

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA



VNIVERSITAT
DE VALÈNCIA

TRABAJO DE FIN DE MÁSTER

EVOLUCIÓN DIARIA DEL MICROBIOMA HUMANO

AUTORA:
M^a TERESA RUBIO
MARTÍNEZ-ABARCA

TUTORES:
CARLOS PEÑA GARAY
VICENTE ARNAU LLOMBART

MAYO, 2017



MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

TRABAJO DE FIN DE MÁSTER

EVOLUCIÓN DIARIA DEL MICROBIOMA HUMANO

AUTORA:
M^a TERESA RUBIO
MARTÍNEZ-ABARCA

Tutores:
CARLOS PEÑA GARAY
VICENTE ARNAU LLOMBART

TRIBUNAL:

PRESIDENTE/A:

VOCAL 1:

VOCAL 2:

FECHA DE DEFENSA:

CALIFICACIÓN:

“Me lo contaron y lo olvidé; lo vi y lo entendí; lo hice y lo aprendí.” – Confucio
“Lo programé en Python y lo disfruté.” – Teresa

Índice general

| | |
|---|------------|
| Índice de figuras | V |
| Índice de tablas | VII |
| Agradecimientos | IX |
| Resumen | XI |
| 1. Introducción | 1 |
| 1.1. Microbiota y metagenómica | 1 |
| 1.2. Pasos en un proyecto metagenómico | 3 |
| 1.3. Series temporales | 5 |
| 1.4. Interacciones | 6 |
| 1.5. Motivación y objetivos del trabajo | 6 |
| 2. Materiales y métodos | 9 |
| 2.1. Preprocesado: FastQC, MultiQC y seq_crumps | 9 |
| 2.2. Clasificación taxonómica: qiime | 10 |
| 2.2.1. Selección de OTUs | 10 |
| Greengenes | 12 |
| Uclust | 13 |
| 2.2.2. Asignación taxonómica | 13 |
| 2.3. Análisis de variabilidad: complexCruncher | 14 |
| 2.3.1. Regresión lineal y exponencial | 14 |
| 2.3.2. Ley de potencia x -ponderada | 14 |
| 2.3.3. Estandarización | 15 |
| 2.3.4. RSI y medidas de variabilidad | 16 |

| | |
|---|-----------|
| 2.4. Estudio de interacciones | 16 |
| 2.4.1. Coeficiente de correlación de Pearson | 16 |
| 2.4.2. Método de búsqueda de comportamientos | 17 |
| 2.4.3. Aproximación para obtener interacciones: LIMITS | 17 |
| 3. Resultados | 21 |
| 3.1. Estado de los datos | 21 |
| 3.2. Preprocesado | 24 |
| 3.3. Clasificación taxonómica | 27 |
| 3.4. Explorando series temporales | 27 |
| 3.4.1. Abundancia de taxones | 28 |
| 3.4.2. Ley de potencias | 30 |
| 3.4.3. Clasificación por rango | 34 |
| 3.5. Correlaciones | 40 |
| 4. Discusión y conclusiones | 45 |
| 4.1. Perspectivas de futuro | 46 |
| Bibliografía | 52 |
| Anexo I: Pipeline de asignación taxonómica | 53 |

Índice de figuras

| | |
|---|----|
| 1.1. Representación esquemática de ARNr 16S. | 2 |
| 1.2. Protocolo de secuenciación Illumina con <i>barcode</i> . | 4 |
| 1.3. <i>Paired end vs. Mate pair</i> | 5 |
| 2.1. Flujo de trabajo de qiime. | 11 |
| 2.2. Procedimiento de LIMITS | 19 |
| 3.1. Control de calidad de los datos crudos | 23 |
| 3.2. Control de calidad tras el filtro de calidad | 25 |
| 3.3. Abundancia absoluta saliva A | 28 |
| 3.4. Abundancia absoluta intestino A | 29 |
| 3.5. Abundancia absoluta intestino B | 29 |
| 3.6. Ajuste a la ley de potencias x-ponderada. | 32 |
| 3.7. V y β en saliva | 33 |
| 3.8. V y β en intestino | 33 |
| 3.9. V y β resumen de saliva e intestino | 34 |
| 3.10. Matriz de rango: saliva A | 36 |
| 3.11. Matriz de rango: intestino A | 38 |
| 3.12. Matriz de rango: intestino B | 39 |
| 3.13. Correlaciones saliva A | 41 |
| 3.14. Abundancia relativa del grupo 2 en saliva | 42 |
| 3.15. Correlaciones intestino A | 43 |
| 3.16. Correlaciones intestino B | 44 |
| 4.1. Correlación Pearson vs. LIMITS en saliva A | 48 |

Índice de tablas

| | |
|---|----|
| 2.1. Tabla de grupos de comportamiento. | 18 |
| 3.1. Tabla de ficheros eliminados. | 26 |
| 3.2. Resumen de subperiodos temporales. | 31 |

Agradecimientos

La realización de este trabajo fin de máster ha sido posible gracias a la dedicación como tutores de Carlos Peña Garay y Vicente Arnau Llombart. Gracias a ambos. Gracias también a todos los profesores del máster en Bioinformática de la Universidad de Valencia por la formación recibida y a todas aquellas personas que siempre me han apoyado. Especial mención a la ayuda recibida de Jose Manuel Martí y Daniel Martínez por todo lo que he aprendido con vosotros.

También quiero agradecer al grupo la oportunidad de haber podido colaborar en el proceso de elaboración de un artículo científico. Es para mi una gran satisfacción tener una publicación como segunda autora siendo aún estudiante de máster. Muchas gracias a todos, he crecido como investigadora y como persona con vuestra ayuda.

Resumen

En este proyecto se analizó el microbioma de intestino y saliva de dos individuos durante todo un año. Se trabajó con datos de secuenciación de la región V4 del gen que codifica el ARN ribosomal 16S de los microorganismos presentes. Además, se recopilaron datos del estilo de vida de los donantes tales como dieta, ejercicio o enfermedad. Los datos procedían del estudio David *et al.* [1] y ocuparon un volumen total de 64,8 GiB aproximadamente (21,2 Gpb analizadas). En cuanto a los sujetos, el primero de ellos realizó un viaje al extranjero durante el estudio y el segundo tuvo una infección de *Salmonella*, lo que nos permitió ver cómo varía la dinámica de la microbiota en estas situaciones.

En primer lugar, se comprobó la calidad y longitud de la secuenciación para realizar un filtro previo. A continuación, se agruparon las *reads* en OTUs al 97% de similitud de secuencia y se asignó la taxonomía a nivel de género. Con estos datos se obtuvieron unas tablas que incluyen la abundancia absoluta de cada taxón, en cada día y en cada individuo.

En segundo lugar, se hizo un estudio de variabilidad temporal de los microorganismos presentes a partir de su abundancia. Se buscaba si estos datos se ajustan a algún modelo y se encontró que siguen la ley de Taylor, lo cual permitió determinar la microbiota de cada individuo con tan solo dos variables. Además, se comprobó la variabilidad haciendo un análisis ordenando por abundancia total a lo largo de los días y haciendo también un ranking de los microorganismos. Estos resultados fueron incorporados al artículo recientemente publicado de Martí *et al.* [2] como la serie temporal más larga analizada por los autores.

Por último, se calcularon las correlaciones de abundancia entre géneros. Se realizó una clasificación previa en grupos de comportamiento en base a una perturbación y se comprobó la correlación entre y dentro de grupos. El uso de correlaciones no supone una medida real de las interacciones y, aunque existen modelos que pueden explicarlas mejor, aún queda un largo camino que recorrer para configurar todos los acontecimientos que se producen en nuestra microbiota.

1 Introducción

1.1. Microbiota y metagenómica

A lo largo de la evolución, los microorganismos han convivido con el ser humano y son fundamentales para su salud. Algunos lo han hecho en simbiosis, siendo esenciales para mantener un estado saludable, pero otros son patógenos y pueden provocar diferentes enfermedades. De forma general, se denomina microbiota al conjunto de microorganismos que se encuentran frecuentemente en distintos sitios del cuerpo de individuos sanos. Hasta hace poco se creía que el ser humano poseía un mayor número de microorganismos que de células propias, sin embargo, este paradigma está cambiando y actualmente se estima que el número de bacterias en nuestro cuerpo es aproximadamente del mismo orden que el número de células humanas [3]. Adquirimos los microorganismos a partir del nacimiento, durante el parto y la lactancia, salvo casos especiales como algunos patógenos capaces de traspasar la barrera de la placenta. A lo largo de nuestra vida van colonizando nuestra piel, mucosas y, sobre todo, el tubo intestinal con clara preferencia por el intestino grueso. Los obtenemos de los alimentos, el agua y el contacto con otras personas. La importancia de la microbiota humana radica en sus numerosas funciones: regulación de procesos digestivos, producción de sustancias bacteriostáticas, producción de antibióticos naturales contra patógenos e, incluso, entrenamiento del sistema inmune para reconocer invasores dañinos. Numerosos estudios demuestran que el cambio en la composición de la microbiota está relacionado con estados de enfermedad, planteando la posibilidad de manipular estas comunidades como posible tratamiento. Los conocimientos en este campo han sido escasos hasta el año 2008 cuando se inició el Proyecto Microbioma Humano. Su misión es caracterizar el microbioma humano, entendiendo como tal el conjunto de genes de nuestra microbiota, y el análisis de su papel en la salud y la enfermedad humana.

La forma tradicional de estudiar microorganismos en cualquier ambiente ha sido mediante técnicas de cultivo. Se obtiene la muestra del medio natural (suelo, agua o heces), se cultiva en un medio previamente definido con condiciones óptimas y cuando los microorganismos crecen formando colonias, se extrae su material genético para analizarlo. A día de hoy, esta técnica se sigue llevando a cabo cuando se pretende estudiar el genoma de un organismo concreto ya que es muy sencilla y barata. Sin embargo, no se conocen las condiciones óptimas de crecimiento para todos los microorganismos presentes en las muestras ambientales. A estos microorganismos se les denomina comúnmente no cultivables, aunque sí son cultivables pero se desconocen los requisitos específicos para su cultivo y en algunas ocasiones requieren de la existencia de otros microorganismos para sobrevivir.

1.1. MICROBIOTA Y METAGENÓMICA

La metagenómica surge como alternativa a las técnicas de cultivo para explicar cuáles son los microorganismos presentes en cualquier muestra y estudiar todo el ADN genómico presente en dicha muestra. En este caso, se obtiene la muestra natural, se aislan los microorganismos presentes, se extrae su material genético y se secuencia todo el genoma o la región de interés. Ejemplos históricos en metagenómica son los estudios de Tyson *et al.*, 2004 [4] y Venter *et al.*, 2004 [5]. El primero se lleva a cabo en ambientes extremos donde se encuentra un grupo relativamente pequeño de microorganismos y el segundo, sin embargo, se realiza en el océano donde hay una enorme variedad de especies. Ambos ponen de manifiesto el esfuerzo que requiere la secuenciación de estas muestras y su posterior análisis. Gracias a la exponencial evolución hacia la secuenciación de siguiente generación, que ha permitido reducir costos y realizar proyectos más robustos, se pueden realizar estos estudios a gran escala. Y gracias al desarrollo de la bioinformática se han obtenido herramientas potentes y sofisticadas para poder analizar esa inmensa cantidad de datos generada.

Cuando se pretende caracterizar la estructura taxonómica de una comunidad microbiana se puede utilizar la secuenciación aleatoria, también conocida como *shotgun*, o bien un gen marcador como el que codifica el ARNr ribosómico (ARNr) 16S. La primera aproximación trata de romper un genoma en fragmentos al azar, secuenciar cada uno de ellos y luego organizar estas partes superpuestas para guiar el ensamblaje; se puede realizar por clonación utilizando vectores, o por secuenciación directa. Los datos en los que se centra este estudio se obtienen por la segunda aproximación ya que el 16S es considerado como el cronómetro molecular en procariotas. Es un polirribonucleótido de unas 1542 pares de bases, codificado por el gen *rss* y se encuentra en la subunidad pequeña 30S del ribosoma procariota, orgánulo encargado de la síntesis celular de proteínas. Se utiliza como marcador porque se encuentra presente en todos los microorganismos y contiene regiones conservadas de forma universal, mientras que otras regiones son variables (Figura 1.1) y esto es lo que hace posible la identificación a un nivel taxonómico suficientemente informativo. Para conseguir una identificación lo suficientemente específica, hay que tener en cuenta la cobertura de secuenciación (para detectar microorganismos que se encuentren en una menor concentración), la longitud de las *reads* (secuencias más largas permiten una asignación taxonómica más precisa) y la tasa de error en la secuenciación (puede generar una asignación incorrecta al enmascarar las regiones variables).

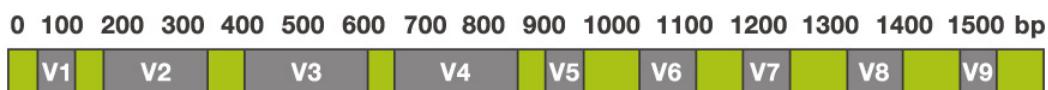


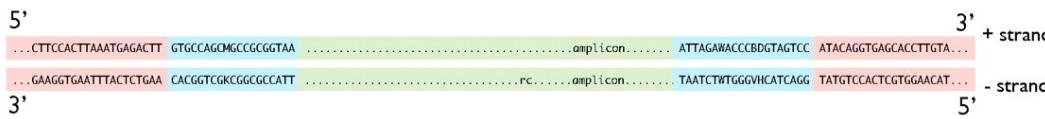
Figura 1.1: Representación esquemática de ARNr 16S en su estructura primaria. En verde se colorean las regiones conservadas (C), que son universales, y en gris las regiones variables (V), que permiten agrupar específicamente a nivel taxonómico.

1.2. Pasos en un proyecto metagenómico

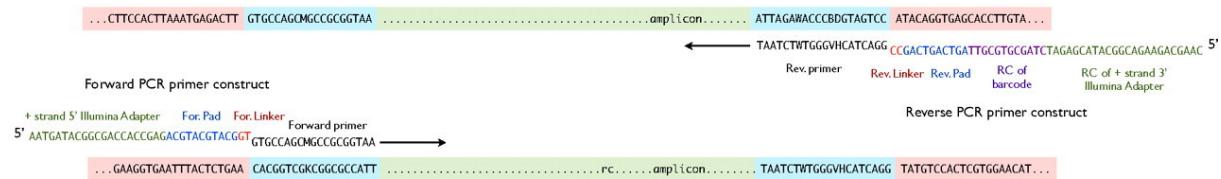
1. Extracción del ADN. Se basa en la lisis de las células mediante compuestos químicos (EDTA, lisozima, detergentes...), seguida de la eliminación de los componentes celulares y finalmente se precipita el material genético para obtener ADN puro en alta concentración. Se utilizan protocolos genéricos o kits comerciales aunque para muestras complicadas, se llega a prescindir de ellos elaborando protocolos propios.
2. Creación una genoteca de 16S para la secuenciación:
 - Amplificación por reacción en cadena de la polimerasa (*polymerase chain reaction*[PCR]) de la región a ser analizada. Los métodos de secuenciación no son aplicables a moléculas individuales, por lo que es necesario sintetizar múltiples copias para obtener una lectura. En cada ciclo de PCR se llevan a cabo 3 etapas: desnaturalización (donde las cadenas de ADN se separan calentándolas), hibridación (se utilizan cebadores o *primers* que hibridan con su secuencia complementaria por el extremo 3') y extensión (etapa donde actúa la Taq polimerasa sobre el *primer* y agrega bases complementarias para crear cadenas completas de ADN). El tipo de PCR a utilizar puede ser PCR en emulsión o PCR puente.
 - *Barcodeing* del ADN para multiplexar el ensayo. Este método permite secuenciar varias muestras en el mismo *run* de secuenciación, mejorando así el cost-eficacia y el tiempo de obtención de los resultados. En el proceso de PCR, además de los cebadores, se añaden por ligación otros trozos de secuencia que no son complementarios a la región diana y por tanto quedan “colgando”. Esos trozos incluyen el adaptador para la plataforma de secuenciación (Illumina en este caso), un *linker* y un código de barras (*barcode*) específico para cada secuencia. El producto de amplificación final es el que se muestra en la fig. 1.2.
3. Secuenciación. El desarrollo de la secuenciación de nueva generación (NGS – del inglés *next generation sequencing*) permite obtener millones de fragmentos de ADN de forma paralela, logrando que el número de bases a secuenciar por unidad de precio haya crecido exponencialmente en los últimos años. Existen distintas plataformas de segunda generación que se pueden clasificar según su método de secuenciar:
 - Síntesis: se basa en el proceso de síntesis de ADN usando la enzima ADN polimerasa para identificar las bases presentes en la molécula complementaria de ADN. A su vez se agrupa en otros tres tipos atendiendo al sistema de detección: pirosecuenciación basada en la detección quimioluminiscente de pirofosfato liberado (Ej.: Roche/454), fluorescencia cuando los nucleótidos están marcados fluorescentemente (Ej.: Illumina/MiSeq, HiSeq) y secuenciación no óptica que mide la liberación de protones (Ej.: ThermoFisher Scientific/Ion Torrent).

1.2. PASOS EN UN PROYECTO METAGENÓMICO

Target gene:



Amplification primers with annealing sites:



Amplification products:



Figura 1.2: Protocolo de secuenciación Illumina con *barcode*. 1^a parte: el gen diana es la región V4 del gen que codifica el ARNr 16S (se colorean de azul las regiones conservadas y en verde la región adecuada para la clasificación taxonómica). 2^a parte: se procede a la amplificación por PCR utilizando cebadores de PCR (negro) homólogos a la secuencia diana a los que se les añadió un *linker* (rojo y azul), un *barcode* (violeta), y el adaptador de secuenciación Illumina (verde). Lo único que queda unido por hibridación al molde es la parte negra y el resto queda “colgando”. 3^a parte: producto de amplificación.

- Ligación: se emplea una ADN ligasa en lugar de una polimerasa para identificar la secuencia objetivo (Ej.: SOLiD).
- Hibridación: un método no enzimático que utiliza una única muestra de ADN marcada fluorescentemente y se hibrida con una colección de secuencias conocidas (chip de ADN). Si la muestra hibrida fuertemente en un punto dado, se deduce que esa es su secuencia.

En general, los nuevos secuenciadores generan lecturas a partir de los dos extremos de un fragmento de ADN, dando lugar a lecturas apareadas con una distancia conocida entre ellas. Para ello utilizan dos estrategias diferentes (Figura 1.3): *paired end* que proporcionan rangos de tamaños de inserto más estrechos y *mate pair* que cubren tamaños mayores. Las lecturas de tipo *paired end* se generan mediante la fragmentación del ADN en pequeños segmentos de los cuales se secuencia el final de ambos extremos. Por contra, los *mate pairs* se crean a partir de fragmentos de ADN de tamaño conocido, que se circularizan y se ligan usando un adaptador interno. Estos fragmentos circularizados se trocean al azar para luego purificar los segmentos que contienen el adaptador a partir del cual se secuenciará.

Actualmente, se encuentran en desarrollo las tecnologías de tercera generación que tienen como objetivo la secuenciación de moléculas individuales para eliminar el paso de amplificación de ADN, que suele introducir errores y/o sesgos. Además, se ha conseguido alargar la longitud de secuencia obtenida, lo que facilita el ensamblaje.

1.3. SERIES TEMPORALES

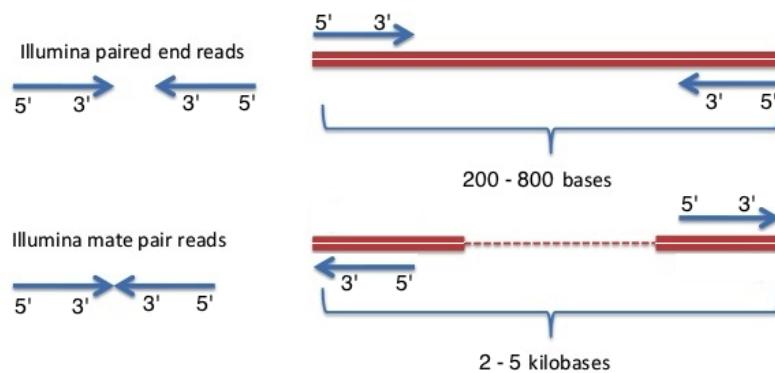


Figura 1.3: *Paired end vs. Mate pair.* El rectángulo rojo simboliza un fragmento de ADN de doble cadena y las flechas azules indican las lecturas que se obtienen tras la secuenciación. En la parte superior se esquematiza la estrategia *paired end*, que permite un rango de inserto de 200-800 pb, mientras la parte inferior representa la estrategia *mate pair*, que llega a 2-5 kb de inserto (o incluso superior).

