

Dedicado a

...

Índice general

Índice de figuras	IV
Agradecimientos	IX
Resumen	XI
1. Introducción	1
2. Materiales y métodos	7
2.1. Preprocesado: FastQC, MultiQC y seq_crumbs	7
2.2. Clasificación taxonómica: QIIME	8
2.2.1. Selección de OTUs	8
Uclust	10
Greengenes	10
2.2.2. Asignación taxonómica	11
2.3. Análisis de variabilidad: complexCruncher	11
2.3.1. Regresión lineal y exponencial	11
2.3.2. Ley de potencia x -ponderada	12
2.3.3. Estandarización	12
2.3.4. RSI y medidas de variabilidad	13
2.4. Estudio de interacciones	14
2.4.1. Coeficiente de correlación de Pearson	14
2.5. Método de búsqueda de comportamientos	14
2.6. Aproximación para obtener interacciones: LIMITS	15

3. Resultados	17
3.1. Estado de los datos	17
3.2. Preprocesado	20
3.3. Clasificación taxonómica	22
3.4. Explorando series temporales	23
3.4.1. Abundancia de taxones	23
3.4.2. Ley de potencias	26
3.4.3. Clasificación por rango	31
3.5. Correlaciones	37
4. Discusión y perspectivas de futuro	43

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.	9
2.2.	Procedimiento de LIMITS.	16
3.1.	Control de calidad de los datos crudos	19
3.2.	Control de calidad tras el filtro de calidad	21
3.3.	Abundancia absoluta saliva A	24
3.4.	Abundancia absoluta intestino A	25
3.5.	Abundancia absoluta intestino B	25
3.6.	Ajuste a la ley de potencias x-ponderada.	28
3.7.	V y β en saliva	29
3.8.	V y β en intestino	30
3.9.	V y β resumen de saliva e intestino	31
3.10.	Matriz de rango: saliva A	33
3.11.	Matriz de rango: intestino A	35
3.12.	Matriz de rango: intestino B	36
3.13.	Correlaciones saliva A	38
3.14.	Abundancia relativa del grupo 2 en saliva	39
3.15.	Correlaciones intestino A	40
3.16.	Correlaciones intestino B	41
4.1.	Correlación Pearson vs. LIMITS en saliva A	44

Índice de tablas

2.1. Tabla de grupos de comportamiento.	15
3.1. Tabla de ficheros eliminados.	22
3.2. Resumen de subperiodos temporales.	27

Agradecimientos

¡Muchas gracias a todos!

Resumen

En este proyecto se analizaron datos del estudio David *et al.* [1] procedentes del microbioma de dos individuos. Se trata de datos de secuenciación de la región V4 del gen que codifica el ARN ribosomal 16S de los microorganismos presentes en intestino y saliva, tomados a lo largo de un año. Además, se recopilaron datos del estilo de vida de los donantes tales como dieta, ejercicio o enfermedad. Esto supone un total aproximado de 64,8 GB de datos (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 del microbioma en estas situaciones.

En primer lugar, se comprobó la calidad y longitud de la secuenciación y se hizo un filtro previo que redujo muy poco el set de datos. 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 busca 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 también un *ranking* de los microorganismos. Estos resultados fueron incorporados al artículo recientemente publicado de Martí *et al.* [11] como la serie temporal más larga analizada por los autores.

Por último, se calcularon las correlaciones de abundancia relativa 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 es 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 nuestro microbioma.

1 Introducción

A lo largo de la evolución, los microorganismos han convivido en simbiosis con el ser humano y son fundamentales para su salud. A este conjunto se le denomina **microbiota humana**. Se estima que el número de bacterias en nuestro cuerpo es aproximadamente del mismo orden que el número de células humanas [?]. Adquirimos los microorganismos a partir del nacimiento, durante el parto y la lactancia, y 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 radica en sus numerosas funciones: regulación de procesos digestivos, producción de sustancias bacteriostáticas y antibióticos naturales contra patógenos, entrenamiento del sistema inmune para reconocer invasores dañinos, etc. 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 cuya misión es generar recursos que permitan la caracterización del microbioma humano y el análisis de su papel en la salud y la enfermedad humana. Por tanto, el conjunto total de los genes de nuestra microbiota es lo que se ha denominado como microbioma.

La forma tradicional de estudiar los microorganismos de 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 las 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 en ocasiones en las que se quiere estudiar el genoma de un organismo cultivable concreto ya que es muy sencilla y barata. Sin embargo, no todos los microorganismos presentes en las muestras ambientales son cultivables. Se estima que sólo el 1 % de las bacterias del suelo y entre el 0,1 - 0,01 % de las bacterias marinas. A estos microorganismos se les denomina **no cultivables** y se debe al desconocimiento de los requisitos específicos del cultivo y a la existencia de grupos de microorganismos que deben mantenerse en equilibrio para sobrevivir.

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

INTRODUCCIÓN

de Tyson *et al.*, 2004 [2] y Venter *et al.*, 2004 [3]. 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 **ARN 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. Es un polirribonucleótido de unas 1542 pares de bases, codificado por el gen *rss* y se encuentra en la subunidad pequeña 30S del ribosoma, 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).

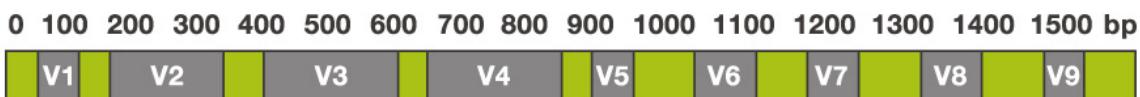


Figura 1.1: Representación esquemática de ARNr 16S en su estructura primaria. En verde se colorean las regiones conservadas (C), que son universales, y en gris las regiones variables (V), que permiten agrupar específicamente a nivel taxonómico.

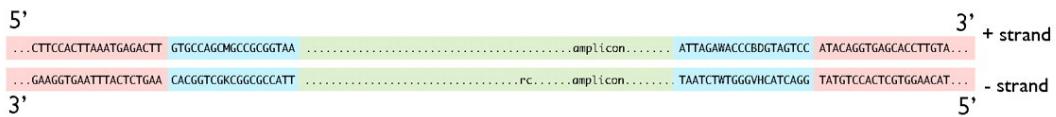
INTRODUCCIÓN

Los pasos en un **proyecto metagenómico** son:

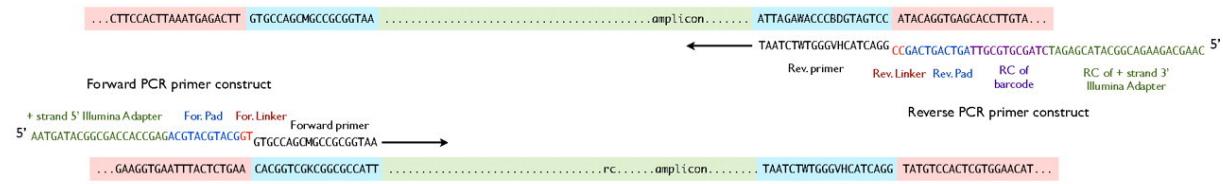
1. Extracción del ADN. Se pueden utilizar distintos kits comerciales cuyo fundamento se basa en la lisis de las células mediante compuestos químicos (EDTA, lisozima, detergentes...), seguida de la eliminación de los componentes celulares y finalmente se precipita el material genético para obtener ADN puro en alta concentración.
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.
 - *Barcoding* del ADN para multiplexar el ensayo. Este método permite secuenciar varias muestras en el mismo *run* de secuenciación, mejorando así el costo-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 Figura 1.2. Éste se utiliza para el siguiente paso.
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 que se pueden 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).

INTRODUCCIÓN

Target gene:



Amplification primers with annealing sites:



Amplification products:

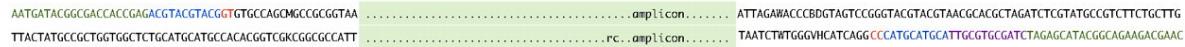


Figura 1.2: Protocolo de secuenciación Illumina con *barcode*. En la primera parte, el gen diana es la región V4 del gen que codifica el ARNr 16S (se colorean de azul las regiones conservadas y en verde la región adecuada para la clasificación taxonómica). En la segunda parte, se procede a la amplificación 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”. En la tercera parte, se observa el 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 tamaño 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 consegui-

INTRODUCCIÓN

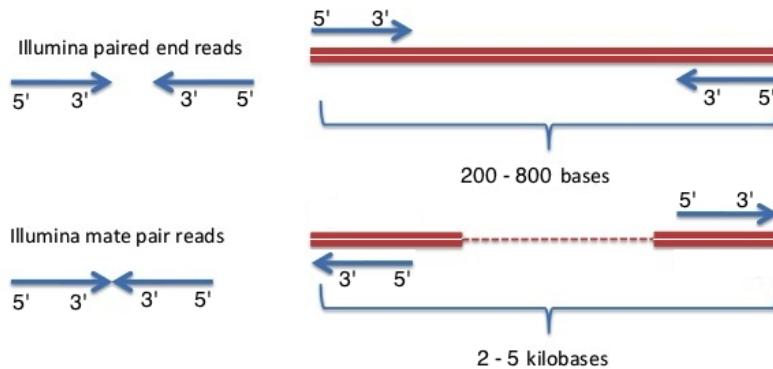


Figura 1.3: *Paired end vs. Mate pair.* El rectángulo rojo simboliza un fragmento de ADN y las flechas azules indican la dirección de las lecturas que se obtienen tras la secuenciación. La parte superior de la figura esquematiza la estrategia *paired end* que permite un rango de inserto de 200-800 pb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

do alargar la longitud de secuencia obtenida, lo que facilita el ensamblaje. Se predice que la plataforma MinION (basada en nanoporos) supondrá una revolución porque incluye las características mencionadas anteriormente y, además, reduce el coste de reactivos y aparatos al tratarse de una técnica no óptica. En julio de 2016 esta tecnología la utilizó 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.

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 [5] basada en la relación entre la media (x) y la varianza (σ) en la abundancia de poblaciones naturales: $\sigma = V \cdot x^\beta$. Donde la constante V indica amplitud de fluctuación y β el tipo de distribución. En 1990, McArdle *et al.* [4] 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 *hotspots* (puntos de crecimiento poblacional súbito) o, comparar poblaciones en base a su ecología sirve

INTRODUCCIÓN

para entender por qué diferentes especies pueden tener dinámicas poblacionales similares.

Dentro del mundo microbiano se han descrito distintas **relaciones entre poblaciones**. Las interacciones entre poblaciones pueden ser positivas, si permiten ocupar nuevos nichos, o negativas, cuando se eliminan las poblaciones poco adaptadas o se protegen de la llegada de especies intrusas. Las relaciones biológicas son muy variadas y pueden generar interdependencias de muy diversa importancia.

