Estudio metagenómico de microbioma oral humano con diferente severidad en diabetes tipo 2, caracterización a nivel de composición y funcional.

Estudiante Noelia Rodríguez Pérez
Plan de Estudios del Estudiante Máster universitario en Bioinformática y bioestadística UOC-UB
Área del trabajo final M0.123-Estadística y bioinformática 24

Consultor/a Modesto Redrejo Rodríguez
Profesor/a responsable de la asignatura José Antonio Morán Moreno

Fecha Entrega 05/06/2018

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Estudio metagenómico de microbioma oral humano con diferente severidad en diabetes tipo 2, caracterización a nivel de composición y funcional.*
Nombre del autor:	*Noelia Rodríguez Pérez*
Nombre del consultor/a:	*Modesto Redrejo Rodríguez*
Nombre del PRA:	*José Antonio Morán Moreno*
Fecha de entrega (mm/aaaa):	06/2018
Titulación::	*Máster universitario en Bioinformática y bioestadística UOC-UB*
Área del Trabajo Final:	*M0.123-Estadística y bioinformática 24*
Idioma del trabajo:	*Castellano*
Palabras clave	*Diabetes tipo 2, periodontitis, 16S rRNA.*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

La finalidad de este trabajo ha sido caracterizar la diabetes tipo 2 con distinto grado de severidad desde un punto de vista metagenómico, realizando un análisis de composición microbiológica y funcional.

Los datos donde se ha aplicado el análisis, provienen de un estudio observacional caso-control realizado en población china, con 31 sujetos clasificados en cuatro grupos distintos según presentasen diabetes tipo 2 y periodontitis o no. En ellos se secuenció el gen que codifica para 16S rRNA con la plataforma 454 GS FLX Titanium. Se llevaron a cabo análisis sobre diversidad alfa y beta, así como análisis de composición taxonómica gracias a la herramienta QIIME2, mientras que el paquete *limma* en Rstudio, permitió el cálculo de funciones diferencialmente expresadas en los grupos de interés, inferidas a partir de los datos del 16S rRNA gracias a PICRUST.

En cuanto a la diversidad alfa, se observó que los grupos con diabetes o periodontitis presentaban menor riqueza que los controles. En diversidad beta, la diabetes y la periodontitis han sido las poblaciones con más distancia respecto al grupo control. Finalmente, los filos Synergistetes y Spirochaetes, y el género *Treponema* de este último, se han identificado como característicos del grupo con diabetes y periodontitis. A nivel funcional, las mayores

diferencias se han presentado entre los grupos con periodontitis, diabéticos y no diabéticos. Ello ha permitido concluir que la condición periodontítica provoca modificaciones en el entorno diabético que la hace adquirir diferencias a nivel de composición microbiológica y funcional, respecto a la diabetes.

Abstract (in English, 250 words or less):

The purpose of this study has been to characterize type 2 diabetes with different severity from a metagenomic point of view, performing a microbiology and functional analysis.

The analysis was applied on data from an observational case-control study performed in Chinese population, with 31 subjects classified into four different groups according to whether they presented type 2 diabetes and periodontitis or not. The gene coding for 16S rRNA was sequenced with the 454 GS FLX Titanium platform. Analysis of alpha and beta diversity as well as taxonomy, were carried out with QIIME2 tool, while the package *limma* on Rstudio, allowed the calculation of differentially expressed functions in groups of interest, inferred from the 16S rRNA data thanks to PICRUST.

Regarding alpha diversity, groups with diabetes or periodontitis had less richness than controls. In beta diversity, diabetes and periodontitis have been the populations with the greatest distance to the control group. Finally, the *Synergistetes* and *Spirochaetes* phyla, and the *Treponema* genus of the latter, have been identified as characteristic of the group with diabetes and periodontitis. At the functional level, the greatest differences have been observed between the diabetic with periodontitis group and the one with periodontitis. This has allowed us to conclude that the periodontitic condition causes changes in the diabetic environment that leads to differences at the microbiological composition and functional level, not present in the diabetic environment itself.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

El presente trabajo de fin de máster pretende caracterizar la diabetes tipo 2 con distinto grado de severidad desde un punto de vista metagenómico. Para ello partimos de un estudio observacional caso-control realizado en población china (EMBL-EBI, Study: PRJNA182759). La cohorte consta de 31 sujetos que fueron clasificados en cuatro grupos distintos según presentasen diabetes tipo 2 y periodontitis o no. Los grupos formados fueron: sujetos no diabéticos sin periodontitis, no diabéticos con periodontitis, diabéticos tipo 2 sin periodontitis y diabéticos tipo 2 con periodontitis. En ellos se extrajo ADN de muestras de la cavidad oral para obtener el metagenoma oral humano a partir de *deep sequencing* (método de secuenciación cultivo-independiente) del gen que codifica para 16S rRNA con la plataforma de pirosecuenciación 454 GS FLX Titanium. Esta tecnología, desarrollada por Roche Diagnostics Corporation, proporciona secuencias con una longitud media de 400 bases de lecturas de un solo extremo (*single-end reads sequences*) [1].

El análisis metagenómico de esta cohorte podría permitir la caracterización de las posibles variaciones en el microboma en muestras con diabetes tipo 2 con distinto grado de severidad, lo cual podría ayudar a un mejor entendimiento de sus mecanismos de acción y por consiguiente podría ser de utilidad en un mejor control de la misma.

La diabetes de tipo 2 es una enfermedad crónica influenciada tanto por aspectos genéticos como ambientales, cuya prevalencia ha aumentado dramáticamente en las últimas décadas, convirtiéndola en un importante problema de salud a lo largo de todo el mundo [2, 3]; de hecho la OMS, prevé que será la séptima causa de mortalidad en 2030 [4]. Adicionalmente su heterogeneidad la ha convertido en blanco de múltiples estudios científicos, encaminados a ilustrar características relacionadas con sus riesgos [5]. Una de esas condiciones relacionadas con la diabetes de tipo 2 es la periodontitis, que consiste en una inflamación gingival y sangrado, recesión gingival y en fases avanzadas puede llevar a la pérdida de dientes [6]. Esta característica clínica a menudo es pasada por alto en pacientes con diabetes tipo 2, pero existen estudios que la asocian a mayor severidad de diabetes tipo 2, correlacionándola con los niveles de hemoglobina glicosilada (HbA1c) [7, 8].

Por otro lado, la implicación de la microbiota (intestinal principalmente) en la patogenia de diversas enfermedades ha cobrado un elevado interés en los últimos años, prueba de ello es el aumento de artículos científicos sobre microbiota publicados en PubMed durante los últimos

17 años [9, 10]. No en vano, el microbioma ha pasado a ser considerado otro órgano humano y como tal se está estudiando en profundidad para entender su fisiología y patología [11]. Un marcador muy ampliamente aceptado y empleado por diversas plataformas para su estudio, es el gen que codifica para el 16S rRNA (ARN ribosomal 16S) que es el componente de la subunidad 30S de los ribosomas procariontes. Su amplio uso en estudios metagenómicos se debe a que está presente en prácticamente la totalidad de las bacterias, presenta bajas tasas de evolución, lo cual le convierte en un buen candidato para estudios filogenéticos, pues cambios aleatorios en su secuencia pueden ser una precisa medida de tiempo, y por último, este gen es suficientemente grande para ser usado con fines informáticos [12]. Dentro de este contexto y gracias al uso de estas nuevas tecnologías de alto rendimiento, enfermedades complejas como la obesidad y la propia diabetes se han asociado con una disbiosis de la microbiota intestinal y con cambios en la composición taxonómica y funcional de la misma [13, 14].

Por tanto, se ha elegido esta área y tema porque entender mejor la diabetes de tipo 2 así como su asociación con otros factores de riesgo desde el punto de vista microbiológico y metagenómico, puede ayudarnos a reducir su comorbilidad y a ahorrar recursos económicos invertidos en su tratamiento. Además establecer riesgos más individualizados de este desorden podría contribuir a encaminarse a la tan ansiada medicina personalizada.

Esta enfermedad está ampliamente extendida entre la población pero su gran heterogeneidad hace que sus diferentes variantes no estén demasiado estudiadas en el contexto metagenómico. La interacción huésped-microbiota puede revelar aspectos aún desconocidos acerca de esta enfermedad. Por otro lado, para abordar el estudio de la microbiota debemos recurrir a técnicas de secuenciación masiva. Estas técnicas de alto rendimiento han supuesto una revolución en el campo de los estudios microbiológicos, pero implican la generación de gran volumen de datos y por tanto han requerido del desarrollo de nuevas maneras de gestionar y manejar estos datos, posibles gracias a la bioinformática. Así mismo han debido aplicarse nuevos métodos estadísticos para analizar dichos datos. Con lo que la aplicación de estas nuevas herramientas bioinformáticas y bioestadísticas permite una mejor comprensión de enfermedades desde un enfoque antes imposible de abordar [15].

1.2 Objetivos del Trabajo

1) Realizar un análisis de la composición microbiológica en muestras de metagenoma oral humano, a partir de sus datos de secuenciación, en el contexto de las categorías aportadas en los metadatos del estudio, para detectar diferencias en riqueza y/o igualdad (*eveness)* y para detectar taxones diferencialmente expresados entre alguno de los grupos de

interés, prestando especial importancia a las diferencias entre el grupo de diabetes sin y con periodontitis.

2) Realizar un análisis funcional para determinar los genes y funciones representativas de los grupos distintos de pacientes, para detectar posibles genes y funciones diferencialmente representadas prestando especial interés en los grupos con diabetes sin y con periodontitis.

Para contestar a las preguntas del primer objetivo general, tendremos que establecer los siguientes objetivos específicos:

1) Preprocesar las secuencias para obtener una tabla de frecuencias de OTUs (*operational taxonomic unit*).
2) Calcular la diversidad alfa y beta de nuestras secuencias.
3) Análisis estadístico de distintas métricas de diversidad alfa y beta en los distintos grupos de estudio.
4) Realizar una asignación taxonómica de la tabla de OTUs.
5) Análisis estadístico de la tabla con asignaciones taxonómicas en los distintos grupos de estudio.

Para determinar el segundo objetivo general y poder contestar a las preguntas, necesitaremos realizar los siguientes objetivos específicos:

6) Normalización de la tabla de OTUs e inferencia de genes y funciones.
7) Estudio estadístico para establecer genes y funciones diferencialmente expresados.

1.3 Enfoque y método seguido

El presente estudio se centrará en análisis de datos de pirosecuenciación en los que se ha empleado como gen marcador el gen que codifica para 16S rRNA. Es cierto que hoy en día han surgido nuevas estrategias de secuenciación como el método *shotgun,* mucho más potentes, pero no exenta de inconvenientes, como sesgo hacia genes altamente expresados que pueden consumir la mayoría de lecturas o el sesgo en el extremo 3', debido a la selección por polyA [16], y por otro lado, la secuenciación del 16S rRNA por pirosecuenciación sigue siendo una técnica metagenómica que permite analizar simultáneamente una misma comunidad desde un punto de vista taxonómico y funcional [17]. Estos datos de partida se encuentran en archivos depositados en el repositorio del European Nucleotide Archive (ENA), bajo el número de acceso PRJNA182759.

El análisis de la calidad de las secuencias de los archivos con los datos se llevará a cabo con FastQC, ya que puede ejecutarse como una aplicación interactiva independiente para el análisis inmediato de un pequeño número de archivos con extensión .fastq como es nuestro caso y permite la generación de informes de control de calidad de las secuencias de forma fácil y rápida [18].

El procesamiento de datos se realizará utilizando la línea de comandos de la terminal de la herramienta QIIME2 (*Quantitative Insights into Microbial Ecology*), instalada en máquina virtual sobre sistema operativo Ubuntu [19]. Esta plataforma bioinformática de nueva generación para el estudio de microbiomas, es un sistema basado en *plugins*, es decir en paquetes creados en lenguaje python3, que nos permitirán procesar los datos en cada fase de nuestro flujo de trabajo. En concreto inicializan un objeto *qiime2.plugin.Plugin* y registran acciones, formatos de datos y/o tipos semánticos que se hacen visibles en QIIME2. Se ha elegido dicha herramienta sobre otras, como MOTHUR, por su gran flexibilidad, ya que en ella se dispone de múltiples *plugins* con *scripts* que permiten realizar diversas tareas, en contraposición con MOTHUR en que cada tarea debe realizarse por separado. Además QIIME2 ofrece excelentes visualizaciones.
También se empleará PICRUST (*Phylogenetic Investigation of Communities by Reconstruction of Unobserved States*), haciendo uso de este paquete de software bioinformático mediante su versión online en la plataforma web de código abierto Galaxy. Se ha optado por esta opción debido a su mayor facilidad de manejo, lo cual la hace más idónea para inexpertos en el campo, como es nuestro caso [20, 21]. Finalmente, realizaremos parte del análisis con R, un entorno de software libre, pues su versatilidad y la gran disponibilidad de librerías ya creadas nos permitirán abordar ciertos análisis estadísticos y la creación de distintos tipos de gráficos. Haremos uso de R a través del entorno de desarrollo integrado RStudio, ya que facilita el flujo de trabajo haciendo más fácil el depurado interactivo para diagnosticar errores y permite la generación de informes con R Markdown [22].

1.4 Planificación del Trabajo

Las tareas propuestas para cada objetivo específico dentro este trabajo de fin de máster son:

1) Preprocesado de las secuencias

- Determinación de la calidad de las secuencias descargadas del ENA.
- Importación de los archivos fastq.gz de partida a QIIME2.
- Análisis de la calidad de las secuencias generadas.
- Eliminación de secuencias ruido, corte de los extremos con baja calidad (*trimming*), filtrado de las secuencias que no tengan la longitud mínima, dereplicado de las secuencias y filtrado de quimeras.
- Visualización del resumen de la tabla de frecuencias y la tabla con secuencias representativas.

2) Cálculo la diversidad alfa y beta de nuestras secuencias.

- Realización de alineamientos múltiples de secuencias *de novo*.
- Filtrado de las posiciones altamente variables de los alineamientos.

- Generación de un árbol filogenético con raíz.
- Cálculo de métricas de diversidad alfa y beta.
- Visualización de métricas diversidad beta.

3) Análisis estadístico de distintas métricas de diversidad alfa y beta en los distintos grupos de estudio.

- Análisis de la composición microbiológica de las muestras.
- Generación de curvas rarefacción alfa.
- Análisis de composición en función del grupo de pertenencia.
- Ordenación restringida.

4) Asignación taxonómica de la tabla de OTUs.

- Entrenar el clasificador Naive Bayes.
- Asignar taxonomía a las secuencias de la tabla de secuencias representativas generada por *dada2*.
- Creación de gráficos de barras de composición taxonómica.

5) Análisis estadístico de la tabla con asignaciones taxonómicas en los distintos grupos de estudio.

- Creación de tabla de taxones colapsada a nivel filo, género y especie.
- Aplicación de test ANCOM a la tabla de taxones según el grupo de pertenencia.

6) Normalización de la tabla de OTUs e inferencia de funciones.

- Selección de taxones de referencia cerrada.
- Exportación en formato .biom.
- Adición de anotaciones taxonómicas de Greengenes.
- Normalización del archivo por número de copia de 16S.
- Obtención de tabla de predicción del metagenoma.
- Obtención de tabla de predicción de funciones.

7) Estudio estadístico para establecer genes diferencialmente expresados.

- Importación de la tabla y comprobación de normalidad.
- Diseño de la matriz de diseño y la matriz de contraste.
- Obtención de tabla con ortólogos KEGG *(Kyoto Encyclopedia of Genes and Genomes)* diferencialmente expresados.
- Visualización gráfica.
- Repetición del proceso con categorías pathways KEGG y COGs *(Clusters of Orthologous Groups)*.

En base a estas tareas y a las fechas de entrega de cada PEC (pruebas de evaluación continua), la planificación temporal propuesta se ha plasmado en un diagrama de Gantt (figura 1), mostrado más abajo.

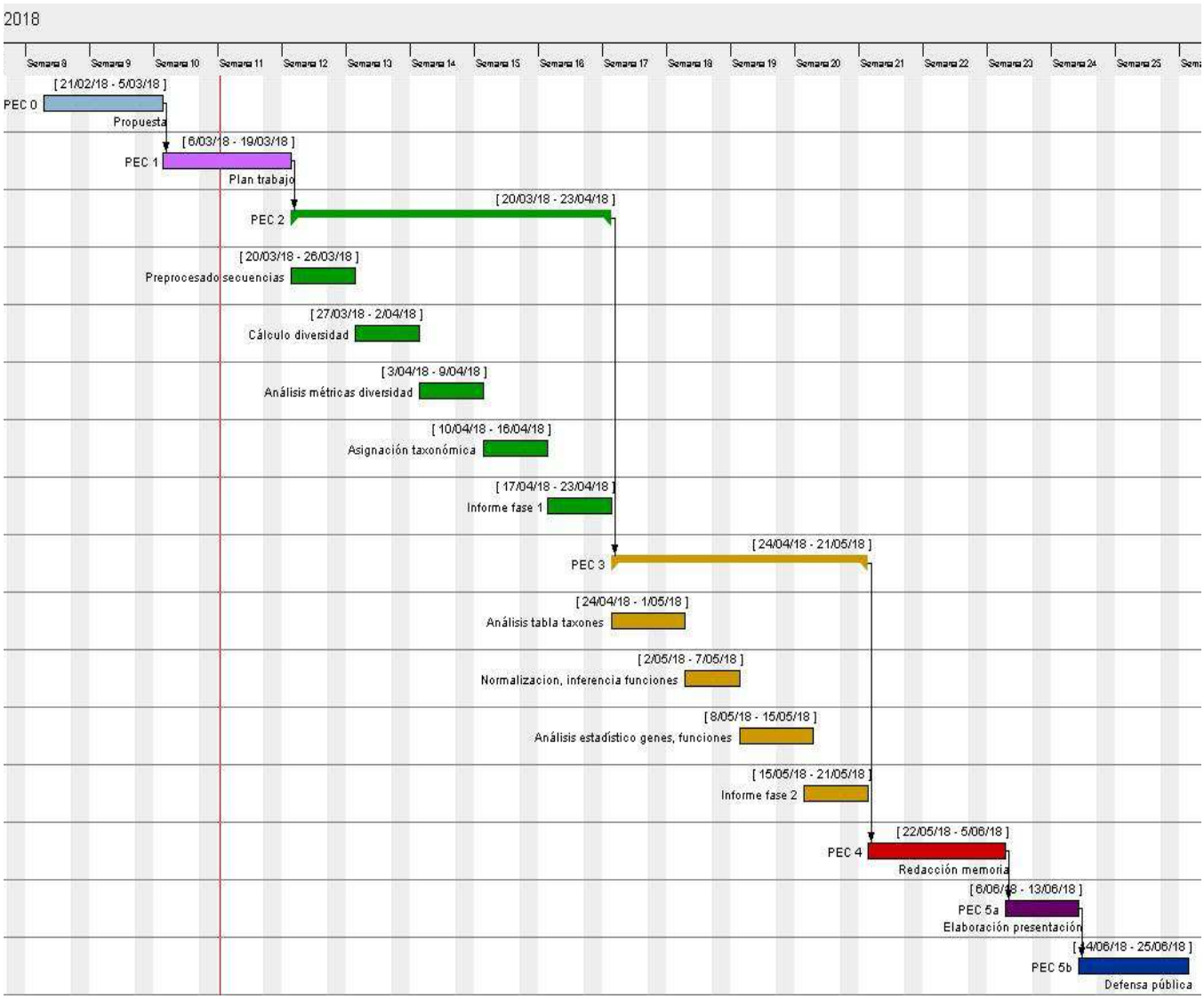

Figura 1. Diagrama de Gantt, donde se muestra la planificación temporal de cada tarea necesaria para realizar el trabajo de fin de máster. Para su realización, se ha empleado el programa de código abierto, escrito en Java, GanttProject.

