 256–260. doi: 10.1016/j.gene.2015.02.037.
- Zhernakova, A. *et al.* (2016) 'Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity', *Science*, 352(6285), pp. 565–569. doi: 10.1126/science.aad3369.

ANEXOS

Tabla A1. Conjunto de datos del **MetaHIT-Eur**. Se muestran las características de 124 individuos con enfermedad inflamatoria intestinal (IBD) y sin ésta, además de sus índices de masa corporal (IMC) (Tomada y modificada de Qin *et al.* 2010).

Nombre	País	Género	Edad	IMC	IBD
MH0001	Dinamarca	Mujer	49	25.55	N
MH0002	Dinamarca	Mujer	59	27.28	N
MH0003	Dinamarca	Hombre	69	33.19	N
MH0004	Dinamarca	Hombre	59	31.18	N
MH0005	Dinamarca	Hombre	64	21.68	N
MH0006	Dinamarca	Mujer	59	22.38	N
MH0007	Dinamarca	Hombre	69	33.6	N
MH0008	Dinamarca	Hombre	59	24.35	N
MH0009	Dinamarca	Hombre	64	29.04	N
MH0010	Dinamarca	Hombre	64	33.27	N
MH0011	Dinamarca	Mujer	0	22.31	N
MH0012	Dinamarca	Mujer	42	32.1	N
MH0013	Dinamarca	Hombre	54	20.46	N
MH0014	Dinamarca	Mujer	54	38.49	N
MH0015	Dinamarca	Hombre	59	25.47	N
MH0016	Dinamarca	Mujer	49	30.5	N
MH0017	Dinamarca	Hombre	64	21.81	N
MH0018	Dinamarca	Hombre	49	31.37	N
MH0019	Dinamarca	Mujer	44	20.01	N
MH0020	Dinamarca	Mujer	63	33.23	N
MH0021	Dinamarca	Mujer	49	25.42	N
MH0022	Dinamarca	Hombre	64	24.42	N
MH0023	Dinamarca	Hombre	69	31.74	N
MH0024	Dinamarca	Mujer	59	22.72	N
MH0025	Dinamarca	Mujer	49	34.2	N
MH0026	Dinamarca	Mujer	49	37.32	N
MH0027	Dinamarca	Mujer	59	23.07	N
MH0028	Dinamarca	Mujer	44	22.7	N
MH0030	Dinamarca	Hombre	59	35.21	N
MH0031	Dinamarca	Hombre	69	22.34	N
MH0032	Dinamarca	Hombre	69	35.28	N
MH0033	Dinamarca	Mujer	59	31.95	N
MH0034	Dinamarca	Hombre	54	39.97	N
MH0035	Dinamarca	Hombre	49	22.66	N

MH0036	Dinamarca	Hombre	64	30.74	N
MH0037	Dinamarca	Hombre	44	24.02	N
MH0038	Dinamarca	Mujer	54	21.97	N
MH0039	Dinamarca	Hombre	58	23.07	N
MH0040	Dinamarca	Mujer	67	20.87	N
MH0041	Dinamarca	Hombre	59	23.17	N
MH0042	Dinamarca	Hombre	49	24.46	N
MH0043	Dinamarca	Hombre	69	23.72	N
MH0044	Dinamarca	Hombre	64	24.48	N
MH0045	Dinamarca	Hombre	59	25.11	N
MH0046	Dinamarca	Hombre	54	23.74	N
MH0047	Dinamarca	Mujer	69	30.4	N
MH0048	Dinamarca	Mujer	54	19.4	N
MH0049	Dinamarca	Mujer	44	35.52	N
MH0050	Dinamarca	Hombre	49	25.08	N
MH0051	Dinamarca	Mujer	69	23.15	N
MH0052	Dinamarca	Mujer	49	33.18	N
MH0053	Dinamarca	Mujer	49	32.7	N
MH0054	Dinamarca	Hombre	49	20.31	N
MH0055	Dinamarca	Hombre	59	30.29	N
MH0056	Dinamarca	Hombre	54	25.35	N
MH0057	Dinamarca	Mujer	54	32.98	N
MH0058	Dinamarca	Mujer	54	22.04	N
MH0059	Dinamarca	Hombre	59	33.27	N
MH0060	Dinamarca	Hombre	54	23.52	N
MH0061	Dinamarca	Mujer	69	30.12	N
MH0062	Dinamarca	Mujer	49	37.54	N
MH0063	Dinamarca	Hombre	59	30.23	N
MH0064	Dinamarca	Mujer	54	23.18	N
MH0065	Dinamarca	Hombre	59	28.23	N
MH0066	Dinamarca	Mujer	44	20.79	N
MH0067	Dinamarca	Hombre	54	21.07	N
MH0068	Dinamarca	Mujer	54	28.97	N
MH0069	Dinamarca	Mujer	59	36.71	N
MH0070	Dinamarca	Hombre	49	22.69	N
MH0071	Dinamarca	Mujer	44	25.37	N
MH0072	Dinamarca	Mujer	64	40.21	N
MH0073	Dinamarca	Hombre	54	32.49	N
MH0074	Dinamarca	Mujer	49	20.46	N
MH0075	Dinamarca	Hombre	64	30.55	N
MH0076	Dinamarca	Mujer	69	34.78	N