- Neutralismo: dos poblaciones se encuentran simultáneamente en el ambiente sin que exista relación entre ellas. Resultado 0/0.
- Comensalismo: la primera población modifica el ambiente y favorece el crecimiento de la segunda. Resultado +/0.
- Sinergismo o protocolooperación: dos comunidades se favorecen mutuamente de forma no obligatoria. Resultado +/++.
- Mutualismo o simbiosis: dos comunidades 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 todas las poblaciones. 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 +/-.

En este estudio se analizan los datos del estudio “Host lifestyle affects human microbiota on daily timescales” de David *et al.* [1] donde aparece una serie temporal de 820 puntos temporales pertenecientes a microbioma intestinal y bucal de dos sujetos, depositada en un repositorio público. La primera parte de este 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.

2 Materiales y métodos

Para analizar la serie temporal anual en la que se centra este estudio, se llevaron a cabo una serie de pasos en el siguiente orden: preprocessado de las secuencias, 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_crumb

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. Permite 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. Así se puede detectar rápidamente cualquier problema que hay que tener en cuenta antes de realizar análisis posteriores.

MultiQC (<http://multiqc.info/>) es una herramienta de código abierto, implementada en Python que da soporte a muchas herramientas bioinformáticas, entre ellas FastQC. Produce un reporte HTML muy parecido pero permite un análisis a lo largo de varias muestras. La visualización de las muestras en conjunto permite realizar comparaciones y también recopila estadísticas numéricas de cada muestra para ver cómo se comportan los datos.

seq_crumb (https://bioinf.comav.upv.es/seq_crumb/) es un software de código abierto implementado en Python que utiliza Biopython e 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 de baja calidad en los extremos (*trimming*), conversión de formatos, etc.

FastQC y MultiQC fueron utilizados por la gran cantidad de información que producen y sus vistosos gráficos. seq_crumb se eligió porque incluye un script específico para filtrar por calidad media bastante fácil de usar y rápido.

2.2. Clasificación taxonómica: QIIME

QIIME (<http://qiime.org/index.html>) [6] 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, aunque no fue utilizado en este trabajo, que realiza el filtrado de *reads* por calidad, longitud y el demultiplexado simultáneamente a partir de un fichero “mapa” con los metadatos. En este proyecto se ha utilizado 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 tres pasos adicionales con sus respectivas visualizaciones que tampoco fueron utilizados: creación de árboles filogenéticos, estudio de diversidad α y β y método de rarefacción.

Existen múltiples herramientas para realizar una clasificación taxonómica pero se eligió QIIME para reproducir y corroborar los resultados obtenidos por los autores de donde se obtuvieron los datos [1]. Se obtienen diferencias taxonómicas a la hora de elegir una herramienta u otra, pero no son muy significativas y ambos métodos son 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 morfológicos. Puede considerarse OTU un individuo, una población, una especie o cualquier otro taxón. QIIME ofrece tres estrategias de selección diferentes para este paso:

- **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*.

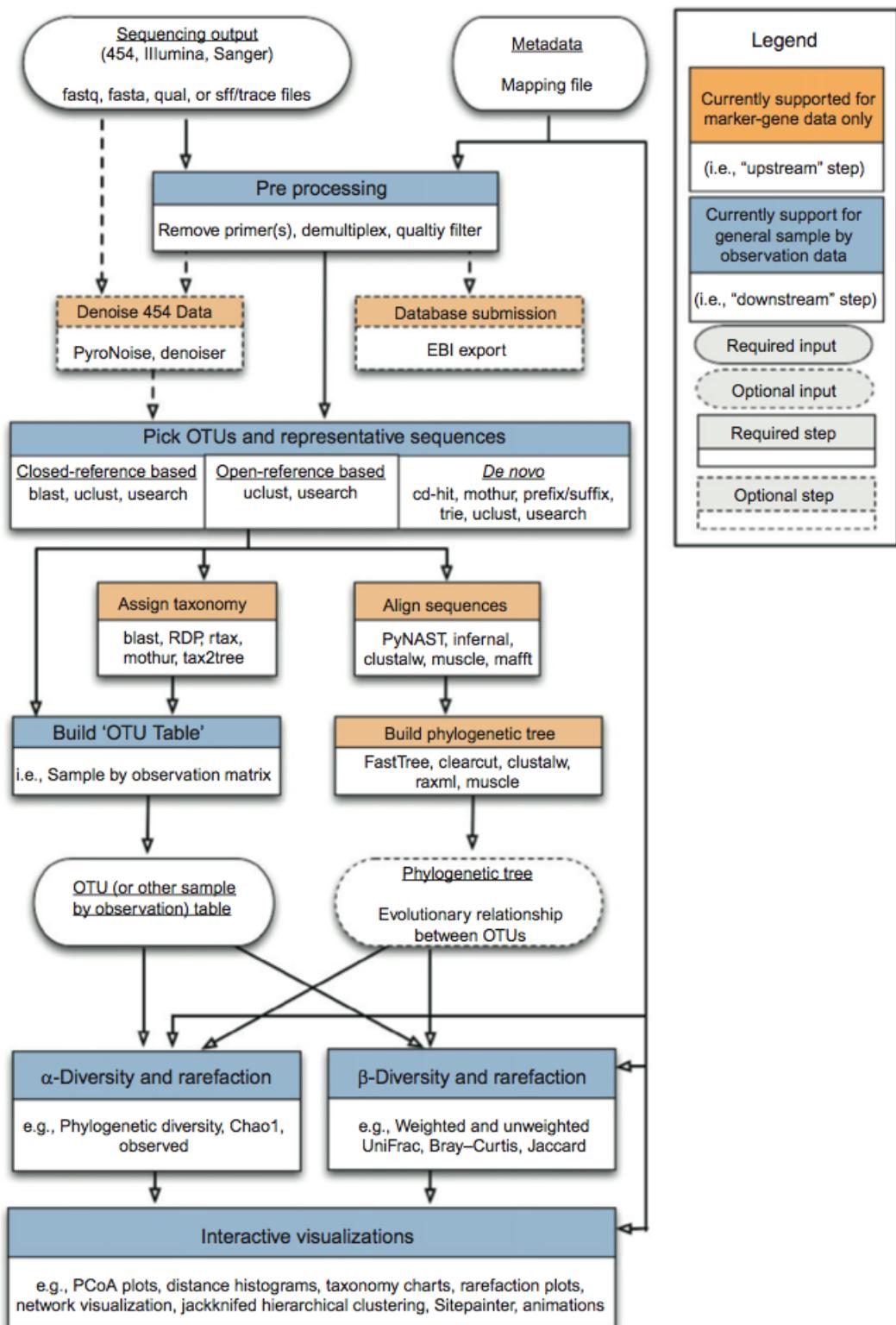


Figura 2.1: Flujo de trabajo de QIIME. Se representan esquemáticamente todas las opciones de QIIME. Cada una de ellas las realiza un *script* diferente, de tal manera que puede iniciarse el análisis en cualquier punto del flujo de trabajo.

2.2. CLASIFICACIÓN TAXONÓMICA: QIIME

- **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 son 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*.

Se seleccionó *Closed-reference* y 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.

Uclust (http://drive5.com/usearch/manual/uclust_algo.html) Fue el método de agrupación utilizado. Es un algoritmo diseñado para agrupar secuencias de nucleótidos o aminoácidos en base a su similitud [7]. 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 con dos centroides diferentes con similitud $> T$. Idealmente, se asignará al centroide más cercano, pero puede haber dos o más a la misma distancia, en cuyo caso la asignación de *cluster* es ambigua y debe hacerse una elección arbitraria. 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 general óptima, 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.

Greengenes (<http://greengenes.secondgenome.com>) Es la base de datos que se eligió como referencia. 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*

2.2.2. Asignación taxonómica

Una vez se ha creado 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 dependerá del formato que se devuelva desde el paso de la asignación de taxonomía (1,2,3,4,5,6 o 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 [8]. Acepta como *input* ficheros excel (en formato xlsx) 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 genera todos los resultados, 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 los datos se ajustan a una recta de regresión, conocida también como recta de mejor ajuste o recta de mínimos cuadrados, que es aquella que mejor se ajusta a los datos lo más estrechamente posible. La idea es que los valores de la recta estén lo más cerca posible de los valores observados pero siempre quedan distancias entre los puntos y la recta, que son los denominados errores residuales. Para realizar el ajuste, se suman todos los cuadrados de los errores residuales para obtener un solo error que se llama suma de los errores al cuadrado (SSE – siglas en inglés de “Sum of Squares Error”) y se elige la recta con el valor más pequeño SSE.

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, se tuvo en cuenta que cada media tiene incertidumbre y se puede estimar, para un tamaño de muestra n , por el SEM (error estándar de la media). En este caso, las incertidumbres afectan a la variable independiente, por lo que el ajuste no fue tan trivial como un ajuste y-ponderado, donde las incertidumbres afectan a la variable dependiente. Un método estándar para realizar este ajuste es (i) invertir las variables antes de aplicar los pesos, (ii) realizar el ajuste ponderado, y (iii) revertir la inversión. Este método es determinista, pero la solución aproximada empeora con coeficientes de determinación más pequeños. Para superar esa limitación, se desarrolló un método estocástico con una estrategia de tipo bootstrap que evitó la inversión y es aplicable independientemente del coeficiente de determinación.

La idea básica de bootstrap es que la inferencia sobre una población a partir de datos de muestra puede ser modelada mediante un nuevo muestreo de los datos de la muestra y realizando la inferencia sobre una muestra a partir de datos remuestreados. Para adaptar esta idea general al problema aquí descrito, se realizan múltiples replicaciones donde se re-muestrea 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

Se pueden 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:

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

2.3. ANÁLISIS DE VARIABILIDAD: COMPLEXCRUNCHER

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 ($\widehat{\sigma}_V$) es:

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

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

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

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 aparece una barra midiendo 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:

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

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 eligió arbitrariamente para aumentar la resolución en la región estable.

Finalmente, debajo de estas mismas matrices de rango, hay un gráfico con dos medidas relevantes para la variabilidad del rango a lo largo del tiempo. Por un lado, RV 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, DV 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 que tenía en el tiempo anterior.

2.4. Estudio de interacciones

2.4.1. Coeficiente de correlación de Pearson

En muchos trabajos se han utilizado 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 ha utilizado en los datos de abundancia. 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).

2.5. 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 desarrolló 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 calculó 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 es lo 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. A continuación se especifican los 7 grupos que se han considerado:

Grupo	Valores	Comportamiento
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 tras 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 y la última columna describe el comportamiento del grupo.