1.5 Breve sumario de productos obtenidos

Gráficos de riqueza por distintas métricas según los cuatro grupos del estudio junto con sus test estadísticos.
Gráficos de igualdad en los cuatro grupos del estudio junto con sus test estadísticos.
Gráficos de ordenación.
Gráficos de barras de los taxones agrupados de acuerdo a los cuatro grupos del estudio, en distintos niveles taxonómicos.
Resumen de los test estadísticos aplicados para detección de los taxones diferencialmente expresados entre el grupo de diabéticos sin y con periodontitis y también entre el grupo de no diabéticos sin y con periodontitis.

Tablas de ortólogos, *pathways* KEGG y COGs expresados en los efectos de interés.
Representación gráfica de los ortólogos, *pathways* KEGG y COGs diferencialmente expresados.
Tablas de OTUs, de secuencias representativas y de taxones.
Tablas con los ortólogos, *pathways* KEGG y COGs diferencialmente expresados en los efectos de interés.
Script con las instrucciones de la línea de comandos en QIIME2 para el análisis de secuencias.
Script en R Markdown para el gráfico de ordenación y los análisis de expresión génica.

1.6 Breve descripción de los otros capítulos de la memoria

Los otros capítulos de la memoria consistirán en un apartado de materiales y métodos (desarrollado a lo largo del capítulo 2), donde se describirán los datos de partida y métodos estadísticos aplicados. También se incluirá un apartado de resultados (capítulo 3), donde se expondrán los resultados gráficos y estadísticos obtenidos relativos a cada uno de los dos grandes objetivos planteados. Finalmente se añadirá un apartado de discusión (capítulo 4), donde los resultados se analizarán realizando una crítica objetiva, argumentada con otros estudios publicados.
Otro capítulo del trabajo final de máster será el de conclusiones (capítulo 5), en él se resumirán las conclusiones obtenidas del estudio. En concreto, se hará un ejercicio de autoevaluación del proyecto, donde se valorará el grado de logro de cada indicador y se explicará críticamente las conclusiones que se han extraído de la actividad, incluyendo posibles acciones futuras en relación al estudio realizado.
El capítulo seis recogerá la bibliografía del trabajo de fin de máster, mientras que el siguiente capítulo incluirá un glosario de términos mencionados en el mismo.
Finalmente en el capítulo ocho de anexos (que se adjuntará por separado), se incluirán los scripts generados, las tablas de OTUs, de secuencias representativas y de taxones y las tablas con los ortólogos, *pathways* KEGG y COGs diferencialmente expresados en los efectos de interés.

2. Material y métodos

Como se ha mencionado anteriormente, los datos referentes a las muestras del estudio se encuentran en el ENA, con el número de acceso PRJNA182759. Los datos clínicos relativos a cada muestra allí presentes, junto con datos referentes a tres variables adicionales: HbA1c (hemoglobina glicosilada), PBS (*2h Postpandrial Blood Sugar*) y AL (*Attachment Loss*), extraídos de un artículo escrito con los datos depositados [23], nos han permitido elaborar un archivo de metadatos para las muestras, que se ha importado a QIIME2.

En primer lugar, se confirmó que los archivos depositados se trataban de archivos fastq con calidades; pues cada lectura consistía en cuatro líneas: un titular que empieza por el carácter @, una secuencia de nucleótidos, un símbolo + y una puntuación de calidad (*quality score*) de la secuencia, para cada posición, codificada por un carácter ASCII imprimible, codificado en escala Phred+33. A continuación, se ha procedido al análisis inicial de los archivos con FastQC, donde se ha determinado que el truncado de las secuencias se haría a una longitud de 367 pares de bases y que se recortaría 9 pares de bases en el inicio de las secuencias.

Los archivos fastq.gz de partida para las lecturas de un solo extremo, son archivos fastq demultiplexados, con una calidad asignada; por lo que para su importación a QIIME2, se ha creado un archivo delimitado por comas (.csv), en formato *fastq manifest*. Dicho archivo ha permitido conectar los identificadores de las muestras con rutas absolutas de archivos fastq.gz que contienen las secuencias y sus datos de calidad de las mismas, también ha indicado la dirección de lectura en cada ruta absoluta de archivos fastq.gz. Esta acción ha permitido continuar el análisis por las tareas posteriores al demultiplexado de secuencias [19].

A continuación se ha usado el *plugin dada2,* en QIIME2, con el método *denoise-single*, para la eliminación de secuencias ruido (*denoise*), corte de los extremos con baja calidad (*trimming*), filtrado de las secuencias que no tuvieran la longitud mínima después del *trimming*, dereplicado de las secuencias y filtrado de quimeras. Cabe destacar que en el derreplicado se combinan todas las lecturas de secuencias idénticas en "secuencias únicas" junto con el número de lectura para esa secuencia única, así se reduce tiempo computacional y se eliminan comparaciones redundantes. Como producto final, este *plugin* ha devuelto una tabla de *amplicon sequence variant* (ASV), que puede considerarse un análogo a las tradicionales tablas OTUs con mayor resolución, por eso a menudo se usa el término tabla de 100% OTUs para referirse a ella. En esta tabla se recoge el número de veces que cada ASV es observado en cada muestra. También se ha generado una tabla con las secuencias representativas de cada ASV [24]. Posteriormente se ha visualizado un

resumen de la tabla de frecuencias y la tabla con secuencias representativas, para tener una idea que cómo fue el proceso. El resumen ha permitido establecer la profundidad de muestreado (*sampling depth*) que deberemos emplear en subsiguientes fases del proceso. La profundidad de muestreo, elegida en base al número de secuencias de la muestra SRR630916, ha sido 1798.

Dado el interés del estudio en determinar si las muestras diferían en su composición de acuerdo a su grupo de pertenencia (diabetes, periodontitis), o diferían en riqueza y en igualdad de acuerdo a este grupo; se ha creado un árbol filogenético para poder obtener medidas de diversidad filogenéticas: *Faith's, weighted UniFrac* y *unweighted UniFrac* [25]. Primero se ha empleado el *plugin alignment* bajo el método *mafft,* para realizar múltiples alineamientos de las secuencias de la tabla de secuencias representativas *de novo* usando MAFFT. Como resultado se ha creado una tabla de secuencias alineadas. Las posiciones altamente variables de los alineamientos (añaden ruido al árbol filogenético) han sido filtradas con el *plugin mask*. Finalmente se ha generado el árbol con esas secuencias filtradas, gracias al plugin *phylogeny* y al método *fasttree*, seguido del método *midpoint-root*, generándose al final del proceso un árbol con raíz [19]. Este trabajo previo, ha permitido el estudio de la diversidad alfa y beta, mediante métricas filogenéticas y no filogenéticas, con el *plugin diversity* empleando la pipeline *core-metrics-phylogenetic*. El método rarifica la tabla de frecuencias de nuestras ASV a la profundidad de 1798 (determinada durante la visualización de la tabla de frecuencias). Esto significa que extraerá muestras al azar sin reemplazo de cada muestra y todas terminarán teniendo la misma profundidad, porque las métricas de diversidad son sensibles a las diferentes profundidades de muestreado a lo largo de las distintas muestras, y computará las siguientes medidas de diversidad alfa y beta.

Diversidad alfa:
Índice de diversidad de Shannon
OTUs observadas
Diversidad de Faith
Igualdad o igualdad de Pielou

Diversidad beta:
Distancia de Jaccard
Distancia de Bray-Curtis
Distancia *unweighted UniFrac*
Distancia *weighted UniFrac*

Se ha analizado la composición microbiológica de las muestras según su grupo de pertenencia (diabetes, periodontitis) con el *plugin diversity,* con el método *diversity alpha-group-significance* sobre las matrices con las medidas de diversidad alfa obtenidas y se han aplicado tests Kruskal Wallis sobre ellas, para determinar si existen diferencias estadísticamente significativas entre los cuatro grupos de sujetos (según la presencia o no de diabetes y periodontitis). A continuación se han

generado las curvas rarefacción alfa, en las que se representa la diversidad alfa en función de la profundidad de muestreo, para comprobar que el muestreo realizado a la profundidad de secuencias indicada fue el idóneo para registrar toda la diversidad de las muestras [19].

Los resultados de beta diversidad generados, han sido visualizados con QIIME2 mediante gráficos generados por análisis de coordenadas principales (PCoA) gracias a EMPEROR [26]. Las matrices con medidas de diversidad beta, se han analizado en busca de diferencias de composición según su grupo de pertenencia mediante PERMANOVA. Dado que también es posible su cálculo en QIIME2 mediante el método *beta-group-significance* del *plugin diversity*, se ha usado para analizar cada una de las métricas obtenidas para la diversidad beta de acuerdo a la columna grupo del archivo con los metadatos [27].

Por otro lado, en RStudio se ha realizado una ordenación restringida, con la ayuda de los paquetes *biomformat* y *vegan*. Para ello se ha cargado el archivo de metadatos y la tabla de frecuencias ASV en RStudio, que ha debido de exportarse de QIIME2. Tras un procesamiento de la tabla (transposición, cálculo de abundancias relativas y selección de las características que aparecen en al menos el 10% de las muestras), se ha empleado la función CCA (*Canonical Correspondence Analysis*) para realizar un análisis de gradiente directo, en el que se ha relacionado las especies con cuatro variables de nuestro archivo de metadatos (HbA1c, PBS, edad y AL). Este análisis se diferencia del análisis de correspondencia en que los *scores* de las muestras están restringidos a ser combinaciones lineares de las variables seleccionadas. Por ello CCA explica menos variación que el análisis de correspondencia [28].

Finalmente, para abordar el análisis de la composición taxonómica de las muestras según nuestros grupos de interés; con QIIME2 se ha asignado taxonomía a las secuencias de la tabla de secuencias representativas generada por *dada2,* gracias a un clasificador Naive Bayes pre-entrenado con nuestras muestras. Para entrenar al clasificador, se han descargado los archivos de la base de datos Greengenes (versión 13_8) de la dirección: https://docs.qiime2.org/2018.2/data-resources/. En concreto se ha importado el archivo fasta con las secuencias de referencia (de los OTUs al 99% de esta base de datos) y el archivo con las clasificaciones taxonómicas correspondientes a las anteriores secuencias. A continuación se ha empleado el *plugin feature-classifier* con el método *extract-reads* para obtener unas secuencias de referencia de la base de datos en función de los *primers* que hemos usado, fwd:AGAGTTTGATCCTGGCTCAG y rev:TTACCGCGGCTGCTGGCAC (que amplifican las regiones hipervariables V1 y V3) y la longitud final de las secuencias(358 pb), para aumentar la precisión del clasificador. Después se ha empleado el método *fit-classifier-naive-bayes* para entrenar el clasificador Naive Bayes con esas secuencias y la taxonomía

de referencia [19]. Con el método *classify-sklearn* del *plugin feature-classifier* empleando nuestro clasificador entrenado y la tabla de secuencias representativas, se ha obtenido una tabla con la asignación taxonómica a cada secuencia junto con la confianza a que fue clasificado. Este archivo ha permitido la creación de gráficos de barras interactivos, empleando la opción *qiime taxa barplot* con los que explorar la composición taxonómica de nuestras muestras.

Para testar si existen taxones diferencialmente abundantes entre los diferentes grupos de estudio, se ha aplicado el test ANCOM (*analysis of composition of microbiomes*), que asume que pocos de los taxones cambiarán entre los grupos (menos de un 25%). Primero se han creado tablas con taxones, con el *plugin taxa* siguiendo el método *collapse*, para agrupar grupos con características que tienen la misma asignación taxonómica a un nivel determinado (2 para filo, 6 para género y 7 para especie). Se han sumado las frecuencias de todas las características de la tabla de ASV al agruparse con las anotaciones taxonómicas creada en la clasificación taxonómica. Este *plugin* se ha aplicado sobre dos tipos de tabla, una que contenía solo población diabética y otra que contenía la población no diabética. Con el *plugin composition add-pseudocount*, se han modificado esas tablas para que reflejen frecuencias de los taxones por muestra, pero para evitar frecuencias de cero se ha añadido un valor de 1, obteniendo una tabla con pseudocuentas. Se ha aplicado el test ANCOM (con el *plugin composition ancom*) sobre cada tabla, seleccionando la columna periodontitis, para determinar qué taxones difieren en abundancia en cada uno de los cuatro grupos [19, 29].

Para la inferencia de funciones mediante la herramienta PICRUST. Primero se han procesado los datos que necesita de entrada, para adaptar los archivos a PICRUST. Se ha realizado una selección de taxones de referencia cerrada usando la versión de Greengenes para la que PICRUST fue entrenado (versión 13_5), para ello se ha empleado el *plugin vsearch*, siguiendo el método *cluster-features-closed-reference* y obteniendo una tabla con Greengenes IDs compatible con PICRUST, que se ha exportado para tener un archivo en formato biom. A este archivo se le han añadido las anotaciones taxonómicas de Greengenes usando *biom add-metadata*. El resto del proceso se ha realizado en Galaxy, donde se ha normalizado el archivo por número de copia de 16S. La tabla resultante se ha usado como input para predecir el metagenoma, con el módulo *predict metagenome,* que ha inferido un metagenoma de abundancias de ortólogos KEGG para cada muestra en la tabla normalizada y ha generado una tabla con las métricas para el metagenóma predicho en valores NSTI (Nearest Sequenced Taxon Index). Después se ha empleado el módulo *categorize by function*, que ha permitido examinar los resultados de KEGG desde un nivel más alto en la jerarquía de la ruta. Pues un ortólogo KEGG puede estar implicado en diferentes rutas y en ese caso el ortólogo KEGG sería contado por cada ruta [20].
La tabla de ortólogos KEGG, se ha usado para realizar un análisis de ortólogos diferencialmente expresados con el paquete *limma* de

Bioconductor en RStudio, empleándose un modelo lineal general, con un método para obtener una estimación mejorada de la varianza. Nos hemos decantado por este paquete porque permite hacer múltiples comparaciones, mientras que otros paquetes como *edgeR* o *DESeq*, están diseñados para una comparación entre dos niveles, como es el caso de control versus tratamiento. Se ha especificado la matriz de diseño y la matriz de contraste. En este caso, para la matriz de diseño se ha optado por un modelo de un factor con cuatro niveles (uno por cada grupo de pertenencia de los sujetos que pertenecieron en el estudio). Las preguntas a contestar (el efecto de la periodontitis en la diabetes en la expresión génica, su efecto en ausencia de diabetes y el efecto de la diabetes en los genes en ausencia de periodontitis) por el modelo se han formulado como contrastes, que han consistido en comparaciones entre los parámetros del modelo. La tabla con los genes diferencialmente expresados, obtenida con método *decidetest*, ha permitido crear un *heatmap* para su visualización gráfica y por otro lado, las tablas de ortólogos *up* y *downregulados* para cada comparación se han guardado. Este análisis se ha repetido para las *pathways* KEGG y para los COGs [30, 31].

3. Resultados

Análisis de diversidad alfa

Diversidad de Faith

Primero se ha decidido confirmar si existía un sesgo en cuanto a género, para esta métrica de diversidad alfa. Para ello se ha representado la distribución que seguían las muestras para esta métrica en el grupo de mujeres y en el grupo de hombres. En la gráfica (figura 2) se aprecia que no hay grandes diferencias en las distribuciones, si bien se observa una tendencia en el grupo de mujeres a presentar una mediana ligeramente superior.

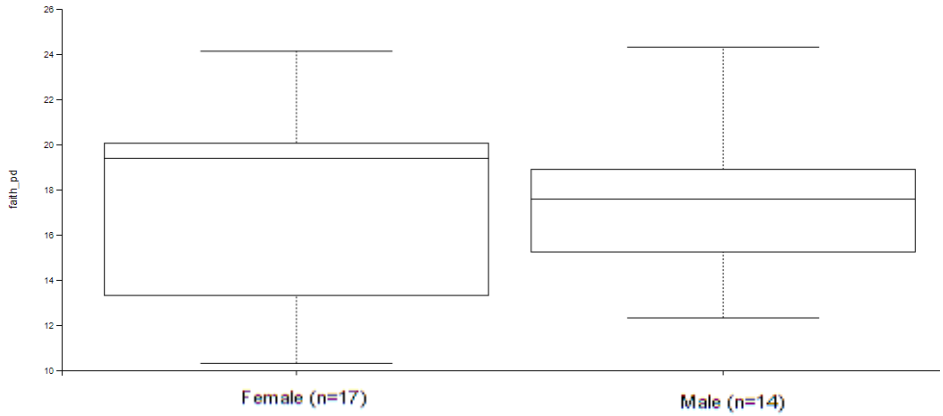

Figura 2. Gráfico de significación de diversidad de Faith dividida según el género de las muestras del estudio.

Los análisis no paramétricos aplicados sobre las muestras divididas en función del género en cuanto a esta métrica de diversidad alfa, han corroborado que no existen diferencias en esta medida de riqueza en cuanto a género.

Resultado Kruskal-Wallis (todos los grupos)	
H	0.0567
p-valor	0.8117

Tabla 1. Test de Kruskal-Wallis para la diversidad de Faith, dividida en dos grupos en función del género de las muestras.

Seleccionando nuestros grupos de interés, en el gráfico (figura 3) sí se pueden observar difererencias. Siendo el grupo con diabetes y sin periodontitis el que menos riqueza filogénetica de Faith presenta y el grupo con dibetes y periodontitis, el que mayor riqueza de esta métrica presenta junto con el grupo control. Por lo que el grupo con diabetes sin periodontitis es el que menor variedad de especies desde el punto de vista filogenético presenta en el grupo, en relación a esta métrica. Cabe destacar que la condición de diabetes y de periodontitis, presentadas por separado, parecen disminuir la riqueza filogenética de Faith respecto al grupo control; mientras que cuando ambas condiciones se presentan conjuntamente la riqueza de esta métrica es equivalente al grupo control.

Los test estadísticos (tabla 2) han confirmado la existencia de diferencias estadísticamente significativas entre el grupo de diabéticos con periodontitis y el de diabéticos sin esta condición.

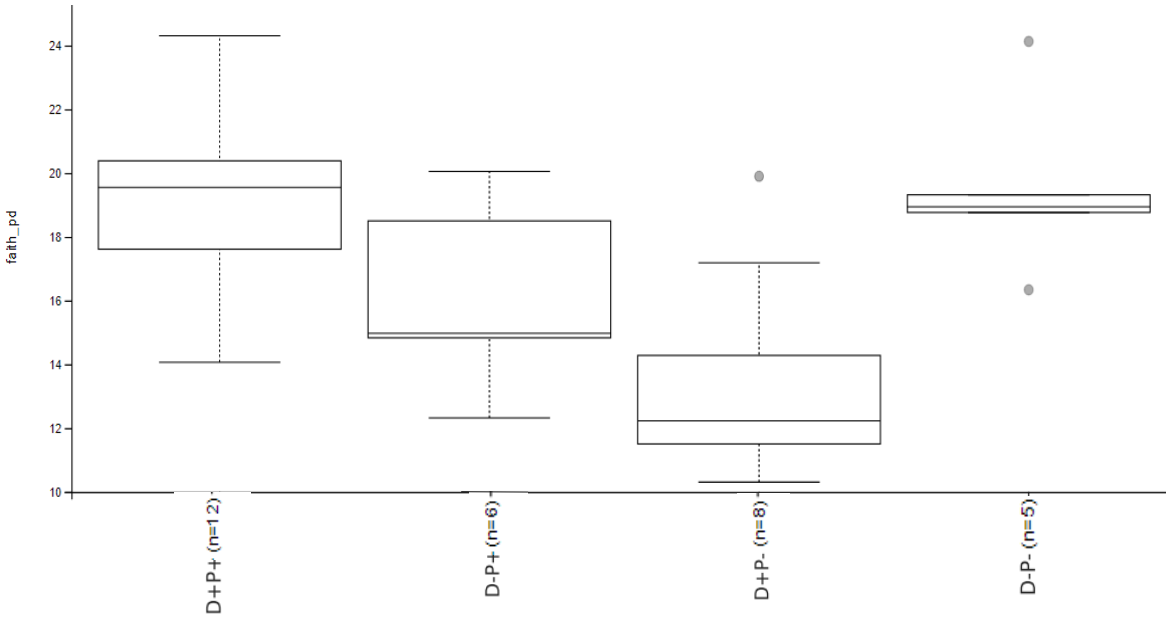

Figura 3. Gráfico de significación de diversidad de Faith dividida según el grupo (diabetes, periodontitis) al que pertenecen las muestras del estudio.