MH0077	Dinamarca	Mujer	49	24.92	N
MH0078	Dinamarca	Mujer	49	36.9	N
MH0079	Dinamarca	Mujer	64	19.97	N
MH0080	Dinamarca	Mujer	59	18.59	N
MH0081	Dinamarca	Mujer	49	37.95	N
MH0082	Dinamarca	Mujer	59	22.56	N
MH0083	Dinamarca	Mujer	54	30.59	N
MH0084	Dinamarca	Hombre	64	31.67	N
MH0085	Dinamarca	Mujer	59	36.46	N
MH0086	Dinamarca	Mujer	59	21.59	N
O2.UC-1	España	Hombre	37	31.02	Y
O2.UC-11	España	Mujer	34	18.68	Y
O2.UC-12	España	Hombre	43	21.6	Y
O2.UC-13	España	Mujer	68	23.38	Y
O2.UC-14	España	Hombre	31	32.65	Y
O2.UC-16	España	Hombre	47	26.42	Y
O2.UC-17	España	Hombre	56	21.87	Y
O2.UC-18	España	Hombre	48	25.72	Y
O2.UC-19	España	Hombre	42	24.15	Y
O2.UC-20	España	Mujer	51	24.03	Y
O2.UC-21	España	Mujer	49	30.46	Y
O2.UC-22	España	Hombre	44	25.39	Y
O2.UC-23	España	Mujer	44	28.16	Y
O2.UC-24	España	Mujer	55	28.76	Y
O2.UC-4	España	Mujer	57	28.53	Y
V1.CD-1	España	Mujer	25	17.93	Y
V1.CD-11	España	Mujer	62	35.46	N
V1.CD-12	España	Mujer	41	20.2	Y
V1.CD-13	España	Hombre	68	25.69	N
V1.CD-14	España	Mujer	41	23.12	N
V1.CD-15	España	Mujer	34	19	Y
V1.CD-2	España	Hombre	49	27.76	N
V1.CD-3	España	Mujer	18	21.51	N
V1.CD-4	España	Mujer	46	29.69	N
V1.CD-6	España	Mujer	36	18.52	Y
V1.CD-8	España	Hombre	51	29.38	N
V1.CD-9	España	Mujer	48	27.55	N
V1.UC-10	España	Hombre	45	27.31	Y
V1.UC-13	España	Mujer	51	28.51	Y
V1.UC-14	España	Mujer	53	20.25	Y
V1.UC-15	España	Mujer	25	22.77	Y

V1.UC-17	España	Mujer	41	24.46	Y
V1.UC-18	España	Mujer	63	28.67	N
V1.UC-19	España	Mujer	37	21.19	N
V1.UC-21	España	Hombre	62	25.21	Y
V1.UC-6	España	Mujer	38	23.18	N
V1.UC-7	España	Mujer	19	23.05	N
V1.UC-8	España	Hombre	22	25.4	N
V1.UC-9	España	Hombre	32	30.37	N

Tabla A2. Conjunto de datos del **MetaHIT-Ch**. Se muestran las características de 145 individuos con diabetes tipo 2 (T2D) y sin ésta, además de sus índices de masa corporal (IMC) (Tomada y modificada de Qin *et al.* 2012).