2.6. Aproximación para obtener interacciones: LIMITS

Trabajos más actuales como el de *Fisher et al.* [9], han cuestionado que las correlaciones midan interacciones reales y, además, han desarrollado un nuevo método que se ha aplicado a los datos para comparar con 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 tiempo discreto (dLV) en la dinámica de los microorganismos.

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.

El algoritmo LIMITS fue implementado en Mathematica (Wolfram Research, Inc.) y es de código abierto. Intenta inferir la matriz de interacciones entre microorganismos

2.6. APROXIMACIÓN PARA OBTENER INTERACCIONES: LIMITS

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.

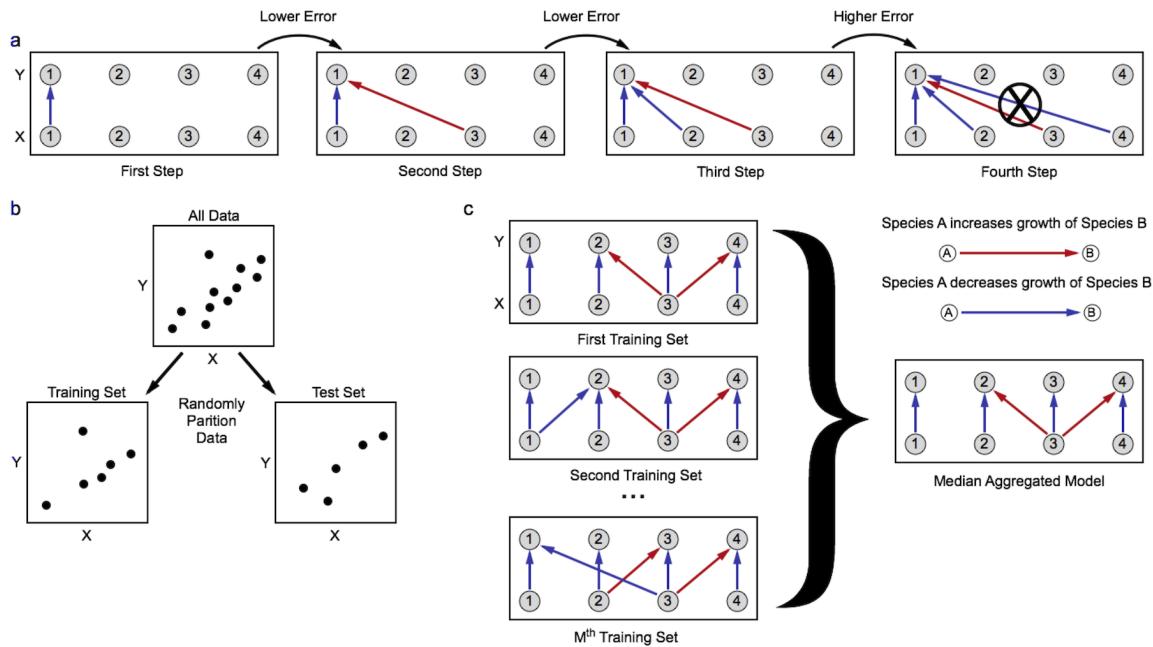


Figura 2.2: Procedimiento de LIMITS. a) Utiliza una regresión por pasos, donde las interacciones se añaden secuencialmente a la regresión si su inclusión reduce el error de predicción por debajo de un umbral predefinido. b) El error se evalúa por agregación bootstrap, que consiste en dividir los datos 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.

Comprenden 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 hay que destacar que en todos los grupos hay 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 tomaron metadatos sobre el estilo de vida mediante una aplicación iOS que utiliza una base de datos SQL donde los sujetos 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 dan dos escenarios de cambio en el ambiente del microbioma debido a que el individuo A hizo 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. El primer obstáculo con los datos se encuentra aquí, ya que en el repositorio no hay dos archivos por muestra como suele ocurrir cuando se trabaja con *paired end* (Materiales y métodos). Existen 820 archivos únicos (uno por día) con *reads* de longitud ≤ 100 pb. Se desconoce el procedimiento llevado a cabo por los autores, aunque puede hipotetizarse que o bien utilizaron *single end* pero han cometido una errata al describir la forma de secuenciación, o bien 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 son necesarias secuencias no apareadas (únicas) para posteriores análisis así que se da por hecho que las secuencias de los ficheros descargados del repositorio son *single* ya que 100 pb son suficientes para una resolución biológicamente significativa si se eligen juiciosamente los cebadores [10].

Se han encontrado otras incidencias en los datos como la falta de metadatos para la muestra “Stool69.1260101.fastq” y la existencia dos muestras para el mismo día (concretamente los días 79, 127, 128, 231, 238, 275, 277, 284 en saliva sujeto A; los días 7, 44, 74, 79, 82, 84, 106, 120, 162, 277 en intestino sujeto A y el día 177 en intestino 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 es muy tedioso ir inspeccionando uno por uno, se utilizó MultiQC para obtener un fichero resumen de todos ellos. El resultado puede observarse en la figura 3.1. El primer plot muestra la calidad medida con *phred score (q)* a lo largo de las bases en las 820 muestras. En general la calidad es buena pues casi todas las bases (a excepción de dos) presentan valores superiores a 20. La calidad tiende a bajar un poco en los extremos de las secuencias, fenómeno que suele darse en secuenciación frecuentemente. 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 tienen calidad q=35 y hay muy pocas de “mala calidad” ($q < 30$) en su extremo izquierdo.

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 es necesario un preprocesado. Del secuenciador se obtiene el fragmento de ADN mencionados en el apartado 1.1.2 – *barcoding* (figura 1.2). A esto se le denomina “multiplex” y a la acción de procesarlo se le denomina “demultiplexar”. Además, hemos visto que la plataforma de secuenciación no es perfecta y en ocasiones se obtienen calidades no deseadas.

En este caso no es necesario realizar un demultiplexado ya que los autores han realizado este paso previamente y los datos que se encuentran en el repositorio ya están libres de adaptadores. Se comprobó buscando la secuencia del cebador de PCR directo (GTGC-CAGCMGCCGCGGTAA) 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 deduce que ya fueron eliminados.

En general las calidades de secuenciación eran buenas como se ha visto en el apartado anterior, pero aún pueden eliminarse algunas secuencias que tenían peor calidad. El valor q al cual filtrar es arbitrario, 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 pierden pocas lecturas y se obtiene una buena calidad: habían 211.731.053 *reads* de partida y tras el filtrado quedan 208.266.760 *reads*, así que se ha eliminado el 1,636 % de las lecturas al filtrar. Las secuencias fueron filtradas con seq_crumps para eliminar todas aquellas con calidad media $q < 30$. Los resultados pueden observarse en la figura 3.2. Como se ha filtrado 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 ve muy claro que el programa ha eliminado todas las *reads* inferiores a 30, perdiendo esa cola que rozaba la franja sombreada en naranja.

Otros elementos que afectan a la calidad son 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 208.266.760 *reads* y tras este paso se obtienen 206.928.490 *reads*, por lo que el 0,64 % de las secuencias eran quimeras.

El número de lecturas que se obtienen 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 salió bien en la secuenciación ese día. Por tanto, se ha realizado un paso más de preprocesado de los datos, eliminando aquellos ficheros que contenían un número de *reads* inferior a 10.000. Este valor también es arbitrario en base al compromiso cantidad-calidad de información, esto es, si elimino muchas secuencias me quedaré sin información pero si incluyo los ficheros con



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.

3.3. CLASIFICACIÓN TAXONÓMICA

pocas *reads* estaré introduciendo errores a mi análisis. Los días eliminados del estudio aparecen detallados en la tabla 3.1. De 208.266.760 *reads* que llegan sin quimeras, ahora obtenemos 208.231.302 *reads* totales con las que se va a realizar todo el análisis.

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	2746	Sujeto A
Stool4.1260013	3553	Sujeto A
Stool343.1259705	4004	Sujeto A
Stool454.1260333	4491	Sujeto B
Stool382.1260123	6395	Sujeto B
Stool345.1259808	7420	Sujeto A
TOTAL	35458	

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 quedan a nivel de familia, orden o incluso clase. En estos casos, la nomenclatura adoptada por QIIME es “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_*”.

3.4. EXPLORANDO SERIES TEMPORALES

Luego está el grupo “Unassigned” donde se guardan todas aquellas secuencias que no ha sido capaz de clasificar taxonómicamente a ningún nivel. Adquiere la siguiente nomenclatura: “***Unassigned;Other;Other;Other;Other;Other;Other***”

En este procedimiento se parte de un gran número de ficheros que contienen secuencias de ADN tomadas a lo largo de un año y se obtiene una gran tabla que resume la abundancia absoluta de OTUs (filas) que había cada uno de los días de ese año (columnas). Se genera 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 queda detallado en el *pipeline* del anexo 1, 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 requiere un excel por individuo con una tabla donde aparezcan todos días como columnas adyacentes).

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 generadas las tablas de abundancia absoluta en el apartado anterior, se representa mediante un histograma la abundancia total cada día para dar una idea global de los datos. En la muestra de saliva (figura 3.3) existe, en general, una abundancia alta a excepción de algunos días donde se aprecia un claro descenso. En el caso de intestino A (figura 3.4) hay más días de muestra, lo que dificulta un poco su visualización pero se aprecia que hay muchos días con abundancias muy bajas. Por último, la muestra de intestino B (figura 3.5) tiene abundancias elevadas durante los primeros días pero a partir de aproximadamente el día 20, casi todas las abundancias son inferiores. Puede apreciarse que en ninguno de los tres casos hay abundancias inferiores a 10.000 por el filtro realizado durante el preprocesado.

3.4. EXPLORANDO SERIES TEMPORALES

Los conteos absolutos están llenos de errores sistemáticos que se deben tanto al proceso de secuenciación como a la asignación taxonómica. Los cambios en abundancia estarían enmascarados por esos errores, así que se trabaja con abundancia relativa de los taxones para ver la variabilidad temporal. ComplexCruncher realiza internamente el cambio de abundancia absoluta a abundancia relativa y realiza todos los cálculos posteriores con estos valores.