Se predice que la plataforma MinION (basada en nanoporos) supondrá una revolución porque produce secuencias mucho más largas (miles de pares de bases) y, además, reduce el coste de reactivos y aparatos al tratarse de una técnica no óptica. Sin embargo, todavía está lejos de las tasas de error que obtienen otras tecnologías y aún no se puede emplear a gran escala. En julio de 2016 esta tecnología fue utilizada por la NASA para identificar un hongo que está creciendo en la Estación Espacial Internacional (ISS) e investigar la posible existencia de vida fuera del planeta Tierra.

4. Análisis bioinformático. Existen varias herramientas pero las más utilizadas por su fácil manejo y su amplia documentación son qiime y mothur. Ambos *pipelines* incorporan algoritmos para llevar a cabo el control de calidad, agrupación de secuencias, asignación de taxonomía, cálculo de diversidad y visualización de resultados.

1.3. Series temporales

Gracias a estas tecnologías se puede conocer la composición microbiana de cualquier ambiente en un determinado momento. Pero esto solo ofrece una fotografía estática del sistema, se puede ir un poco más allá analizando series temporales para observar la dinámica de estas poblaciones. Existe en ecología un modelo universal que describe las distribuciones espaciales resultantes de muestreos poblacionales. Se conoce como Ley de potencia de Taylor y refleja el grado de correlación entre individuos de una población. Es una ley empírica propuesta por Taylor en 1961 [6] basada en la relación entre la media (x) y la varianza (σ^2) en la abundancia de poblaciones naturales: $\sigma^2 = V \cdot x^\beta$. Donde la constante V indica amplitud de fluctuación y β el tipo de distribución. En 1990, McArdle *et al.* [7] propusieron utilizar esta ley con otra medida de la variabilidad poblacional, reemplazando la varianza por el coeficiente de variabilidad. Esto permite detectar patrones de variabilidad en una población e incluso entre distintas poblaciones. Por ejemplo, comparando la variabilidad de una población con su densidad media se pueden detectar puntos de

1.4. INTERACCIONES

crecimiento poblacional súbito o, comparar poblaciones en base a su ecología sirve para entender por qué diferentes especies pueden tener dinámicas poblacionales similares.

1.4. Interacciones

Las relaciones biológicas son muy variadas y pueden generar interdependencias de muy diversa importancia. Es muy complejo definir detalladamente las interacciones entre poblaciones pero, en general, son positivas si permiten ocupar nuevos nichos, o negativas cuando se eliminan las poblaciones poco adaptadas o se protegen de la llegada de especies intrusas. En el mundo biológico se han descrito distintas relaciones entre individuos:

- Neutralismo: dos individuos se encuentran simultáneamente en el ambiente sin que exista relación entre ellos. Resultado 0/0.
- Comensalismo: el primero modifica el ambiente y favorece el crecimiento del segundo. Resultado +/0.
- Sinergismo o protocolooperación: los individuos se favorecen mutuamente de forma no obligatoria. Resultado +/++.
- Mutualismo o simbiosis: los individuos se favorecen mutuamente de forma obligatoria y adquieren nuevas propiedades. Resultado +/+.
- Competencia: cuando los recursos del ecosistema en que se desarrollan son insuficientes para suplir las necesidades de todos los seres. Resultado -/-.
- Amensalismo: un organismo se ve perjudicado en la relación y el otro no experimenta ninguna alteración. Resultado -/0.
- Parasitismo: el parásito depende del hospedador y obtiene algún beneficio. Resultado -/+.
- Depredación: el depredador caza a una presa para subsistir. Resultado +/-.

1.5. Motivación y objetivos del trabajo

En este trabajo se analiza una serie temporal de 820 puntos temporales pertenecientes al microbioma intestinal y bucal de dos sujetos, pertenecientes al estudio de David *et al.* [1] y depositados en un repositorio público. El interés en estos datos se centra en que el grupo de investigación donde se ha realizado el proyecto, lleva trabajando desde 2012 con series temporales de la microbiota. Motivados por los datos de una tesis doctoral sobre individuos con colon irritable y sanos a lo largo del tiempo, e influenciados por el trabajo “Dinámica de procesos de clasificación en sistemas complejos” de Blumm *et al.* [8], surge la idea de aplicar dinámica de clasificación a la microbiota. Tras muchas horas de trabajo, se creó la herramienta complexCruncher que sirve para el análisis de este tipo de datos.

1.5. MOTIVACIÓN Y OBJETIVOS DEL TRABAJO

Una vez automatizado el proceso, ya se puede aplicar sobre distintas series temporales y esto constituye la creación del artículo Martí *et al.* [2] donde se condensan todas ellas. Las series temporales de ese estudio se obtienen de diversos tipos de microbioma humano: niños con la enfermedad kwashiorkor, dietas vegetarianas frente a no vegetarianas, sujetos con obesidad, ingesta de antibióticos y síndrome de colon irritable. Todos estos datos son procedentes de repositorios públicos o cedidos por los autores. Sin embargo, se trata de series temporales no muy largas, por lo que el grupo se estaba muy interesado en analizar los datos anuales presentes en este estudio. La principal causa por la que surge este trabajo es precisamente agregar esta serie al estudio global.

Al inicio del trabajo se fijaron unos objetivos muy claros, siguiendo unas pautas previamente establecidas debido a que el grupo de investigación ya ha realizado análisis similares anteriormente. Esos objetivos son dos esencialmente:

- Análisis de secuencias de múltiples genomas en la misma muestra (metagenómica) y su correspondiente clasificación taxonómica.
- Análisis de series temporales: caracterizar el comportamiento global del sistema, de las correlaciones entre OTUs (a nivel de género o familia) e identificar modos, si existen, vía descomposición de Fourier.

La motivación personal de realizar este trabajo es aprender las herramientas e indagar en el campo de la metagenómica, que no está cubierto por el máster. Resulta muy útil e interesante elaborar un *pipeline* de principio a fin que parte de secuencias crudas y llega a proporcionar parámetros capaces de indicar el estado de disbiosis de la microbiota así como aportar algo de luz sobre las relaciones entre los taxones encontrados. Un estudio de esta envergadura ofrece la oportunidad de abarcar un amplio conjunto de destrezas:

- Integrar diferentes códigos bioinformáticos y de computación intensiva en un único *pipeline*.
- Emplear múltiples formatos de archivos bioinformáticos.
- Profundizar en Python en su vertiente de *scripting* como potente nexo de unión entre las distintas etapas del *pipeline*.
- Manejar paquetes de Python que ofrecen procesamiento de datos muy extendido, como es el caso de pandas.
- Disponer de servidores de *high performance computing* basados en GNU Linux con la distribución Scientific Linux con un sistema de colas en un *cluster* de computación científica.
- Utilizar un sistema de control de versiones distribuido como es git y su conocida aplicación, github.
- Aprender L^AT_EX, el sistema preferido por científicos e ingenieros para edición profesional de documentos técnicos.

2 Materiales y métodos

La serie temporal anual en la que se centra este estudio, se divide en 3 bloques de muestras: saliva del sujeto A, intestino del sujeto A e intestino del sujeto B. Durante el periodo de estudio el donante A realiza un viaje al extranjero y el donante B sufre una salmonelosis. Las muestras son tomadas por los propios sujetos y posteriormente se secuencian en laboratorio mediante la plataforma Illumina GAIIX. Para el análisis completo, se llevan a cabo una serie de pasos en el siguiente orden: preprocesado de los datos, clasificación taxonómica, análisis de la variabilidad y estudio de interacciones. En el presente capítulo, se exponen los distintos materiales y métodos utilizados a lo largo del trabajo y la justificación de su elección.

2.1. Preprocesado: FastQC, MultiQC y seq_crumbs

FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) es una herramienta de control de calidad para datos de secuenciación, de código abierto e implementada en Java. Acepta un fichero de entrada en distintos formatos (fastq, SAM o BAM) y produce un fichero de salida en formato HTML con gráficos y tablas que permiten evaluar los datos. Proporciona mucha información sobre una única muestra: estadísticas numéricas (codificación de calidad según la plataforma utilizada, número total de secuencias...), *score* de calidad, contenido en GC, distribución de longitud de secuencias, etc. Esta herramienta permite detectar rápidamente cualquier problema que exista antes de realizar análisis posteriores.

MultiQC (<http://multiqc.info/>) es una herramienta de código abierto, implementada en Python y que da soporte a muchas herramientas bioinformáticas, entre ellas FastQC. Produce un reporte HTML muy parecido al anterior pero recoge la visualización de múltiples muestras en conjunto, permitiendo realizar comparaciones. También recopila estadísticas numéricas de cada muestra para ver cómo se comportan los datos.

seq_crumbs (https://bioinf.comav.upv.es/seq_crumbs/) es un software de código abierto, implementado en Python, que incluye utilidades para procesar secuencias. Toma un fichero de secuencias como entrada y crea un nuevo fichero de salida con las secuencias procesadas. Dentro de sus muchas funciones, caben destacar: filtrado de secuencias por calidad media, filtrado por longitud según un umbral máximo y mínimo, eliminación de regiones con baja calidad en los extremos (*trimming*), etc.

2.2. CLASIFICACIÓN TAXONÓMICA: QIIME

Se utilizan FastQC y MultiQC por la gran cantidad de información que producen, siendo especialmente útil en comprobar el comportamiento de la calidad de secuenciación. Por otro lado, se elige seq_scrubs porque incluye un *script* específico para filtrar por calidad media bastante rápido y fácil de usar. Además, todas ellas se han seleccionado por ser de código abierto y estar disponibles para la comunidad tanto para su uso como para su adaptación, si fuera necesario.

2.2. Clasificación taxonómica: qiime

Qiime (<http://qiime.org/index.html>) [9] son la siglas en inglés de *Quantitative Insights Into Microbial Ecology*. Es un *pipeline* bioinformático de código abierto para realizar análisis de microbiomas a partir de datos de secuenciación. Fue construido utilizando el lenguaje de programación Python con una implementación modular en forma de *scripts* para poder usar cualquier punto dentro de su flujo de trabajo (figura 2.1) de manera independiente.

Qiime acepta ficheros de entrada en formato fastq, fasta+qual o sff. Incorpora su propio método de preprocesado que realiza el filtrado de *reads* por calidad, longitud y el demultiplexado simultáneamente a partir de un fichero “mapa” con los metadatos (aunque no se usa en este trabajo). En este proyecto se utiliza qiime para la selección de OTUs y para la asignación de taxonomía, que son los dos siguientes pasos que incorpora el flujo de trabajo. Incluye, además, tres pasos adicionales con sus respectivas visualizaciones que tampoco se utilizan aquí: creación de árboles filogenéticos, estudio de diversidad α y β y método de rarefacción.

Existen múltiples herramientas para realizar una clasificación taxonómica pero se prefiere qiime para poder reproducir y corroborar los resultados obtenidos por los autores del trabajo donde se obtienen los datos [1]. Se generan diferencias taxonómicas a la hora de elegir una herramienta u otra, pero no son muy significativas y tanto qiime como mothur son métodos robustos.

2.2.1. Selección de OTUs

OTU (del inglés *Operational Taxonomic Unit*) es una unidad taxonómica operativa, es decir, una unidad de clasificación elegida por el investigador para individualizar los objetos de su estudio sin juzgar si se corresponden a una entidad biológica particular. Se aplica cuando se tienen datos de secuencias de ADN o datos morfológicos. Puede considerarse OTU un individuo, una población, una especie o cualquier otro taxón. Qiime ofrece tres estrategias de selección diferentes para este paso:

2.2. CLASIFICACIÓN TAXONÓMICA: QIIME

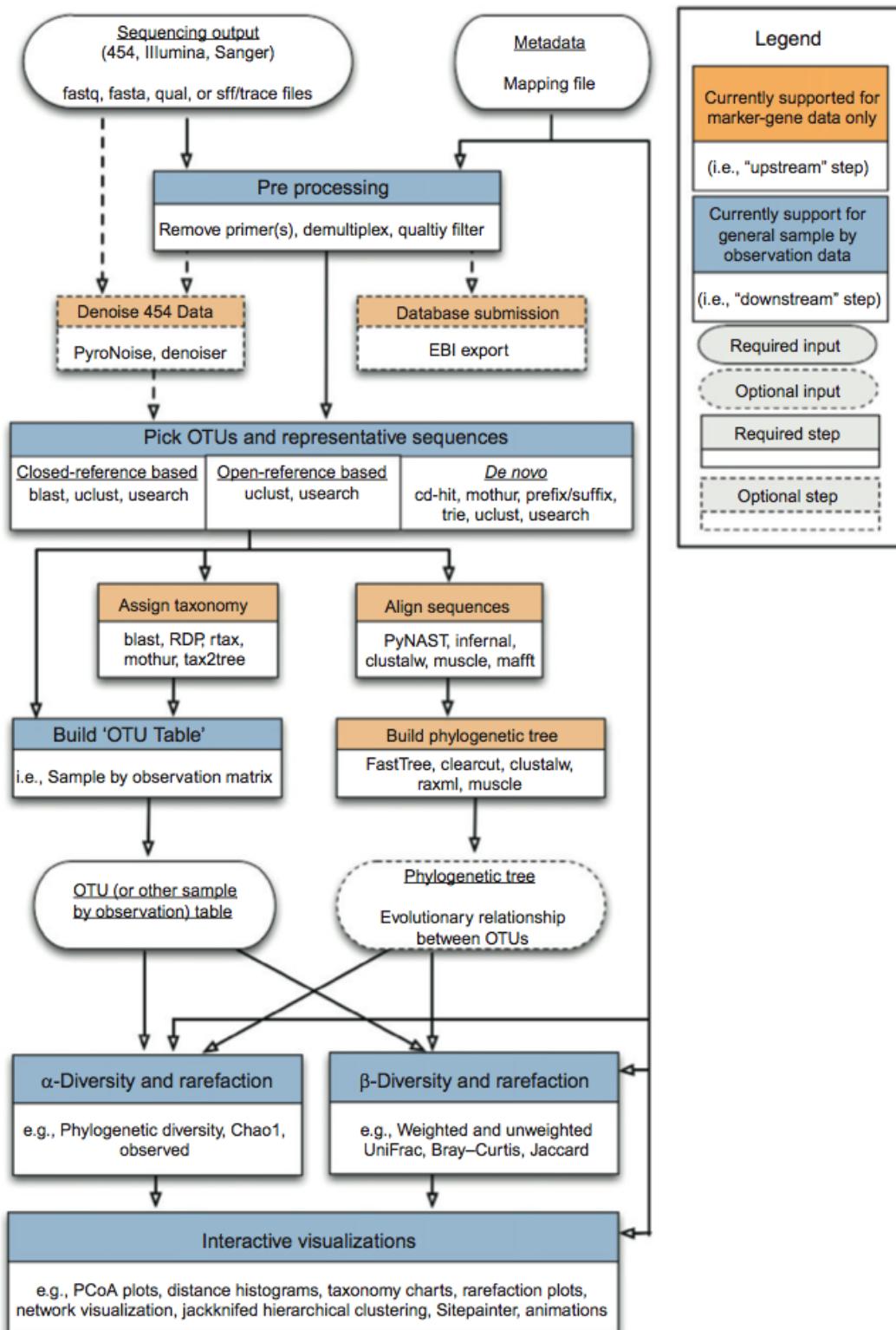


Figura 2.1: Flujo de trabajo de qiime. Se representan esquemáticamente todas las opciones de qiime. Cada una de ellas las realiza un *script* diferente, de tal manera que puede iniciarse el análisis en cualquier punto del flujo de trabajo.

2.2. CLASIFICACIÓN TAXONÓMICA: QIIME

- Closed-reference: Las lecturas son agrupadas contra una colección de secuencias referencia y las que no agrupan son excluidas del análisis. Es el método más rápido, al ser muy paralelizable, y se obtienen mejores taxonomías porque son OTUs definidas previamente. Sin embargo, no permite detectar nuevas OTUs así que depende mucho de lo bien caracterizada que esté la base de datos. Los métodos de agrupación que se pueden utilizar son: *blast*, *uclust* y *usearch*.
- De novo: Las lecturas se agrupan por similitud unas contra otras, sin ningún tipo de referencia externa. El beneficio es que todas las *reads* son agrupadas pero no es paralelizable por lo que sería un proceso muy lento para grandes sets de datos. Los métodos de agrupación son: *uclust* y *usearch*.
- Open-reference: Las lecturas son agrupadas contra la referencia y las que no se encuentran en la referencia, son agrupadas posteriormente *de novo*. Presenta la ventaja de que todas las *reads* quedan agrupadas y además es paralelizable la mitad del proceso. Suele ser la estrategia preferida aunque no es recomendable utilizarla en datos con pocas referencias porque el proceso puede tardar días en completarse. Los métodos de agrupación son: *cd-hit*, *mothur*, *prefix/suffix*, *trie*, *uclust* y *usearch*.

En este proyecto se selecciona *open-reference* como estrategia de selección para que todas las lecturas queden agrupadas incluso aunque no se encuentren en la base de datos, ya que nos interesa obtener toda la información posible de los datos. Al final del proceso, se obtiene una tabla cuya primera columna es el identificador de OTU y la segunda columna son los conteos, que pueden generarse en valor absoluto o relativo.

Greengenes (<http://greengenes.secondgenome.com>) Es la base de datos que se elige como referencia en este caso. Contiene taxonomía de 16S de calidad controlada, basada en una filogenia *de novo* que proporciona conjuntos de OTUs estándar. Está bajo la licencia Creative Commons BY-SA 3.0. Incluye los siguientes niveles de taxonomía:

- Nivel 1 = Reino (por ejemplo, *Bacteria*),
- Nivel 2 = Filo (por ejemplo, *Firmicutes*),
- Nivel 3 = Clase (por ejemplo, *Bacilli*),
- Nivel 4 = Orden (por ejemplo, *Lactobacillales*),
- Nivel 5 = Familia (por ejemplo, *Streptococcaceae*),
- Nivel 6 = Género (por ejemplo, *Streptococcus*),
- Nivel 7 = Especies (por ejemplo, *pneumoniae*).

Un ejemplo del formato sería:

k_Bacteria; p_Firmicutes; c_Bacilli; o_Lactobacillales;
f_Streptococcaceae; g_Streptococcus; s_pneumoniae

2.2. CLASIFICACIÓN TAXONÓMICA: QIIME

Uclust (http://drive5.com/usearch/manual/uclust_algo.html) Es el método de agrupación utilizado en estos datos. Es un algoritmo diseñado para agrupar secuencias de nucleótidos o aminoácidos en base a su similitud [10]. Cada grupo o *cluster* está definido por una secuencia representativa conocida como “centroide”. Sigue dos criterios de agrupamiento simples, con respecto a un umbral de similitud (T) dado: (1) todas las secuencias dentro de un *cluster* tienen similitud $\geq T$ con la secuencia centroide y (2) todos los centroides tienen similitud $< T$ entre ellos. Hay que tener en cuenta que una secuencia puede coincidir en similitud con dos centroides diferentes. Idealmente, se asignará al centroide más cercano, pero puede haber dos o más centroides a la misma distancia, en cuyo caso la asignación de *cluster* es ambigua y se debe tomar un criterio de elección arbitrario. La similitud se calcula utilizando alineamiento global. Además, se trata de un algoritmo voraz (también conocido como *greedy*) que es aquél que elige la opción óptima en cada paso local esperando llegar a una solución óptima global, por lo que es importante el orden en que van entrando las secuencias. Si la secuencia entrante coincide con un centroide existente, se asigna a ese grupo y si no coincide, se convierte en el centroide de un nuevo grupo. Esto significa que las secuencias deben estar ordenadas para que los centroides más adecuados tiendan a aparecer más temprano. Dado que las lecturas más abundantes tienen más probabilidades de ser secuencias de amplicones correctas, y por tanto son más probables de ser verdaderas secuencias biológicas, considera las secuencias de entrada en orden de disminución de la abundancia.