Resultado Kruskal-Wallis (todos los grupos)	
H	11.4042
p-valor	0.0097

Kruskal-Wallis (por pares)		**H**	**p-valor**	**q-valor**
Group 1	**Group 2**			
D+P+	**D-P+**	3.166667	0.0752	0.1399
	D+P-	8.595238	0.0034	0.0202
	P- D-	0.044444	0.8330	0.8330
D-P+	**D+ P-**	2.816667	0.0933	0.1399
	D- P-	1.633333	0.2012	0.2415
D+P-	**D- P-**	4.821429	0.0281	0.0843

Tabla 2. Test de Kruskal-Wallis para la diversidad de Faith según el grupo (diabetes, periodontitis) al que pertenecen las muestras del estudio.

OTUs observadas

Al igual que en la anterior métrica, la agrupación por género no es estadísticamente significativa (no mostrada). En cuanto a los grupos de interés (figura 4), la tendencia es similar a la anterior métrica. Observándose menor riqueza en el grupo de diabéticos sin periodontitis y mayor en el de diabéticos con periodontitis, siendo ésta última más similar a la del grupo control, aunque en el gráfico se observa una tendencia en el grupo control a presentar más riqueza de OTUs.

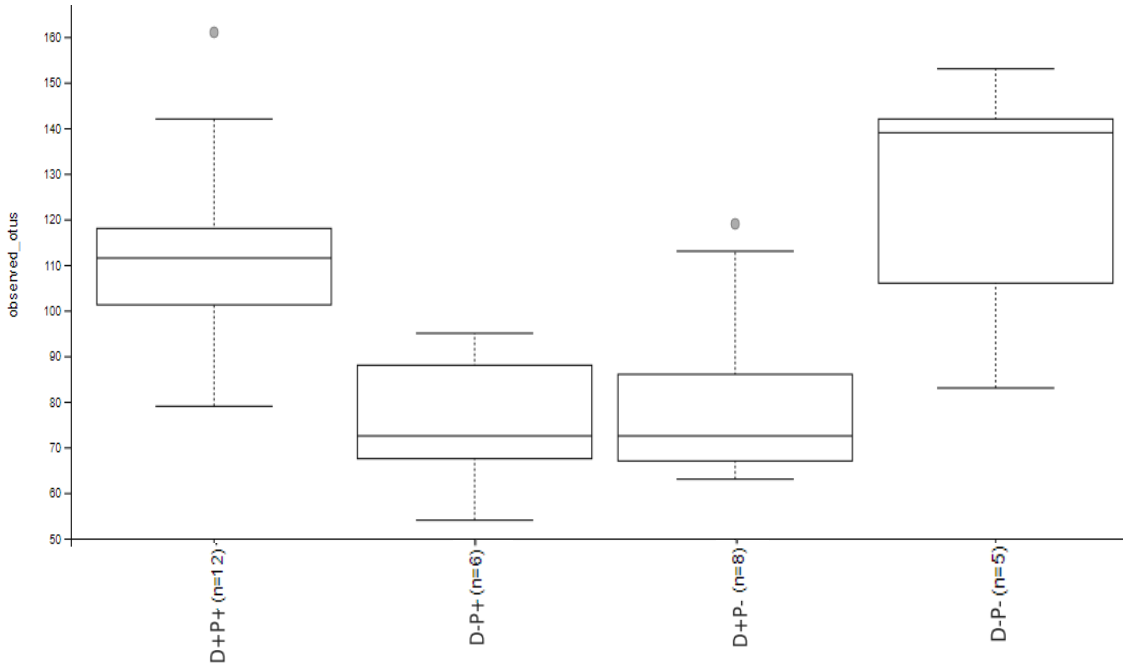

Figura 4. Gráfico de significación de diversidad de OTUs observadas dividida según el grupo (diabetes, periodontitis) al que pertenecen las muestras del estudio.

Resultado Kruskal-Wallis (todos los grupos)	
H	14.5362
p-valor	0.0023

Kruskal-Wallis (por pares)		H	p-value	q-value
Grupo 1	**Grupo 2**			
D+P+	**D- P+**	9.8545	0.0017	0.0102
	D+ P-	6.0952	0.0136	0.0288
	D- P-	0.5458	0.4600	0.5521
D-P+	**D+ P-**	0.0667	0.7963	0.7963
	D- P-	5.633333	0.0176	0.0288
D+P-	**D- P-**	5.485714	0.0192	0.0288

Tabla 3. Test de Kruskal-Wallis para la diversidad de OTUs observadas según el grupo (diabetes, periodontitis) al que pertenecen las muestras del estudio.

El test estadístico de Kruskal-Wallis (tabla 3), ha confirmado la existencia de diferencias estadísticamente significativas entre el grupo de diabéticos con periodontitis frente al de diabéticos sin esta condición y frente al grupo sin diabetes y con periodontitis. Por otra parte, también se han encontrado diferencias significativas entre el grupo control y el grupo de diabéticos sin periodontitis y el grupo control y el grupo de no diabéticos con periodontitis.

Índice de diversidad de Shannon

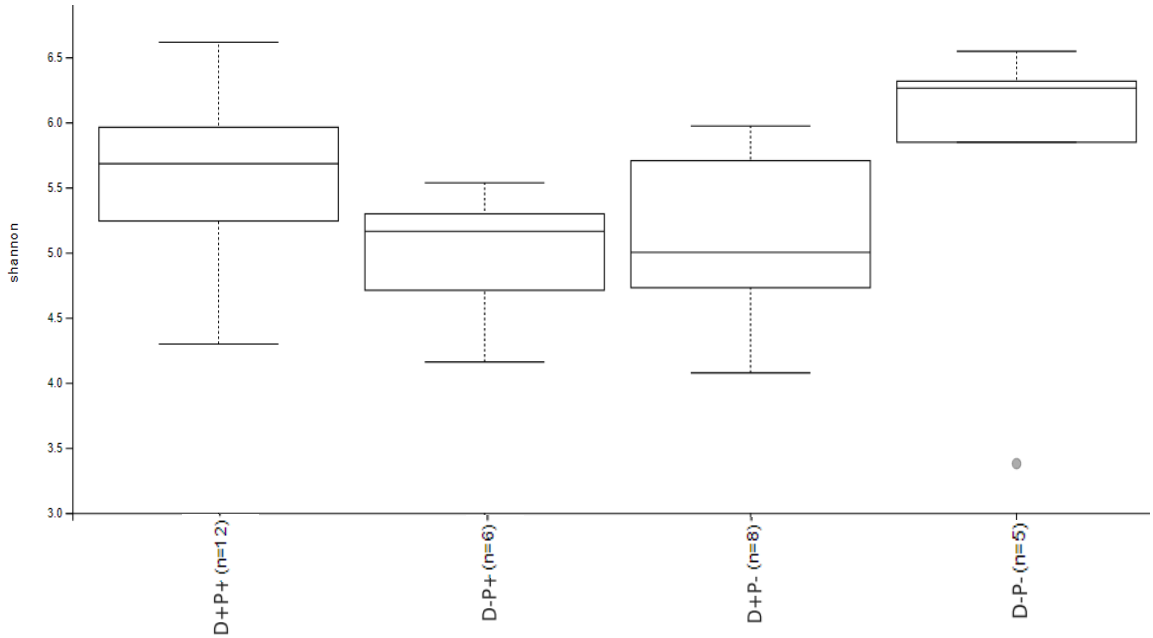

Figura 5. Gráfico de significación de diversidad de Shannon dividida según el grupo (diabetes, periodontitis) al que pertenecen las muestras del estudio.

Resultado Kruskal-Wallis (todos los grupos)	
H	6.3013
p-valor	0.0978

Tabla 4. Test de Kruskal-Wallis para la diversidad de Shannon según el grupo (diabetes, periodontitis) al que pertenecen las muestras del estudio.

A diferencia de los anteriores, esta métrica de diversidad alfa no muestra diferencias significativas entre los distintos grupos (tabla 4), aunque los diagramas de cajas muestran una tendencia similar al de los anteriores gráficos (figura 5).

Igualdad de Pielou

Tanto el gráfico de la distribución de las muestras para la métrica de igualdad de Pielou dividida en función del género (figura 6), como el test estadístico aplicado (tabla 5), muestran que las muestras sí presentan diferencias en igualdad en cuanto a género. Presentando los varones mayor igualdad. Aunque el efecto de la diferente igualdad entre géneros, se vería controlado por el diseño experimental, en el que se ha incluido prácticamente el mismo número de mujeres que de hombres por cada grupo de interés, según presente o no diabetes y periodontitis.

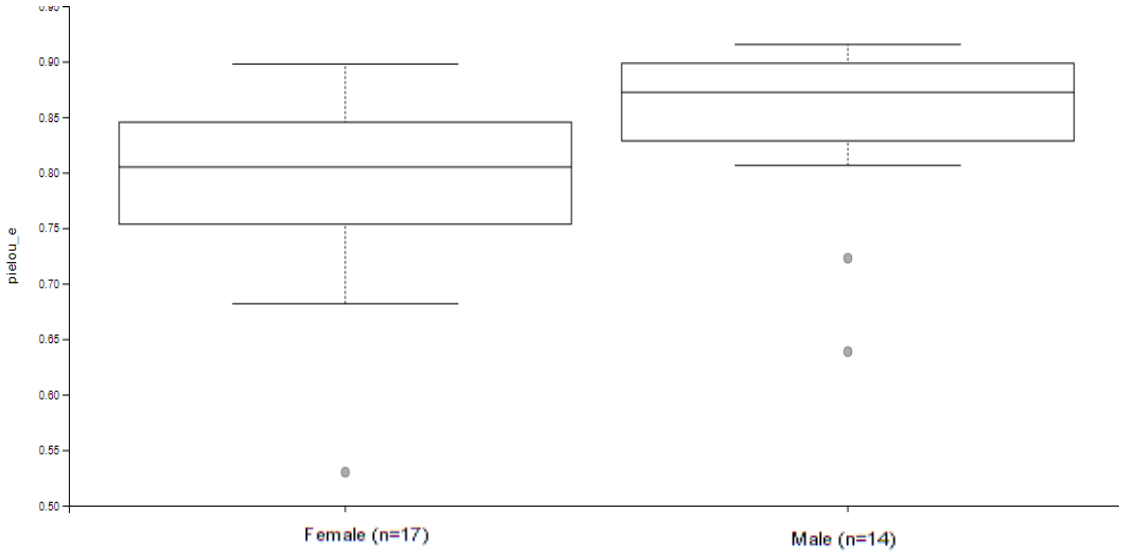

Figura 6. Gráfico de igualdad de Pielou dividida según el género de las muestras del estudio.

Resultado Kruskal-Wallis (todos los grupos)	
H	6.6570
p-valor	0.0099

Tabla 5. Test de Kruskal-Wallis para la igualdad de Pielou dividida en dos grupos en función del género de las muestras.

Sin embargo, cuando se han dividido las muestras en función de los cuatro grupos de interés, no se han detectado diferencias en igualdad respecto a los mismos (figura 7, tabla 6).

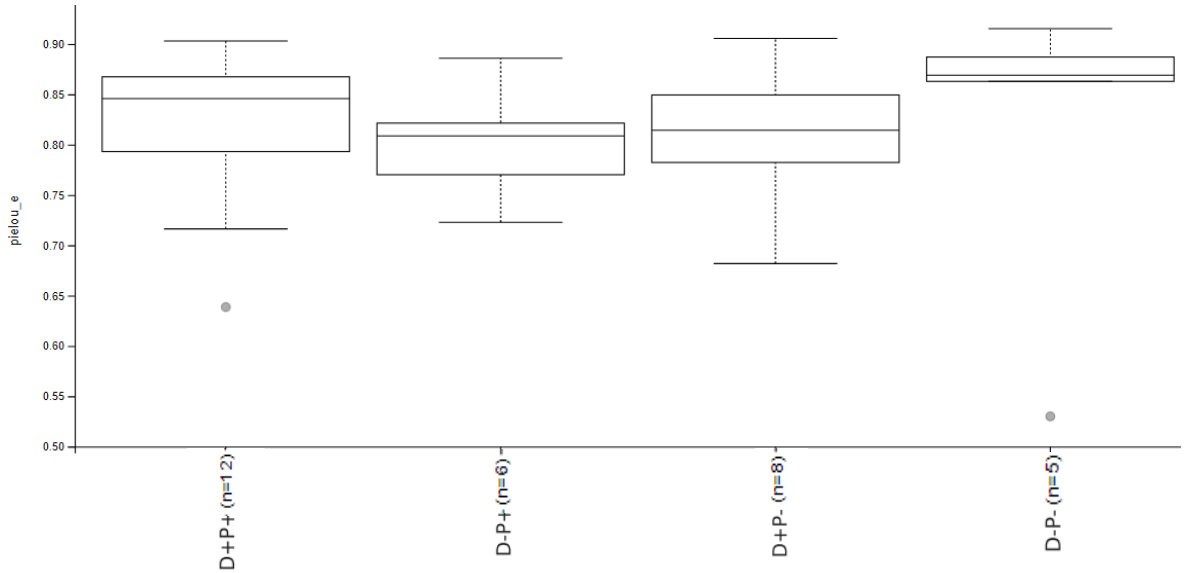

Figura 7. Gráfico de igualdad de Pielou dividida según el grupo (diabetes, periodontitis) al que pertenecen las muestras del estudio.

Resultado Kruskal-Wallis (todos los grupos)	
H	2.4421
p-valor	0.4858

Tabla 6. Test de Kruskal-Wallis para la igualdad de Pielou según el grupo (diabetes, periodontitis) al que pertenecen las muestras del estudio.

Alfa rarefacción

Los gráficos de rarefacción para las tres métricas diferentes de riqueza (figuras 8, 9,10), parecen indicar que la profundidad de muestreo escogida fue correcta, pues mayor profundidad no resultaría en una observación de características adicionales ya que la riqueza de las muestras al encontrarse en la fase *plateau*, se ha observado plenamente.

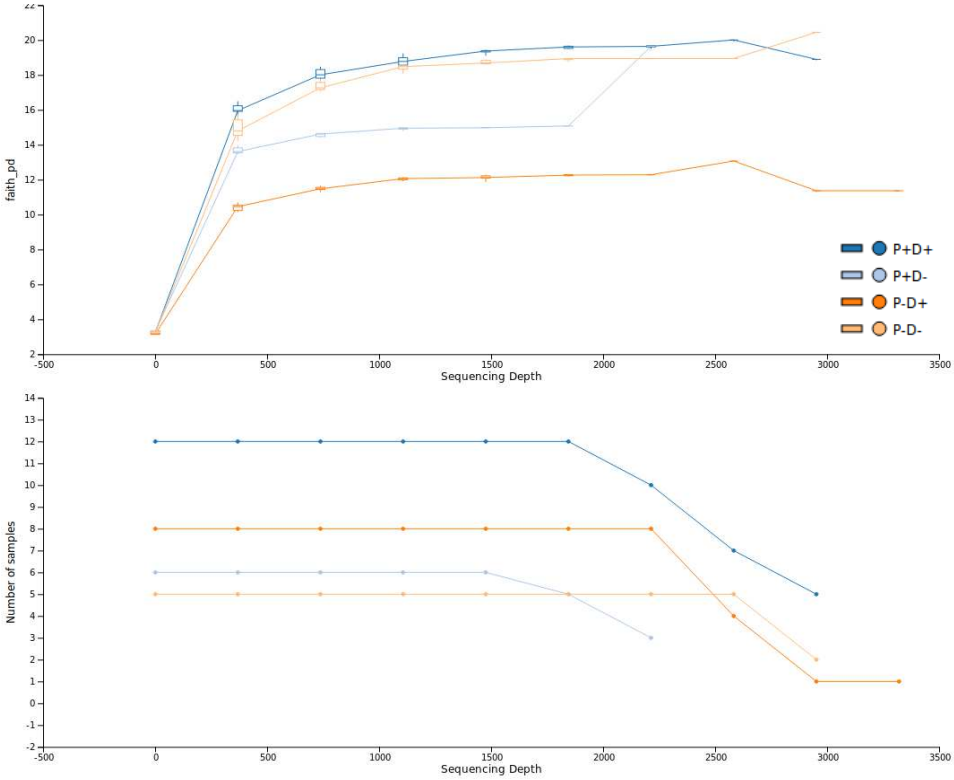

Figura 8. Gráficos de rarefacción. El gráfico superior muestra la diversidad de Faith para cada uno de los grupos de interés (diabetes, peiodontitis) en función de la profundidad de muestreo, mientras que el gráfico inferior muestra el número de muestras en función de la profundidad de muestreo.

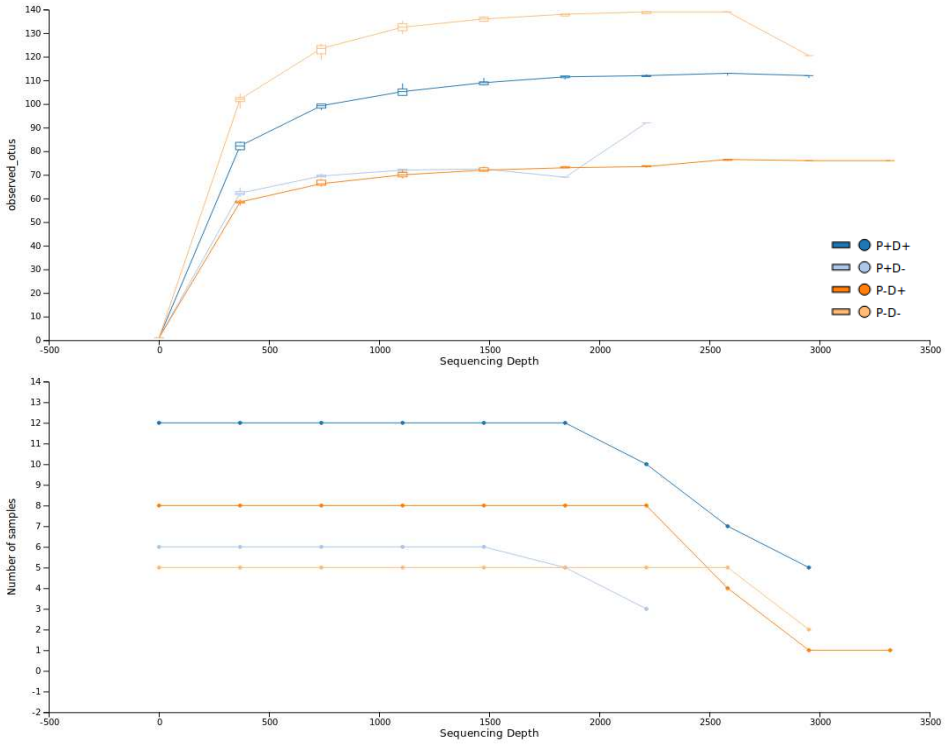

Figura 9. Gráficos de rarefacción. El gráfico superior muestra los OTUs observados para cada uno de los grupos de interés (diabetes, periodontitis) en función de la profundidad de muestreo, mientras que el gráfico inferior muestra el número de muestras en función de la profundidad de muestreo.