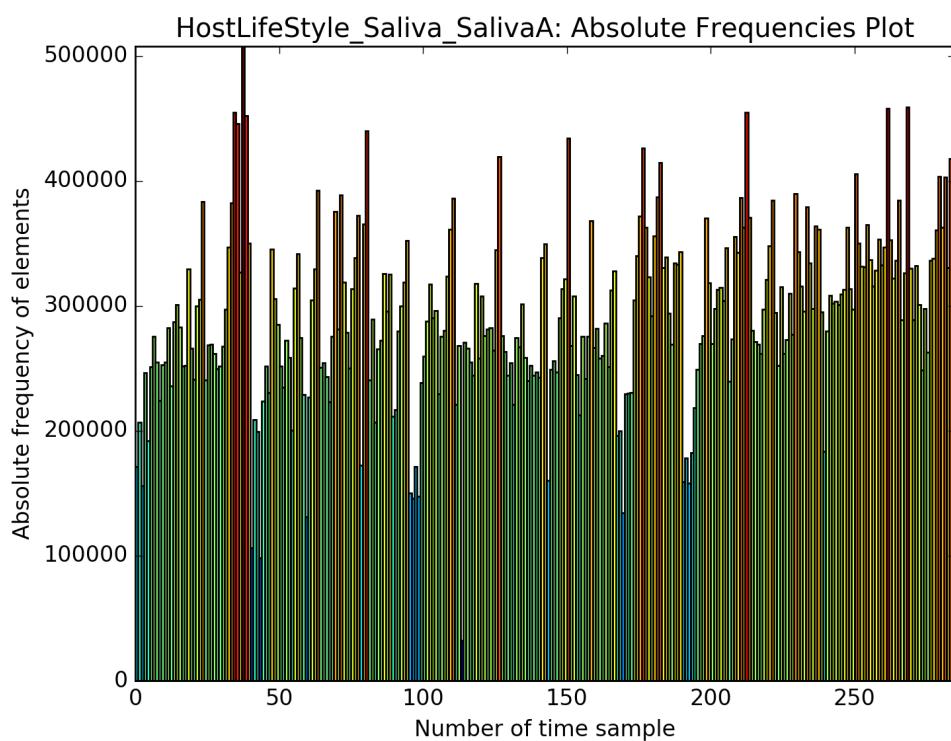


Figura 3.3: Histograma de la muestra saliva A que representa la abundancia total de los géneros en frecuencia absoluta a lo largo del tiempo (285 días).

Los colores altos indican altas abundancias y los colores fríos, bajas abundancias.

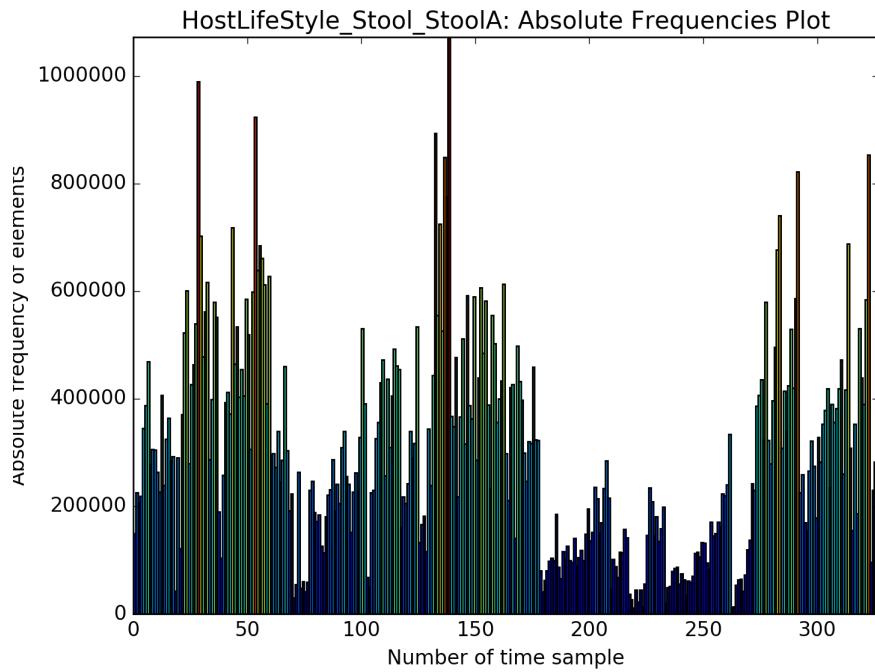


Figura 3.4: Histograma de la muestra intestino A que representa la abundancia total de los géneros en frecuencia absoluta a lo largo del tiempo (329 días). Los colores altos indican altas abundancias y los colores fríos, bajas abundancias.

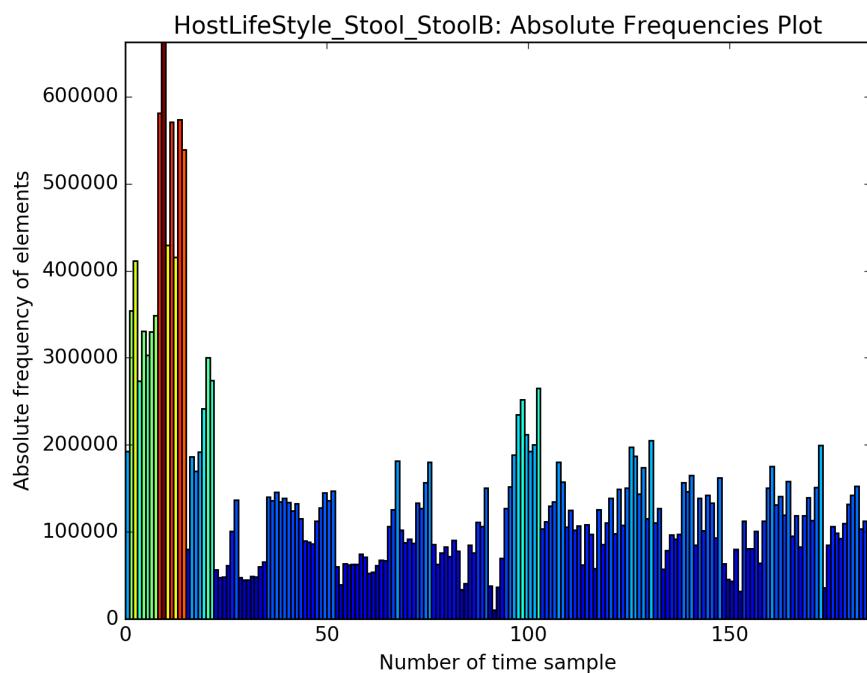


Figura 3.5: Histograma de la muestra intestino B que representa la abundancia total de los géneros en frecuencia absoluta a lo largo del tiempo (186 días). Los colores altos indican altas abundancias y los colores fríos, bajas abundancias.

3.4.2. Ley de potencias

Con las tablas de abundancia relativa, complexCruncher comprueba si los datos se ajustan a un modelo lineal, no lineal o mixto (Materiales y métodos). En este estudio se encuentra que las fluctuaciones de abundancia relativa en los taxones siguen la ley de potencias de Taylor en todos los casos como se muestra en la figura 3.6. Se representa el ajuste exponencial de los datos representado en escala logarítmica para facilitar la visualización. El ajuste es robusto porque todos los casos presentan un coeficiente de determinación alto ($R^2 > 0.9$). Dentro de la ecuación, V corresponde a la ordenada en el origen y β a la pendiente. Estos dos parámetros están relacionados con la estabilidad del sistema, es decir, describen 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 obtiene siempre una $\beta < 1$, lo cual indica que los taxones dominantes son menos susceptibles a las perturbaciones que el resto. Por otro lado, V representa la máxima amplitud de las fluctuaciones, esto es, la variación máxima teórica correspondiente a un género hipotético de abundancia relativa 1. 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 es menor en saliva que en intestino.

Los parámetros de Taylor, V y β , están relacionados con el estado de salud del hospedador [11]. 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 hemos visto los valores de V y β generales de los sujetos de este estudio pero ahora hay que analizarlos aprovechando las perturbaciones que causan el viaje y la infección de los individuos. Es un hecho empírico que los sujetos están enfermos durante el viaje y la infección por lo que la variabilidad temporal en el microbioma será mayor durante las perturbaciones pero, ¿cómo se encuentra el microbioma una vez que han pasado estos sucesos? Nuestra hipótesis es que se recupera la variabilidad inicial tras la perturbación. ComplexCruncher está preparado para hacer el cálculo de V y β en todos estos casos, ya que permite introducir un fichero excel por individuo con varias hojas: la primera hoja puede ser el periodo anual completo, la segunda hoja puede incluir el periodo previo a la perturbación, la tercera el periodo de la perturbación y la cuarta el periodo tras la perturbación. El programa distingue entre periodos sanos y de perturbación cuando el usuario se lo indica. Este paso se realizó aprovechando el mismo *script* en Python que formateaba los datos para pasar de QIIME a complexCruncher. El *script* genera las subtablas y produce 3 ficheros excel para introducir a complexCruncher:

3.4. EXPLORANDO SERIES TEMPORALES

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 muestran 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 nos 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 ha producido un aumento de la variabilidad en el microbioma del individuo pero al regresar a su rutina habitual, se recuperan valores similares a los iniciales.

El resultado obtenido para intestino es 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 son los valores generales que se vieron en la figura 3.6B y 3.6C. Si nos centramos en el sujeto A, vemos en violeta el periodo anterior al viaje, en amarillo el periodo del viaje y en negro y rojo el periodo tras el viaje. Se observa, al igual que en saliva, que los valores aumentan mucho durante el viaje y se recuperan casi como al estado inicial a la vuelta. 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 son mayores durante la infección y se recuperan hacia valores similares al estado inicial (aunque de forma más dispersa que en el donante A). Se corrobora así la hipótesis de que se recupera un estado similar, aunque no igual, al de partida tras una perturbación.