El motivo por el que se selecciona el método *uclust* y la base de datos *greengenes* es porque se pretende reproducir lo que los autores de los datos obtienen originalmente y ellos emplean estas opciones. Tanto los métodos como las bases de datos actuales están bastante optimizados y decantarse por uno u otro implica obtener diferencias, aunque son poco significativas.

2.2.2. Asignación taxonómica

Una vez creada la tabla OTU, qiime permite asignar una taxonomía a cada secuencia representativa. Actualmente los métodos implementados son BLAST, clasificador RDP, RTAX, mothur y *uclust*. Después, se realiza un resumen de la representación de los grupos taxonómicos dentro de cada muestra según un nivel elegido por el usuario. Ese nivel depende del formato que se devuelve desde el paso de la asignación de taxonomía (nivel 1 a 7). La salida de este proceso es una tabla donde la primera columna es la taxonomía y en la segunda columna se mantienen los conteos en valor absoluto o relativo.

2.3. Análisis de variabilidad: complexCruncher

Es un software de código abierto implementado en Python que sirve para el estudio de la variabilidad en series temporales [11]. Acepta como *input* ficheros excel (en formato xlsx) o ficheros de texto, y genera como *output* una serie de gráficos (eps, svg, png, pdf o ps) y tablas (tex o xlsx). Se puede utilizar en modo automático, que analiza en paralelo todos los conjuntos de datos de entradas, o interactivo, que genera los resultados que el usuario pide. En los siguientes apartados se especifica la forma de generar cada resultado.

2.3.1. Regresión lineal y exponencial

Primero se comprueba si se ajusta a una recta de regresión la desviación estándar de los datos frente a su media. La idea es que los valores observados se sitúen lo más cerca posible de la recta de ajuste, minimizando sus distancias a la misma. Esas distancias se denominan errores residuales o, simplemente, residuos. Para realizar el ajuste, se suman todos los cuadrados de los errores residuales para obtener un solo error que se conoce como suma de los errores al cuadrado (SSE – del inglés “Sum of Squares Error”) y se elige la recta con el valor SSE más pequeño.

En ocasiones los datos no se ajustan a una recta sino a una curva exponencial de la forma $y = A \cdot r^x$. En estos casos, se convierte la curva exponencial en recta por medio de logaritmos y se aplica el ajuste visto anteriormente. Aplicando propiedades de logaritmos quedaría de la forma:

$$\log y = \log A + x \cdot \log r \quad (2.1)$$

donde la pendiente es $\log r$ y la intersección con el eje de ordenadas es $\log A$.

2.3.2. Ley de potencia x -ponderada

Cuando se ajusta la ley de potencia de desviación estándar frente a la media, hay que tener en cuenta que cada media tiene incertidumbre y se puede estimar, para un tamaño de muestra n , por el error estándar de la media (SEM – del inglés *Standard Error of the Mean*). En este caso, las incertidumbres afectan a la variable independiente, por lo que el ajuste no es tan trivial como un ajuste y-ponderado, donde las incertidumbres afectan a la variable dependiente. Un método estándar para realizar este ajuste es (1) invertir las variables antes de aplicar los pesos, (2) realizar el ajuste ponderado, y (3) revertir la inversión. Este método es determinista, pero la solución aproximada empeora con coeficientes de determinación más pequeños. Para superar esa limitación, se desarrolla un método estocástico con una estrategia de tipo *bootstrap* que evita la inversión y es aplicable independientemente del coeficiente de determinación.

La idea básica del *bootstrap* es que la inferencia sobre una población de datos muestreados puede ser modelada mediante un nuevo muestreo de los datos y realizando la inferencia a partir de datos remuestreados. Para adaptar esta idea general al problema aquí descrito, se realizan múltiples replicaciones donde se remuestrea la matriz de datos x utilizando su matriz de errores. Es decir, se calcula cada vez una nueva matriz de datos x sobre la base de $x_i^* = x_i + v_i$, donde v_i es una variable aleatoria gaussiana con media $\mu_i = 0$ y desviación estándar $\sigma_i = \text{SEM}_i$. En cada repetición, se realiza un ajuste completo de la ley de potencia no ponderada. Los parámetros de la x-ponderación se estiman promediando a través de todos los ajustes de repetición realizados, y sus errores se determinan mediante el cálculo de la desviación estándar para todos los ajustes.

2.3.3. Estandarización

Sirve para visualizar varios estudios en un diagrama compartido con unidades de desviación estándar de los parámetros Taylor en sus ejes. Para ello, se estandarizan V y β utilizando el grupo de sujetos sanos de cada estudio individualmente.

Para el parámetro V , la estimación de la media (\hat{V}) del grupo de sanos, compuesta por h individuos, es:

$$\hat{V} = \frac{1}{W_1} \sum_{i=1}^h V_i \omega_i = \sum_{i=1}^h V_i \omega_i \quad (2.2)$$

con $W_1 = \sum_i^h \omega_i = 1$, donde ω_i son los pesos normalizados calculados como:

$$\omega_i = \frac{\frac{1}{\sigma_{V_i}^2}}{\sum_i^h \frac{1}{\sigma_{V_i}^2}} \quad (2.3)$$

donde σ_{V_i} es una estimación de la incertidumbre en V_i obtenida junto con V_i de la ley de potencia x -ponderada (descrita en el apartado anterior) para sujetos sanos.

Del mismo modo, la estimación de la desviación estándar para la población sana ($\hat{\sigma}_V$) es:

$$\hat{\sigma}_V = \sqrt{\frac{1}{W_1 - \frac{W_2}{W_1}} \sum_{i=1}^h \left[\omega_i (V_i - \hat{V})^2 \right]} \quad (2.4)$$

con $W_2 = \sum_i^h \omega_i^2$, que finalmente queda como:

$$\hat{\sigma}_V = \sqrt{\frac{1}{1 - \sum_i^h \omega_i^2} \sum_{i=1}^h \left[\omega_i (V_i - \hat{V})^2 \right]} \quad (2.5)$$

2.3.4. RSI y medidas de variabilidad

ComplexCruncher genera unas matrices de rango para los 50 taxones más abundantes (figuras 3.10, 3.11 y 3.12) que muestran el puesto de un taxón en el ranking de abundancia. En la parte derecha de estas matrices puede observarse una barra que mide el índice de estabilidad de rango (RSI – siglas en inglés de *Rank Stability Index*) en porcentaje. RSI puede oscilar entre 0 y 1, siendo estrictamente 1 para un elemento cuyo rango nunca cambia con el tiempo y 0 para un elemento cuyo rango oscila entre los extremos. Por tanto, RSI se calcula para cada elemento como:

$$\text{RSI} = \left(1 - \frac{\text{saltos de rango reales}}{\text{saltos de rango posibles}}\right)^p = \left(1 - \frac{D}{(N-1)(t-1)}\right)^p \quad (2.6)$$

donde D es el número total de saltos de rango dados por el elemento estudiado, N es el número de elementos que han sido clasificados, y t es el número de muestras temporales. El índice de potencia, $p = 4$, se elige arbitrariamente para aumentar la resolución en la región estable.

Finalmente, bajo las matrices de rango, hay un gráfico con dos medidas relevantes para la variabilidad del rango a lo largo del tiempo. Por un lado, la variabilidad de rango (RV – siglas en inglés de *Rank variability*) se calcula como un promedio para todos los taxones del valor absoluto de la resta entre el rango de cada taxón en el tiempo que se calcula y el rango global de cada taxón. Y por otro lado, las diferencias de variabilidad (DV – del inglés *Differences Variability*) se calculan como un promedio para todos los taxones del valor absoluto de la resta entre el rango de cada taxón en el tiempo que se calcula y el rango que tiene en el tiempo anterior.

2.4. Estudio de interacciones

2.4.1. Coeficiente de correlación de Pearson

En muchos trabajos se utilizan las correlaciones como medida de interacción entre taxones, de tal manera que si dos géneros aparecen y desaparecen de forma similar a lo largo del tiempo quiere decir que están interaccionando. Existen diversos coeficientes que miden el grado de correlación, adaptados a la naturaleza de los datos, pero el más conocido es el coeficiente de correlación de Pearson y es el que se aplica a los datos de abundancia absoluta de este estudio.

2.4. ESTUDIO DE INTERACCIONES

Este coeficiente mide el grado de relación lineal entre dos variables cuantitativas (x e y) sobre una población y se calcula con la siguiente expresión:

$$\rho_{x,y} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \quad (2.7)$$

donde σ_{xy} es la covarianza de (x, y) , σ_x es la desviación típica de x y σ_y es la desviación típica de y .

El resultado numérico fluctua entre el rango $[-1, +1]$. Una correlación de $+1$ significa que existe una relación lineal directa perfecta (positiva) entre las dos variables estudiadas. Una correlación de -1 significa es una relación lineal inversa perfecta (negativa). Y una correlación de 0 se interpreta como que no existe una relación lineal (pero pueden darse otras).

2.4.2. Método de búsqueda de comportamientos

Se pretende agrupar taxones en grupos que reflejen su comportamiento frente a una perturbación (viaje o salmonelosis) para reordenar la matriz de correlaciones. Para ello, se ha diseñado un método sencillo dividiendo las tablas de abundancia relativa en 3 períodos: antes (a), durante (d) y tras (r) el punto en cuestión. Para cada taxón, se calcula la mediana de su abundancia en cada uno de estos períodos (M_a , M_d , M_t). Por último, se calculan los incrementos $\Delta M_1 = M_a - M_d$ y $\Delta M_2 = M_t - M_d$. Si ΔM_1 es positivo, quiere decir que la abundancia ha disminuido con la perturbación y si el incremento es negativo, quiere decir que ha aumentado. En ΔM_2 se da el caso contrario, valores positivos indican disminución de abundancia y valores negativos indican aumento. Además, se escoge un valor de $0,01$ para que el incremento se considere significativamente grande como para que haya cambio en la abundancia. En la tabla 2.1, se especifican los 7 grupos que se han considerado.

2.4.3. Aproximación para obtener interacciones: LIMITS

Trabajos más actuales como el de Fisher *et al.* [12] han cuestionado que las correlaciones midan interacciones reales y, además, han desarrollado un nuevo método que se ha aplicado a los datos para comparar frente a correlaciones. El algoritmo se denomina LIMITS (siglas de *Learning Interactions from Microbial Time Series*) y utiliza una regresión lineal con agregación *bootstrap* para inferir el modelo Lotka-Volterra de tiempo discreto (dLV) en la dinámica de los microorganismos.

2.4. ESTUDIO DE INTERACCIONES

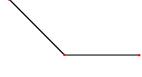
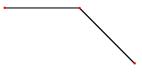
| Grupo | Valores | Comportamiento | Descripción |
|---------|---|---|---|
| Grupo 1 | $\Delta M_1 < 0,01$ y $\Delta M_2 > 0,01$ |  | Aumenta tras la perturbación. |
| Grupo 2 | $\Delta M_1 > 0,01$ y $\Delta M_2 > 0,01$ |  | Disminuye durante la perturbación y recupera el estado inicial. |
| Grupo 3 | $\Delta M_1 > 0,01$ y $\Delta M_2 < 0,01$ |  | Disminuye con la perturbación. |
| Grupo 4 | $\Delta M_1 < -0,01$ y $\Delta M_2 > -0,01$ |  | Aumenta con la perturbación. |
| Grupo 5 | $\Delta M_1 < -0,01$ y $\Delta M_2 > -0,01$ |  | Aumenta durante la perturbación y recupera el estado inicial. |
| Grupo 6 | $\Delta M_1 > -0,01$ y $\Delta M_2 < -0,01$ |  | Disminuye tras la perturbación. |
| Grupo 7 | $-0,01 < \Delta M_1 < 0,01$ y $-0,01 < \Delta M_2 < 0,01$ |  | Sin variación o con variación independiente de la perturbación. |

Tabla 2.1: Tabla de grupos de comportamiento de los microorganismos encontrados tras una perturbación. La primera columna recoge los nombres de los grupos y su color identificativo, la segunda columna incluye los valores de incremento de la mediana que debe tener cada taxón para pertenecer al grupo, la tercera columna describe el comportamiento del grupo y la última especifica la descripción del comportamiento.

El modelo dLV, también conocido como modelo de Ricker, es un modelo clásico de población discreta que relaciona la abundancia de una especie i a tiempo $t+1$ ($x_i(t+1)$) con la abundancia de todas las especies del ecosistema en el tiempo t ($\vec{x} = \{x_1(t), \dots, x_N(t)\}$). Las interacciones se calculan a través del coeficiente de interacción, c_{ij} , que describe la influencia que la especie j tiene sobre la abundancia de la especie i . La dinámica se modela con la ecuación:

$$x_i(t+1) = \eta_i(t) \cdot x_i(t) \cdot \exp \left(\sum_j c_{ij} (x_j(t) - \langle x_j \rangle) \right) \quad (2.8)$$

donde $\eta_i(t)$ es el ruido multiplicativo con distribución log-normal y $\langle x_j \rangle$ es la abundancia de equilibrio de las especies j y se establece por la capacidad de carga del entorno. Aplicando logaritmos se pueden obtener los coeficientes de interacción.

2.4. ESTUDIO DE INTERACCIONES

El algoritmo LIMITS está implementado en Mathematica (Wolfram Research, Inc.). Intenta inferir la matriz de interacciones entre microorganismos a partir de la abundancia absoluta de los microbios en el ecosistema. El procedimiento utiliza regresión por pasos y *bootstrap* que están esquematizados en la figura 2.2. La matriz se inicializa con valores en la diagonal (c_{ii}) distintos de cero porque se sabe que cada especie tiene que interaccionar consigo misma. En cada subsecuente iteración, se añade una interacción adicional c_{ij} al modelo escaneando el resto de especies y eligiendo la que produce el menor error en el grupo test.

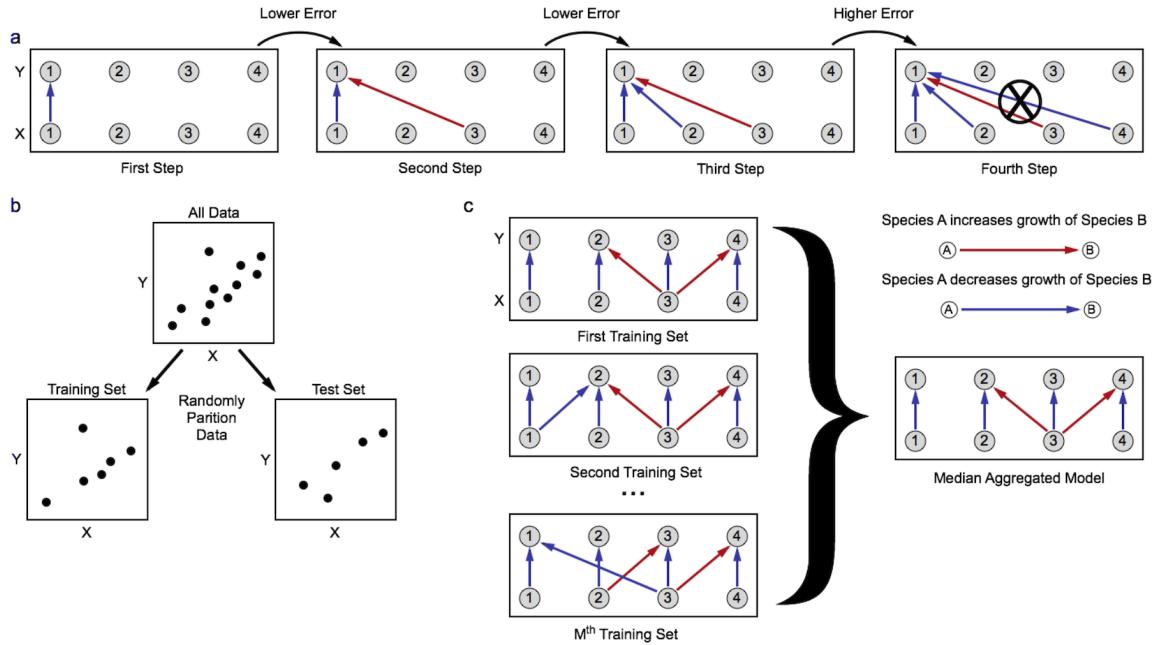


Figura 2.2: Procedimiento de LIMITS. a) Utiliza una regresión por pasos, donde las interacciones se añaden secuencialmente a la regresión si su inclusión reduce el error de predicción por debajo de un umbral predefinido. b) El error se evalúa por agregación bootstrap, que consiste en dividir los datos de forma aleatoria en dos conjuntos: uno de entrenamiento utilizado para la regresión y otro de sondeo para evaluar el error. c) La regresión por pasos se aplica repetidas veces para construir varios conjuntos de entrenamiento que finalmente se simplifican en uno solo aplicando la mediana a todos ellos.

3 Resultados

3.1. Estado de los datos

Los datos utilizados en el presente trabajo proceden del estudio David *et al.* [1] y se encuentran en el repositorio EBI (*European Bioinformatics Institute*) ENA (*European Nucleotide Archive*) con el número de acceso ERP006059.

Abarcan un total de 820 ficheros en formato fastq cada uno de los cuales corresponde a un día de toma de muestra y secuenciación. Los donantes fueron dos varones de 26 y 36 años denominados sujeto A y B, respectivamente. Las muestras fueron tomadas de saliva y heces para analizar el microbioma de boca e intestino, generando tres grupos de estudio:

- Boca del donante A: muestras recogidas entre los días 26-364 que comprenden un total de 286 ficheros.
- Intestino del donante A: muestras recogidas entre los días 0-364 que comprenden un total de 342 ficheros.
- Intestino del donante B: muestras recogidas entre los días 0-318 que comprenden un total de 192 ficheros.

No se tomaron muestras de saliva del sujeto B. Como puede observarse en los grupos anteriores, la saliva comenzó a recolectarse más tarde y cabe que destacar que en todos los grupos hubieron algunos días sin muestra (razones sin especificar). Las muestras las tomaban los propios donantes en casa guardándolas temporalmente a -20°C hasta que se transportaban al laboratorio donde se almacenaban a -80°C.

También se recabaron metadatos sobre el estilo de vida de los sujetos mediante una aplicación iOS que utiliza una base de datos SQL donde anotaban diariamente 13 categorías: alimentación, movimientos intestinales, notas, dieta, ejercicio, aptitud física, cambio de ubicación, medicación, estado de ánimo, higiene bucal, sueño, micción y consumo de vitaminas. Además, se produjeron dos escenarios de cambio en el ambiente de la microbiota debido a que el individuo A realizó un viaje entre los días 71-122 desde América (residencia habitual) al sureste de Asia donde presentó episodios de diarrea por el cambio de dieta/entorno (entre los días 80-85 y 104-113) y el individuo B sufrió salmonelosis durante los días 151-159 pero no tomó antibióticos durante la infección.

3.1. ESTADO DE LOS DATOS

Respecto a la identificación de microorganismos, se decantaron por secuenciar la región V4 del ARN ribosomal 16S con la plataforma Illumina GAIIX. El ADN fue amplificado utilizando *barcoding* y secuenciando lecturas *paired end* de 100 pb (materiales y métodos). El primer obstáculo con los datos se produjo aquí, ya que en el repositorio no se encontraron dos archivos por muestra como suele ocurrir cuando se trabaja con *paired end*. En su lugar, fueron hayados 820 archivos únicos (uno por día) con *reads* de longitud ≤ 100 pb. Se desconoce el procedimiento llevado a cabo por los autores a pesar de escribirles para aclarar este y otros aspectos, pero no se llegó a recibir respuesta. Por tanto, se hipotetiza (1) que utilizaron *single end* pero han cometido una errata al describir la forma de secuenciación, o (2) que utilizaron *paired end* pero posteriormente los solaparon creando lecturas *single* de 100 pb con mejor calidad o incluso que solo hayan utilizado uno de los dos pares (5' o 3'). Para este trabajo eran necesarias secuencias no apareadas (únicas) para posteriores análisis, así que se dio por hecho que las secuencias de los ficheros descargados del repositorio eran *single*, ya que 100 pb son suficientes para una resolución biológicamente significativa si se eligen juiciosamente los cebadores [13].