Muestra ID	País	Género	Edad	IMC	Diabético (Sí= Y No= N)
DLF001	China	Mujer	59	19.83	Y
DLF002	China	Mujer	43	17.58	Y
DLF003	China	Mujer	46	20.31	Y
DLF004	China	Mujer	25	19.78	Y
DLF005	China	Mujer	60	18.43	Y
DLF006	China	Mujer	71	16.00	Y
DLF007	China	Mujer	62	19.69	Y
DLF008	China	Mujer	54	19.40	Y
DLF009	China	Mujer	53	18.49	Y
DLF010	China	Mujer	51	19.29	Y
DLF012	China	Mujer	52	19.47	Y
DLF013	China	Mujer	49	19.63	Y
DLF014	China	Mujer	61	19.98	Y
DLM001	China	Hombre	46	20.24	Y
DLM002	China	Hombre	29	17.73	Y
DLM003	China	Hombre	41	20.80	Y
DLM004	China	Hombre	38	20.20	Y
DLM005	China	Hombre	35	19.72	Y
DLM006	China	Hombre	28	18.59	Y
DLM007	China	Hombre	43	20.05	Y
DLM008	China	Hombre	38	19.59	Y
DLM009	China	Hombre	30	20.05	Y
DLM010	China	Hombre	37	19.03	Y
DLM011	China	Hombre	49	20.05	Y
DLM012	China	Hombre	55	20.57	Y
DLM013	China	Hombre	51	18.37	Y
DLM014	China	Hombre	48	20.31	Y
DLM015	China	Hombre	56	21.11	Y
DLM016	China	Hombre	70	20.90	Y

DLM017	China	Hombre	52	18.83	Y
DLM018	China	Hombre	49	20.07	Y
DLM019	China	Hombre	47	20.08	Y
DLM020	China	Hombre	41	20.62	Y
DLM021	China	Hombre	58	18.21	Y
DLM022	China	Hombre	53	19.84	Y
DLM023	China	Hombre	45	20.81	Y
DLM024	China	Hombre	39	20.57	Y
DLM027	China	Hombre	70	21.60	Y
DLM028	China	Hombre	63	20.42	Y
DOF002	China	Mujer	56	25.30	Y
DOF003	China	Mujer	59	26.71	Y
DOF004	China	Mujer	63	25.39	Y
DOF006	China	Mujer	51	26.04	Y
DOF007	China	Mujer	70	25.78	Y
DOF008	China	Mujer	40	25.84	Y
DOF009	China	Mujer	68	27.64	Y
DOF010	China	Mujer	62	27.78	Y
DOF011	China	Mujer	48	27.06	Y
DOF012	China	Mujer	61	27.15	Y
DOF013	China	Mujer	53	25.97	Y
DOF014	China	Mujer	63	29.79	Y
DOM001	China	Hombre	59	25.26	Y
DOM003	China	Hombre	53	25.16	Y
DOM005	China	Hombre	49	25.31	Y
DOM008	China	Hombre	75	25.39	Y
DOM010	China	Hombre	74	25.75	Y
DOM012	China	Hombre	57	25.69	Y
DOM013	China	Hombre	36	26.82	Y
DOM014	China	Hombre	71	25.39	Y
DOM015	China	Hombre	68	25.35	Y
DOM016	China	Hombre	39	27.58	Y
DOM017	China	Hombre	41	27.72	Y
DOM018	China	Hombre	68	25.50	Y
DOM019	China	Hombre	53	25.31	Y
DOM020	China	Hombre	43	25.65	Y
DOM021	China	Hombre	72	25.71	Y
DOM022	China	Hombre	31	27.43	Y
DOM023	China	Hombre	39	25.51	Y
DOM024	China	Hombre	42	25.39	Y
DOM025	China	Hombre	62	25.39	Y
DOM026	China	Hombre	36	27.28	Y
NLF001	China	Mujer	31	17.94	N
NLF002	China	Mujer	34	17.97	N
NLF005	China	Mujer	44	19.33	N
NLF006	China	Mujer	27	17.80	N
NLF007	China	Mujer	39	20.57	N
NLF008	China	Mujer	26	18.03	N
NLF009	China	Mujer	27	18.99	N