3.4. EXPLORANDO SERIES TEMPORALES

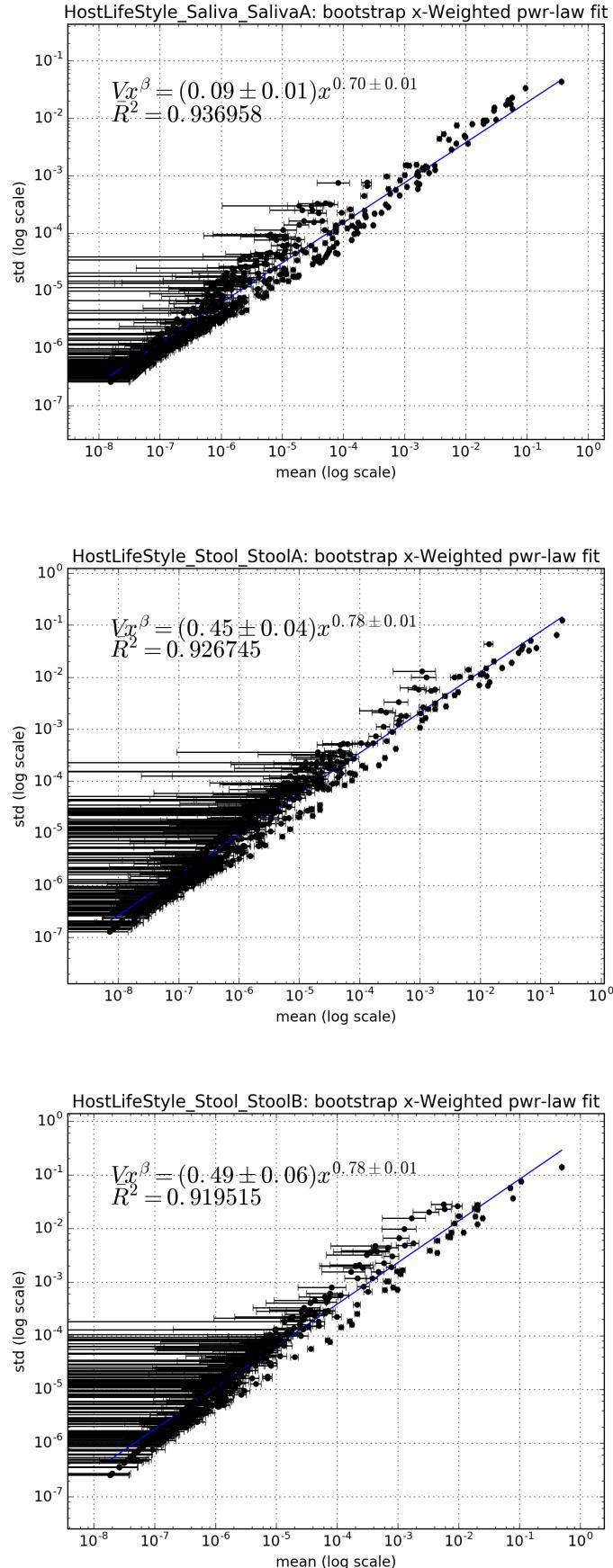


Figura 3.6: Ley de potencias x-ponderada de la desviación estándar (SD) frente a la media de los valores de cada género monitorizados a lo largo del tiempo. El primer ajuste corresponde a la muestra de saliva A, el segundo al intestino A y el tercero al intestino B. Y corresponde a la intersección con el eje y, y β corresponde a la pendiente de la recta. Las barras de error corresponden al error estándar de la media.

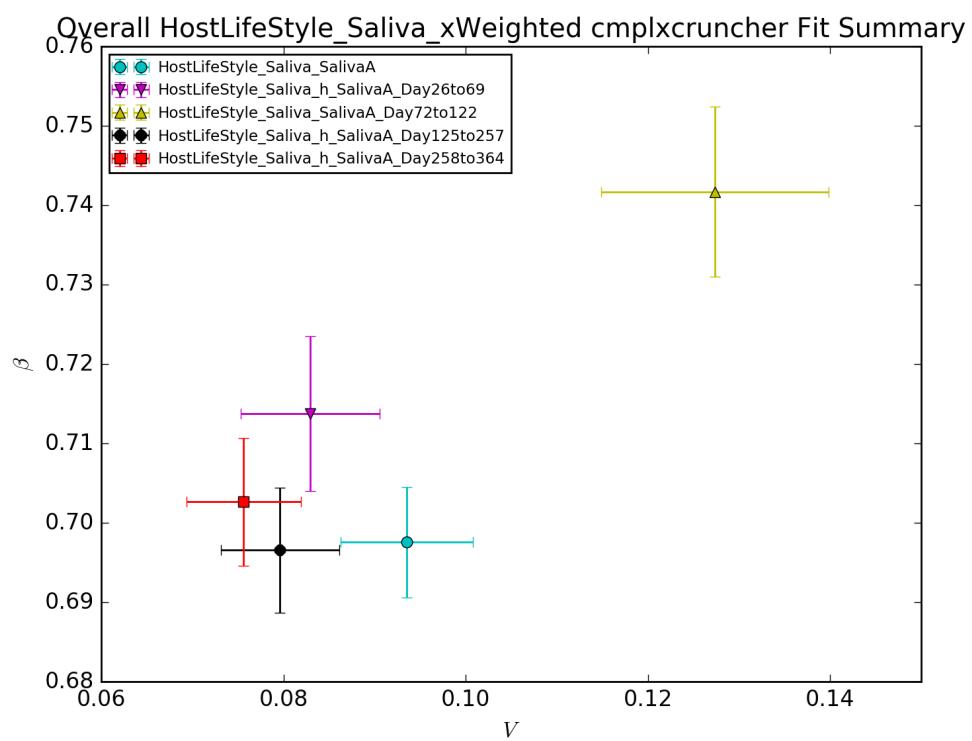


Figura 3.7: Se representan los parámetros de Taylor, V y β , correspondientes a muestras de saliva en distintos períodos: durante todo el año, antes del viaje, durante el viaje y tras el viaje (dividido en dos subperiodos). Los errores fueron calculados por el método bootstrap.

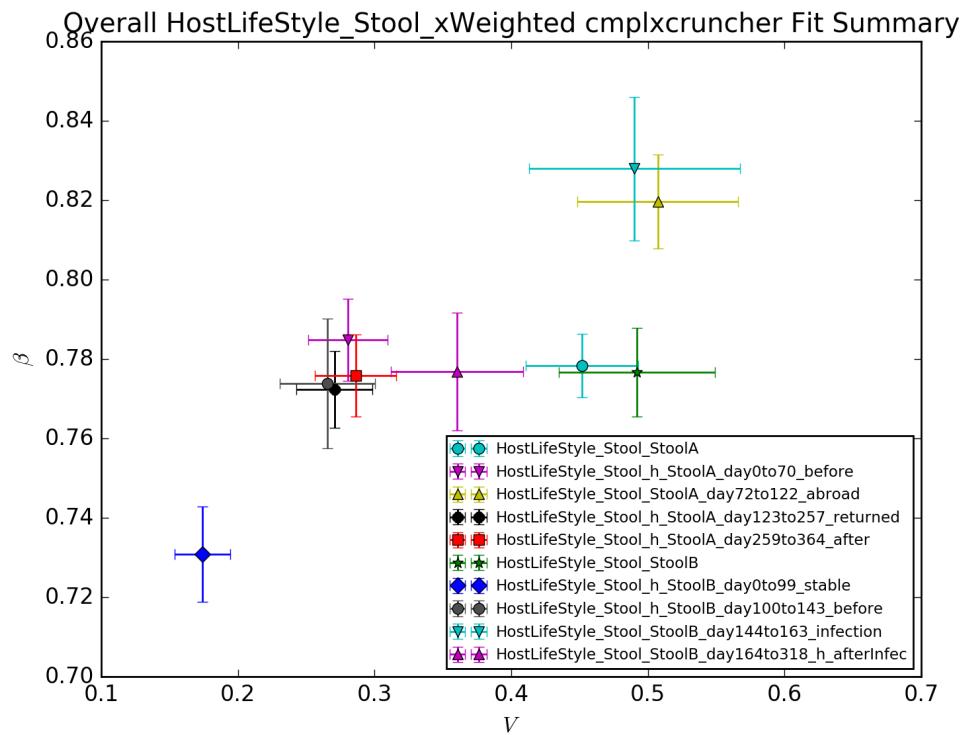


Figura 3.8: Se representan los parámetros de Taylor, V y β , correspondientes a muestras de intestino en distintos períodos. Para el sujeto A: durante todo el año, antes del viaje, durante el viaje y tras el viaje (dividido en dos subperiodos). Para el sujeto B: durante todo el año, antes de la infección (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 genera un plot resumen para comparar entre muestras. Hemos visto que tanto en saliva como en intestino aumentan los valores de V y β al producirse una alteración, pero los ejes presentan escalas diferentes. Para visualizarlas conjuntamente se normalizan 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 [?]). Así, se define un área dentro de la cual quedan los puntos correspondientes a los períodos sanos (antes y después de la perturbación). Quedan fuera de este área los puntos correspondientes al periodo del viaje y a la infección además de las dos series anuales completas.

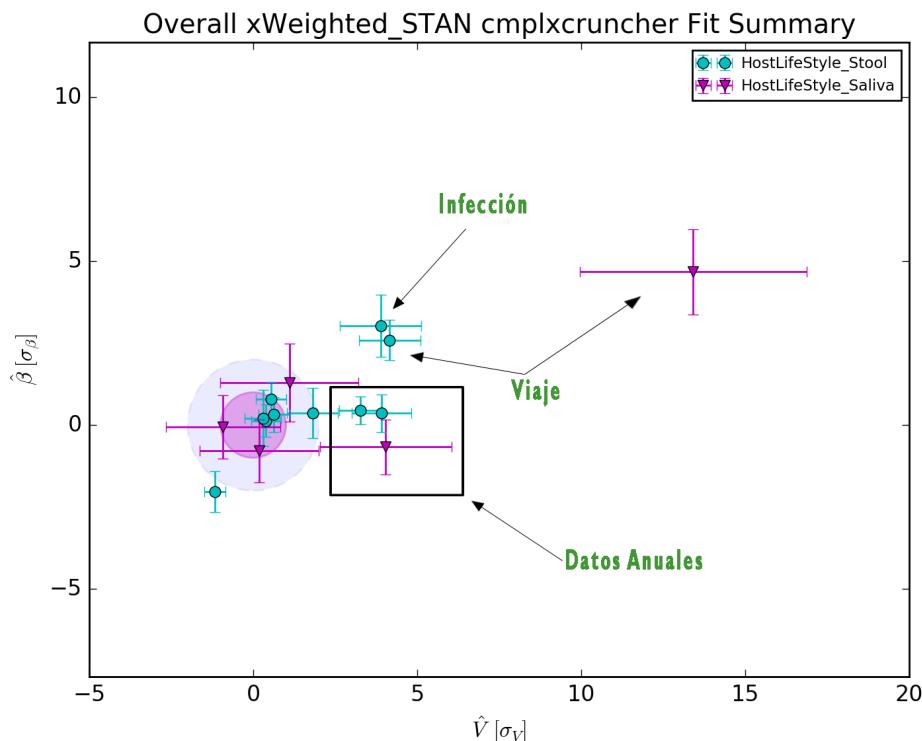


Figura 3.9: Se representan conjuntamente los parámetros de Taylor, V y β , correspondientes a muestras de intestino (azul) y saliva en distintos períodos (violeta). El área sombreada en rosa corresponde a la zona sana definida con la estandarización. El resto de puntos incluyen las perturbaciones y se 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 del microbioma humano, 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. En esta matriz se recogen 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 representan 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 se corresponde con el número 1 y violeta oscuro con el número 573 (que es el último taxón). En general, se aprecia que los géneros más abundantes suelen ser los más estables. El género más abundante es *Streptococcus* y ocupa la primera posición en abundancia a lo largo de todos los días del año (ningún género lo supera en abundancia nunca). Otro género que llama mucho la atención es *Chryseobacterium*, ya que los primeros días del año es muy poco abundante y a partir del día 50 aproximadamente, aumenta su abundancia. Luego hay otros patrones intermitentes, que aparecen y desaparecen 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 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 es 100 (máximo RSI) y violeta oscuro es 82.4 (mínimo RSI). En el primer tercio de los 50 taxones, se observan valores de RSI elevados remarcando que los taxones más abundantes tienen más estabilidad. Sin embargo, en ocasiones se encuentran 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 mantienen estables en su rango a lo largo del tiempo. Algunos ejemplos en la muestra de saliva son *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 (siglas en inglés de *Rank Variability*) respecto al rango global y DV (siglas en inglés de *Difference variability*) respecto al rango del día anterior (detallado en Materiales y métodos). En la figura 3.10 se muestra que hay un pequeño aumento en ambas medidas de variabilidad durante los días 40-75 que se corresponden a los días que el sujeto estuvo de viaje (nótese que el periodo de viaje comprende los días 71-122 pero se corresponde al intervalo 40-75 en este gráfico porque se representan los 285 días que hubo muestra consecutivamente, obviando aquellos días en los que no hubo muestra). Especialmente se ven dos picos máximos de DV en los días 60 y 80 aproximadamente. Esta variación en la variabilidad es muy pequeña para considerarse significativa, por lo cual se

3.4. EXPLORANDO SERIES TEMPORALES

deduce que el viaje no ocasionó demasiado cambio en la variabilidad del microbioma de saliva.