Se encontraron otras incidencias en los datos como la falta de metadatos para la muestra “Stool69.1260101.fastq” y la existencia de dos muestras para el mismo día (concretamente los días 79, 127, 128, 231, 238, 275, 277, 284 en saliva del sujeto A; los días 7, 44, 74, 79, 82, 84, 106, 120, 162, 277 en intestino del sujeto A y el día 177 en intestino del sujeto B).

Para hacerse una idea de la calidad de los datos, se utilizó FastQC generando un informe para cada uno de los 820 ficheros. Como era muy tedioso ir inspeccionando uno por uno, se utilizó MultiQC para obtener un fichero resumen de todos ellos. El resultado obtenido se recoge en la figura 3.1. El primer gráfico muestra la calidad medida con *phred score (q)* a lo largo de las bases en las 820 muestras. FastQC informó de que los datos tienen la codificación de calidad Sanger/Illumina 1.9, lo cual es importante tener en cuenta. En general, se puede observar que la calidad fue buena pues casi todas las bases (a excepción de dos) presentaron valores superiores a 20. La calidad tendía a bajar un poco en los extremos de las secuencias, fenómeno que suele producirse en secuenciación frecuentemente. Este comportamiento de la calidad remarcaba que las lecturas fueran *single*. Si fueran secuencias solapadas, la figura sería más o menos simétrica respecto a la posición 50, estando las peores calidades entorno a esta posición, pero no se apreciaría un extremo con calidades claramente mejores que el otro, como se aprecia en la figura. El segundo *plot* muestra la calidad en base al número de secuencias. Se forma algo parecido a una campana de Gauss, mostrando que la mayoría de las secuencias tenían calidad q=35 y había muy pocas de “mala calidad” ($q < 30$) en el extremo izquierdo de la campana.

3.1. ESTADO DE LOS DATOS



Figura 3.1: Control de calidad de los datos crudos. El primer plot muestra la calidad medida con *phred score* (q) a lo largo de las bases en las 820 muestras. El segundo plot muestra la calidad en base al número de secuencias. Ambos fueron generados con MultiQC.

3.2. Preprocesado

Antes de realizar cualquier análisis era necesario un preprocesado. Del secuenciador se obtuvo el fragmento de ADN mencionado en el apartado 1.2 (figura 1.2). A esto se le denomina “multiplex” y a la acción de procesarlo se le denomina “demultiplexar”. Además, ya se ha mencionado que la plataforma de secuenciación no es perfecta y en ocasiones se obtienen calidades no deseadas.

En este caso, no fue necesario realizar un demultiplexado porque los autores ya habían realizado este paso previamente y los datos del repositorio ya estaban libres de adaptadores. Se comprobó buscando la secuencia del cebador de PCR directo (GTGCCAGCM-GCCGCGGTAA) y el cebador reverso (GGACTACHVGGGTWTCTAAT) en las *reads* pero no fueron hayadas (ni tampoco las secuencias reversa, complementaria y reversa-complementaria a los cebadores). Por ello, se dedujo que ya fueron eliminados.

En general, las calidades de secuenciación eran buenas como se vio en el apartado anterior, pero aún podían eliminarse algunas secuencias que tenían peor calidad. El valor q al cual filtrar fue de elección arbitraria, siempre hay que llegar a un compromiso entre quedarse con lecturas de buena calidad pero sin perder demasiada información. En este caso, con un valor $q = 30$ se perdieron pocas lecturas y se obtuvo una buena calidad: habían $211,7 \cdot 10^6$ *reads* de partida y tras el filtrado quedaron $208,3 \cdot 10^6$ *reads*, así que se eliminó el 1,64% de las lecturas al filtrar. Las secuencias fueron filtradas con seq_crumps (materiales y métodos) para eliminar todas aquellas con calidad media $q < 30$. Los resultados pueden observarse en la figura 3.2. Como se filtró por calidad media, no se observa ningún cambio en la primera gráfica de calidad a lo largo de las bases. Sin embargo, en el segundo gráfico se comprueba que el programa ha eliminado todas las *reads* inferiores a 30, perdiendo esa cola izquierda de la campana de Gauss que rozaba la franja sombreada en naranja.

Otros elementos que afectaban a la calidad fueron las quimeras, que son combinaciones de dos o más secuencias producidas durante el proceso de PCR como un artefacto. Para eliminarlas se utilizó qiime v1.9.1 (materiales y métodos). Tras el filtro de calidad, se convirtieron los ficheros fastq en fasta que es el formato de entrada que acepta qiime. A continuación, se eliminaron quimeras con los scripts *identify_chimeric_seqs.py* y *filter_fasta.py* como se detalla en el Anexo I. Después de filtrar por calidad quedaban $208,3 \cdot 10^6$ *reads* y tras este paso se obtuvieron $206,9 \cdot 10^6$ *reads*, por lo que el 0,64% de las secuencias eran quimeras.



Figura 3.2: Control de calidad tras el filtro de calidad. El primer plot muestra la calidad medida con *phred score* (q) a lo largo de las bases en las 820 muestras. El segundo plot muestra la calidad en base al número de secuencias. Ambos fueron generados con MultiQC.

3.2. PREPROCESADO

El número de lecturas que se generaban cada día también es un factor importante. Si en un fichero aparecen tan solo 2 o 3 *reads* es un indicativo de que algo no funcionó bien en la secuenciación ese día. Por tanto, se realizó un paso más de preprocesado de los datos, eliminando aquellos ficheros que contenían un número de *reads* inferior a 10.000. Este valor también fue arbitrario en base al compromiso cantidad-calidad de información: si se eliminan muchas secuencias, no hay suficiente información pero si se incluyen los ficheros con pocas *reads*, se introducen errores en el análisis. Los días eliminados del estudio aparecen detallados en la tabla 3.1. De $206,9 \cdot 10^6$ *reads* que llegaron sin quimeras, se eliminaron el 0,02 %, con las que se realizó todo el análisis.

Las cifras globales son:

- Número de *reads* iniciales: $211,7 \cdot 10^6$ *reads*
- Número reads al final de todo el filtrado: $206,9 \cdot 10^6$ *reads*
- Porcentaje global de reducción: 2,28 %

En el Anexo I se detalla todo el preprocesado de los datos con los *scripts* utilizados y sus opciones.

| ID muestra | Número de <i>reads</i> | Donante |
|-------------------|------------------------|----------|
| Stool448.1259730 | 1 | Sujeto B |
| Stool196.1259770 | 2 | Sujeto A |
| Stool13.1259916 | 4 | Sujeto A |
| Saliva267.1260193 | 5 | Sujeto A |
| Stool85.1260354 | 8 | Sujeto A |
| Stool217.1260272 | 8 | Sujeto A |
| Stool63.1259769 | 29 | Sujeto A |
| Stool120.1259849 | 31 | Sujeto A |
| Stool147.1260039 | 39 | Sujeto A |
| Stool36.1259652 | 54 | Sujeto A |
| Stool453.1260253 | 1006 | Sujeto B |
| Stool92.1259811 | 1423 | Sujeto A |
| Stool452.1259809 | 1738 | Sujeto B |
| Stool384.1259728 | 2501 | Sujeto B |
| Stool340.1260381 | 2772 | Sujeto A |
| Stool4.1260013 | 3554 | Sujeto A |
| Stool343.1259705 | 4026 | Sujeto A |
| Stool454.1260333 | 4493 | Sujeto B |
| Stool382.1260123 | 6395 | Sujeto B |
| Stool345.1259808 | 7462 | Sujeto A |
| TOTAL | 35551 | |

Tabla 3.1: Esta tabla recoge los 20 ficheros eliminados del estudio por tener un número bajo de lecturas, ordenados de menor a mayor. La primera columna muestra el nombre de la muestra, la segunda el número de lecturas que contiene el fichero y la tercera el donante al cual pertenece dicha muestra.

3.3. Clasificación taxonómica

Para este paso se utilizó también qiime v1.9.1. La selección de OTUs se llevó a cabo con la estrategia *open-reference* al 97 % de similitud, con la base de datos Greengenes y con el método UCLUST. Por último, se asignó la taxonomía resumiendo los taxones a nivel de género (L6). Cabe destacar que un gran número de géneros no pudieron ser clasificados y se quedaron a nivel de familia, orden o incluso clase. En estos casos, la nomenclatura adoptada por qiime fue “others” o “g_”. Por ejemplo: “*k_Bacteria; p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Streptococcaceae; Others*” o “*k_Bacteria; p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Streptococcaceae; g_*”.

También se obtuvo el grupo “Unassigned” donde se guardaron todas aquellas secuencias que el programa no fue capaz de clasificar taxonómicamente a ningún nivel. Presentaba la siguiente nomenclatura: “**Unassigned;Other;Other;Other;Other;Other**”

Este procedimiento se inició de un gran número de ficheros que contenían secuencias de ADN tomadas a lo largo de un año y se creó una gran tabla en formato xlsx (fichero Excel) que resume la abundancia absoluta de OTUs (filas) que había cada uno de los días de ese año (columnas). Se generó una tabla de abundancia por cada muestra y sujeto con las siguientes dimensiones:

- Saliva del donante A: 573 (OTUs) x 285 (días).
- Heces del donante A: 582 (OTUs) x 329 (días).
- Heces del donante B: 402 (OTUs) x 186 (días).

Todo este proceso también queda detallado en el *pipeline* del Anexo I, en el que pueden encontrarse todos los *scripts* de qiime utilizados con la explicación de cada opción. Es reproducible e incluye además un *script* de creación propia implementado en Python, que formatea los ficheros de salida de qiime (un .txt por día) en el fichero de entrada de la siguiente herramienta de análisis, complexCruncher (que acepta de entrada un excel por individuo con una tabla donde aparezcan todos días como columnas adyacentes). Las tablas finales pueden encontrarse en el material suplementario y en la siguiente dirección: <https://github.com/TeresaRubio/TFM/tree/a/Tablas>.

3.4. Explorando la variabilidad temporal

Para extraer las propiedades globales del sistema, se utilizó el software complexCruncher v.1.1rc12. Se utilizó en modo automático para generar todos los análisis que incluye de forma simultánea (ver materiales y métodos). A continuación se detallan todos los resultados obtenidos.

3.4.1. Abundancia de taxones

Una vez que se generaron las tablas de abundancia absoluta del apartado anterior, se representó mediante un histograma la abundancia total cada día para dar una idea global de los datos. En la muestra de saliva (figura 3.3) existía, en general, una abundancia alta a excepción de algunos días donde se apreciaba un claro descenso. En el caso de intestino A (figura 3.4) se tenían más días de muestra, lo que dificultaba un poco su visualización, pero se aprecia que hubieron muchos días con abundancias muy bajas. En el caso de intestino B (figura 3.5) presentaba abundancias elevadas durante los primeros días pero a partir de aproximadamente el día 20, casi todas las abundancias fueron inferiores. Puede apreciarse que en ninguno de los tres casos se conservaron abundancias inferiores a 10.000 por el filtro realizado durante el preprocesado.

Se tuvo en cuenta que los conteos absolutos estaban llenos de errores sistemáticos debidos tanto al proceso de secuenciación como a la asignación taxonómica. Los cambios en abundancia estarían enmascarados por esos errores, así que se trabajó con abundancia relativa de los taxones para ver la variabilidad temporal. ComplexCruncher realizó internamente el cambio de abundancia absoluta a abundancia relativa y elaboró todos los cálculos posteriores con estos valores.

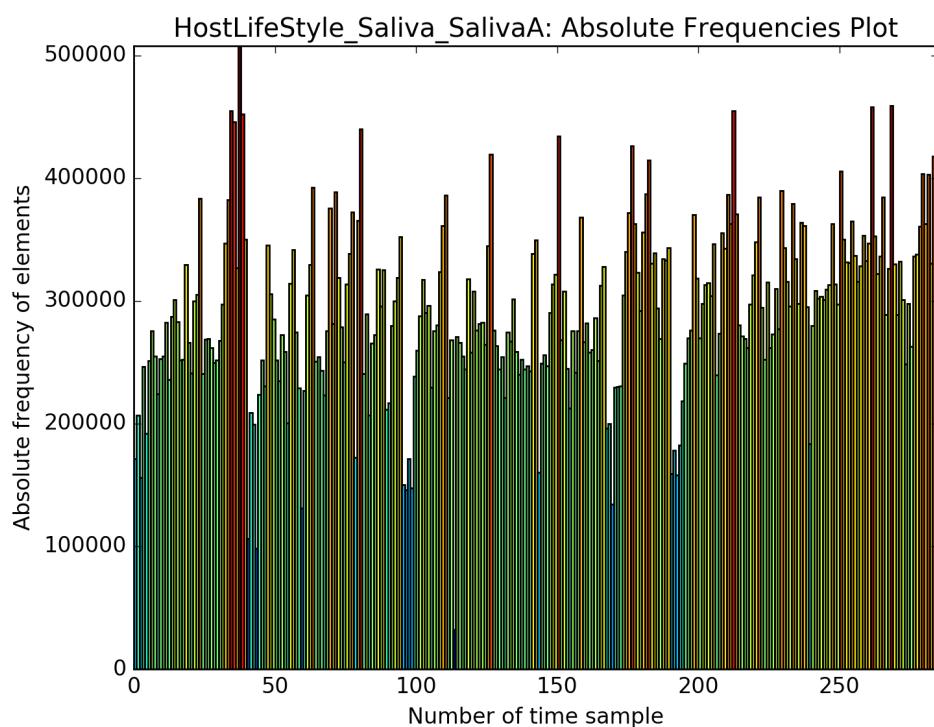


Figura 3.3: Histograma de la muestra saliva A que representa la abundancia total de los géneros en frecuencia absoluta a lo largo del tiempo (285 días). Los colores altos indican altas abundancias y los colores frios, bajas abundancias.

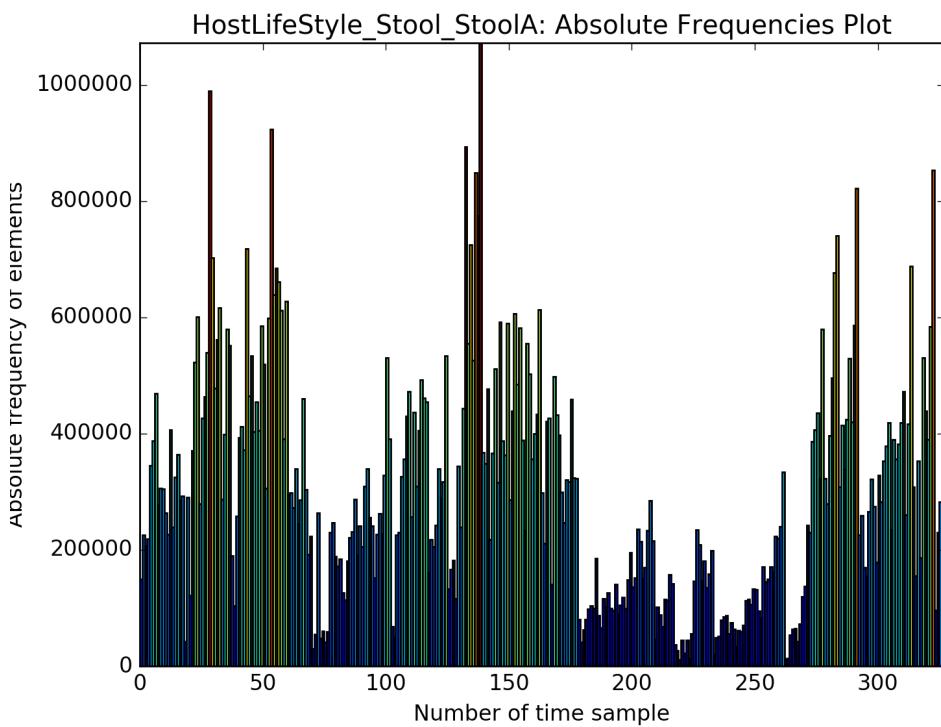


Figura 3.4: Histograma de la muestra intestino A que representa la abundancia total de los géneros en frecuencia absoluta a lo largo del tiempo (329 días). Los colores altos indican altas abundancias y los colores frios, bajas abundancias.

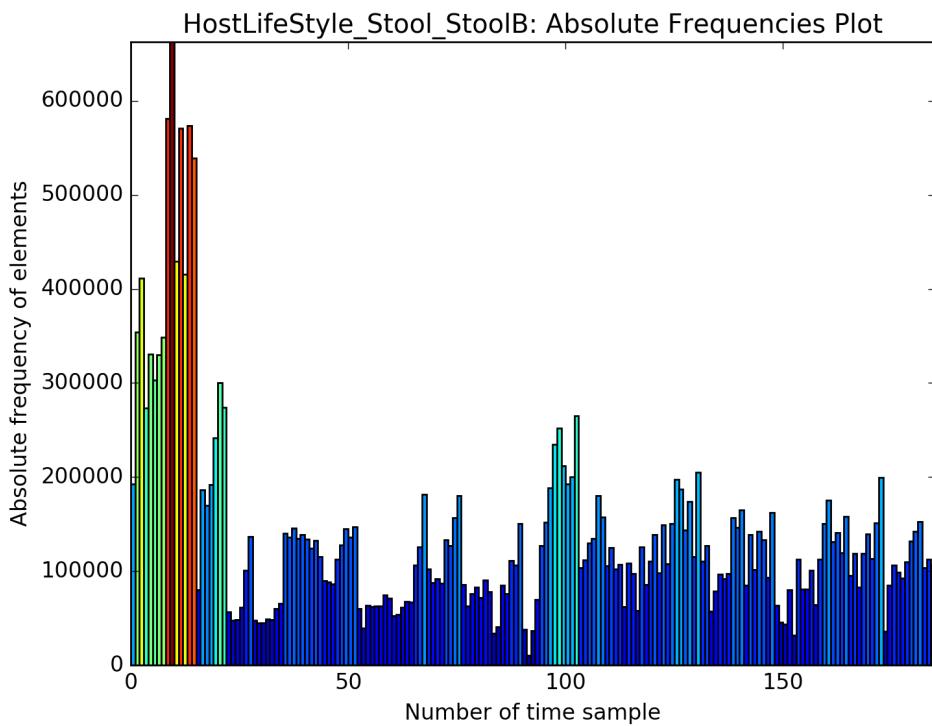


Figura 3.5: Histograma de la muestra intestino B que representa la abundancia total de los géneros en frecuencia absoluta a lo largo del tiempo (186 días). Los colores altos indican altas abundancias y los colores frios, bajas abundancias.

3.4.2. Ley de potencias

Con las tablas de abundancia relativa, complexCruncher comprobó si los datos se ajustaban a un modelo lineal, no lineal o mixto (materiales y métodos). En este estudio se encontró que las fluctuaciones de abundancia relativa en los taxones seguían la ley de potencias de Taylor en todos los casos, como muestra la figura 3.6. Se representa el ajuste exponencial de los datos representado en escala logarítmica para facilitar la visualización. El ajuste fue robusto porque todos los casos presentaban un coeficiente de determinación alto ($R^2 > 0.9$). Dentro de la ecuación, V correspondía a la ordenada en el origen y β a la pendiente. Estos dos parámetros estaban relacionados con la estabilidad del sistema, es decir, describían la variabilidad temporal del microbioma. En metagenómica existen, en general, dos tipos de propiedades estadísticas: $\beta = 0.5$ (distribución de Poisson) y $\beta = 1$ (distribución exponencial). En estos resultados se alcanzó siempre una $\beta < 1$, lo cual indicaba que los taxones dominantes eran menos susceptibles a las perturbaciones que el resto. Por otro lado, V representaba la máxima amplitud de las fluctuaciones, esto es, la variación máxima teórica correspondiente a un género hipotético de abundancia relativa 1. También puede interpretarse como la variabilidad en el origen de β , es decir, cuando $\beta = 0$. Si V es pequeña, la variabilidad de abundancia a lo largo del tiempo sería pequeña y si V es grande, la variabilidad sería grande. Puede observarse que la variabilidad fue menor en saliva que en intestino.