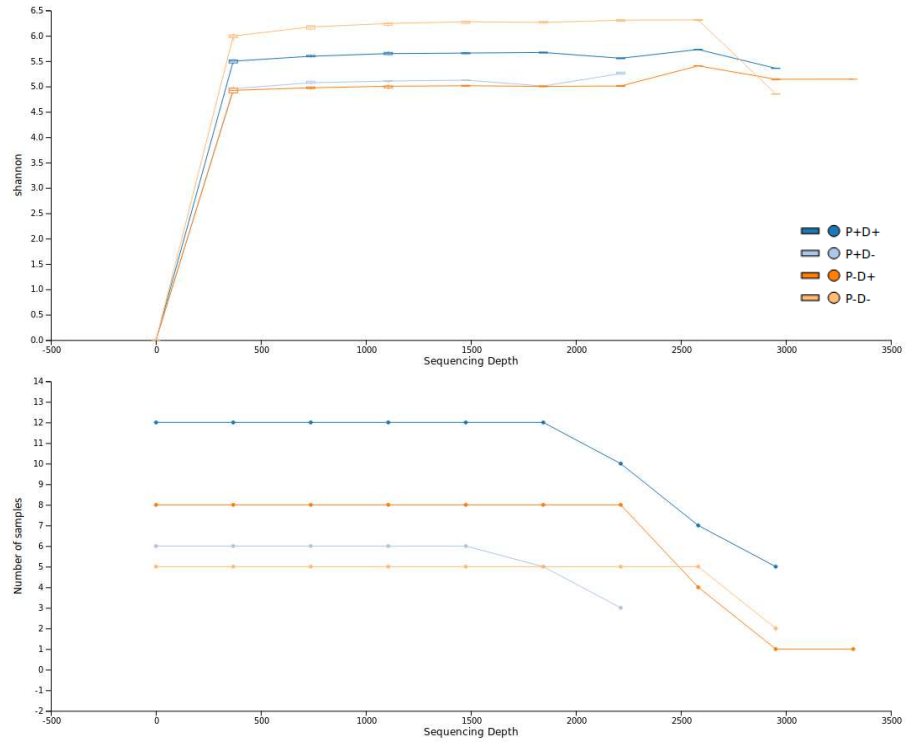

Figura 10. Gráficos de rarefacción. El gráfico superior muestra la diversidad de Shannon para cada uno de los grupos de interés (diabetes, periodontitis) en función de la profundidad de muestreo, mientras que el gráfico inferior muestra el número de muestras en función de la profundidad de muestreo.

Diversidad beta

Además de la diversidad alfa, también se han analizado diferentes métricas de diversidad beta, puesto que este tipo de diversidad nos informa del grado de diferenciación entre comunidades biológicas [32]. En este estudio, será de gran utilidad para saber cuánto de distintos son los cuatro grupos de interés.

Distancia de Jaccard

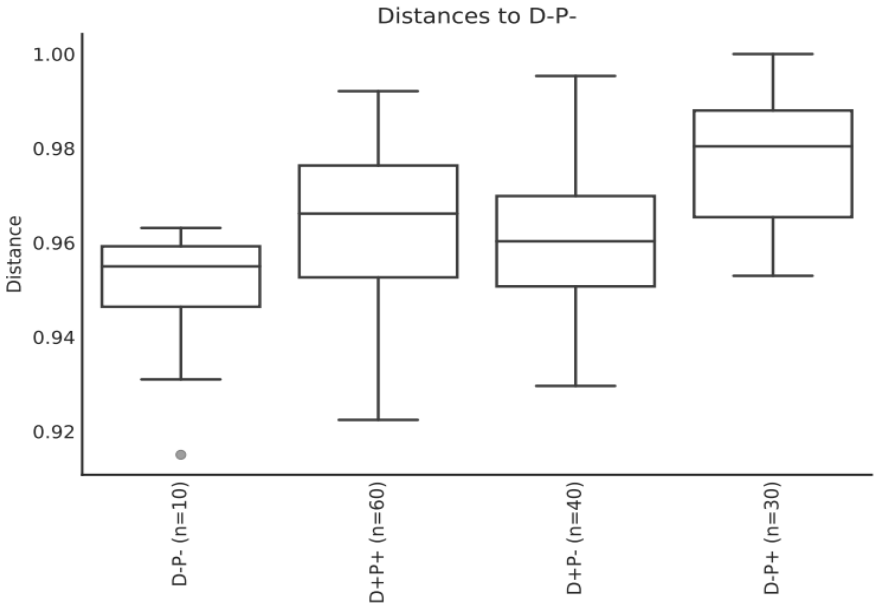

Figura 11. Gráfico que muestra la distancia de Jaccard entre cada grupo de interés en el estudio (presencia o no de diabetes y periodontitis).

En el gráfico de la figura 11, podemos observar como las muestras diabéticas con periodontitis y no diabéticas con periodontitis son las que muestran más distancia respecto al grupo control. Por otro lado, el grupo con diabétes y periodontitis presenta también mayor distancia de Jaccard respecto al grupo que presenta exclusivamente diabetes, aunque en menor medida. El test estadístico aplicado para confirmar si estas diferencias observadas son significativas ha sido el test PERMANOVA (tabla 7).

Método	PERMANOVA
Nombre test estadístico	pseudo-F
Test estadístico	1.17798
p-valor	0.001
Nº permutaciones	999

Resultados permanova por pares		Tamaño muestral	Permutaciones	pseudo-F	p-valor	q-valor
Grupo 1	**Grupo 2**					
D+P+	**D+P-**	20	999	1.32802	0.001	0.006
	D-P+	18	999	1.06254	0.038	0.038
	D-P-	17	999	1.08297	0.034	0.038
D+P-	**D-P+**	14	999	1.30928	0.003	0.009
	D-P-	13	999	1.10815	0.036	0.038
D-P+	**D-P-**	11	999	1.15740	0.006	0.012

Tabla 7. Test de permanova para la distancia Jaccard entre los grupos de interés en el estudio

En base a la distancia de Jaccard, el test PERMANOVA confirma que todos los grupos difieren en composición. Presentando más disimilaridad (desde el punto de vista cualitativo de esta métrica) el grupo de diabéticos con periodontitis que el de diabéticos sin esta condición y siendo estos más diversos que el grupo control. Aunque la mayor distancia fue presentada por el grupo de pacientes con periodontitis no diabéticos, que son los más disimilares respecto al grupo control.

Distancia de Bray-Curtis

Esta métrica de diversidad beta, proporciona una medida cuantitativa de las diferencias en diversidad entre los cuatro grupos de interés en este estudio.

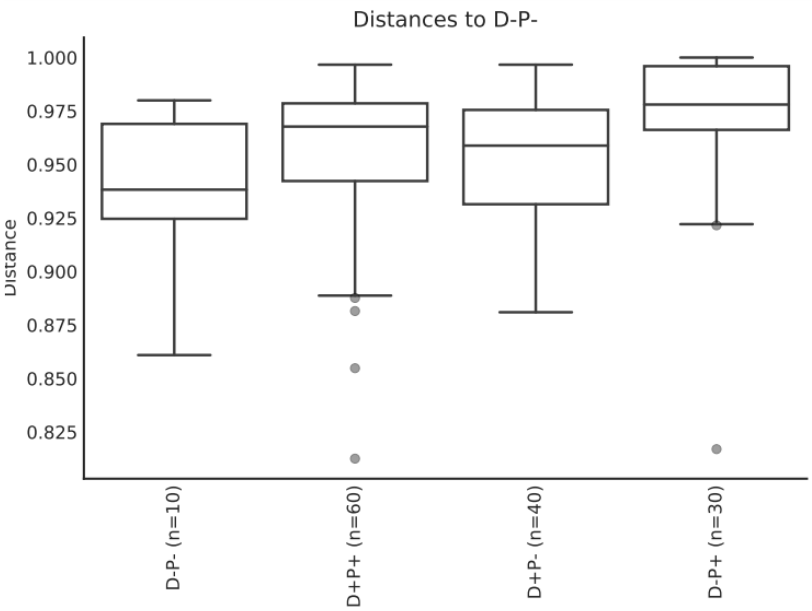

Figura 12. Gráfico que muestra la distancia de Bray-Curtis entre cada grupo de interés en el estudio (presencia o no de diabetes y periodontitis).

Método	PERMANOVA
Nombre test estadístico	pseudo-F
Test estadístico	1.25251
p-valor	0.001
Nº permutaciones	999

Resultados permanova por pares		**Tamaño muestral**	**Permutaciones**	**pseudo-F**	**p-valor**	**q-valor**
Grupo 1	**Grupo 2**					
D+P+	**D+P-**	20	999	1.63320	0.001	0.0060
	D-P+	18	999	0.95165	0.686	0.6860
	D-P-	17	999	1.06733	0.183	0.2196
D+P-	**D-P+**	14	999	1.46939	0.004	0.0120
	D-P-	13	999	1.17692	0.058	0.0870
D-P+	**D-P-**	11	999	1.16458	0.046	0.0870

Tabla 8. Test de PERMANOVA para la distancia Bray-Curtis entre los grupos de interés en el estudio

Como se puede observar en la figura 12, desde el punto de vista de la métrica de Bray Curtis, las diferencias no son tan marcadas. Aunque el grupo de diabetes con periodontitis presenta una mayor distancia, respecto a esta métrica, que el grupo de diabéticos sin periodontitis. El test

PERMANOVA (tabla 8), confirma que esta diferencia es estadísticamente significativa. También confirma que hay una diferencia significativa entre el grupo de diabéticos sin periodontitis y el de no diabéticos con periodontitis, indicativo de que el número de organismos compartidos entre ambas comunidades será menor.

Distancia *unweighted UniFrac*

UniFrac es una medida de β-diversidad que usa información filogenética para comparar muestras pertenecientes a los cuatro grupos de interés en este caso. La versión *unweighted* es cualitativa, porque se basa en la presencia o ausencia en los ASV observados [25].

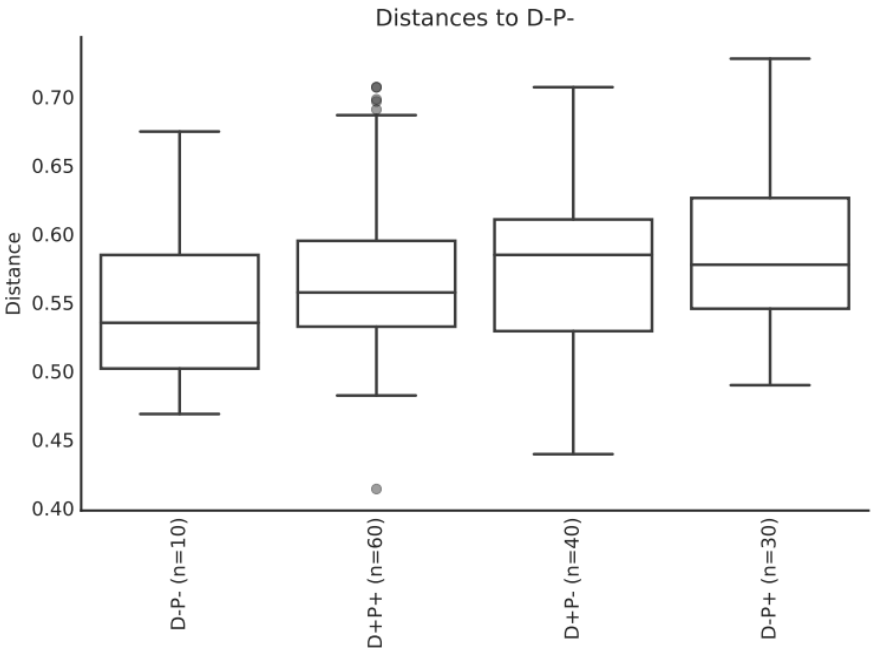

Figura 13. Gráfico que muestra la distancia *unweighted UniFrac* entre cada grupo de interés en el estudio (presencia o no de diabetes y periodontitis).

Método	PERMANOVA
Nombre test estadístico	pseudo-F
Test estadístico	2.37754
p-valor	0.001
Nº permutaciones	999

Resultados permanova por pares		**Tamaño muestral**	**Permutaciones**	**pseudo-F**	**p-valor**	**q-valor**
Grupo 1	**Grupo 2**					
D+P+	**D-P+**	18	999	1.281114	0.077	0.077
	D+P-	20	999	3.569410	0.002	0.004
	D-P-	17	999	1.562596	0.016	0.024
D-P+	**D+P-**	14	999	3.749903	0.002	0.004
	D-P-	11	999	2.251405	0.002	0.004
D+P-	**D-P-**	13	999	1.842904	0.020	0.024

Tabla 9. Test de PERMANOVA para la distancia *unweighted UniFrac* entre los grupos de interés en el estudio

Como resultado de los test de PERMANOVA (tabla 9) sobre la distancia un*weighted UniFrac,* podemos afirmar que existen diferencias en composición entre todos los grupos, a excepción del grupo con diabetes y periodontitis y el grupo sin diabetes con periodontitis. El grupo con diabetes presenta mayor diferencia en composición respecto al grupo control, pero esta diferencia disminuye cuando la diabetes se presenta con periodontitis, aunque sigue siendo estadísticamente significativa.

Distancia *weighted UniFrac*

La versión *weighted de UniFrac,* es una medida cuantitativa de β-diversidad, ya que tiene en cuenta la abundancia de los ASV observados y usa información filogenética para comparar las muestras pertenecientes a los cuatro grupos de interés, en el caso de este estudio [25].

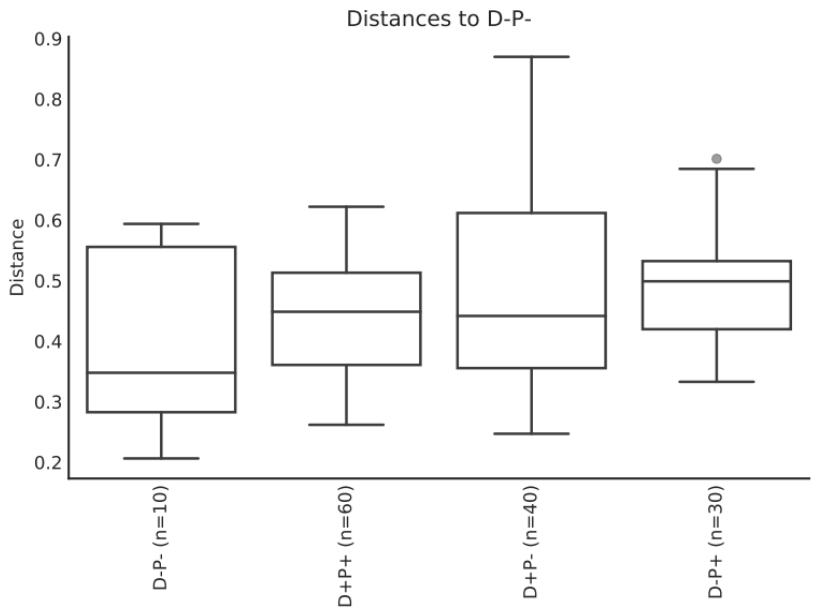

Figura 14. Gráfico que muestra la distancia *weighted UniFrac* entre cada grupo de interés en el estudio (presencia o no de diabetes y periodontitis).

Método	PERMANOVA
Nombre test estadístico	pseudo-F
Test estadístico	5.97549
p-valor	0.001
Nº permutaciones	999

Resultados permanova por pares		**Tamaño muestral**	**Permutaciones**	**pseudo-F**	**p-valor**	**q-valor**
Grupo 1	**Grupo 2**					
D+P+	**D+P-**	20	999	10.686014	0.001	0.0060
	D-P+	18	999	1.285356	0.228	0.2280
	D-P-	17	999	2.671687	0.011	0.0195
D+P-	**D-P+**	14	999	11.869136	0.002	0.0060
	D-P-	13	999	2.628219	0.034	0.0408
D-P+	**D-P-**	11	999	4.791233	0.013	0.0195

Tabla 10. Test de PERMANOVA para la distancia *weighted UniFrac* entre los grupos de interés en el estudio

Los test de PERMANOVA (tabla 10) sobre la distancia *weighted UniFrac* muestran menores diferencias. Aunque se mantienen diferencias en

composición entre el grupo de diabetes con periodontitis y sin periodontitis, que a su vez muestran diferencias significativas en composición (respecto a esta métrica) con el grupo control, siendo ambas mayores.

En general, las diferencias cualitativas de las métricas de diversidad beta entre grupos son más marcadas que las cuantitativas. Una manera de ver esto resumido de forma visual es con los gráficos de ordenación, en este caso de coordenadas principales (PCoA).

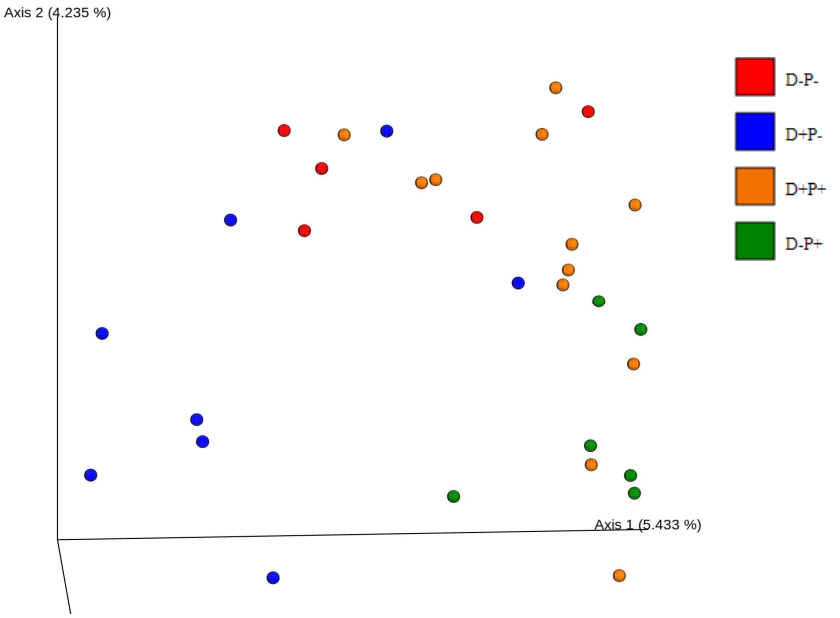

Figura 15. Gráfico de PCoA con la métrica de beta diversidad de Jaccard.