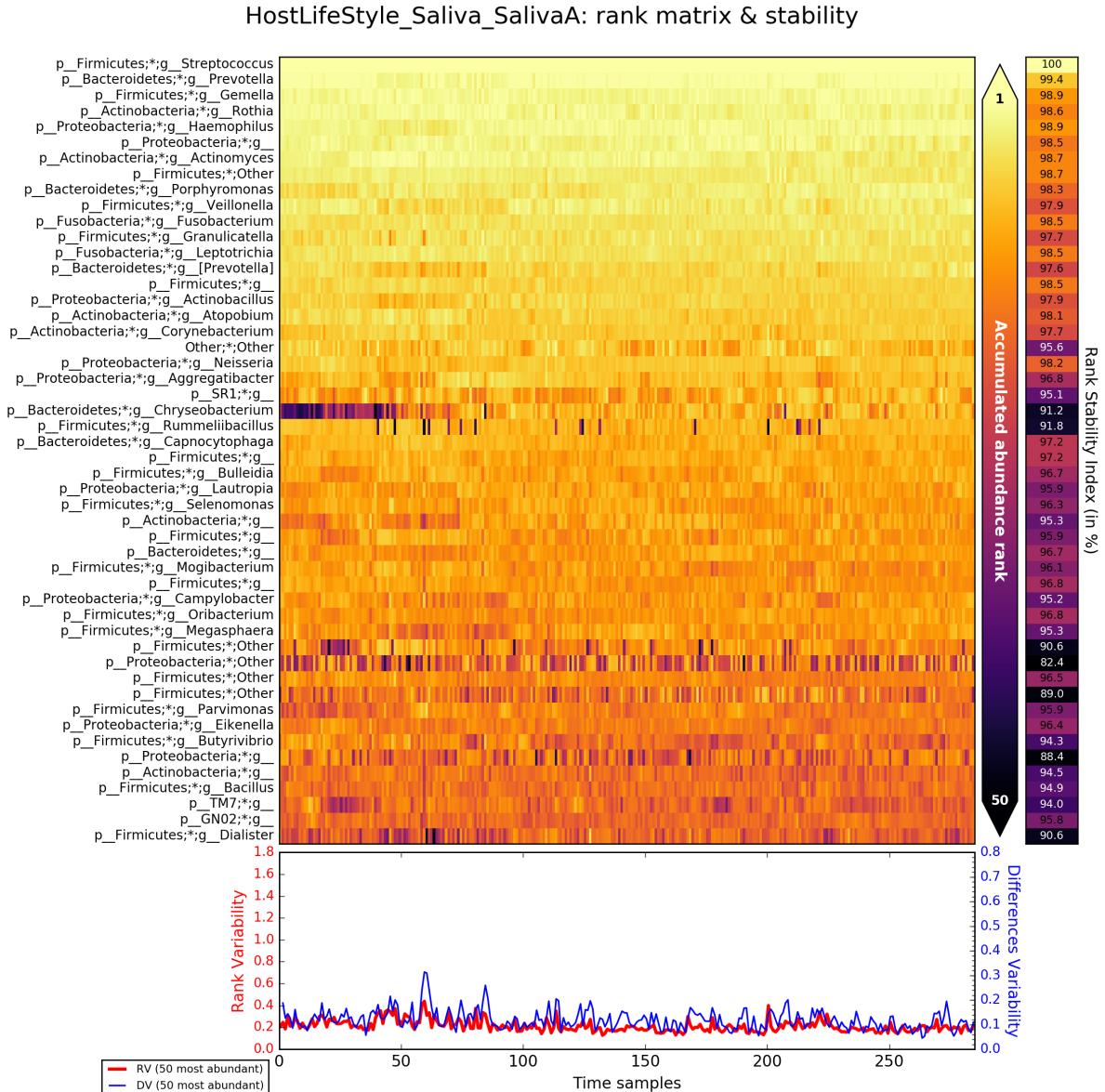


Figura 3.10: Matriz de rango correspondiente a la muestra de saliva A. En filas aparecen los 50 taxones más abundantes ordenados descendenteamente, en columnas se representan los días del año donde hubo muestra y el color determina el rango de mayor (amarillo) a menor (violeta). En la parte derecha aparecen 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 está resumida la dinámica en la estabilidad de la muestra de intestino perteneciente al sujeto A. Aquí también se aprecia que los géneros más abundantes suelen ser los más estables. Sin duda lo que más llama la atención de esta figura es el desorden de rangos que se produce entre los puntos 71-122, correspondientes a los días en los que el sujeto estuvo de viaje en el extranjero (destacar que en estos datos también existen días sin muestra pero son a partir del regreso del viaje, por lo que se corresponden bien durante el mismo). Existen comportamientos muy interesantes como el género *Anaerostipes* que era muy abundante antes del viaje pero durante el viaje disminuye y a la vuelta recupera su abundancia inicial. También los géneros *Plesiomonas* y *Fusobacterium* que no son nada abundantes pero durante el viaje se dan las condiciones óptimas para su crecimiento. Y por último, el orden *Methylphilales* (etiqueta “p_Proteobacteria;*;g_”) que no le afecta el viaje pero aumenta drásticamente su abundancia al rededor del día 230 por motivos desconocidos. Si nos fijamos en el RSI de este último orden, con valor 96.2 %, podría considerarse la isla de estabilidad más llamativa en este sujeto.

En cuanto a la medida de variabilidad, se observa que ambas medidas se disparan durante el viaje. Además, cabe destacar que a partir del punto 100, RV baja simulando una exponencial lo que supone una recuperación al estado inicial muy rápida.

En la figura 3.12 está resumida la dinámica en la estabilidad de la muestra de intestino perteneciente al donante B. De nuevo los géneros más estables se corresponden con los más abundantes en general. Además, también se aprecia un cambio brusco de rango en el punto 120 aproximadamente que se corresponde al día de inicio de la salmonelosis (apreciar que existen días donde no hubo muestra y en este gráfico se representan 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 poco abundantes que son oportunistas y aumentan su abundancia (como *Dialister*). Luego existen comportamientos ajenos como *Bacteroides* y el orden *YS2* (etiqueta “p_Cyanobacteria;*;g_”) que son muy estables en su alta y baja abundancia respectivamente. De hecho, el segundo es una clara isla de estabilidad con RSI = 97.7 cuando ocupa el puesto 43 en abundancia.

Fijándonos en el gráfico inferior, la variabilidad sufre aumentos de un día a otro en varias ocasiones pero los picos más relevantes se obtienen a partir de la infección. Tanto RV como DV bajan casi exponencialmente como en el caso anterior pero parece que no recupera el estado inicial sino que más bien llega a un nuevo estado de variabilidad.

3.4. EXPLORANDO SERIES TEMPORALES

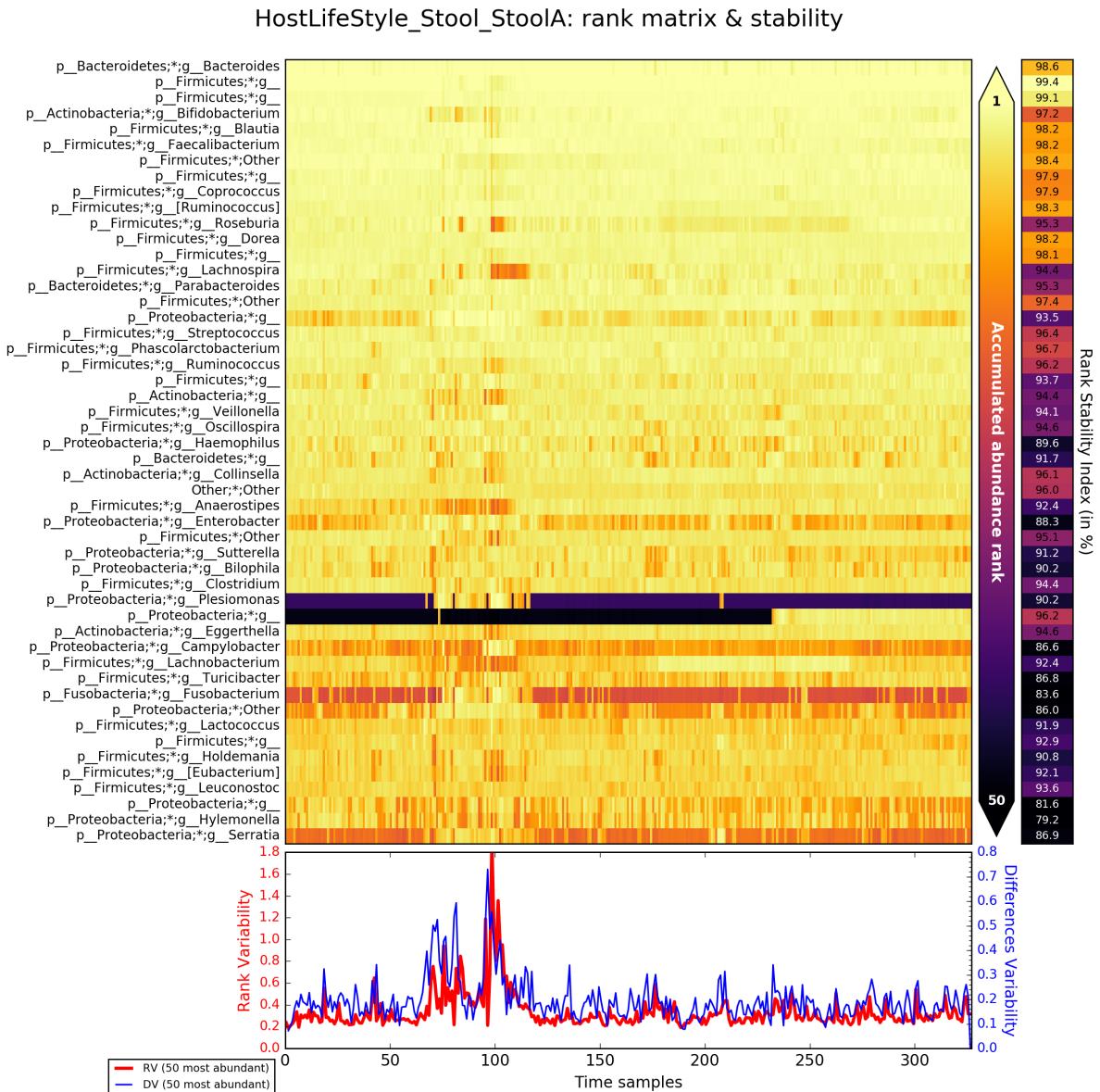


Figura 3.11: Matriz de rango correspondiente a la muestra de intestino A.