Los parámetros de Taylor, V y β , se relacionan con el estado de salud del hospedador [2]. En general, se considera un estado sano del hospedador cuando el microbioma es estable a lo largo del tiempo y un estado de enfermedad cuando presenta variabilidad temporal. Existen excepciones como por ejemplo en niños, donde el microbioma está en continuo cambio hasta que se desarrolla por completo y entonces aquí el concepto se invierte, se considera sano un microbioma variable y enfermo un microbioma estable. Ya se han comprobado los valores de V y β generales de los sujetos de este estudio pero también fueron analizados aprovechando las perturbaciones que causaban el viaje y la infección en los individuos. Un hecho empírico fue que los sujetos estaban enfermos durante el viaje y la infección, por lo que presentaron una mayor variabilidad temporal del durante las perturbaciones pero, ¿cómo se encontraba una vez pasado estos sucesos? Nuestra hipótesis fue que se recupera la variabilidad inicial tras la perturbación. Para demostrarlo se realizó el cálculo de V y β en todos estos casos con complexCruncher, ya que permitía introducir un fichero excel por individuo con varias hojas: la primera hoja incluía el periodo anual completo, la segunda hoja el periodo previo a la perturbación, la tercera el periodo de la perturbación y la cuarta el periodo tras la perturbación. El programa distinguió entre periodos sanos y de perturbación porque fueron previamente indicados. Este paso de división en periodos temporales se realizó aprovechando el mismo *script* en Python que formateaba los datos para pasar de qiime a complexCruncher.

3.4. EXPLORANDO SERIES TEMPORALES

El *script* generó las subtablas y produjo 3 ficheros excel para introducir a complex-Cruncher:

| Muestra | Días | Periodo |
|-------------|-----------|-------------------------|
| Saliva A | 26 - 364 | Datos anuales |
| | 26 - 69 | Antes del viaje |
| | 72 - 122 | Durante el viaje |
| | 125 - 257 | Después del viaje |
| | 258 - 364 | Después del viaje |
| Intestino A | 0 - 364 | Datos anuales |
| | 0 - 70 | Antes del viaje |
| | 72 - 122 | Durante el viaje |
| | 123 - 257 | Después del viaje |
| | 259 - 364 | Después del viaje |
| Intestino B | 0 - 318 | Datos anuales |
| | 0 - 99 | Antes de la infección |
| | 100 - 143 | Antes de la infección |
| | 144 - 163 | Infección |
| | 164 - 318 | Después de la infección |

Tabla 3.2: Se resume la división de las tablas de abundancias anuales en 4 subperiodos temporales para cada muestra. La primera columna indica la muestra, la segunda el intervalo de días que comprende el periodo y la tercera información acerca de cada periodo.

El resultado obtenido para saliva de sujeto A se muestra en la figura 3.7. Se observan los valores de V y β enfrentados para los 5 intervalos de datos introducidos. El punto azul representa los valores generales que ya habíamos visto en la figura 3.6 A. El punto violeta muestra los valores antes del viaje, el amarillo durante el viaje y el negro y rojo a la vuelta del viaje. Se aprecia claramente que el viaje produjo un aumento de la variabilidad en el microbioma del individuo pero al regresar a su rutina habitual, se recuperaron unos valores similares a los iniciales.

El resultado obtenido para intestino fue aún más interesante. En la figura 3.9 se muestran combinados los valores V y β para el sujeto A y B. El círculo turquesa y la estrella verde simbolizan los valores generales que se vieron en la figura 3.6B y 3.6C. Analizando el sujeto A, se representa en violeta el periodo anterior al viaje, en amarillo el periodo del viaje y en negro y rojo el periodo tras el viaje. Se aprecia, al igual que en saliva, que los valores aumentaron mucho durante el viaje y se recuperaron a la vuelta alcanzando casi el estado inicial. Respecto al sujeto B, se colorea en azul oscuro y gris el periodo antes de la infección, en turquesa con forma de triángulo el periodo de infección y en fucsia el periodo tras la infección. De nuevo, los valores fueron mayores durante la infección y se recuperaron hacia valores similares al estado inicial (aunque de forma más dispersa que en el donante A). Se corroboró así la hipótesis de que tras una perturbación se recuperaba un estado similar, aunque no igual, al de partida.

3.4. EXPLORANDO SERIES TEMPORALES

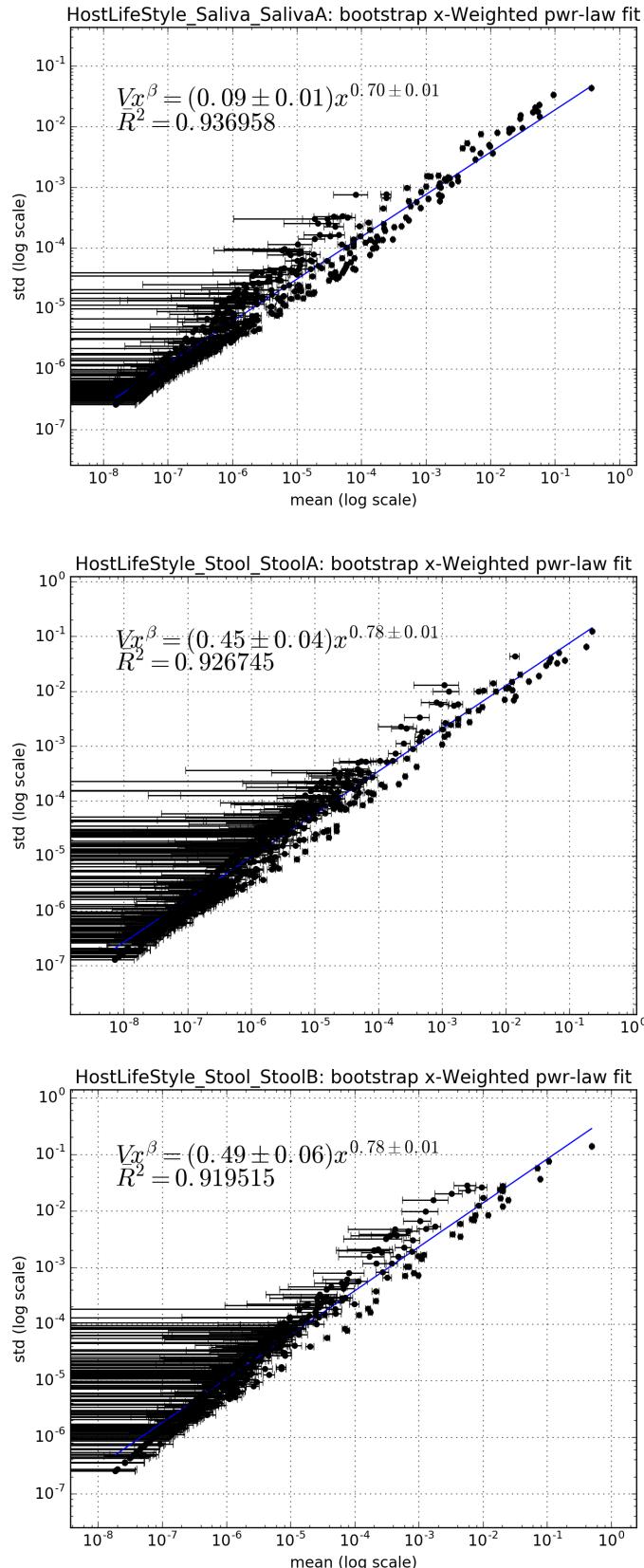


Figura 3.6: Ley de potencias x-ponderada de la desviación estándar (SD) frente a la media de los valores de cada género monitorizados a lo largo del tiempo. El primer ajuste corresponde a la muestra de saliva A, el segundo al intestino A y el tercero al intestino B.

V corresponde a la intersección con el eje y , y β corresponde a la pendiente de la recta.

Las barras de error corresponden al error estándar de la media.

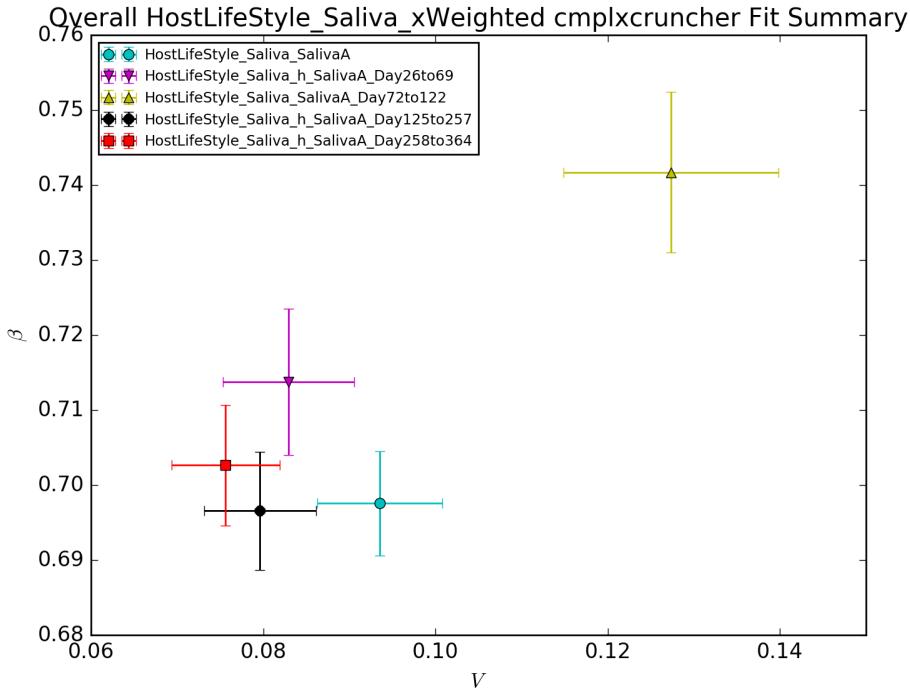


Figura 3.7: Se representan los parámetros de Taylor, V y β , correspondientes a muestras de saliva en distintos períodos: durante todo el año, antes del viaje, durante el viaje y tras el viaje (dividido en dos subperiodos). Los errores fueron calculados por el método bootstrap.

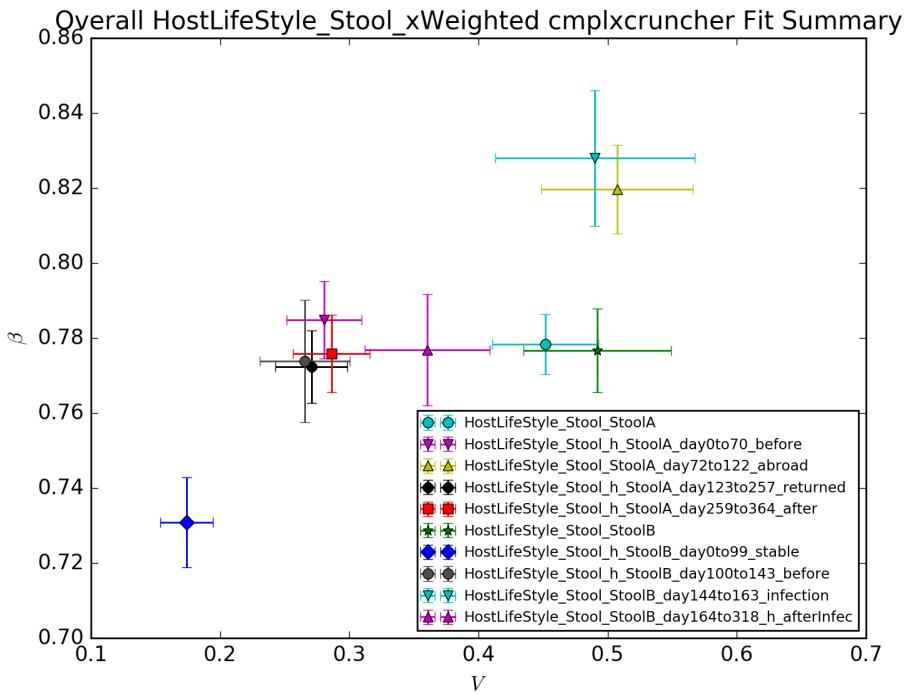


Figura 3.8: Se representan los parámetros de Taylor, V y β , correspondientes a muestras de intestino en distintos períodos. Para el sujeto A: durante todo el año, antes del viaje, durante el viaje y tras el viaje (dividido en dos subperiodos). Para el sujeto B: durante todo el año, antes de la infección (dividido en dos subperiodos), durante la infección y tras la infección. Los errores fueron calculados por el método bootstrap.

3.4. EXPLORANDO SERIES TEMPORALES

Se generó también un *plot* resumen para comparar entre muestras. Anteriormente se había comprobado que tanto en saliva como en intestino aumentan los valores de V y β al producirse una alteración, pero los ejes presentaban escalas diferentes. Para poder visualizarlas conjuntamente se normalizaron los datos (ver apartado normalización de materiales y métodos) restando a cada parámetro el valor medio y dividiendo el resultado por la desviación estándar del grupo de sujetos sanos para cada estudio independientemente (figura 3.9). Así, se definió un área dentro de la cual quedan los puntos correspondientes a los períodos sanos (antes y después de la perturbación). Quedaron fuera de este área los puntos correspondientes al periodo del viaje y a la infección junto con las dos series anuales completas.

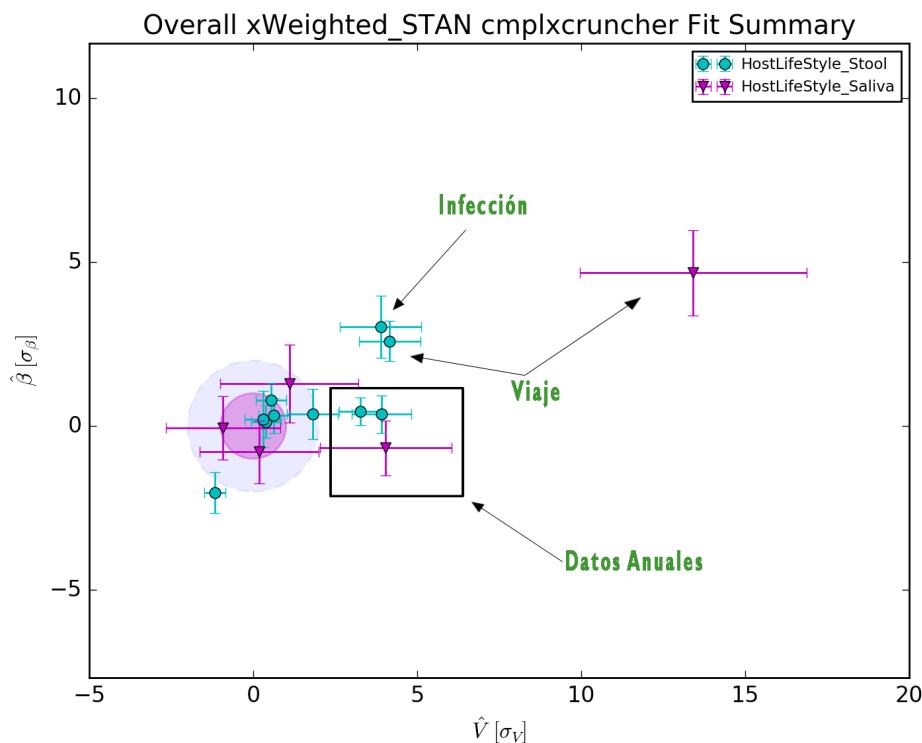


Figura 3.9: Se representan conjuntamente los parámetros de Taylor, V y β , correspondientes a muestras de intestino (azul) y saliva en distintos períodos (violeta). El área sombreada en rosa corresponde a la zona sana definida con la estandarización. El resto de puntos incluyen las perturbaciones y se encuentran a distintas σ de distancia de la zona sana.

3.4.3. Clasificación por rango

Existe una dinámica en la estabilidad de los taxones. Imaginemos un día puntual en la vida de la microbiota humana, supongamos que el taxón X es el más dominante ese día. Al día siguiente resulta que el taxón Y ha aumentado, por los motivos que sean, y ahora es el más dominante dejando al taxón X en segunda posición. Y al tercer día, el taxón Y vuelve a disminuir dejando al taxón X de nuevo en primera posición de abundancia.

3.4. EXPLORANDO SERIES TEMPORALES

Una forma de representar este *ranking* de taxones se plasma en la figura 3.10. Esta matriz recoge en filas los 50 géneros más abundantes ordenados por abundancia y en columnas los días de toma de muestra a lo largo de un año. Nótese que no hay 356 días como corresponde a un año, ya que existen días que tuvieron que ser eliminadas por bajo número de lecturas o incluso algunos días en los que directamente no hubo muestra. Se representaron consecutivamente para facilitar la visualización. El color de cada celda representa el rango, esto es, el orden en ranking de cada taxón, siendo amarillo la representación del primer puesto y violeta oscuro el último puesto. Por ejemplo, en el caso de saliva (figura 3.10): amarillo correspondía al número 1 y violeta oscuro al número 573 (que fue el último taxón). En general, se encontró que los géneros más abundantes suelen ser los más estables. El género más abundante fue *Streptococcus* y ocupaba la primera posición en abundancia a lo largo de todos los días del año (ningún género lo superó nunca en abundancia). Otro género destacable es *Chryseobacterium*, ya que los primeros días del año fue muy poco abundante y a partir del día 50 aproximadamente, aumentó su abundancia. También se presentaron otros patrones intermitentes, que aparecían y desaparecían en pocos días como, por ejemplo, el caso de *Rummeliibacillus*.

En la parte derecha de la figura se muestra el cálculo de RSI (discutido en materiales y métodos) cuyo valor es 100 % para un elemento que nunca cambia en el ranking con el tiempo y 0 % para un elemento que oscila continuamente entre la primera posición y la última de un día a otro. El color en esta columna muestra a su vez una ordenación en base al RSI, es decir, amarillo será el máximo valor de RSI en los 50 taxones y violeta oscuro será el mínimo valor de RSI. Por ejemplo, en el caso de saliva (figura 3.10): amarillo fue 100 (máximo RSI) y violeta oscuro fue 82.4 (mínimo RSI). En el primer tercio de los 50 taxones, se observaron valores de RSI elevados remarcando que los taxones más abundantes presentaban más estabilidad. Sin embargo, en ocasiones se encontraron RSI elevados en el segundo o tercer tercio de los datos, generando las denominadas “islas de estabilidad”. Serían géneros que a pesar de no ser los más abundantes, se mantuvieron estables en su rango a lo largo del tiempo. Algunos ejemplos en la muestra de saliva fueron *Parvimonas* y *Eikenella*.

Por último, en la parte inferior de la matriz se muestra un gráfico con el estudio de la variabilidad a lo largo del tiempo. Se trata de dos medidas de variabilidad que aportan matices distintos: RV respecto al rango global y DV respecto al rango del día anterior (detallado en materiales y métodos). En la figura 3.10 se muestra que hubo un pequeño aumento en ambas medidas de variabilidad durante los días 40-75, correspondientes a los días que el sujeto estuvo de viaje (nótese que el periodo de viaje comprendía los días 71-122 pero en este gráfico se corresponde al intervalo 40-75 porque se representaron consecutivamente los 285 días donde hubo muestra, obviando aquellos días en los que no hubo).

3.4. EXPLORANDO SERIES TEMPORALES

Especialmente se obtuvieron dos picos máximos de DV en los días 60 y 80 aproximadamente. Este incremento en la variabilidad fue muy pequeño para considerarlo significativo, con lo que se dedujo que el viaje no ocasionó demasiado cambio en la variabilidad del microbioma de saliva.