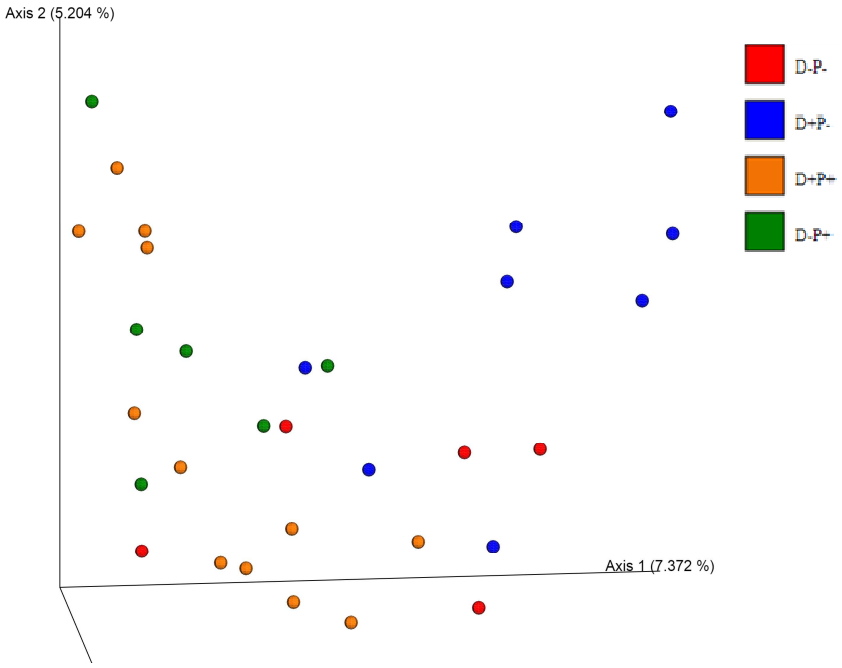

Figura 16. Gráfico de PCoA con la métrica de beta diversidad de Bray Curtis.

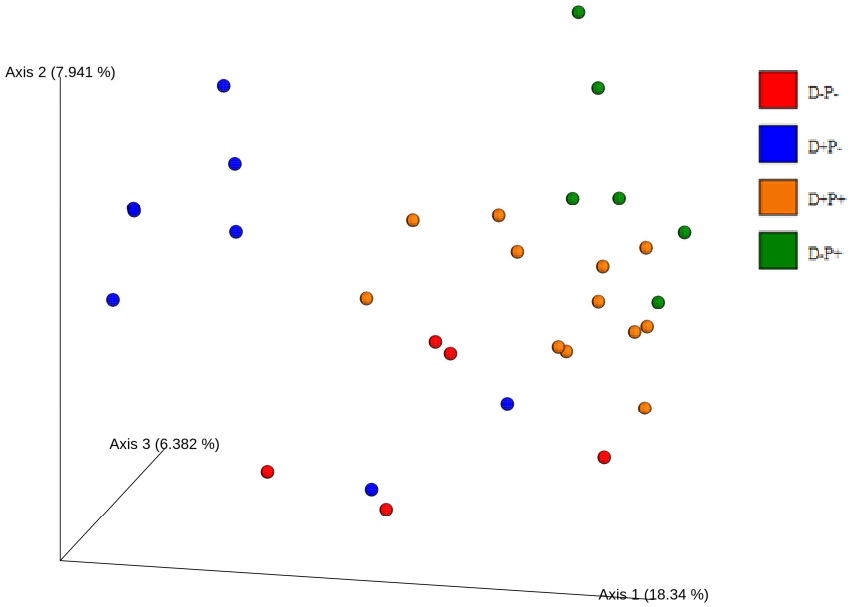

Figura 17. Gráfico de PCoA con la métrica de beta diversidad de un*weighted UniFrac*.

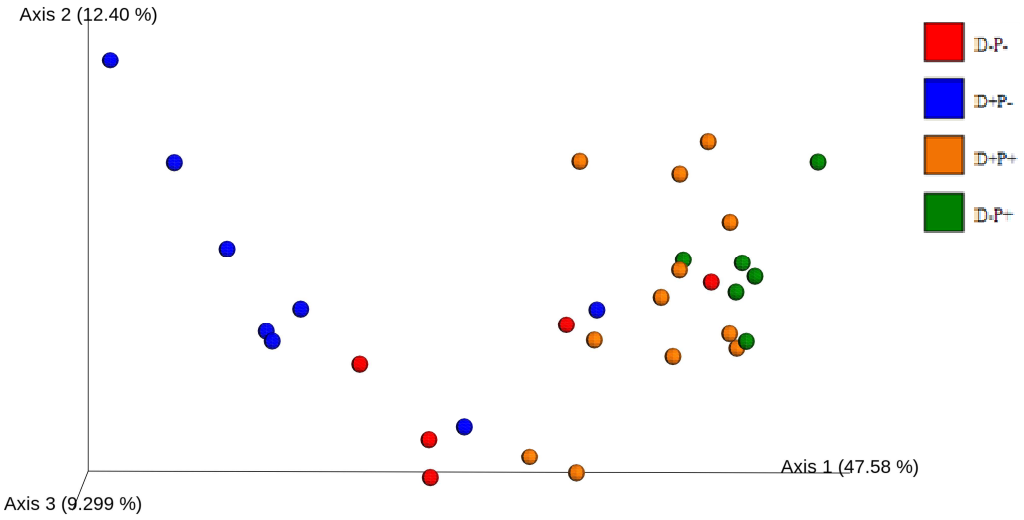

Figura 18. Gráfico de PCoA con la métrica de beta diversidad de *weighted UniFrac*.

En todos los gráficos (figuras 15 a 18) observamos que la comunidad más separada del resto es el grupo de pacientes diabéticos sin periodontitis (en azul). Aunque los cuatro grupos se separan mejor de acuerdo a las métricas de beta diversidad de distancia *UniFrac*, siendo la métrica *weighted UniFrac* la que mejor explica la variabilidad de los datos. Un 47.58% de la variabilidad de los datos es explicada por el primer eje, que separa razonablemente bien los cuatro tipos de grupo creados de acuerdo a la presencia o no de diabetes y periodontitis. También podemos confirmar que en general los gráficos respaldan los resultados obtenidos con los tests PERMANOVA.

Ordenación restringida

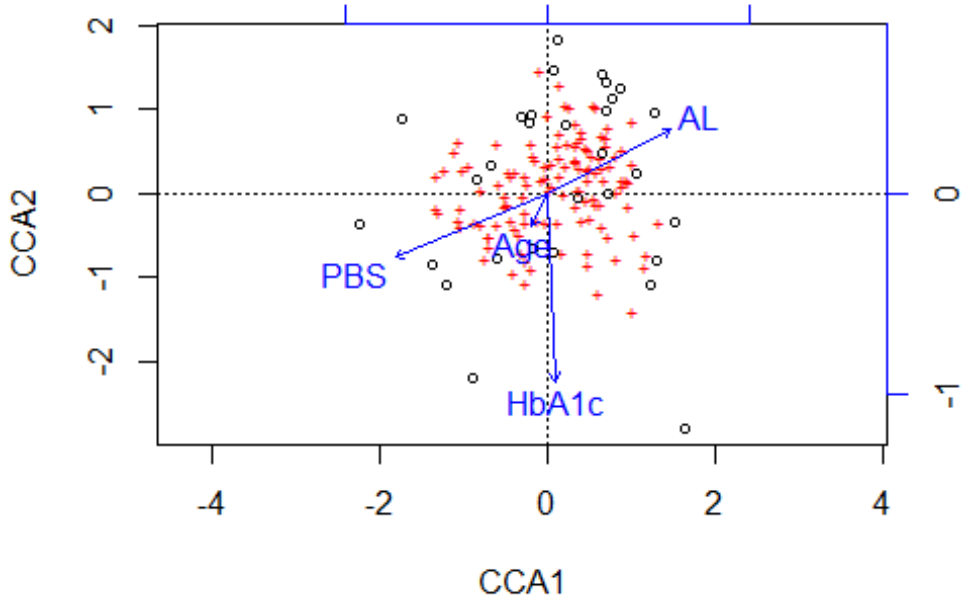

Figura 19. Gráfico de ordenación restringida en función de las variables: *age* (edad), PBS (*Postpandrial Blood Sugar*), HbA1c (hemoglobina glicosilada) y AL (*attachment loss*).

En el gráfico de ordenación restringida, se muestra los ASVs asociados con las dos variables asociadas con la condición de diabetes: la hemoglobina glicosilada y el azúcar en sangre postpandrial (a las dos horas de haber comido) y una variable asociada a la condición de la periodontitis, la pérdida de fijación de los dientes. También se ha decidido tener en cuenta otra posible variable confusora, la edad. Se observa que la periodontitis aumenta a medida que nos desplazamos hacia la derecha mientras que la condición de la diabetes parece aumentan en dirección casi opuesta pues HbA1c y PBS no aumentan exactamente en la misma dirección. Por otro lado, los ASVs más implicados en edad se asocian más con el estado de la diabetes, pues se aprecia que aumenta en la misma dirección que las dos variables asociadas con diabetes, aunque su dimensión es menor. La diabetes se podría asociar más al eje 2 mientras que la periodontitis podría encontrarse un poco más asociada al eje 1.

Composición taxonómica

La asignación taxonómica a la tabla de frecuencias ha permitido la creación de gráficos de barras, donde se representan la frecuencia relativa de los taxones, a un determinado nivel taxonómico, a lo largo de las muestras del estudio. Estos gráficos de barras permiten una exploración visual de la composición taxonómica de las muestras.

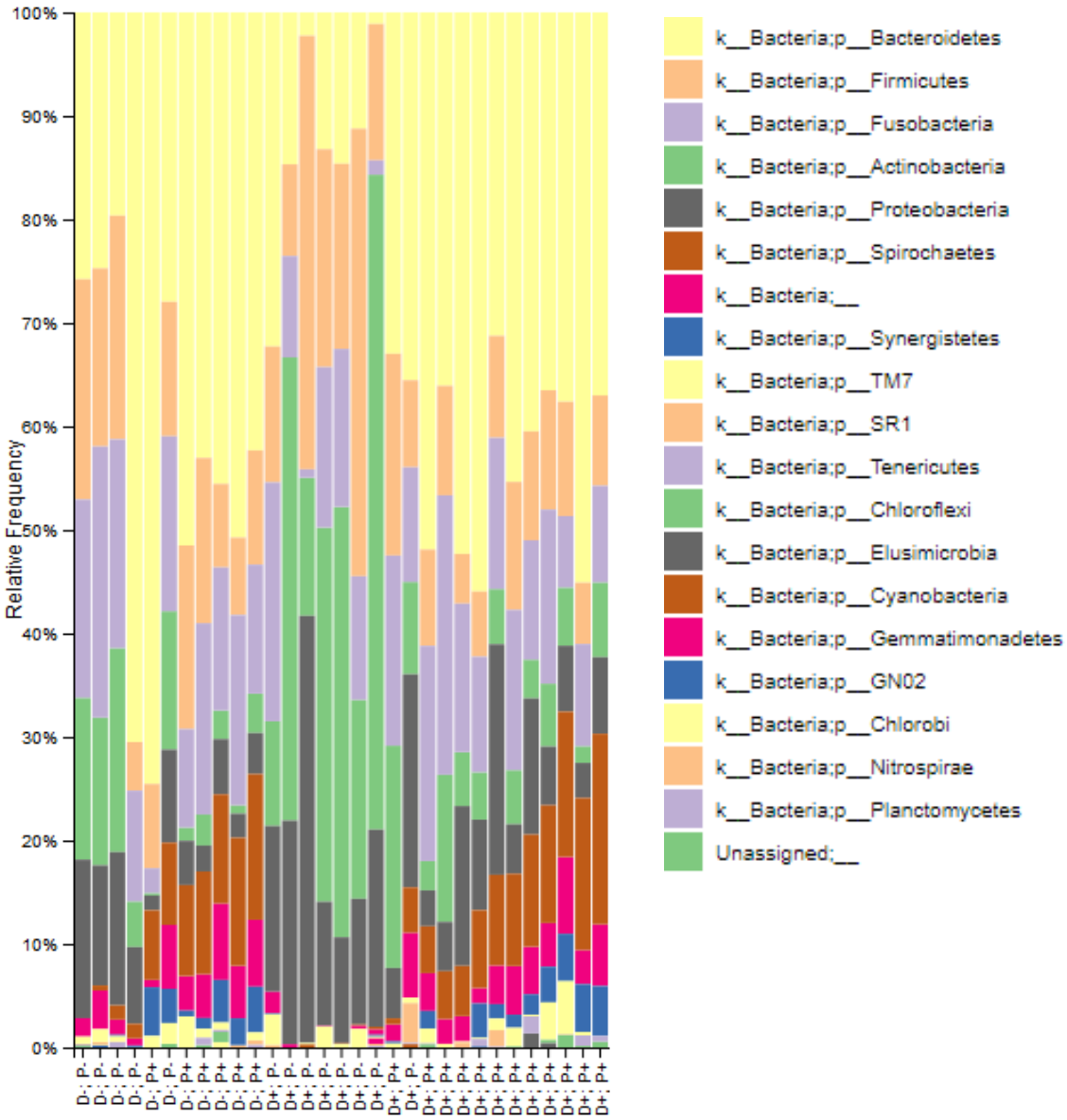

Figura 20. Gráfico de composición taxonómica a nivel de filo. Permite visualizar los filos dominantes en cada grupo.

En la figura 20 podemos apreciar la composición taxonómica de las muestras a nivel de filo, agrupadas según los cuatro grupos de estudio (en función de que presenten o no diabetes y periodontitis). Parece que el filo *Bacteroidetes* es el más abundante y puede destacarse el filo *Spirochaetes* como bastante asociado con la presencia de periodontitis, mientras que el filo *Actinobacteria* puede asociarse con la ausencia de esta condición, pues apacere con mayor frecuencia en muestras sin periodontitis, independientemente de que presenten o no diabetes, aunque parece que aparece con mucha mayor frecuencia en el grupo de diabéticos sin periodontitis. Una situación similar se observa en el filo *Proteobacteria*, que parece presentarse con mayor frecuencia en el grupo con diabetes sin periodontitis.

También se ha explorado la composición taxonómica a distintos niveles, como puede apreciarse en las figuras 21 y 22.

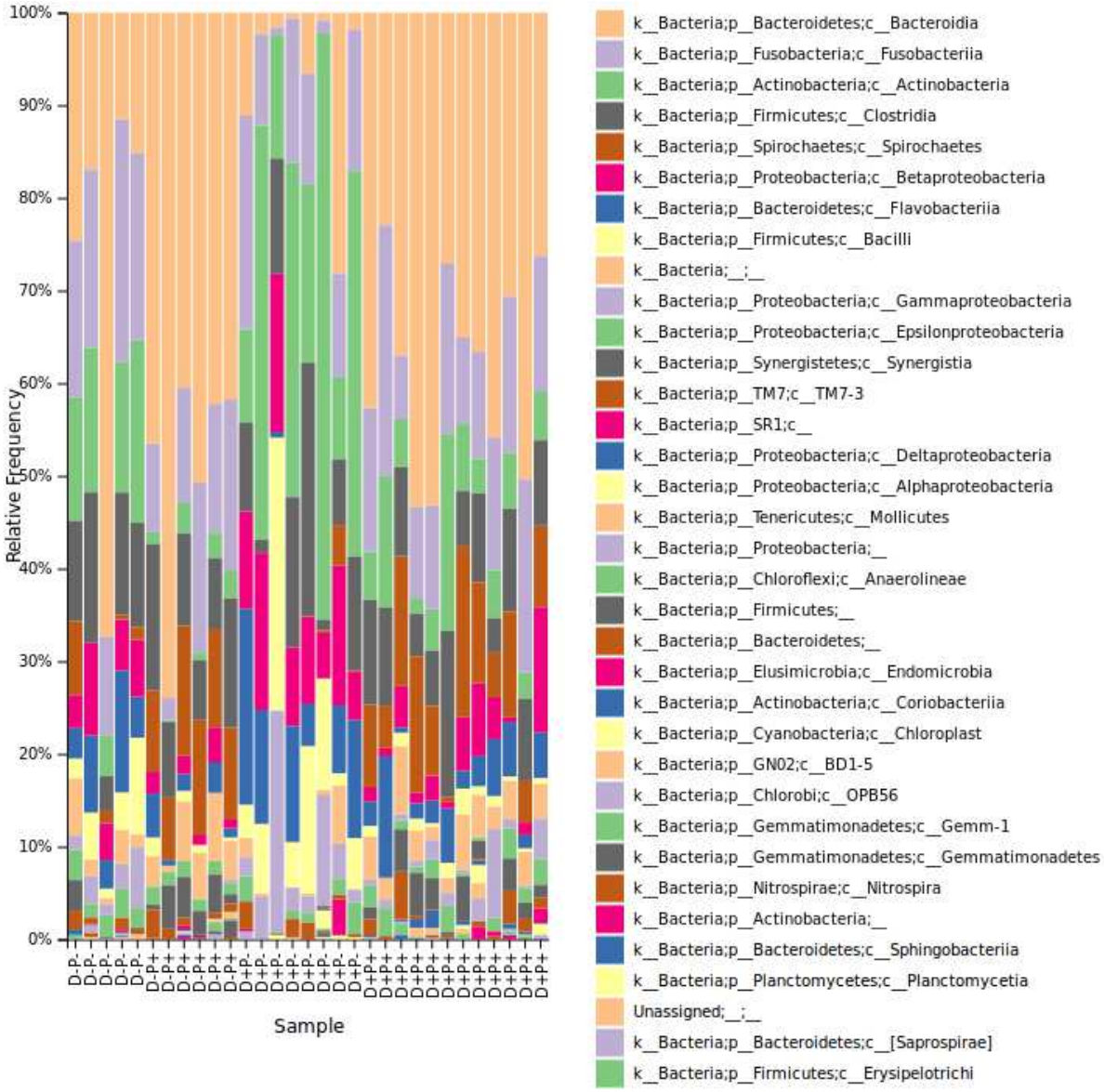

Figura 21. Gráfico de composición taxonómica a nivel de clase.

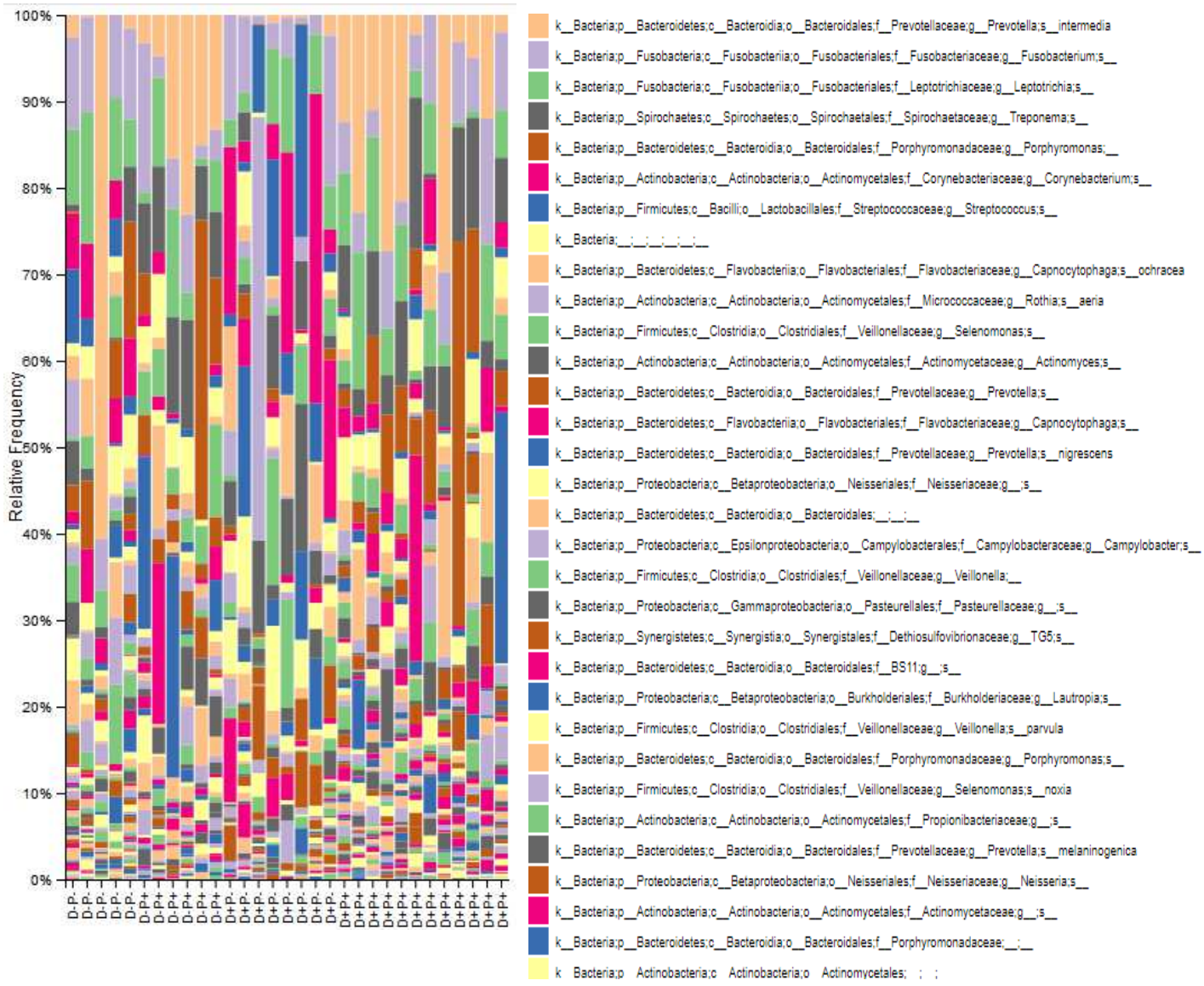

Figura 22. Gráfico de composición taxonómica a nivel de especie.