En filas aparecen los 50 taxones más abundantes ordenados descendientemente, 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

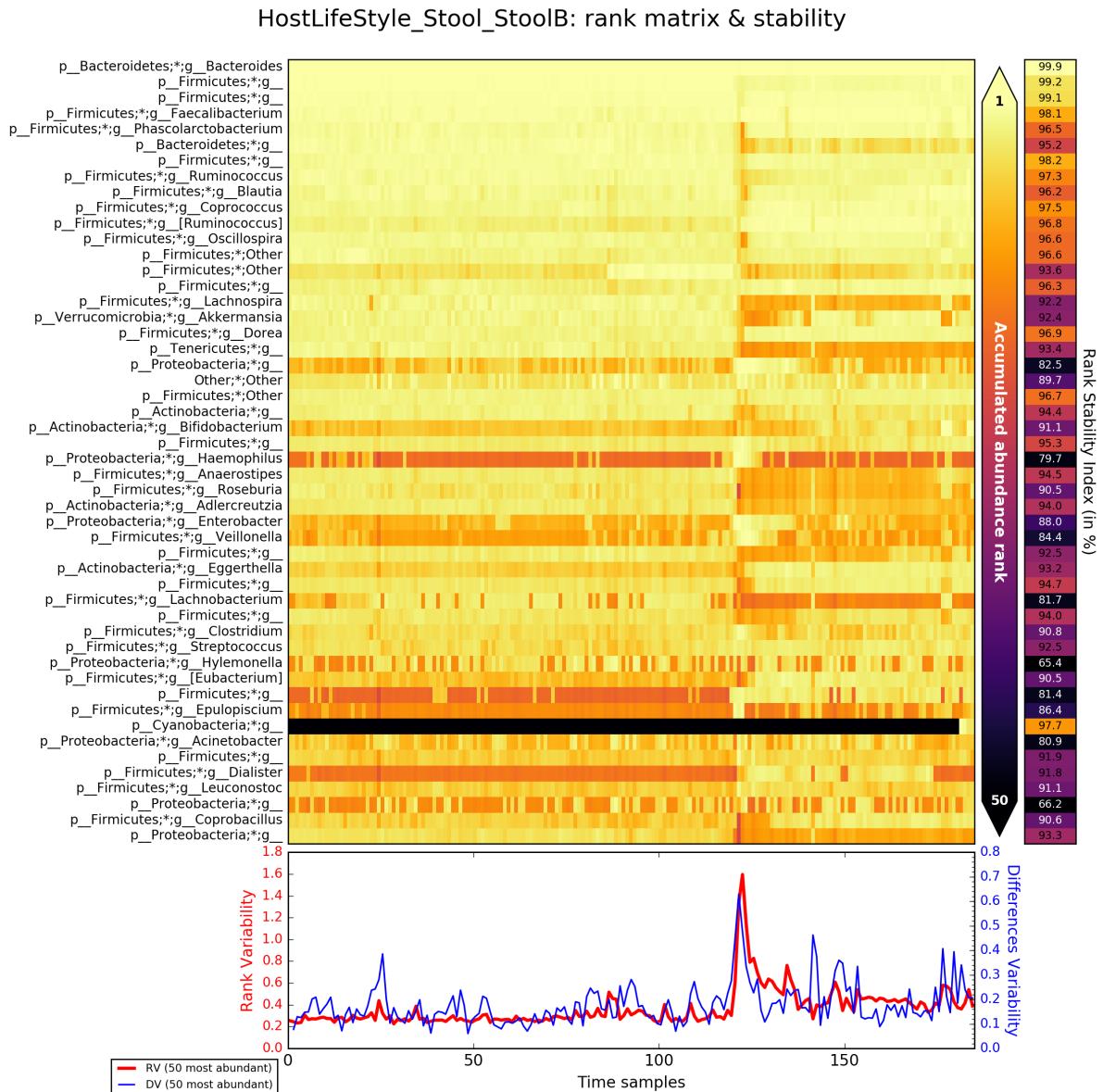


Figura 3.12: Matriz de rango correspondiente a la muestra de intestino B.

En filas aparecen los 50 taxones más abundantes ordenados descendientemente, 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.

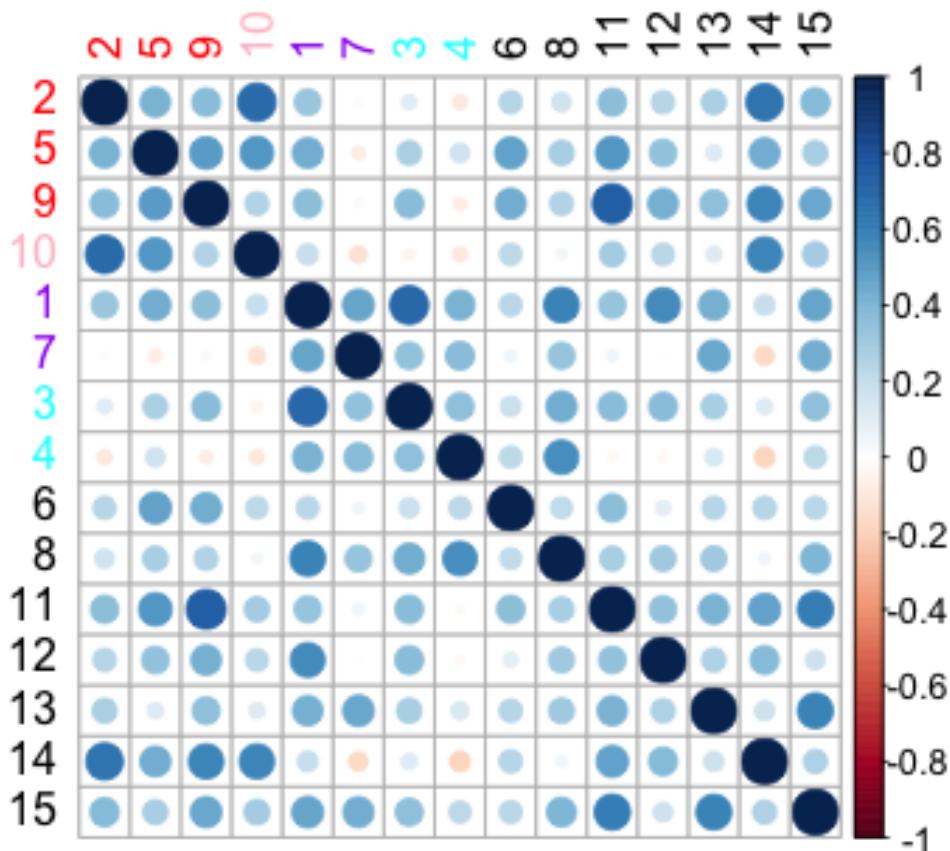
3.5. Correlaciones

Las correlaciones de abundancia en especies microbianas se utilizan en muchos trabajos para indicar interacción entre ellas. La correlación positiva indicaría una interacción mutualista, y la correlación negativa una interacción competitiva. Trabajar con abundancia relativas produce estimaciones sesgadas porque, como deben sumar a 1, las fracciones no son independientes y tienden a tener una correlación negativa independiente de la verdadera correlación entre las abundancias absolutas subyacentes. Se han desarrollado algoritmos como SparCC [12] para mitigar estos problemas.

En este proyecto también se han calculado 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 sirve para reorganizar la matriz de correlaciones y comprobar las correlaciones que se dan entre estos grupos de comportamiento. En la figura 3.13 vemos el resultado para la muestra de saliva. Los géneros 4, 7 y 14 muestran correlación negativa en algunas ocasiones. Los grupos presentan correlaciones positivas entre sus miembros aunque las correlaciones máximas se dan fuera de grupos (como el género 2 con 10 y 14, o el género 9 con 11). Esto quiere decir que los microorganismos que tienen la una respuesta parecida al viaje, no son los que más correlacionan. En la figura 3.14 se muestra el comportamiento de los géneros que componen el grupo 2. Se representa la abundancia relativa frente al tiempo de 3 géneros y se puede comprobar que todos disminuyen su abundancia los días del viaje (40-75) pero la recuperan a la vuelta. En concreto, *Prevotella* es el género que más disminuye durante el viaje y también es el más variable a lo largo del tiempo. Se puede observar cómo los géneros correlacionan bien en algunos puntos (como *Haemophilus* y *Porphyromonas* los primeros días) aunque, en general, la correlación no es muy grande dentro del grupo.

Los resultados para intestino del donante A se encuentran en la figura 3.15. Se han encontrado 4 grupos de comportamiento distinto y ninguna anti-correlación. El grupo 2 (rojo) es el que presenta mejores correlaciones entre sus miembros, en la figura ?? se representa la abundancia a lo largo del tiempo para comprobar su comportamiento. Los grupos 1 (verde) y 3 (naranja) solo tienen un miembro por lo que no se puede ver correlación interna de grupo mientras que el grupo 5 (violeta) tiene correlación positiva pero no tan alta como el grupo 2. De nuevo, en algunos casos se dan correlaciones más fuertes entre grupos como es el caso de los géneros 10 y 13 que tienen mucha correlación con el grupo 2.

En la figura 3.16 se muestra el intestino B como último caso. Se han hallado 5 grupos, de los cuales el grupo 2 (rojo) tiene la mayor correlación interna. También hay casos de correlación muy fuerte entre grupos distintos como *Oscillospira* con *Phascolarctobacterium*. La proteobacteria del grupo 5 (violeta) tiene correlación negativa con el resto y la



- 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 círculo corresponde al valor de correlación.

máxima anti-correlación se da entre *Ruminococcus* y el grupo 4 (rosa).