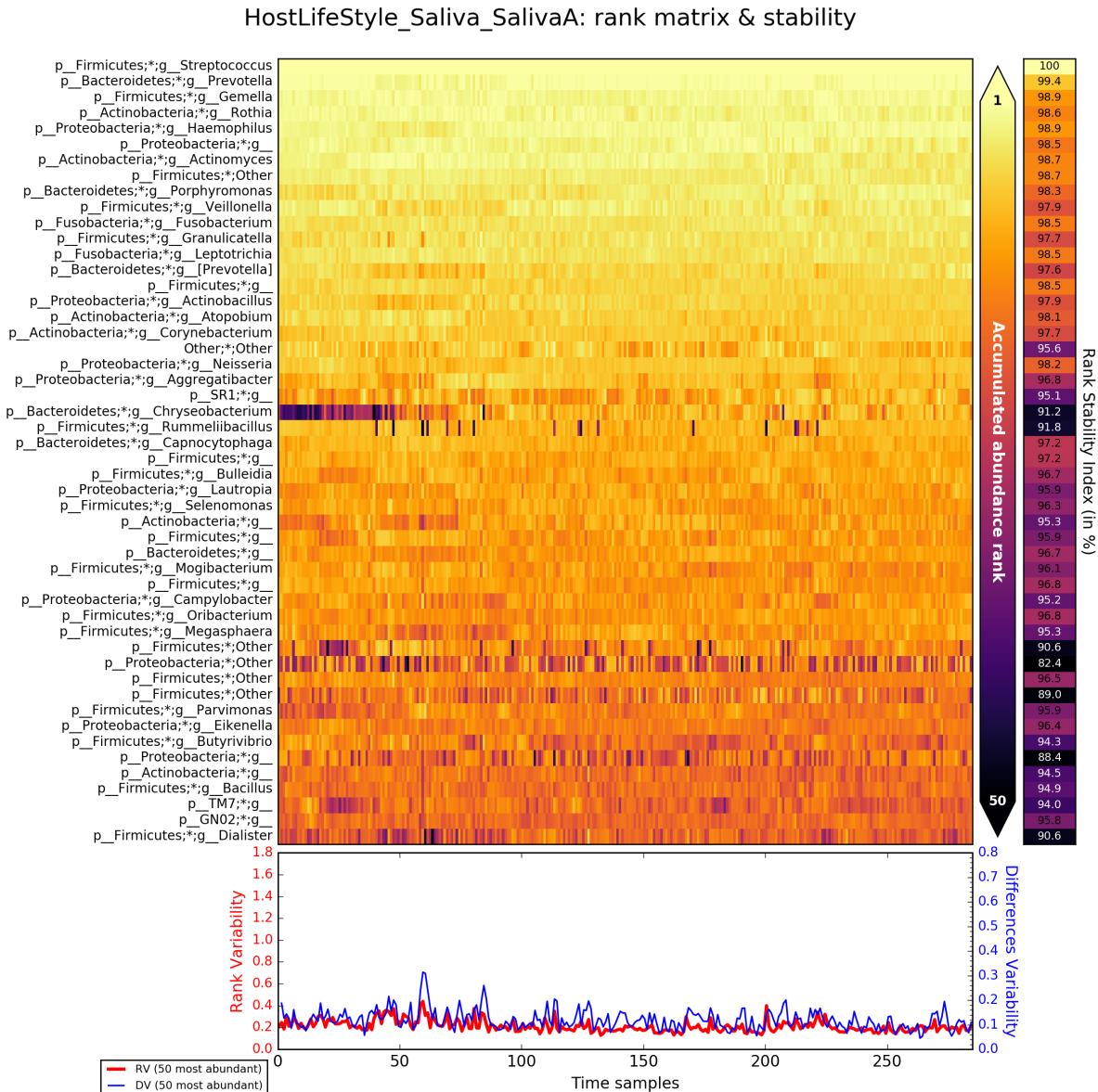


Figura 3.10: Matriz de rango correspondiente a la muestra de saliva A. En filas aparecen los 50 taxones más abundantes ordenados descendenteamente, en columnas se representan los días del año donde hubo muestra y el color determina el rango de mayor (amarillo) a menor (violeta). En la parte derecha aparecen los valores de RSI y en la parte inferior las medidas de RV y DV, todo respecto a los 50 taxones.

3.4. EXPLORANDO SERIES TEMPORALES

En la figura 3.11 queda resumida la dinámica en la estabilidad de la muestra de intestino perteneciente al sujeto A. Aquí también se apreció que los géneros más abundantes solían ser los más estables. Sin duda, lo más llamativo de esta figura es el desorden de rangos que se produjo entre los puntos 71-122, correspondientes a los días en los que el sujeto permaneció de viaje en el extranjero (destacar que en estos datos también existen días sin muestra pero se dan a partir del regreso del viaje, por lo que la correspondencia durante el mismo es correcta). Se dieron comportamientos interesantes como el género *Anaerostipes* que era muy abundante antes del viaje pero durante el viaje disminuyó y a la vuelta recuperó su abundancia inicial. También los géneros *Plesiomonas* y *Fusobacterium* que no fueron nada abundantes pero durante el viaje se dieron las condiciones propicias para su crecimiento. Y por último, al orden *Methylophilales* (etiqueta “p_Proteobacteria;*:g_”) no le afectó el viaje pero, sin embargo, aumentó drásticamente su abundancia al rededor del día 230 por motivos desconocidos. El valor RSI de este último orden fue de 96.2 %, así que fue la isla de estabilidad más llamativa en este sujeto.

En cuanto a la medida de variabilidad, se observó que ambas medidas se dispararon durante el viaje. Además, cabe destacar que a partir del punto 100, RV fue disminuyendo simulando una exponencial lo que supuso una recuperación al estado inicial muy rápida.

En la figura 3.12 queda resumida la dinámica de la estabilidad en la muestra de intestino perteneciente al donante B. De nuevo, los géneros más estables encajaron con los más abundantes de forma general. Además, también se apreció un cambio brusco de rango en el punto 120 aproximadamente que se correspondía al día de inicio de la salmonelosis (apreciar que durante algunos días no hubo muestra y en este gráfico se representaron todos seguidos, sin huecos, como si fuera un muestreo continuo). Con esta turbación del sistema, géneros que eran muy abundantes ahora han disminuido su abundancia (como *Lachnospira*); o al contrario, géneros que eran poco abundantes, aumentan al ser oportunistas (como *Dialister*). También se produjeron comportamientos ajenos como el de *Bacteroides* y el del orden *YS2* (etiqueta “p_Cyanobacteria;*:g_”) que fueron muy estables en su alta y baja abundancia, respectivamente. De hecho, el segundo se consideró como una clara isla de estabilidad con RSI = 97.7 cuando ocupaba el puesto 43 en abundancia.

Examinando el gráfico inferior, la variabilidad sufrió aumentos de un día a otro en varias ocasiones pero los picos más relevantes se obtuvieron a partir de la infección. Tanto RV como DV bajaban de forma parecida a una exponencial como en el caso anterior, pero parece que no se recuperó el estado inicial sino que más bien alcanzó a un nuevo estado de variabilidad.

3.4. EXPLORANDO SERIES TEMPORALES

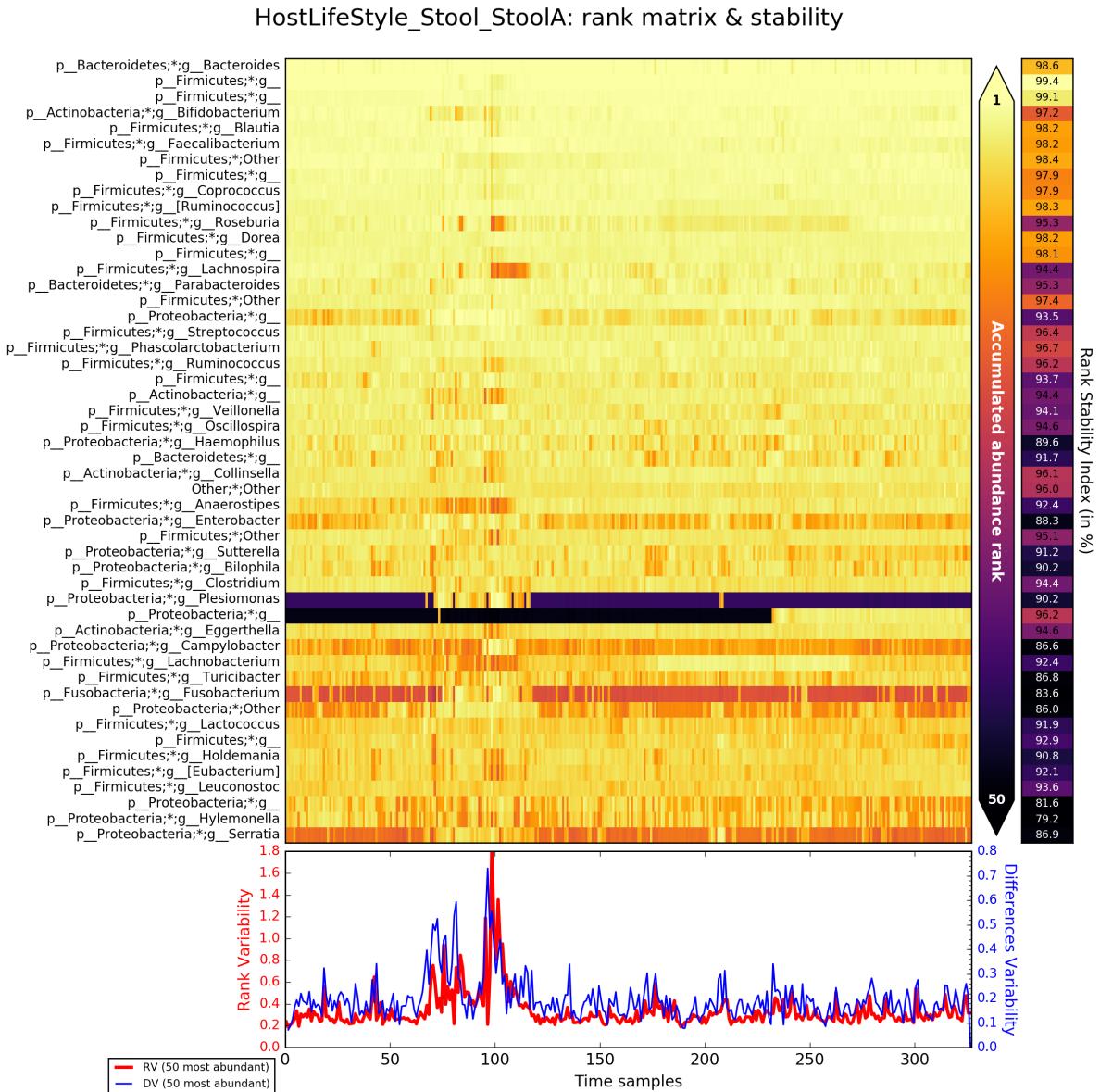


Figura 3.11: Matriz de rango correspondiente a la muestra de intestino A. En filas aparecen los 50 taxones más abundantes ordenados descendenteamente, en columnas se representan los días del año donde hubo muestra y el color determina el rango de mayor (amarillo) a menor (violeta). En la parte derecha aparecen los valores de RSI y en la parte inferior las medidas de RV y DV, todo respecto a los 50 taxones.

3.4. EXPLORANDO SERIES TEMPORALES

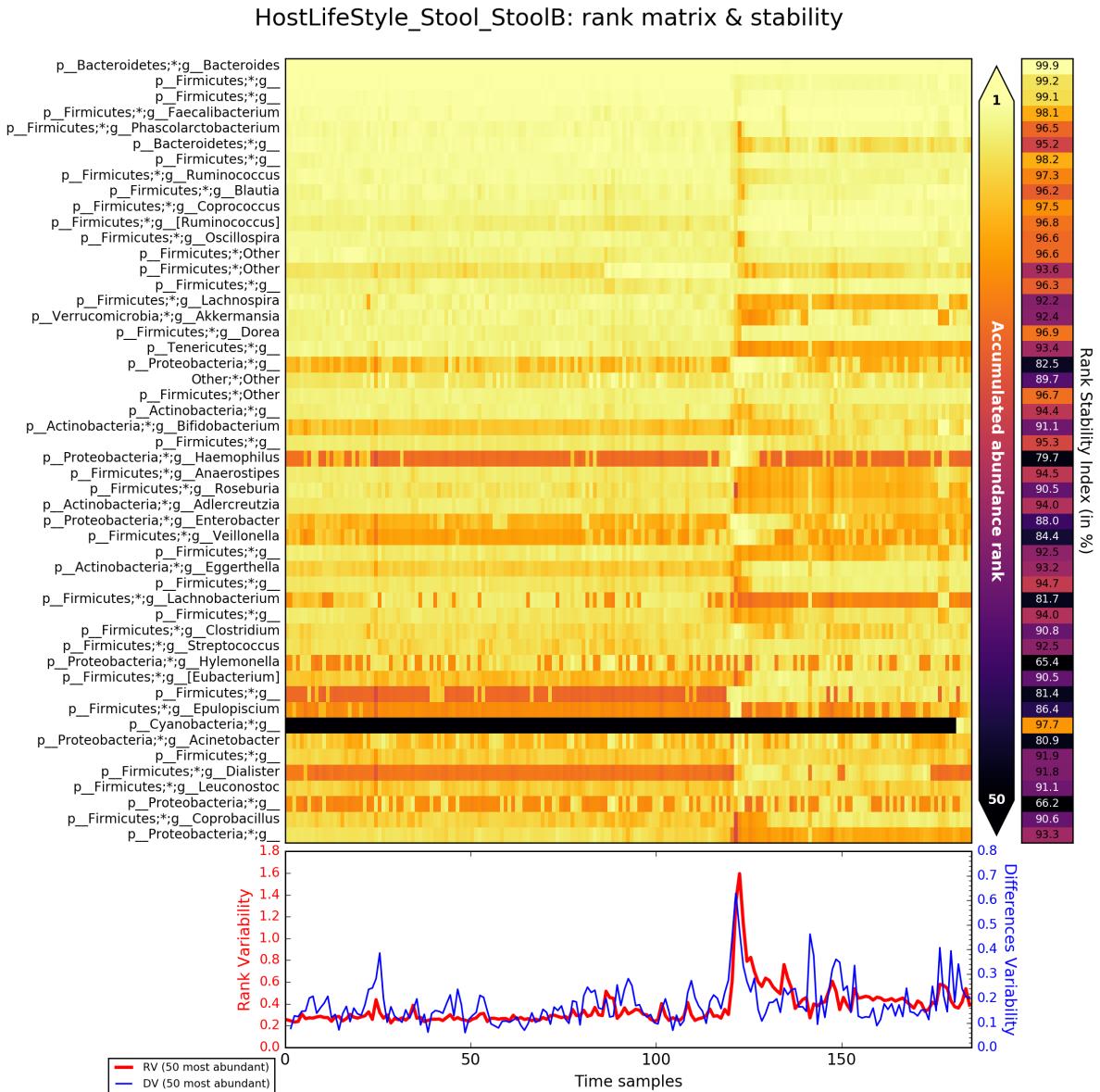


Figura 3.12: Matriz de rango correspondiente a la muestra de intestino B. En filas aparecen los 50 taxones más abundantes ordenados descendenteamente, en columnas se representan los días del año donde hubo muestra y el color determina el rango de mayor (amarillo) a menor (violeta). En la parte derecha aparecen los valores de RSI y en la parte inferior las medidas de RV y DV, todo respecto a los 50 taxones.

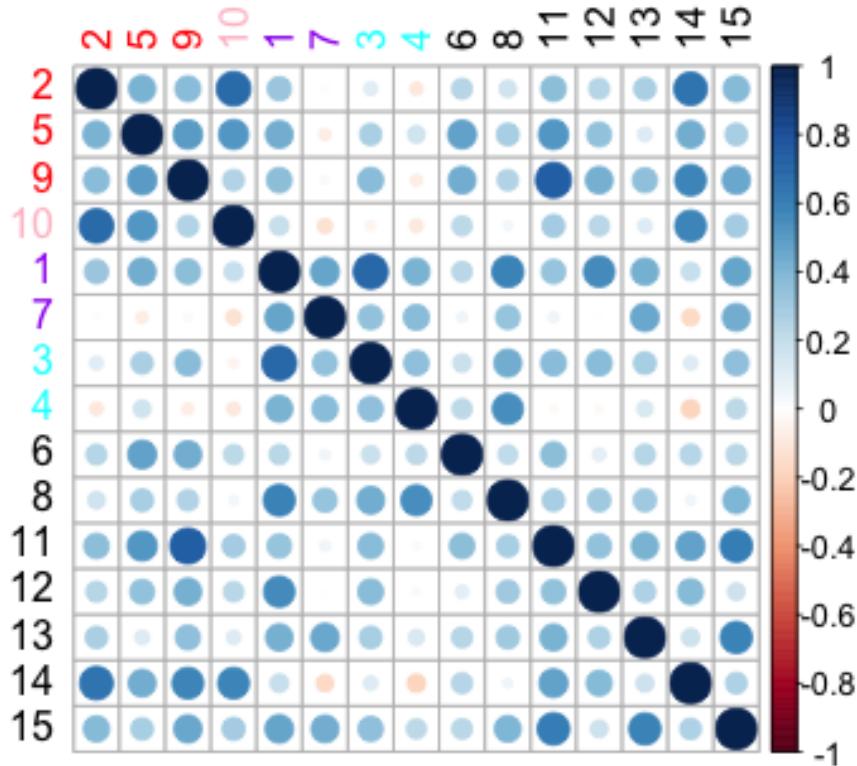
3.5. Correlaciones

Las correlaciones de abundancia en especies microbianas se han utilizado en muchos trabajos para indicar interacción entre ellas. La correlación positiva indicaría una interacción mutualista, y la correlación negativa una interacción competitiva. Trabajar con abundancias relativas produce estimaciones sesgadas porque, como deben sumar a 1, las fracciones no son independientes y tienden a tener una correlación negativa ajena a la verdadera correlación entre las abundancias absolutas subyacentes. Se han desarrollado algoritmos como SparCC [14] para mitigar estos problemas.

En este proyecto también se calcularon las correlaciones aplicando el método de Pearson (materiales y métodos) a las abundancias absolutas de todos los géneros. Además, se ha desarrollado un método de agrupación de los microorganismos más abundantes en base a su respuesta al viaje y a la infección (ver materiales y métodos). Esto sirvió para reorganizar la matriz de correlaciones y comprobar las correlaciones que se dan entre estos grupos de comportamiento. En la figura 3.13 está el resultado para la muestra de saliva. Los géneros 4, 7 y 14 mostraron correlación negativa en algunas ocasiones. Los grupos presentaban correlaciones positivas entre sus miembros aunque las correlaciones máximas se dieron fuera de grupos (como el género 2 con 10 y 14, o el género 9 con 11). Esto quiere decir que los microorganismos que tuvieron una respuesta parecida al viaje, no fueron los que más correlacionan. En la figura 3.14 se muestra el comportamiento que presentaron los géneros del grupo 2. Se representa la abundancia relativa frente al tiempo de 3 géneros y se pudo comprobar que todos disminuyeron su abundancia los días del viaje (40-75) pero la recuperaron a la vuelta. En concreto, *Prevotella* fue el género que más disminuyó durante el viaje y también el más variable a lo largo del tiempo. Se puede observar cómo los géneros correlacionaron bien en algunos puntos (como *Haemophilus* y *Porphyromonas* los primeros días) aunque, en general, la correlación no fue muy grande dentro del grupo.

Los resultados para intestino del donante A se encuentran en la figura 3.15. Se encontraron 4 grupos de comportamiento distinto y ninguna anti-correlación, además el grupo 2 (rojo) es el que presentó mejores correlaciones entre sus miembros. Los grupos 1 (verde) y 3 (naranja) solo incorporaron un miembro por lo que no se pudo apreciar la correlación interna, mientras que el grupo 5 (violeta) presentó correlación positiva pero no tan alta como el grupo 2. De nuevo, en algunos casos se dieron correlaciones más fuertes entre grupos como es el caso de los géneros 10 y 13 que correlacionan mucho con el grupo 2.

En la figura 3.16 se muestra el intestino B como último caso. Fueron hayados 5 grupos, de los cuales el grupo 2 (rojo) presentaba la mayor correlación interna. También hubo casos de correlación muy fuerte entre grupos distintos como *Oscillospira* con *Phascolarctobacterium*. La proteobacteria del grupo 5 (violeta) presentó correlación negativa con el resto y la máxima anti-correlación se dio entre *Ruminococcus* y el grupo 4 (rosa).



- 1: k_Bacteria;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Streptococcaceae;g_Streptococcus
- 2: k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella
- 3: k_Bacteria;p_Firmicutes;c_Bacilli;o_Gemellales;f_Gemellaceae;g_Gemella
- 4: k_Bacteria;p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Micrococcaceae;g_Rothia
- 5: k_Bacteria;p_Proteobacteria;c_Gammaproteobacteria;o_Pasteurellales;f_Pasteurellaceae;g_Haemophilus
- 6: k_Bacteria;p_Proteobacteria;c_Betaproteobacteria;o_Neisseriales;f_Neisseriaceae;g_-
- 7: k_Bacteria;p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Actinomyces
- 8: k_Bacteria;p_Firmicutes;c_Bacilli;o_Lactobacillales;Other;Other
- 9: k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Porphyromonas
- 10: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Veillonella
- 11: k_Bacteria;p_Fusobacteria;c_Fusobacteriia;o_Fusobacteriales;f_Fusobacteriaceae;g_Fusobacterium
- 12: k_Bacteria;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Carnobacteriaceae;g_Granulicatella
- 13: k_Bacteria;p_Fusobacteria;c_Fusobacteriia;o_Fusobacteriales;f_Leptotrichiaceae;g_Leptotrichia
- 14: k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_[Paraprevotellaceae];g_[Prevotella]
- 15: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_-

Figura 3.13: Matriz de correlaciones de los 15 taxones más abundantes correspondientes a la muestra saliva A. En azul se representan las correlaciones positivas y en rojo las negativas. El área de cada circulo corresponde al valor de correlación.