A nivel de especie, resulta imposible detectar cambios de ordenación a nivel visual y se deberá recurrir a los tests estadísticos.

Para los análisis de las diferencias en composición taxonómica según el grupo de pertenencia de los sujetos del estudio, se han empleado tests ANCOM. La razón por la que se ha elegido esta metodología es porque tiene en cuenta las restricciones de composición y esto permite reducir descubrimientos falsos positivos en la detección de taxones diferencialmente abundantes a nivel de ecosistema, a la vez que se mantiene un alto poder estadístico. Esto la diferencia de otras aproximaciones, que descartan la estructura compositiva subyacente en los datos de microbioma o usan modelos de probabilidad como las distribuciones multinomial y Dirichlet-multinomial, que imponen una estructura de correlación no adecuada para datos de microbioma. Como se ha explicado, ANCOM tiene en cuenta la estructura subyacente en los datos y hace esto usando resultados de los análisis de log-ratio, además no

hace asunciones de distribución y puede escalar bien para comparar muestras de miles de taxones [29]. Por ello se han realizado tests ANCOM a tres niveles taxonómicos distintos para comparar la composición de los microbiomas en dos poblaciones de diabéticos diferentes: una con una complicación adicional (periodontitis), asociada con una mayor severidad, y otra sin ella. Adicionalmente se ha pretendido chequear el efecto que esta complicación puede tener en ausencia de diabetes, comparando poblaciones sin diabetes con periodontitis y sin ella; para comprobar que las posibles diferencias halladas en la primera comparación son debidas a un efecto conjunto de ambas condiciones y no exclusivamente al efecto de la periodontitis.

Análisis estadístico de diferencias en composición taxonómica a nivel filo:

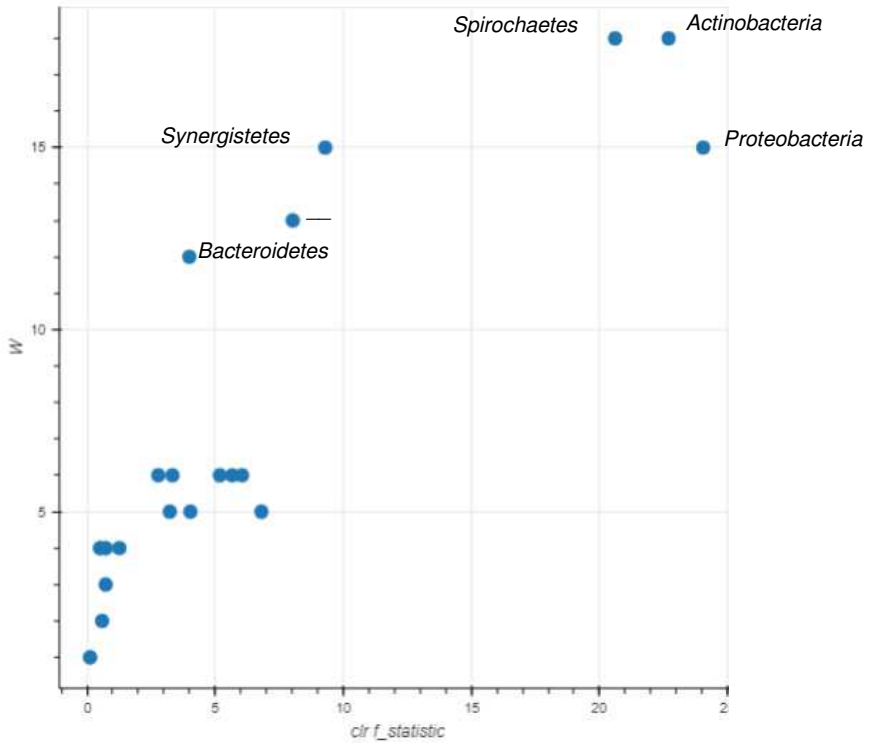

Figura 23. Gráfico volcano generado en el resumen del test ANCOM, donde se representa el valor del estadístico W en el test para cada filo frente al *F-score.*

En el gráfico generado (figura 23), podemos observar seis filos destacados y más abajo, en la tabla 11, podemos observar que cinco de ellos han resultado diferencialmente expresados según nuestros cuatro grupos de estudio: *Actinobacteria, Spirochaetes, Proteobacteria, Synergistetes* y una categoría correspondiente a las características de la tabla de frecuencias a los que no se les pudo asignar un filo de pertenencia.
En la visualización por diagrama de barras de la composición taxonómica de las muestras a nivel de filo, agrupadas según los cuatro grupos de estudio (figura 20), se destacaron el filo *Bacteroidetes*, y en el gráfico volcano comprobamos que pertenece al sexto punto destacado (que no

llega a alcanzar significación estadística) y también llamaron la atención los filos *Spirochaetes, Actinobacteria* y *Proteobacteria*, que tras realizar el test estadístico han resultado ser filos diferencialmente expresados, por lo que los presentes resultados confirman las tendencias detectadas visualmente con el gráfico de barras.

Filos	W
k__Bacteria;__	13
k__Bacteria;p__Actinobacteria	18
k__Bacteria;p__Proteobacteria	15
k__Bacteria;p__Spirochaetes	18
k__Bacteria;p__Synergistetes	15

Tabla 11. Resumen del análisis estadístico por ANCOM según el grupo de pertenencia de las muestras (D-P- , D-P+, D+P- o D+P+).

Respecto a la comparación de la composición de los microbiomas en las dos poblaciones de diabéticos, con complicación adicional (periodontitis) y sin ella, mediante ANCOM:

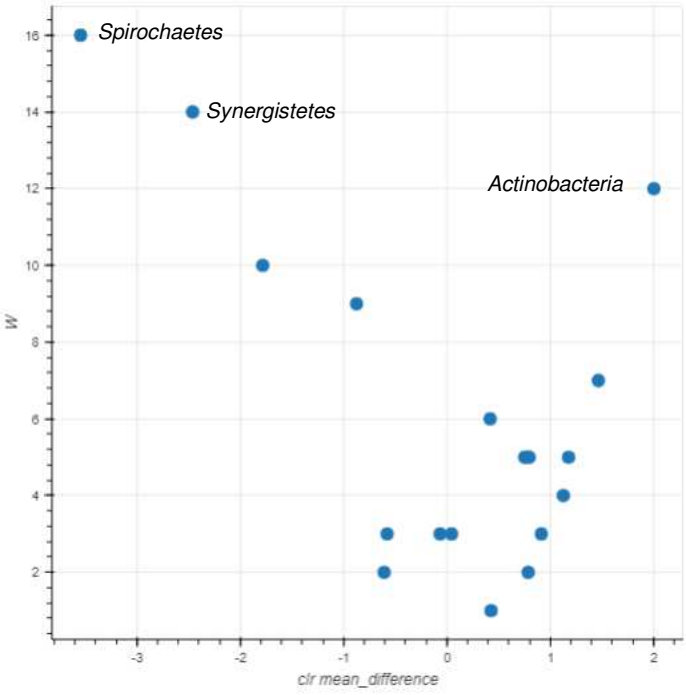

Figura 24. Gráfico generado en el resumen del test ANCOM, en muestras de diabéticos con periodontitis frente a diabéticos sin esta condición. Se representan el valor del estadístico W en el test para cada filo frente a la diferencia de medias.

Filos	W
k__Bacteria;p__Actinobacteria	12
k__Bacteria;p__Spirochaetes	16
k__Bacteria;p__Synergistetes	14

Grupo	D+P+					D+P-				
Percentil	0	25	50	75	100	0	25	50	75	100
p__Actinobacteria	54.0	114.5	141.5	175.3	686.0	227.0	402.0	694.0	1076.5	1865.0
p__Spirochaetes	20.0	108.0	243.0	357.3	554.0	1.0	1.0	2.0	5.0	112.0
p__Synergistetes	1.0	26.0	51.5	113.8	152.0	1.0	1.0	1.0	1.8	6.0

Tabla 12. Resumen del análisis estadístico por ANCOM, sobre la población diabética. En la primera tabla se muestran los filos significativos para el análisis ANCOM, mientras que la segunda tabla muestra las abundancias de los filos significativos por percentiles en los dos grupos estudiados.

Comprobamos que el filo *Actinobacteria* se presenta con más abundancia en la población con diabetes sin periodontitis, mientras que *Spirochaetes* y *Synergistetes*, son más abundantes en la población de diabeticos con periodontitis.

Por otro lado, los análisis ANCOM realizados entre no diabéticos con y sin periodontitis, reveló unicamente una presencia diferencial del filo *Actinobacteria*, más abundante en la población no diabética sin periodontitis. Por lo que la disminución de este filo observada en los sujetos D+P+ no es exclusiva de este grupo, sino que es debida a la presencia de la periodontitis independientemente de la condición diabética. Mientras que los filos *Spirochaetes* y *Synergistetes*, sí pueden considererse exclusivamente aumentados debidos a la presencia de las dos codiciones (diabetes y periodontitis) en los individuos del estudio.

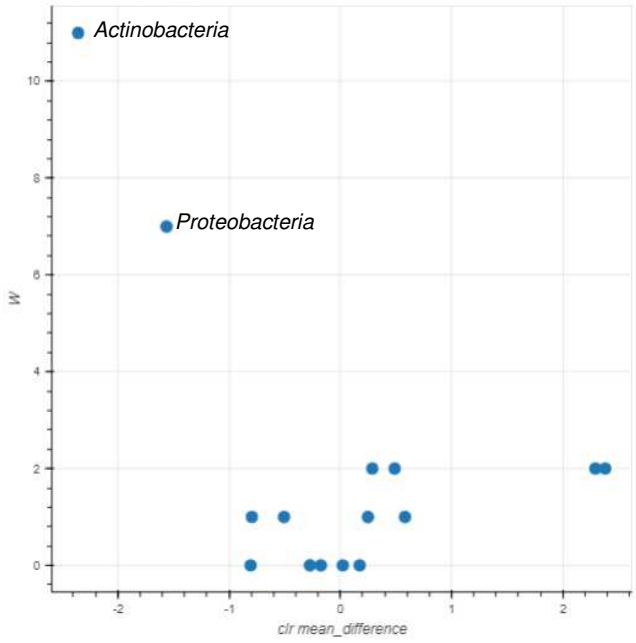

Figura 25. Gráfico generado en el resumen del test ANCOM, en muestras de no diabéticos con periodontitis frente a no diabéticos sin esta condición. Se representan el valor del estadístico W en el test para cada filo frente a la diferencia de medias.

Filo	W
k__Bacteria;p__Actinobacteria	11

Grupo	D-P+					D-P-				
Percentile	0.0	25.0	50.0	75.0	100.0	0.0	25.0	50.0	75.0	100.0
p__Actinobacteria	6.0	20.3	43.0	68.8	95.0	142.0	411.0	428.0	431.0	549.0

Tabla 13. Resumen del análisis estadístico por ANCOM sobre la población no diabética. En la primera tabla se muestran los filos significativos para el análisis ANCOM, mientras que la segunda tabla muestra las abundancias de los filos significativos por percentiles en los dos grupos estudiados.

Diferencias en composición taxonómica a nivel género:

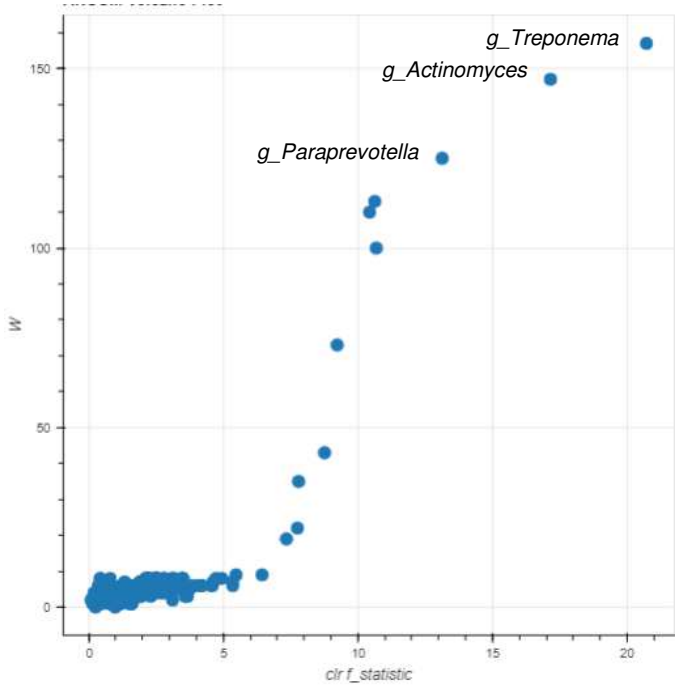

Figura 26. Gráfico generado en el resumen del test ANCOM, donde se representa el valor del estadístico W en el test para cada género frente al *F-score.*

Género	W
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__Actinomyces	147
k__Bacteria;p__Spirochaetes;c__Spirochaetes;o__Spirochaetales;f__Spirochaetaceae;g__Treponema	157

Tabla 14. Resumen del análisis estadístico por ANCOM según el grupo de pertenencia de las muestras (D-P- , D-P+, D+P- o D+P+).

El test ANCOM sobre la tabla de frecuencias colapsada a nivel de género (tabla 14), ha permitido destacar dos géneros: *Actinomyces* (perteneciente al filo *Actinobacteria*) y *Treponema (*perteneciente al filo *Spirochaetes)*.

Los resultados del test ANCOM sobre población diabética para las diferencias en composición taxonómica a nivel de género son:

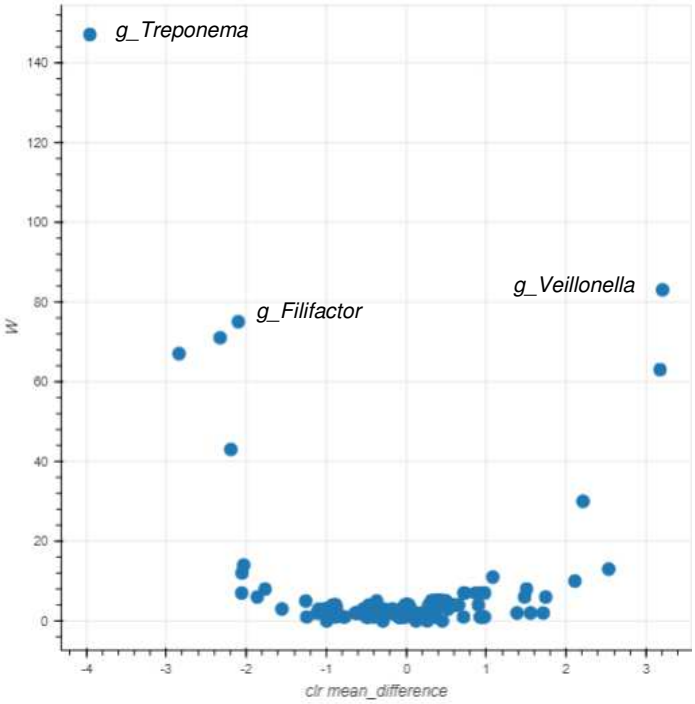

Figura 27. Gráfico generado en el resumen del test ANCOM, en muestras de diabéticos con periodontitis frente a diabéticos sin esta condición. Se representan el valor del estadístico W en el test para cada género frente a la diferencia de medias.

Género	W
k__Bacteria;p__Spirochaetes;c__Spirochaetes;o__Spirochaetales;f__Spirochaetaceae;g__Treponema	147

Grupo	D+P+					D+P-				
Percentil	0.0	25.0	50.0	75.0	100.0	0.0	25.0	50.0	75.0	100.0
g__Treponema	20.0	108.0	243.0	357.25	554.0	1.0	1.0	2.0	5.0	112.0

Tabla 15. Resumen del análisis estadístico por ANCOM, sobre la población diabética.

Los resultados del test ANCOM en población diabética con periodontitis frente a diabética sin periodontitis (tabla 15), muestran una mayor presencia del género *Treponema* en población diabética con periodontitis. Mientras que los resultados del test ANCOM sobre población no diabética para las diferencias en composición taxonómica a nivel de género son:

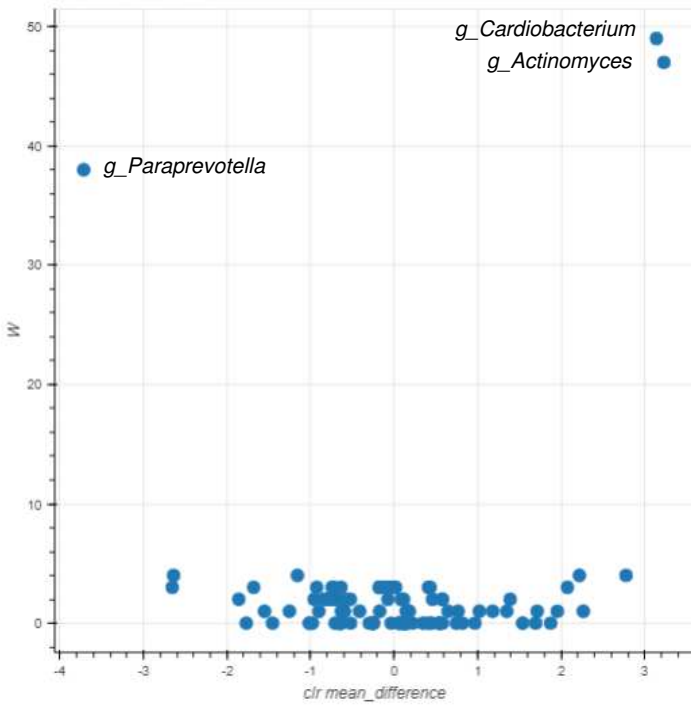

Figura 28. Gráfico generado en el resumen del test ANCOM en muestras de no diabéticos con periodontitis frente a no diabéticos sin esta condición. Se representan el valor del estadístico W en el test para cada género frente a la diferencia de medias.