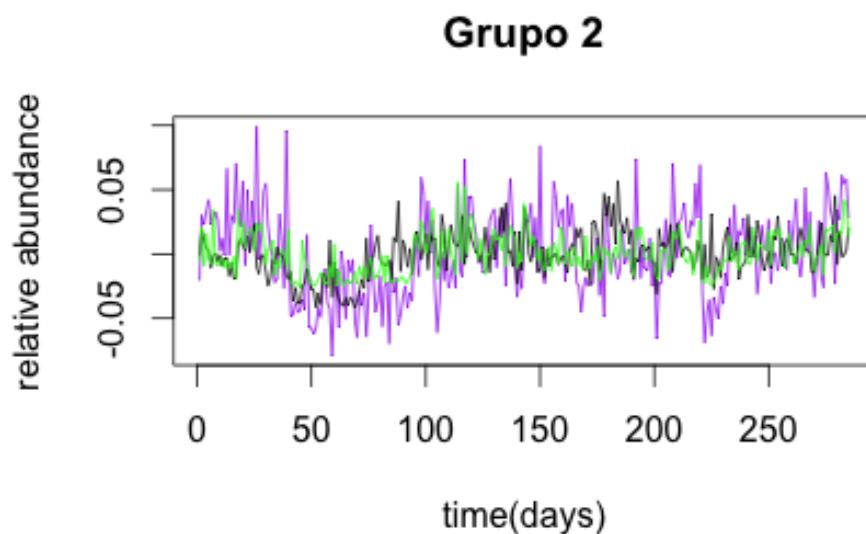
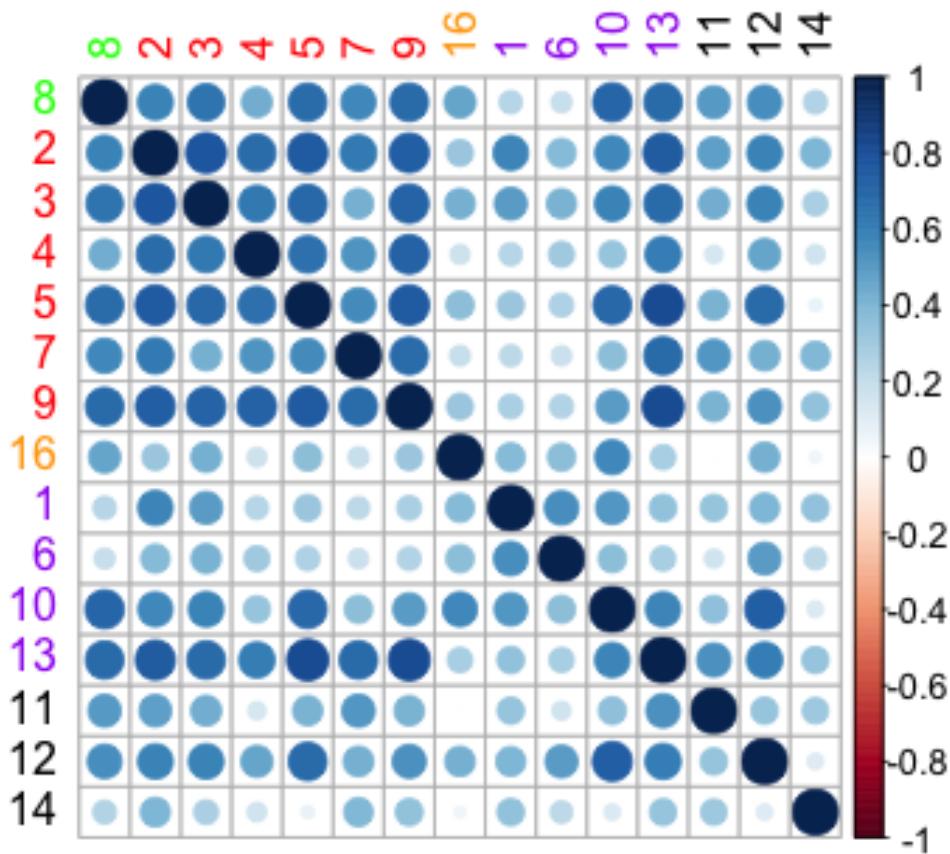
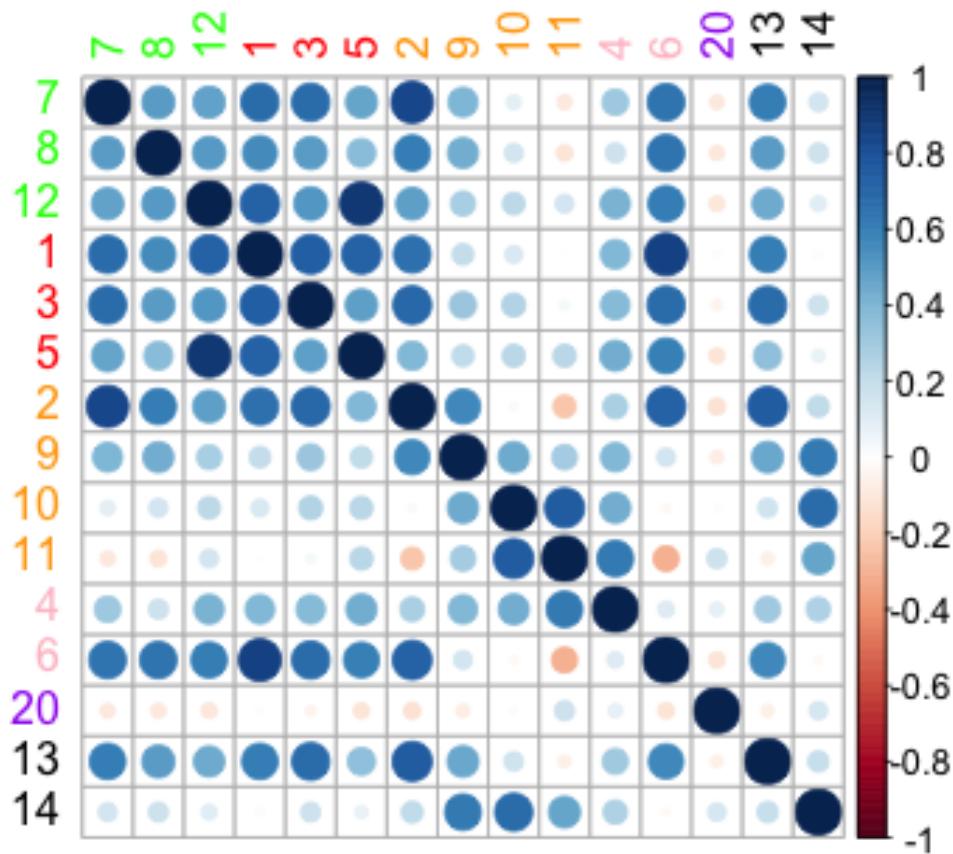


Figura 3.14: Se representa la abundancia relativa a lo largo del tiempo del grupo 2 en la muestra de saliva. Este grupo consta de 3 géneros: *Prevotella* (violeta), *Haemophilus* (negro) y *Porphyromonas* (verde). Se demuestra que no existe una buena relación entre correlaciones y grupos de comportamientos 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_Ruminococcus
- 3: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Blautia
- 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_Rikenella
- 7: k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Bacteroides
- 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_Bacteroides
- 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_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Erysipelotrichi

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 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 el microbioma tiene cierta estabilidad. Sin embargo, cuando el hospedador cambia de ambiente realizando un viaje al extranjero, su flora intestinal se desequilibra debido a la nueva dieta y el tipo de agua del 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, el microbioma recupera la estabilidad inicial. También se ha analizado la micobiota de la cavidad oral pero ésta no sufre tanto la perturbación como el intestino.

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 hubiera sido muy interesante comprobar lo que sucede. Algunas posibilidades pueden ser (i) 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 (ii) 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 que 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 exponerse 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.

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.

Para alcanzar esa meta aún queda un largo camino que recorrer. 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 [9]. 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). 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.

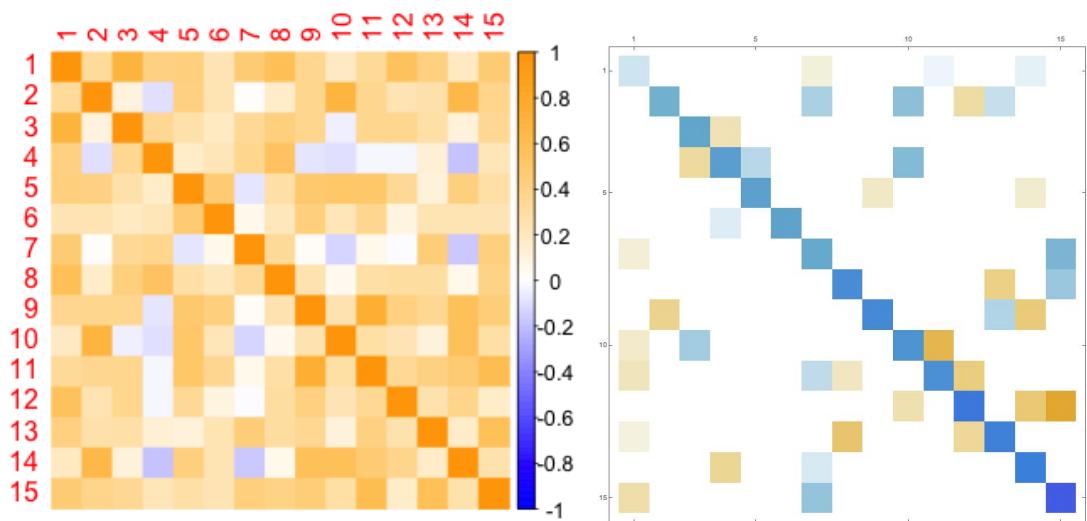


Figura 4.1: Correlación Pearson vs. LIMITS en saliva A

DISCUSIÓN Y PERSPECTIVAS DE FUTURO

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 [13]. 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 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.
- [2] 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/10.1101/036103>
- [3] 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.
- [4] 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.
- [5] 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.
- [6] Taylor, L.R. 1961. Aggregation, Variance and the mean. *Nature* **189**, pp. 732-35.
- [7] 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.
- [8] Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**(19), pp. 2460-2461.
- [9] Martí, J.M., and Garay, C.P. 2017. ComplexCruncher: dynamics of ranking processes toolkit. In preparation.
- [10] 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.

BIBLIOGRAFÍA

- [11] 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.
- [12] 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.
- [13] Friedman, J., and Alm, E.J. 2012. Inferring correlation networks from genomic survey data. PLoS Comput Biol, **8**(9), p. e1002687.
- [14] 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>