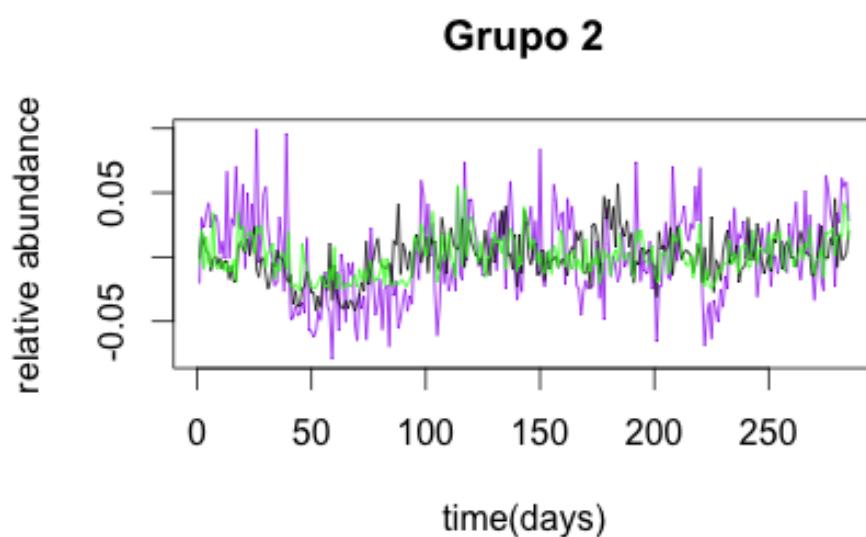
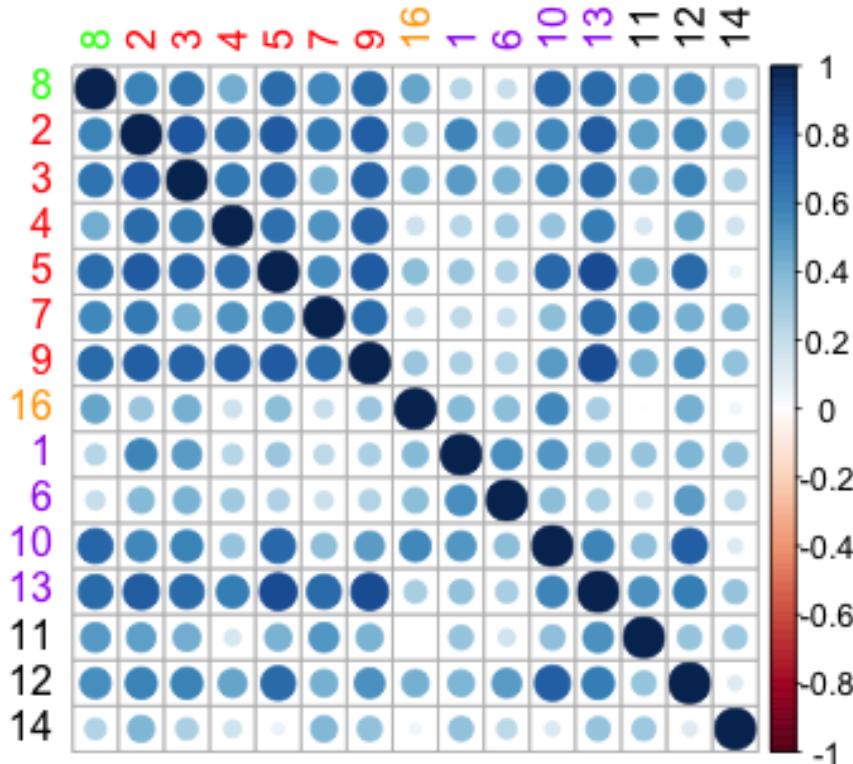
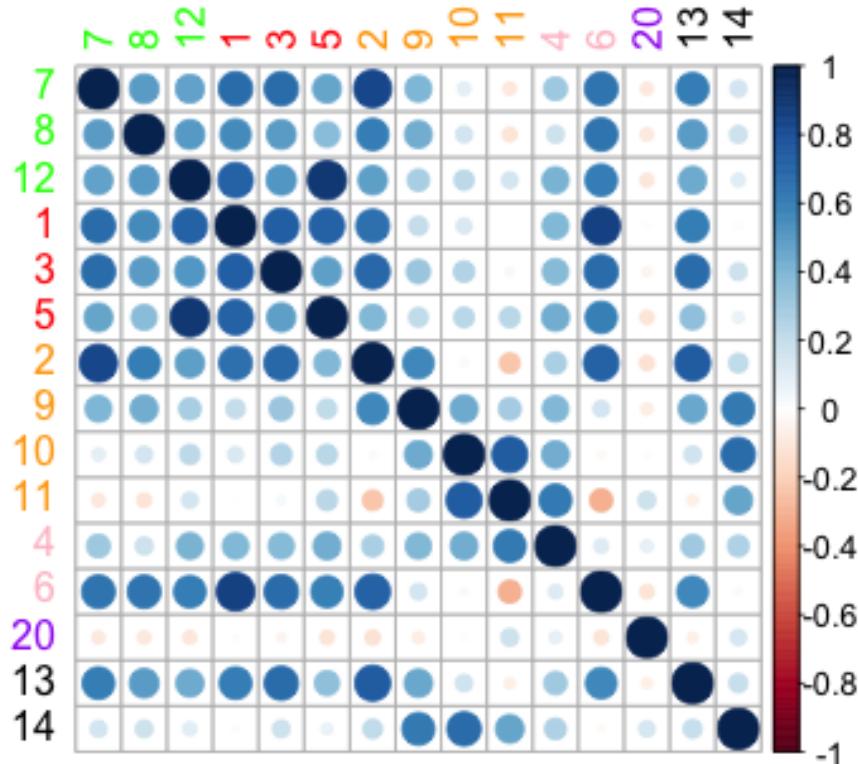


Figura 3.14: Se representa la abundancia relativa a lo largo del tiempo del grupo 2 en la muestra de saliva. Este grupo consta de 3 géneros: *Prevotella* (violeta), *Haemophilus* (negro) y *Porphyromonas* (verde). Se demuestra que no existe una buena relación entre correlaciones y grupos de comportamientos frente al viaje.



- 1: k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides
- 2: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g__
- 3: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g__
- 4: k_Bacteria;p_Actinobacteria;c_Actinobacteria;o_Bifidobacteriales;f_Bifidobacteriaceae;g_Bifidobacterium
- 5: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia
- 6: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium
- 7: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;Other
- 8: k_Bacteria;p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g__
- 9: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus
- 10: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_[Ruminococcus]
- 11: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Roseburia
- 12: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Dorea
- 13: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f__;g__
- 14: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Lachnospira
- 15: k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Parabacteroides

Figura 3.15: Matriz de correlaciones de los 15 taxones más abundantes correspondientes a la muestra intestino A. En azul se representan las correlaciones positivas y en rojo las negativas. El área de cada circulo corresponde al valor de correlación.



- 1: k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae;g_Bacteroides
- 2: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g__
- 3: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g__
- 4: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium
- 5: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Phascolarctobacterium
- 6: k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae;g__
- 7: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f__;g__
- 8: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus
- 9: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia
- 10: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Coprococcus
- 11: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_[Ruminococcus]
- 12: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira
- 13: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;Other
- 14: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;Other
- 15: k_Bacteria;p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g__

Figura 3.16: Matriz de correlaciones de los 15 taxones más abundantes correspondientes a la muestra intestino B. En azul se representan las correlaciones positivas y en rojo las negativas. El área de cada circulo corresponde al valor de correlación.

4 Discusión y conclusiones

El grueso de este proyecto se desglosa en tres bloques principales, que coinciden con los objetivos marcados desde el principio. La primera parte del trabajo, se centra en reproducir la identificación taxonómica de los microorganismos presentes en las muestras. En la segunda parte, se encuentra que los datos siguen la ley de Taylor, lo que permite explorar la estabilidad temporal de la microbiota en diferentes condiciones para entender la relación con el estado de salud de los sujetos. Por último, se hace un estudio de las correlaciones entre microorganismos y se abren las puertas al uso de alternativas para medir interacciones, que es un campo donde aún queda mucho por explorar debido a su complejidad.

El primer objetivo incluye la clasificación taxonómica de diversos genomas en una misma muestra. La idea original de reproducir el proceso llevado a cabo por los autores de los datos originales, se cumple satisfactoriamente. A pesar de emplear una versión más reciente tanto del software como de la base de datos, los resultados son similares (se comprueba por comparación). Esto sirve para testar la reproducibilidad de los métodos empleados que en determinadas ocasiones es importante. Por otro lado, el nivel taxonómico alcanzado no es el más deseable. Se prefiere poder caracterizar los microorganismos a nivel de especie, o incluso de cepa, para poder dar un significado biológico a lo que ocurre en la microbiota. A nivel de género es muy difícil especificar una sola función, porque se trata de un grupo de especies que pueden englobar tanto organismos beneficiosos como patógenos. La aproximación mediante 16S es poco precisa y para un estudio más profundo es necesario el uso de la secuenciación del genoma completo (*shotgun*). Esta última permite identificar las diferencias entre microorganismos aislados y conocer la información funcional y genética, llegando incluso a detectar mutaciones en el genoma. Permite mostrar el contenido de genes de la comunidad, lo que es muy útil para definir las capacidades de la comunidad comparando con bases de datos para conocer las funciones de esos genes.

Respecto al segundo objetivo, abarca todo el análisis de series temporales y se obtienen varios resultados interesantes. Caracterizar el comportamiento global del sistema no es trivial, pero gracias a la herramienta complexCruncher se puede extraer bastante información. Esta es la parte más matemática del proyecto, pero intenta explicar la biología subyacente. Principalmente se consiguen distintos enfoques de la variabilidad del microbioma, lo que se relaciona con la salud del hospedador. Otra idea un poco diferente planteada en esta parte es identificar modos vía descomposición de Fourier. Esto sirve para encontrar posibles periodicidades en el microbioma, es decir, comprobar si se repite

4.1. PERSPECTIVAS DE FUTURO

algún patrón a lo largo del tiempo (semanas, meses...). Los resultados no se añaden a la memoria porque no se encuentran modos elementales en esta serie temporal. Quizás porque son necesarios más puntos temporales o quizás porque simplemente no existen estos patrones.

Por último, el tercer objetivo es inferir la dinámica del sistema. Se logran medir los comportamientos paralelos entre grupos, aunque no se ha llegado a definir cada una de las interacciones que se dan entre microorganismos. Conseguir una red coloreada por bloques de especies que interactúan entre sí será el siguiente objetivo a cumplir y sería ideal que además se pudiera visualizar a través del tiempo. Como es un campo en auge, se abren así las puertas a futuros trabajos.

Como conclusiones finales al estudio, remarcar:

- ARNr 16S es una buena estrategia para la clasificación taxonómica pero no se obtienen buenos resultados más allá de género.
- El estudio de series temporales en la microbiota permite visualizar la dinámica del sistema.
- Los parámetros de Taylor calculados para el microbioma, están relacionados con el estado de salud del hospedador.
- Queda mucho camino que recorrer en el campo en la biología de sistemas para desentrañar las interacciones que ocurren entre nuestros huéspedes microscópicos.

Además, la reciente publicación del artículo Martí *et al.* [2] recalca la envergadura de los datos utilizados y pone de manifiesto la utilidad del proyecto.

4.1. Perspectivas de futuro

El estudio de series temporales ofrece una perspectiva dinámica de cualquier sistema. En concreto, analizar el microbioma humano a lo largo del tiempo nos permite dilucidar el comportamiento de las bacterias que viven con nosotros, las cuales influyen directamente en la salud humana. En este estudio, se demuestra que en condiciones normales la microbiota tiene cierta estabilidad. Sin embargo, cuando el hospedador cambia de ambiente realizando un viaje al extranjero, su flora intestinal se desequilibra debido al nuevo entorno que ofrece el lugar de destino. Este suceso hace que aparezcan nuevos géneros y desaparezcan otros existentes. Cuando el sujeto regresa a su país de origen y a sus hábitos cotidianos, la microbiota recupera la estabilidad inicial. También se ha analizado el microbioma de la cavidad oral pero éste no sufre tanto la perturbación como el intestino.

4.1. PERSPECTIVAS DE FUTURO

Una cuestión interesante que surge de este estudio es, ¿qué le pasa al microbioma de una persona emigrante? Si el sujeto hubiera permanecido más tiempo en el extranjero, habría sido muy interesante comprobar lo que sucede. Algunas posibilidades pueden ser (1) que alcance un nuevo estado de equilibrio, ya sea uno igual al estado inicial o uno nuevo (lo que sería interesante poder demostrar cuánto tiempo se necesita para adquirir el equilibrio) y (2) que nunca alcance el estado de equilibrio (lo cual sería poco probable ya que no existen casos de diarrea crónica causadas por un viaje). La hipótesis más probable es la primera porque la variabilidad disminuye de forma exponencial a partir del día 100, de lo que se deduce que ya había alcanzado el equilibrio antes de su regreso. Otra cuestión interesante es, ¿por qué no hay un nuevo desequilibrio a la vuelta del viaje? El microbioma parece que tiene memoria y no sufre tanto al reexpresarse a un ambiente que ya le es conocido.

En este proyecto también se ha estudiado lo que ocurre en el caso de que un sujeto tenga una infección intestinal causada por un patógeno. De nuevo, la estabilidad de su microbioma se rompe y se puede medir el tiempo que tarda en recuperarse. Se demuestra que una infección supone el cambio del microbioma a un equilibrio nuevo. Este mismo efecto ocurre con la ingesta de antibióticos, que afecta a la mayoría de microorganismos y los oportunistas ocupan esos nichos conformando un nuevo equilibrio.

Los trabajos hasta la fecha han supuesto un enorme avance y además se ha logrado en un periodo de tiempo relativamente corto. Como ya se ha comentado, tanto las tecnologías de secuenciación como los sistemas de clasificación no son perfectos todavía e introducen errores en los resultados. Además, se ha demostrado que los intentos de explicar las relaciones entre microorganismos mediante correlaciones no muestran las verdaderas interacciones [12]. En ese mismo artículo, los autores proponen una aproximación capaz de superar todos esos obstáculos a la que han llamado LIMITS (detallada en materiales y métodos). Para comprobar su potencial y comparar los resultados expuestos en apartados anteriores, se aplicó LIMITS a los datos del estudio. En la figura 4.1 se puede observar un ejemplo que compara la matriz de correlaciones para los 15 géneros más abundantes en saliva A con la matriz de interacciones propuesta por LIMITS para la misma muestra. Se puede observar que hay muchas más correlaciones que interacciones y, además, no se corresponden en la mayoría de los casos. Los elementos de la diagonal obtenidos en la matriz de interacción son todos negativos. La biología subyacente a este resultado debe significar que cada una de estas especies llegaría finalmente a la capacidad de carga incluso en ausencia de otras especies. Otra diferencia importante es que la primera matriz presenta simetría, mientras que la segunda es asimétrica (lo que se ajusta mejor a la realidad debido a que un género puede interaccionar con otro pero éste no necesariamente debe interaccionar con el primero).

4.1. PERSPECTIVAS DE FUTURO

En resumen, LIMITS es una buena aproximación, pero hay que asumir que está ocurriendo el modelo en el que se basa y solamente ofrece dos tipos de interacción (competición o cooperación). Por tanto, ofrece una de las posibles soluciones aunque puede que no sea la correcta ni la única, ya que el mundo de las interacciones es bastante más complejo.

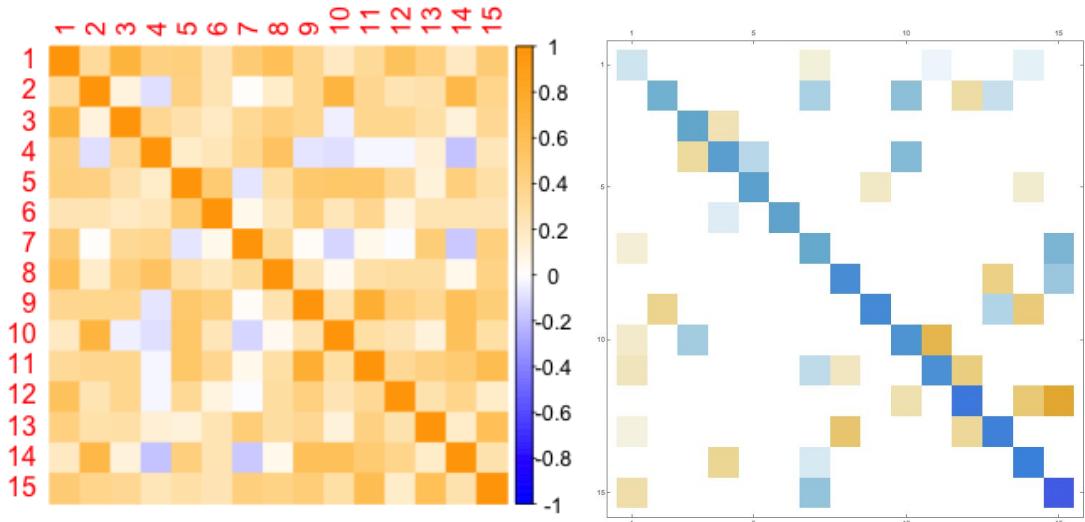


Figura 4.1: Correlación Pearson vs. LIMITS en saliva A

La aplicación de este tipo de estudios radica en monitorizar la administración de probióticos para prevenir y tratar enfermedades como la obesidad o la diabetes. Se requiere de forma paralela un análisis metabolómico para conocer la composición funcional de la flora intestinal y el estado inmunológico del hospedador. Con estas investigaciones, se pueden dilucidar los mecanismos moleculares del microbioma que influyen en las enfermedades y hará posible adoptar un nuevo enfoque en el desarrollo de terapias aprovechando los beneficios de la modulación de la microbiota intestinal sobre el metabolismo.

Hay una clara necesidad de una teoría que identifique patrones y principios generales del microbioma. En el campo de la ecología se han utilizado a lo largo de la historia modelos de red que están específicamente diseñados para tratar con comunidades grandes y complejas. Podría utilizarse una red multicapa que incorpore diferentes instancias de la misma especie en diferentes lugares en vez de una matriz de interacciones para una sola comunidad en el espacio [15]. Una red multicapa consiste en: (1) “nodos físicos” que representan entidades (por ejemplo, géneros); (2) capas, que agrupan los nodos según alguna característica común (por ejemplo, dependencia del tiempo); (3) “nodos de estado”, cada uno de los cuales corresponde a la manifestación de un nodo físico en una capa específica; y (4) aristas (ponderadas o no ponderadas) para conectar los nodos de estado entre sí. Es un nuevo marco que permite considerar múltiples tipos de interacciones y sus ejes permiten investigar cómo las especies se mueven entre las comunidades locales. Generar una red multicapa presenta algunas limitaciones como un esfuerzo adicional en la toma de

4.1. PERSPECTIVAS DE FUTURO

muestras porque se requieren muchos datos de múltiples lugares, múltiples tiempos y/o diferentes métodos de observación. La ventaja es que estudiando la modularidad de las redes, se puede comprobar las variaciones temporales en el tamaño y la composición de los módulos, lo que puede ser relevante para fenómenos como estabilidad de la comunidad, coevolución y coexistencia de especies. Esta nueva perspectiva ofrece una visión teórica y empírica de la dinámica en los sistemas ecológicos.