Género	W
p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__Actinomyces	47
p_Proteobacteria;c__Gammaproteobacteria;o__Cardiobacteriales;f__Cardiobacteriaceae;g__Cardiobacterium	49

Grupo	D-P+					D-P-				
Percentil	0.0	25.0	50.0	75.0	100.0	0.0	25.0	50.0	75.0	100.0
g__Actinomyces	1.0	1.0	1.0	1.0	1.0	9.0	20.0	43.0	68.0	142.0
g__Cardiobacterium	1.0	1.0	1.0	1.0	1.0	23.0	23.0	30.0	53.0	57.0

Tabla 16. Resumen del análisis estadístico por ANCOM sobre la población no diabética.

Los resultados de este tercer análisis a nivel de género (tabla 16), destacan una expresión diferencial del género *Actinomycetes* y del género *Cardiobacterium* (perteneciente al filo *Proteobacteria*) en población no diabética, siendo más abundantes ambos en los sujetos no diabéticos sin periodontitis (es decir, en los controles). Esto confirma que la presencia de la periodontitis en un ambiente diabético aporta un efecto en la abundancia de géneros que es diferente al ejercido cuando no se presenta esta condición. Por lo tanto, la interacción de ambas condiciones en un mismo individuo propicia la presencia de un perfil taxonómico de ciertos taxones distinto de cuando se presenta la condición diabética o de periodontitis en exclusiva.

Diferencias en composición taxonómica a nivel especie:

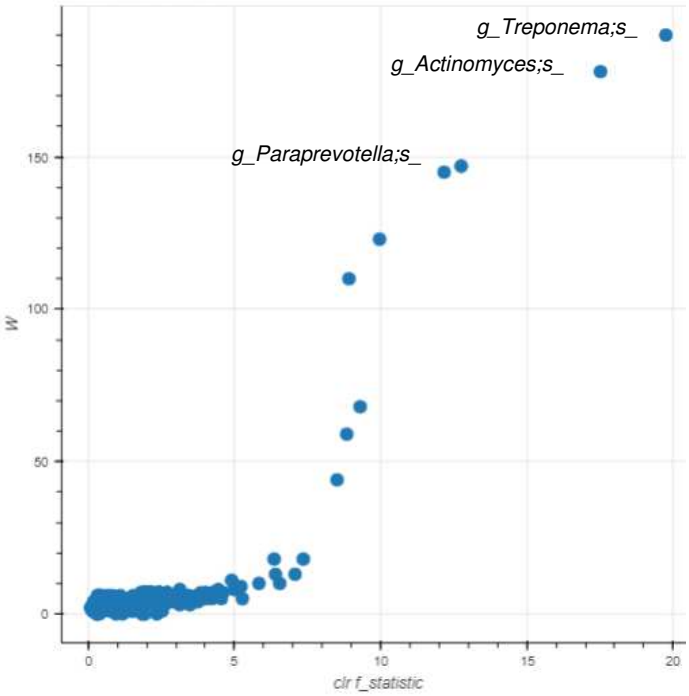

Figura 29 Gráfico generado en el resumen del test ANCOM, donde se representa el valor del estadístico W en el test para cada especie frente al *F-score.*

Especie	W
p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__Actinomyces;s__	178
p__Spirochaetes;c__Spirochaetes;o__Spirochaetales;f__Spirochaetaceae;g__Treponema;s__	190

Tabla 17. Resumen del análisis estadístico por ANCOM a nivel de especie.

En este caso, el análisis por grupos a nivel de especie no aporta ninguna información adicional respecto al análisis a nivel de género, puesto que no es capaz de identificar a nivel de especie en ninguno de los géneros detectados diferencialmente (*Actionomyces* y *Treponema*).
Abajo se presentan los test ANCOM para contrastar las diferencias en población diabética con y sin periodontitis y los tests para contrastar las diferencias en población no diabética con y sin periodontitis (tablas 18 y 19).

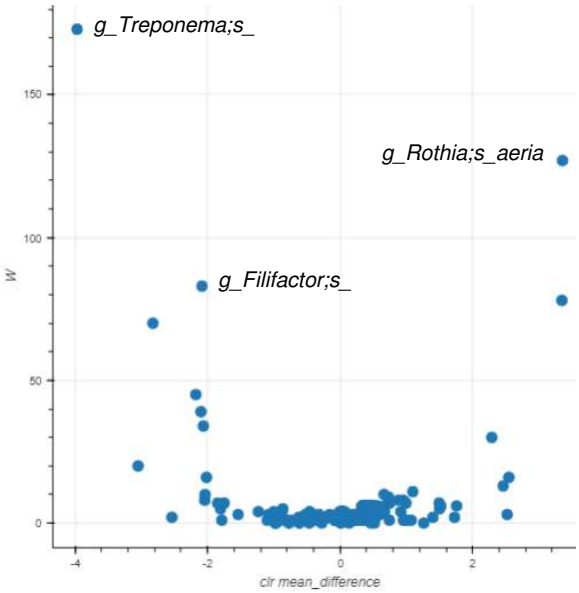

Figura 30 Gráfico generado en el resumen del test ANCOM en muestras de diabéticos con periodontitis frente a diabéticos sin esta condición. Se representan el valor del estadístico W en el test para cada especie frente a la diferencia de medias.

Especie	W
p__Spirochaetes;c__Spirochaetes;o__Spirochaetales;f__Spirochaetaceae;g__Treponema;s__	173

Grupo	D+P+					D+P-				
Percentil	0.0	25.0	50.0	75.0	100.0	0.0	25.0	50.0	75.0	100.0
g__Treponema;s__	20.0	85.25	210.0	324.0	528.0	1.0	1.0	1.0	5.0	85.0

Tabla 18. Resumen del análisis estadístico por ANCOM a nivel de especie, sobre la población diabética.

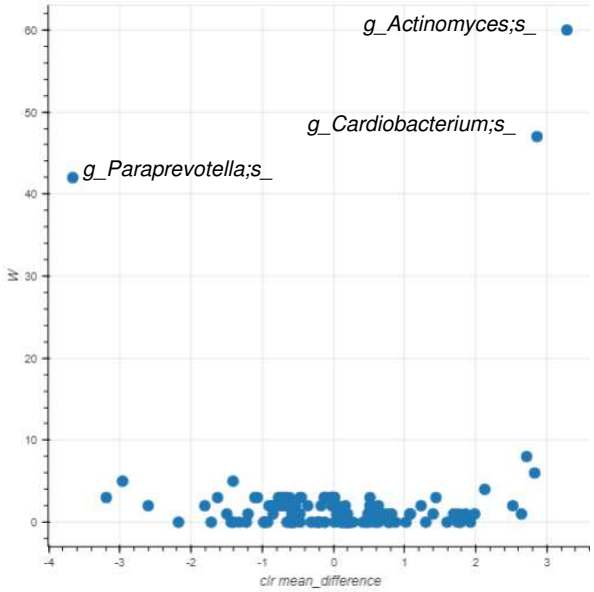

Figura 31 Gráfico generado en el resumen del test ANCOM en muestras de no diabéticos con periodontitis frente a no diabéticos sin esta condición. Se representan el valor del estadístico W en el test para cada especie frente a la diferencia de medias.

Especie	W
p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__Actinomyces;s__	60

Grupo	D-P+					D-P-				
Percentile	0.0	25.0	50.0	75.0	100.0	0.0	25.0	50.0	75.0	100.0
g__Actinomyces;s__	1.0	1.0	1.0	1.0	1.0	9.0	20.0	43.0	68.0	142.0

Tabla 19. Resumen del análisis estadístico por ANCOM sobre la población no diabética.

En resumen, con el análisis de la composición taxonómica de las muestras, se puede apreciar que la diabetes con periodontitis presenta un efecto sobre la abundancia de ciertos taxones diferente del que puede presentar una forma menos extrema de diabetes (sin periodontitis) y este efecto se debe a una interacción de ambas condiciones porque los taxones expresados diferencialmente en presencia de periodontitis cuando los sujetos presentan diabetes es diferente al causado por la periodontitis en ausencia de esta condición.

Selección de genes diferencialmente expresados

La selección de ortólogos KEGG diferencialmente expresados se ha basado en un análisis con un modelo lineal general. Se ha creado la matriz de diseño como un modelo de un factor con cuatro niveles (D-P-/D-P+/D+P-/D+P+) y se ha estudiado: el efecto que la presencia de diabetes con periodontitis, diabetes sin periodontitis y periodontitis cuando no hay diabetes, pueden tener sobre la expresión diferencial de ciertos genes.

Análisis de ortólogos KEGG

Exploración de datos

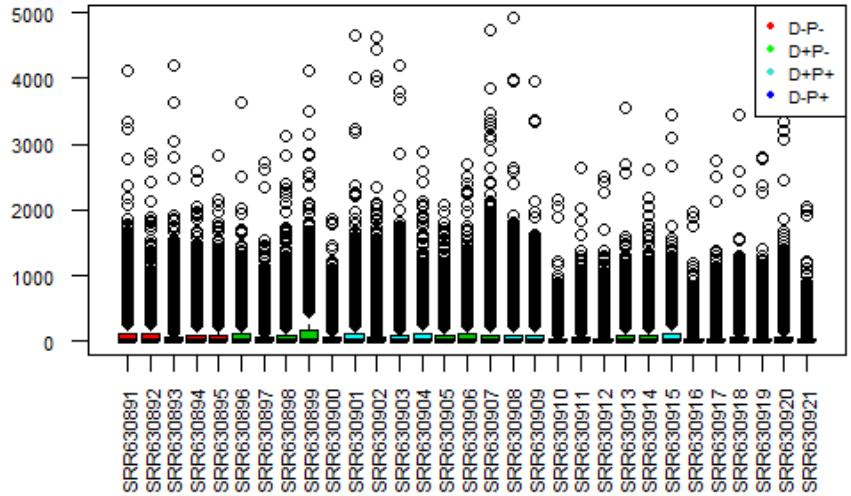

Figura 32 Gráfico *boxplot*, que muestra la distribución de las cuentas por muestra de la matriz con los datos del metagenoma predicho por PICRUST.

En el gráfico anterior (figura 32), observamos que la mayoría de los datos, procedentes de la predicción del metagenoma obtenida con PICRUST, se sitúan en torno al cero con colas largas hacia la derecha, lo que sugiere que los valores presentan una distribución binomial negativa.

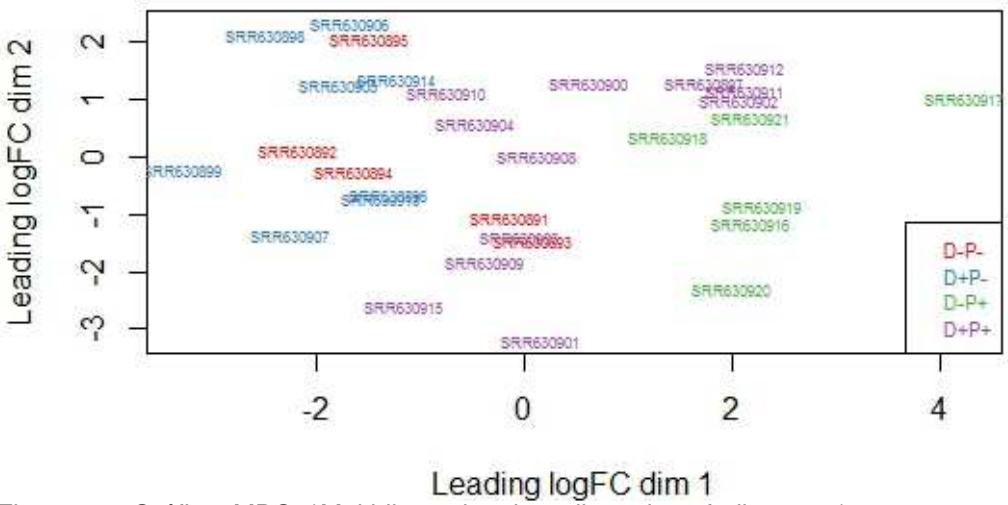

Figura 33 Gráfico MDS (*Multidimensional scaling plot of distances),* que muestra la ordenación de las muestras en las dos primeras dimensiones tras el filtrado y la normalización de las cuentas por el tamaño de las librerías y la transformación logarítmica de los datos. Las muestras pertenecientes al grupo D+P- aparecen en azul, las pertenecientes al grupo D-P- en rojo, el grupo D+P+ se representa en morado y el grupo D-P+ en verde.

Este tipo de gráficos muestra la relación entre las muestras del estudio y realiza un análisis similar a un análisis de componentes principales. En concreto, en el gráfico anterior los grupos de interés se separan en torno a la primera dimensión, pero no hay una clara separación entre grupos; eso nos ofrece una anticipación de lo que serán los resultados de los genes diferencialmente expresados, en el sentido de que con este gráfico de ordenación, no esperamos que haya grandes diferencias entre los grupos.

Efecto de la diabetes con periodontitis en los ortólogos KEGG

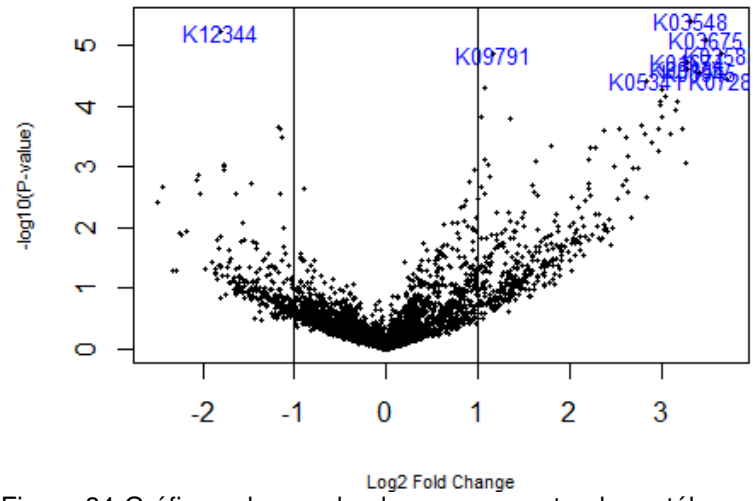

Figura 34 Gráfico volcano, donde se representan los ortólogos KEGG del contraste sobre el efecto de la diabetes con periodontitis. En abcisas se represantan los cambios de expresión en escala logarítmica y en ordenadas el menos logaritmo del p-valor.

Los ortólogos cuyo *–log10(P-value)* se sitúa en torno a 4 o más y cuyo *log2 Fold Change* es mayor a 1 en valor absoluto, son candidatos a estar diferencialmente expresados. En el gráfico se ha destacado las identidades de ortólogos KEGG con p-valores más bajos en color azul. Parece que la diabetes con periodontitis puede inducir más a una *upregulación* de ciertos ortólogos KEGG.
El mismo procedimiento se ha realizado para evaluar el efecto de los otros dos contrastes restantes.

Efecto de la diabetes sin periodontitis en los ortólogos KEGG

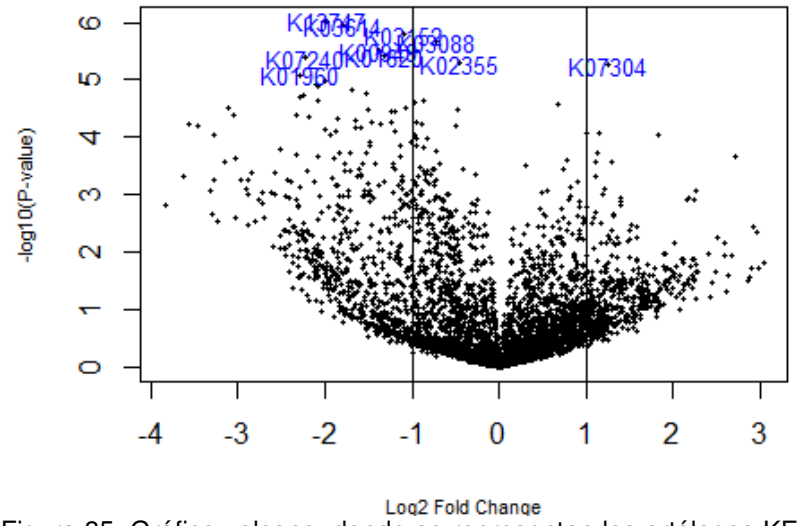

Figura 35. Gráfico volcano, donde se representan los ortólogos KEGG del contraste sobre el efecto de la diabetes sin periodontitis. En abcisas se represantan los cambios de expresión en escala logarítmica y en ordenadas el menos logaritmo del p-valor.

Parece que la diabetes sin periodontitis puede tener más efecto en la *downregulación* de ciertos ortólogos KEGG.

Efecto de la periodontitis cuando no hay diabetes en los ortólogos KEGG

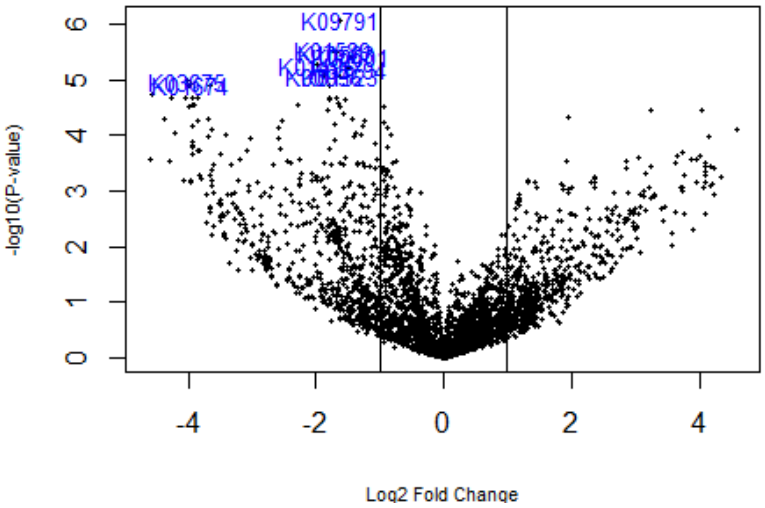

Figura 36. Gráfico volcano, donde se representan los ortólogos KEGG del contraste sobre el efecto de la periodontitis cuando no hay diabetes. En abcisas se represantan los

cambios de expresión en escala logarítmica y en ordenadas el menos logaritmo del p-valor.

En este caso, parece también que la periodontitis por sí sola, puede tener más efecto en la *downregulación* de ciertos ortólogos KEGG.

Comparaciones múltiples

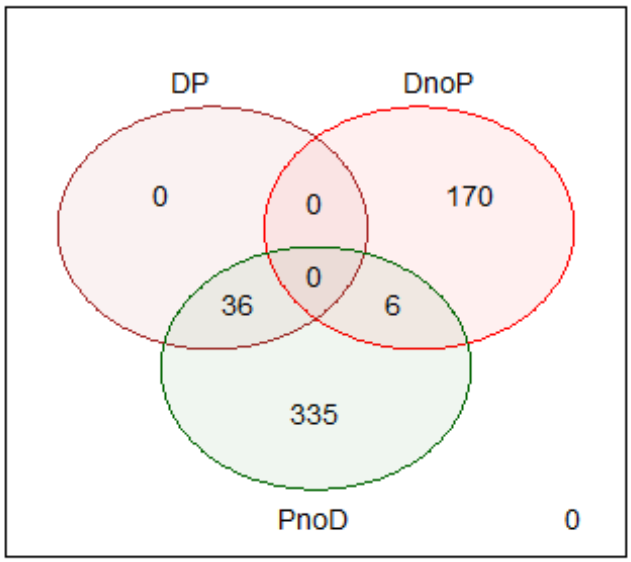

Figura 37 Diagrama de Venn, con el número de ortólogos KEGG diferencialmente expresados en cada comparación y comunes a las tres comparaciones o a las comparaciones dos a dos.