NLF010	China	Mujer	35	20.17	N
NLF011	China	Mujer	25	18.52	N
NLF012	China	Mujer	27	19.53	N
NLF013	China	Mujer	32	18.20	N
NLF014	China	Mujer	37	18.67	N
NLF015	China	Mujer	59	19.15	N
NLM001	China	Hombre	32	19.14	N
NLM002	China	Hombre	23	18.72	N
NLM003	China	Hombre	23	19.49	N
NLM004	China	Hombre	24	19.35	N
NLM005	China	Hombre	42	19.13	N
NLM006	China	Hombre	24	19.03	N
NLM007	China	Hombre	22	19.92	N
NLM008	China	Hombre	22	19.57	N
NLM009	China	Hombre	24	18.37	N
NLM010	China	Hombre	23	19.96	N
NLM015	China	Hombre	23	18.07	N
NLM016	China	Hombre	23	19.47	N
NLM017	China	Hombre	23	19.16	N
NLM021	China	Hombre	25	18.59	N
NLM022	China	Hombre	29	19.84	N
NLM023	China	Hombre	38	19.38	N
NLM024	China	Hombre	23	19.38	N
NLM025	China	Hombre	19	17.29	N
NLM026	China	Hombre	24	20.66	N
NLM027	China	Hombre	22	19.36	N
NLM028	China	Hombre	21	19.05	N
NLM029	China	Hombre	25	19.44	N
NLM031	China	Hombre	33	20.31	N
NLM032	China	Hombre	39	19.49	N
NOF001	China	Mujer	45	25.51	N
NOF002	China	Mujer	51	26.56	N
NOF004	China	Mujer	33	25.07	N
NOF005	China	Mujer	32	25.39	N
NOF006	China	Mujer	35	26.13	N
NOF007	China	Mujer	32	24.89	N
NOF008	China	Mujer	30	26.71	N
NOF009	China	Mujer	26	25.24	N
NOF010	China	Mujer	40	24.89	N
NOF011	China	Mujer	30	25.39	N
NOF012	China	Mujer	38	25.81	N
NOF013	China	Mujer	41	25.48	N
NOF014	China	Mujer	42	25.59	N
NOM001	China	Hombre	48	26.45	N
NOM002	China	Hombre	68	28.30	N
NOM004	China	Hombre	32	29.06	N
NOM005	China	Hombre	46	25.40	N
NOM007	China	Hombre	72	23.66	N
NOM008	China	Hombre	54	26.57	N

NOM009	China	Hombre	33	25.18	N
NOM010	China	Hombre	30	25.71	N
NOM012	China	Hombre	32	25.71	N
NOM013	China	Hombre	42	26.22	N
NOM014	China	Hombre	41	25.10	N
NOM015	China	Hombre	32	25.73	N
NOM016	China	Hombre	37	24.77	N
NOM017	China	Hombre	39	27.24	N
NOM018	China	Hombre	43	25.00	N
NOM019	China	Hombre	26	25.47	N
NOM020	China	Hombre	24	25.21	N
NOM022	China	Hombre	23	26.73	N
NOM023	China	Hombre	46	27.12	N
NOM025	China	Hombre	28	25.59	N
NOM026	China	Hombre	52	25.71	N
NOM027	China	Hombre	38	25.71	N
NOM028	China	Hombre	47	26.57	N
NOM029	China	Hombre	14	32.58	N