Bibliografía

- [1] David, L.A., Materna, A.C., Friedman, J., Campos-Baptista, M.I., Blackburn, M.C., Perrotta, A., Erdman S.E., and Alm, E.J. 2014. Host lifestyle affects human microbiota on daily timescales. *Genome biology*, **15**(7), p. R89. <http://dx.doi.org/10.1186/gb-2014-15-7-r89>
- [2] Martí, J.M., Martínez-Martínez, D., Rubio, T., Gracia, C., Peña, M., Latorre, A., Moya, A. and Garay, C.P. 2017. Health and disease imprinted in the time variability of the human microbiome. *mSystems*, **2**(2), pp. e00144-16. <http://dx.doi.org/10.1128/mSystems.00144-16>
- [3] Sender, R., Fuchs, S., and Milo, R. 2016. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol*, **14**(8), p. e1002533. <http://dx.doi.org/10.1371/journal.pbio.1002533>
- [4] Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin E.M., Rokhsar D.S., and Banfield, J.F. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**(6978), pp.37-43. <http://dx.doi.org/10.1038/nature02340>
- [5] Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W. and Fouts, D.E. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**(5667), pp.66-74. <http://dx.doi.org/10.1126/science.1093857>
- [6] Taylor, L.R. 1961. Aggregation, Variance and the mean. *Nature* **189**, pp. 732-35. <http://dx.doi.org/10.1038/189732a0>
- [7] McArdle, B.H., Gaston, K.J., and Lawton, J.H. 1990. Variation in the size of animal populations: patterns, problems and artefacts. *The Journal of Animal Ecology*, **59**(2), pp. 439-454. <http://dx.doi.org/10.2307/4873>
- [8] Blumm, N., Ghoshal, G., Forró, Z., Schich, M., Bianconi, G., Bouchaud, J. P., and Barabási, A. L. 2012. Dynamics of ranking processes in complex systems. *Physical review letters*, **109**(12), p.128701. <http://dx.doi.org/10.1103/PhysRevLett.109.128701>

BIBLIOGRAFÍA

- [9] Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., and Hutley, G. A. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, **7**(5), pp.335-336. <http://dx.doi.org/10.1038/nmeth.f.303>
- [10] Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**(19), pp. 2460-2461. <http://dx.doi.org/10.1093/bioinformatics/btq461>
- [11] Martí, J.M., and Garay, C.P. 2017. ComplexCruncher: dynamics of ranking processes toolkit. In preparation.
- [12] Fisher, C.K., and Mehta, P. 2014. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PloS one*, **9**(7), p. e102451. <http://dx.doi.org/10.1371/journal.pone.0102451>
- [13] Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D., and Knight, R. 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic acids research*, **35**(18), p. e120. <http://dx.doi.org/10.1093/nar/gkm541>
- [14] Friedman, J., and Alm, E.J. 2012. Inferring correlation networks from genomic survey data. *PLoS Comput Biol*, **8**(9), p. e1002687. <http://dx.doi.org/10.1371/journal.pcbi.1002687>
- [15] Pilosof, S., Porter, M.A., Pascual, M., and Kéfi, S. 2017. The multilayer nature of ecological networks. *Nature Ecology & Evolution*, **1**. <https://doi.org/10.1038/s41559-017-0101>

Anexo I: *Pipeline* de asignación taxonómica

Software utilizado: fastQC (0.11.5), multiQC (0.9), seq_crumb (0.1.9), qiime (1.9.1).

Documentación:

- <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- <http://multiqc.info>
- https://bioinf.comav.upv.es/seq_crumb
- <http://qiime.org>

1. Descargar los datos en formato fastq desde ENA (EBI)

En cada punto del *pipeline* se crea un directorio para guardar los ficheros de salida de cada apartado. Como se genera una gran cantidad de datos, hay directorios que pueden ser borrados por el usuario si no son necesarios.

```
$ mkdir 01_Data_fastq  
$ cd 01_Data_fastq
```

Los datos se encuentran depositados en *European Bioinformatics Institute (EBI) European Nucleotide Archive (ENA)* con el código de acceso ERP006059. Son 820 archivos con formato fastq. Se descargan los datos via FTP:

```
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/ERA318/ERA318858/fastq/seqs_Saliva* #  
    Todas las muestras de saliva  
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/ERA318/ERA318858/fastq/seqs_Stool* #  
    Todas las muestras de intestino
```

Descomprimir los datos:

```
$ gunzip *.gz
```

También se descarga un fichero con los metadatos (<http://www.biomedcentral.com/content/supplementary/gb-2014-15-7-r89-S18.csv>), que serán necesarios para identificar las muestras que corresponden a cada individuo y a cada día.

1.1 Comprobar calidad de los datos con fastQC y multiQC

FastQC es un software que permite hacer un control de calidad de forma sencilla. Con la opción “*” se detectan todos los archivos con formato “.fastq” del directorio actual.

```
$ fastqc *
```

El reporte de fastQC informa que los datos tienen la codificación de calidad Sanger/Illumina 1.9. Con el avance de las tecnologías de secuenciación, ha ido mejorando enormemente la calidad de las lecturas y cada plataforma tiene su propia codificación de calidad. El software tiene que ir cambiando de forma paralela a la tecnología, para que uno no avance sin el otro. Los programas que se utilizaron en este trabajo tienen en cuenta el tipo de codificación que hay en los datos para saber los valores reales de calidad que se tienen.

Como hay 820 archivos, se utiliza multiQC para hacer un resumen de los datos analizados por fastQC. Con la opción “.” se detectan los archivos necesarios:

```
$ multiqc .
```

2. Filtro de calidad utilizando seq_crumb 0.1.9

Software desarrollado en Valencia que incorpora utilidades para procesar secuencias.

```
$ cd ..
$ mkdir 02_Quality_filter
$ cd 02_Quality_filter
```

Como se trabaja con 820 ficheros, hay que lanzar varios trabajos a la vez. En este grupo de investigación existen cuatro servidores llamados SOM (1-4) y utilizan un sistema de colas para no colapsar las máquinas por todos los usuarios. Modo de uso:

clusterlauncher -N ejemplo -n 1 -s som1 /bin/sleep 240 (lanza 1 proceso secuencial al sistema de colas llamado “ejemplo” para ejecutarlo en som1)

Opciones:

- N : nombre del trabajo.
- n : número de procesos paralelos.
- s : servidor/es donde lanzar el trabajo.

Para filtrar por calidad se utiliza el *script filter_by_quality* de seq_crumb:

- q, --threshold : umbral de calidad.
- o, --outfile : nombre del fichero de salida.

```
$ for i in ../01_Data_fastq/seqs-S*.fastq;
do clusterlauncher -N ${i:22} -n 1 -s som1 filter_by_quality $i -q 30 -o
  q30-${i:22};
done
```

3. De .fastq a .fasta con qiime

```
$ cd ..
$ mkdir 03_Data_fasta
$ cd 03_Data_fasta
```

Qiime utiliza el formato fasta como entrada, así que incorpora un *script* para cambiar de formatos. Éste produce 2 ficheros: uno con secuencias (.fna) y otro con calidades (.qual).

Para pasar a formato fasta se utiliza *convert_fastaqual_fastq.py*:

- c, --conversion_type : tipo de conversión (*fastaqual_to_fastq* o *fastq_to_fastaqual*)
- f, --fasta_file_path : fichero de entrada (fasta o fastq)
- a, --ascii_increment : número a sumar (restar si se parte desde FASTQ) a la puntuación de calidad para obtener el carácter ASCII
- full_fasta_headers : para incluir las cabeceras de los archivos FASTA en el fichero de salida (en lugar de simplemente la etiqueta de secuencia).

```
$ for i in ../02_Quality_filter/q30-S*.fastq;
do clusterlauncher -N ${i:21} -n 1 -s som1 convert_fastaqual_fastq.py -c
  fastq_to_fastaqual -f $i -a 64 --full_fasta_headers ;
done
```

Nótese que se utiliza la opción “-a” porque los datos de partida tienen la codificación de calidad Sanger/Illumina 1.9.

4. Eliminar quimeras

```
$ cd ..
$ mkdir 04_Chimeras
$ cd 04_Chimeras
```

Las quimeras son artefactos producidos durante el proceso de PCR. Se trata de secuencias de ADN que contienen mezclas de otras secuencias. Qiime permite eliminarlas mediante dos pasos: primero identificar y luego limpiar las quimeras.

ANEXO I

1) Identificar quimeras con *identify_chimeric_seqs.py*:

- i, --input_fasta_fp : fichero de entrada en formato fasta.
- m, --chimera_detection_method : método de detección de quimeras (*blast_fragments*, *ChimeraSlayer* o *usearch61*).
- r, --reference_seqs_fp : ruta a la base de datos de referencia.
- o, --output_fp : nombre del directorio de salida.

```
$ for i in ../03_Data.fasta/q30_*.fna;
do
clusterlauncher -N ${i:17} -n 1 identify_chimeric_seqs.py -i $i -m
    usearch61 -r /software/databases/gg_13_8_otus/rep_set/97_otus.fasta -o
        chimeras_${i:17};
done
```

2) Limpiar secuencias quimera con *filter.fasta.py*:

- f, --input_fasta_fp : fichero de entrada en formato fasta.
- o, --output_fasta_fp : fichero de salida en formato fasta.
- s, --seq_id_fp : lista de identificadores de secuencias que deben retenerse.
- n, --negate : desecha los identificadores de secuencia pasados en lugar de mantenerlos.

```
$ for i in ../03_Data.fasta/q30_*.fna;
do
clusterlauncher -N non_chimeric_${i:17} -n 1 filter.fasta.py -f $i -o
    non_chimeric_${i:17} -s chimeras_${i:17}/chimeras.txt -n;
done
```

NOTA: es muy importante en este paso usar “-n” porque le dice al *script* que descarte todas las secuencias que le hemos pasado a través de la opción “-s”, es decir, que elimine las quimeras. Esto se hace así porque el *script* *identify_chimeric_seqs.py* ofrece una lista de las secuencias químéricas y no de las secuencias no químéricas.

4.1 Eliminar ficheros con bajo número de lecturas

Se hace un conteo del número de *reads* por fichero tras todo el preprocesado y se guarda en un fichero adicional:

```
$ for i in non_chimeric_q30_S*.fna; do r=$(awk '{s++}END{ print s/2}' $i) ;
    echo $r $i | column -t; done > numero_reads.txt
```

Se muestran las 25 primeras líneas del fichero ordenado:

```
1      non_chimeric_q30_Stool448.1259730.fastq.fna
2      non_chimeric_q30_Stool196.1259770.fastq.fna
3      non_chimeric_q30_Stool13.1259916.fastq.fna
4      non_chimeric_q30_Saliva267.1260193.fastq.fna
5      non_chimeric_q30_Stool217.1260272.fastq.fna
8      non_chimeric_q30_Stool185.1260354.fastq.fna
8      non_chimeric_q30_Stool163.1259769.fastq.fna
7
```

```

8 31      non_chimeric_q30_Stool120.1259849.fastq.fna
9 39      non_chimeric_q30_Stool147.1260039.fastq.fna
10 54     non_chimeric_q30_Stool136.1259652.fastq.fna
11 1006   non_chimeric_q30_Stool1453.1260253.fastq.fna
12 1423   non_chimeric_q30_Stool192.1259811.fastq.fna
13 1738   non_chimeric_q30_Stool1452.1259809.fastq.fna
14 2501   non_chimeric_q30_Stool1384.1259728.fastq.fna
15 2772   non_chimeric_q30_Stool1340.1260381.fastq.fna
16 3554   non_chimeric_q30_Stool14.1260013.fastq.fna
17 4026   non_chimeric_q30_Stool1343.1259705.fastq.fna
18 4493   non_chimeric_q30_Stool1454.1260333.fastq.fna
19 6395   non_chimeric_q30_Stool1382.1260123.fastq.fna
20 7462   non_chimeric_q30_Stool1345.1259808.fastq.fna
21 11248  non_chimeric_q30_Stool1455.1260157.fastq.fna
22 11459  non_chimeric_q30_Stool1372.1259688.fastq.fna
23 11676  non_chimeric_q30_Stool1326.1260401.fastq.fna
24 11859  non_chimeric_q30_Stool138.1260296.fastq.fna
25 13512  non_chimeric_q30_Stool1373.1259958.fastq.fna

```

Se eliminan manualmente los 20 ficheros que contienen menos de 10000 lecturas.

5. Seleccionar OTUs

```

$ cd ..
$ mkdir 05_OTUs
$ cd 05_OTUs

```

OTU son las siglas en inglés de *Operational Taxonomic Unit*. Es una unidad de clasificación para individualizar los taxones del estudio.

Para seleccionar OTUs se utiliza *pick_open_reference_otus.py*:

- i, --input_fps : fichero de secuencias de entrada.
- o, --output_dir : directorio de salida.
- p : parámetros para introducir en los distintos pasos del *script*.

En este caso se utiliza un fichero auxiliar llamado “uc_fast_params.txt” con los parámetros que tiene el siguiente contenido:

```

1 pick_otus:enable_rev_strand_match True

```

Esta instrucción sirve para que tenga en cuenta ambos sentidos de lectura de cada secuencia cuando haga el “pick otus”.

```

$ for i in ../04_Chimeras/non_chimeric_*.fna;
do
clusterlauncher -N otus_${i:15} -s som1 -n 1 pick_open_reference_otus.py -i
    $i -o "otus_"${i:15} -p uc_fast_params.txt;
done

```

6. Resumir taxones

```
$ cd ..
$ mkdir 06_Summarize_taxa
$ cd 06_Summarize_taxa
```

El *script summarize_taxa.py* proporciona información resumida de la representación de los grupos taxonómicos dentro de cada muestra:

- i, --otu_table_fp : fichero de entrada (tabla OTU que contiene información taxonómica).
- o, --output_dir : directorio de salida.
- L, --level : nivel taxonómico para el que se proporciona la información resumida. [Niveles:
1 (Reino), 2 (Fílum), 3 (Clase), 4 (Orden), 5 (Familia), 6 (Género), 7 (Especie)]
- suppress_biom_table_output : la tabla de taxones con formato BIOM no se creará en el directorio de salida.
- a, --absolute_abundance : para obtener como resultado una tabla con la abundancia absoluta de cada grupo taxonómico. Por defecto (si no se pone “-a”), este *script* utiliza abundancia relativa.

```
$ for i in ../05_OTUs/otus_non_chimeric_q30_*/otu_table_mc2_w_tax.biom;
do
clusterlauncher -N ${i:11:38} -n 1 summarize_taxa.py -i $i -o absolute_L6_$
{i:11:-25} -L 6 --suppress_biom_table_output -a;
done
```

7. Datos para complexCruncher

Agrupar las tablas de abundancia en 3 directorios según el tipo de muestra: donante A saliva, donante A intestino y donante B intestino. Para muestras de saliva es fácil porque se distinguen los ficheros por su nombre:

```
$ mkdir Saliva_DonorA
$ for i in *q30_Saliva*; do
  mv $i/otu_table_mc2_w_tax_L6.txt Saliva_DonorA/$i;
  rm -r $i;
done
```

Para muestras de intestino hay que separar entre sujeto A y B. Para ello se crea un fichero auxiliar que contiene los nombres de las tablas de abundancia del sujeto B. Ese fichero se obtiene a partir de los metadatos simplemente ordenando la tabla por la columna 18 “Description” y guardando los elementos de la primera columna que tengan la descripción “DonorB Stool”. Es necesario que el fichero auxiliar se encuentre en el directorio actual de trabajo.

ANEXO I

```
$ mkdir Stool_DonorB
$ for i in *q30_Stool*;
do
    for line in $(cat Stool_DonorB.txt);
    do if [ $line = ${i:34} ]; then
        mv $i/otu_table_mc2_w_tax_L6.txt Stool_DonorB/$i;
        rm -r $i;
    fi ;
done;
done
```

El resto de tablas que quedan son de intestino de Sujeto A. Así que se mueven a su carpeta con la misma estrategia utilizada en saliva:

```
$ mkdir Stool_DonorA
$ for i in *q30_Stool*; do
    mv $i/otu_table_mc2_w_tax_L6.txt Stool_DonorA/$i;
    rm -r $i;
done
```

En este punto se tienen 3 directorios con múltiples ficheros, cada uno de ellos es una tabla de abundancia para cada día. La idea es juntar todos esas tablas pequeñas en una tabla grande con entradas únicas que resuma todas las abundancias a lo largo del año.

Además, para pasos posteriores se necesita subdividir esa tabla grande final en subtablas que abarcan distintos intervalos de tiempo. Todo ello lo hace un *script* implementado en Python. Se muestra como ejemplo el *script* de saliva A:

```
1 import glob
2 import pandas as pd
3
4
5 # Fichero auxiliar con el nombre de los ficheros a añadir
6 sup_f = open("Saliva_samples.txt","r")
7 sup_f_text = sup_f.read().split("\n")
8 sup_f.close()
9
10 # Generar lista de tablas por día
11 list_of_dic = []
12 for i in sup_f_text:
13     try:
14         # Ficheros de abundancia absoluta:
15         file = "absolute_L6_" + i
16         f_in = open(file, "r")
17         text = f_in.read()
18         f_in.close()
19
20         # Quitar cabecera
21         sentences = text.split("\n")
22         sentences = "\t".join(sentences)
23         sentences = sentences.split("\t")
24
25         sample = sentences[2]
26         sentences = sentences[3:]
27
```

ANEXO I

```
28     # Crear un diccionario que sirve como tabla final
29     otu = []
30     abundance = []
31
32     for j in range(len(sentences)):
33         if j%2 == 0:
34             otu.append(sentences[j])
35         else:
36             abundance.append(float(sentences[j]))
37
38     raw_data = {
39         "otu_id": otu,
40         sample: abundance}
41
42     df_a = pd.DataFrame(raw_data, columns = ["otu_id", sample])
43     list_of_dic.append(df_a)
44 except:
45     print("No se encuentra el fichero" + str(i))
46
47 # Crear una tabla única final
48 f_table = list_of_dic[0]
49
50 for k in range(1, len(list_of_dic)):
51     f_table = pd.merge(f_table, list_of_dic[k], on="otu_id", how=
52                         "outer")
53
54 # Dividir tabla final en intervalos de tiempo y guardar cada
55 # subtabla en una hoja excel
56 df2 = f_table.iloc[:,1:40]
57 df2.insert(0, "otu_id", value=f_table.iloc[:,0])
58 df3 = f_table.iloc[:,40:75]
59 df3.insert(0, "otu_id", value=f_table.iloc[:,0])
60 df4 = f_table.iloc[:,75:203]
61 df4.insert(0, "otu_id", value=f_table.iloc[:,0])
62 df5 = f_table.iloc[:,203:]
63 df5.insert(0, "otu_id", value=f_table.iloc[:,0])
64
65 # Formatear salida
66 pd.set_option("expand_frame_repr", False)
67
68 # Escribir cada dataframe en una hoja Excel diferente
69 writer = pd.ExcelWriter('HostLifeStyle_SalivaA_absoluta.xlsx')
70 f_table.to_excel(writer, sheet_name="SalivaA", index = True,
71                  na_rep = 0)
72 df2.to_excel(writer, sheet_name="h_SalivaA_Day26to69", index =
73                  False, na_rep = 0)
74 df3.to_excel(writer, sheet_name="SalivaA_Day72to122", index =
75                  False, na_rep = 0)
76 df4.to_excel(writer, sheet_name="h_SalivaA_Day123to257", index =
77                  False, na_rep = 0)
78 df5.to_excel(writer, sheet_name="h_SalivaA_Day258to364", index =
79                  False, na_rep = 0)
80
81 # Cerrar el escritor Pandas Excel y guardar el fichero
82 writer.save()
```

Los *scripts* para intestino A e intestino B pueden encontrarse en el material suplementario o en la dirección <https://github.com/TeresaRubio/TFM/tree/a/Scripts> con los nombres “merge&split_stoolA.py” y “merge&split_stoolB.py”, respectivamente.

ANEXO I

Se ejecuta un *script* por directorio y se obtienen las 3 tablas (que pueden encontrarse en el material suplementario o en la dirección <https://github.com/TeresaRubio/TFM/tree/a/Tablas>:

```
$ cd Saliva_DonorA  
$ python merge&split_salivaA.py
```

```
$ cd ..  
$ cd Stool_DonorA  
$ python merge&split_stoolA.py
```

```
$ cd ..  
$ cd Stool_DonorB  
$ python merge&split_stoolB.py
```