En el diagrama de Venn, vemos que la diabetes con periodontitis no comparte ningún ortólogo KEGG, de entre sus diferencialmente expresados, con el otro tipo de diabetes. Por el contrario, comparte todos sus ortólogos diferencialmente expresados con la periodontitis, pero esta condición, cuando se presenta en exclusiva, presenta por separado y en exclusiva mucho mayor número de ortólogos diferencialmente expresados.

Perfiles de expresión de ortólogos KEGG diferencialmente expresados

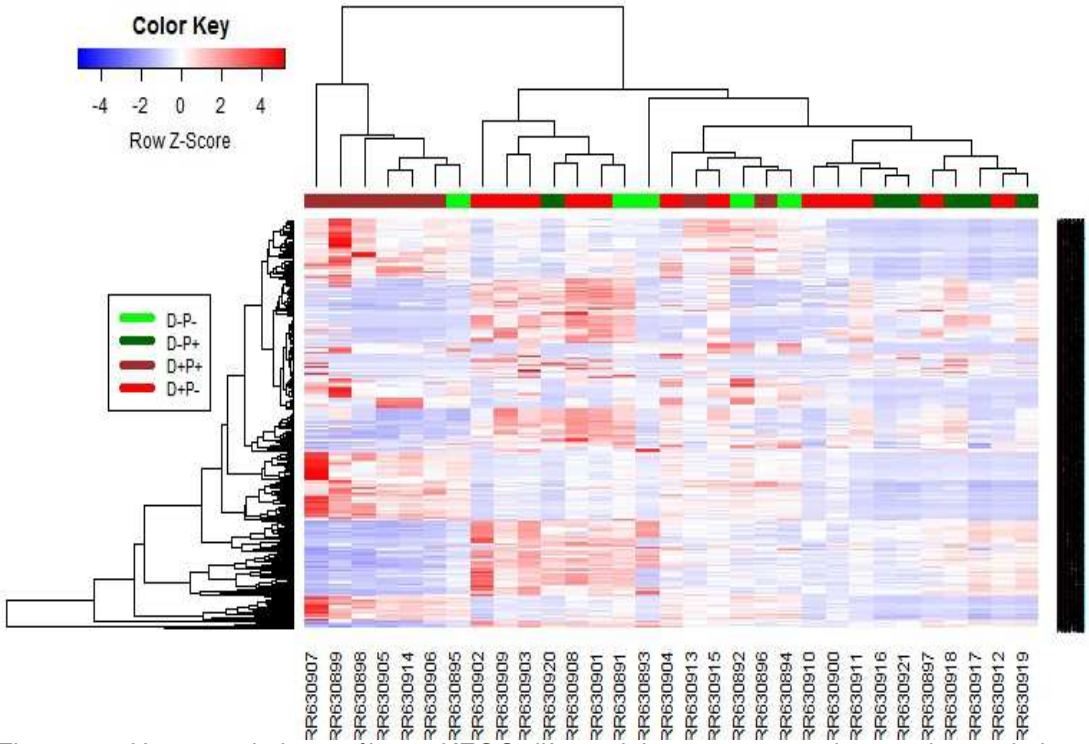

Figura 38. *Heatmap* de los ortólogos KEGG diferencialmente expresados en alguno de los cuatro grupos.

En la figura 38, podemos observar las expresiones de cada ortólogo KEGG diferencialmente expresado agrupándas para destacar los ortólogos que se encuentran *upregulados* (en rojo) o *downregulados* (en azul) simultáneamente, constituyendo perfiles de expresión. Comprobamos que no existen patrones de expresión muy marcados, pero cabe destacar que se observan más diferencias entre el grupo de diabéticos con periodontitis, que presenta en general grupos de ortólogos KEGG más expresados, con la periodontitis en exclusiva, caracterizada por una mayor *downregulación* general y entre medias de estos grupos más marcados, se presenta una transición gradual de patrones de expresión formada por muestras pertenecientes a distintos grupos.

A continuación se presentan tres fragmentos de tablas con los ortólogos diferencialmente expresados de cada grupo (se han seleccionado los tres más *upregulados* y más *downregulados*), junto con las *pathways* en que están implicados, su logFC y p valor ajustado:

	Nombre KO	Pathway	logFC	adj.P.Val
K07280	hypothetical protein	"Unclassified", "Cellular Processes and Signaling", "Membrane and intracellular structural molecules"	3.71	0.02
K03583	exodeoxyribonuclease V gamma subunit	"Genetic Information Processing", "Replication and Repair", "Homologous recombination"	3.66	0.01
K03675	glutaredoxin 2	"Genetic Information Processing", "Folding, Sorting and Degradation", "Chaperones and folding catalysts"	3.49	0.01
K12344	3-oxo-5-alpha-steroid 4-dehydrogenase 2	"Metabolism", "Lipid Metabolism", "Steroid hormone biosynthesis", "Human Diseases", "Cancers", "Prostate c	-1.81	0.01
K14591	protein AroM	"Unclassified", "Poorly Characterized", "Function unknown"	-1.17	0.04
K15039	3-hydroxypropionate dehydrogenase (NADP+)	None	-1.16	0.04

Tabla 20. Tabla con anotaciones de ortólogos KEGG. Se muestran las tres identidades más *upreguladas* y las tres más *downreguladas* en el grupo con diabetes y periodontitis.

	Nombre KO	Pathway	logFC	adj.P.Val
K05820	MFS transporter, PPP family, 3-phenylpropionic acid transporter	"Environmental Information Processing" "Membrane Transport", "Transporters"	2.71	0.01
K03212	RNA methyltransferase, TrmA family	"Genetic Information Processing", "Translation", "Ribosome Biogenesis"	2.26	0.03
K11737	D-serine/D-alanine/glycine transporter	"Unclassified", "Cellular Processes and Signaling", "Other ion-coupled transporters"	2.24	0.03
K07482	transposase, IS30 family	"Unclassified", "Genetic Information Processing", "Replication, recombinatior and repair proteins"	-3.84	0.03
K06019	pyrophosphatase PpaX	"Metabolism", "Energy Metabolism", "Oxidative phosphorylation"	-3.62	0.02
K13990	glutamate formiminotransferase/ formiminotetrahydrofolate cyclodeaminase	"Metabolism", "Metabolism of Cofactors and Vitamins", "One carbon pool by folate",	-3.57	0.01

Tabla 21. Tabla con anotaciones de ortólogos KEGG. Se muestran las tres identidades más *upreguladas* y las tres más *downreguladas* en el grupo con diabetes y sin periodontitis.

	Nombre KO	Pathway	logFC	adj.P.Val
K10671	sarcosine reductase	"Unclassified", "Metabolism", "Others"	4.58	0.01
K07276	hypothetical protein	"Unclassified", "Cellular Processes and Signaling", "Membrane and intracellular structural molecules"	4.35	0.02
K09700	hypothetical protein	"Unclassified", "Poorly Characterized", "Function unknown"	4.24	0.01
K01055	3-oxoadipate enol-lactonase	"Metabolism", "Xenobiotics Biodegradation and Metabolism", "Benzoate degradation"	-4.61	0.01
K07280	hypothetical protein	"Unclassified", "Cellular Processes and Signaling", "Membrane and intracellular structural molecules"	-4.56	0.00
K03834	tyrosine-specific transport protein	"Unclassified", "Cellular Processes and Signaling", "Other ion-coupled transporters"	-4.36	0.01

Tabla 22. Tabla con anotaciones de ortólogos KEGG. Se muestran las tres identidades más *upreguladas* y más *downreguladas* en el grupo con periodontitis y sin diabetes.

Análisis de *pathways* de KEGG diferencialmente expresadas

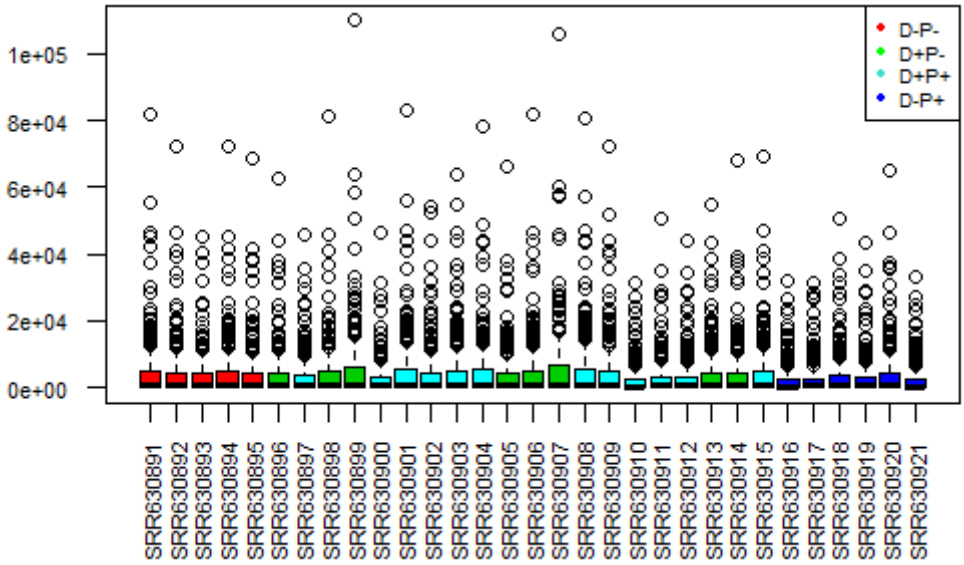

Figura 39. Gráfico que muestra la distribución de las cuentas por muestra de la matriz con los datos del módulo *categrize by function* de PICRUST.

Estos datos permiten examinar los resultados desde un nivel más alto de jerarquía de la ruta, ya que un ortólogo KEGG puede estar implicado en diferentes rutas. Al igual que en el anterior análisis, la mayoría de los datos se sitúan en torno al cero con colas largas hacia la derecha, lo que sugiere que los valores presentan una distribución binomial negativa.

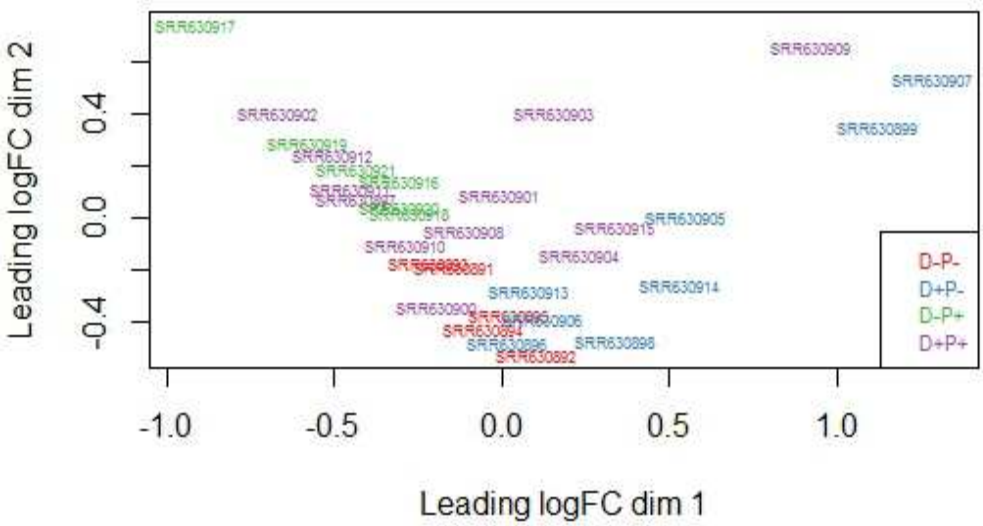

Figura 40. Gráfico MDS que muestra la ordenación de las muestras en las dos primeras dimensiones tras el filtrado y la normalización de las cuentas por el tamaño de las librerías y la transformación logarítmica de los datos. Las muestras pertenecientes al grupo D+P- aparecen en azul, las pertenecientes al grupo D-P- en rojo, el grupo D+P+ se representa en morado y el grupo D-P+ en verde.

En este análisis no se observa separación entre los grupos de estudio.

Efecto de la diabetes con periodontitis en las *pathways*

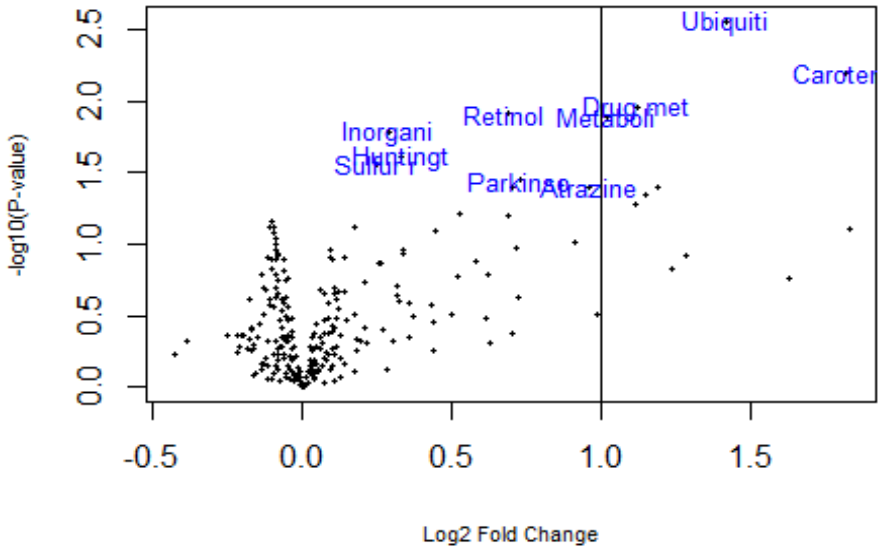

Figura 41. Gráfico volcano, donde se representan las *pathways* KEGG del contraste sobre el efecto de la diabetes con periodontitis. En abcisas se represantan los cambios de expresión en escala logarítmica y en ordenadas el menos logaritmo del p-valor.

El gráfico (figura 41) da idea de que serán pocas las *pathways* expresadas diferencialmente por esta condición y el efecto principal será de *upregulación* de las mismas.

Efecto de la diabetes sin periodontitis en las *pathways*

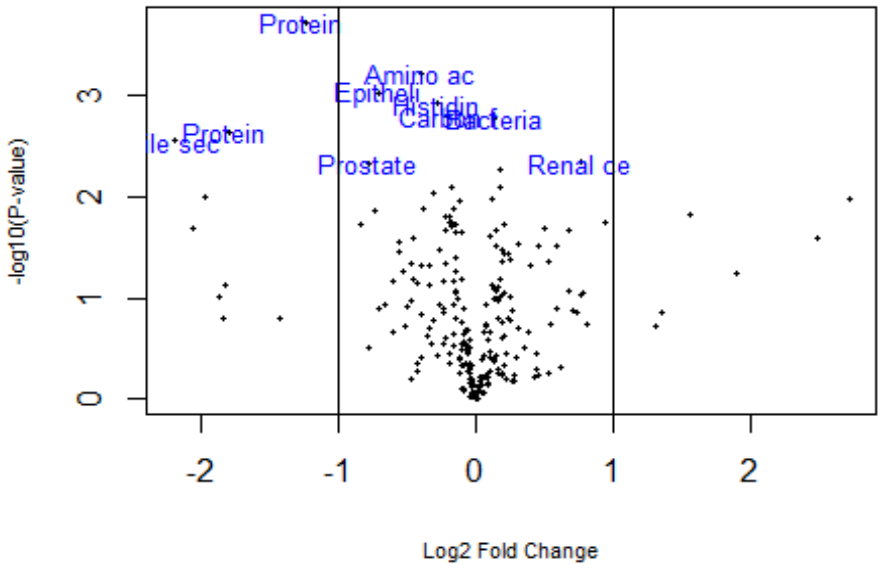

Figura 42. Gráfico volcano, donde se representan las *pathways* KEGG del contraste sobre el efecto de la diabetes sin periodontitis. En abcisas se represantan los cambios de expresión en escala logarítmica y en ordenadas el menos logaritmo del p-valor.

Con el segundo efecto, se repite la misma situación que con el primero, aunque el gráfico (figura 42) hace intuir que el efecto causado por esta condición sobre las *pathways* sería de *downregulación*.

Efecto de la periodontitis cuando no hay diabetes en las *pathways*

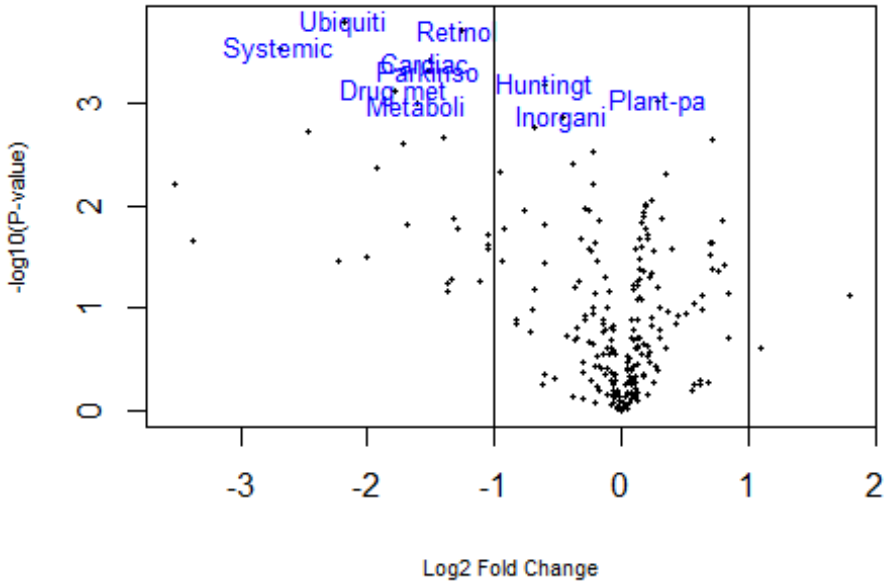

Figura 43. Gráfico volcano, donde se representan las *pathways* KEGG del contraste sobre el efecto de la periodontitis cuando no hay diabetes. En abcisas se represantan los cambios de expresión en escala logarítmica y en ordenadas el menos logaritmo del p-valor.

A este tercer gráfico (figura 43) pueden aplicarse los comentarios realizados en el anterior.

Comparaciones múltiples

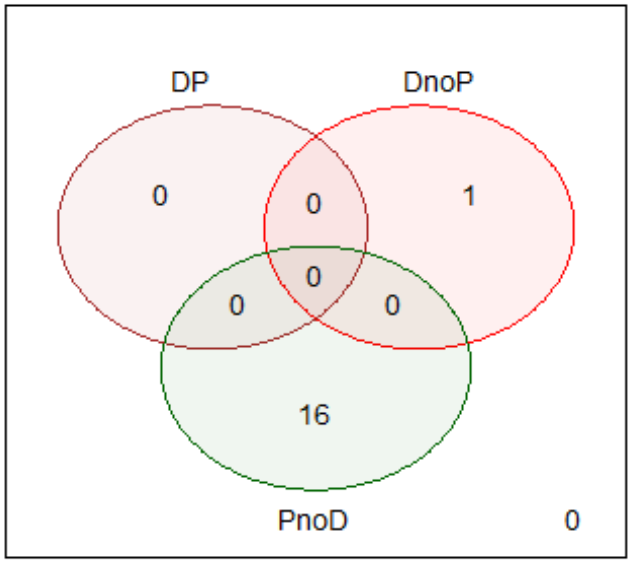

Figura 44. Diagrama de Venn, con el número de *pathways* KEGG diferencialmente expresadas en cada comparación y comunes a las tres comparaciones o a las comparaciones dos a dos.

Se observa que no hay ninguna *pathway* en común entre las comparaciones y la periodontitis, presentada en exclusiva, presenta 16 *pathways* diferencialmente expresadas exclusivamente en esta condición, mientras que la diabetes sin periodontitis presenta una.

Perfiles de expresión de *pathways* diferencialmente expresadas

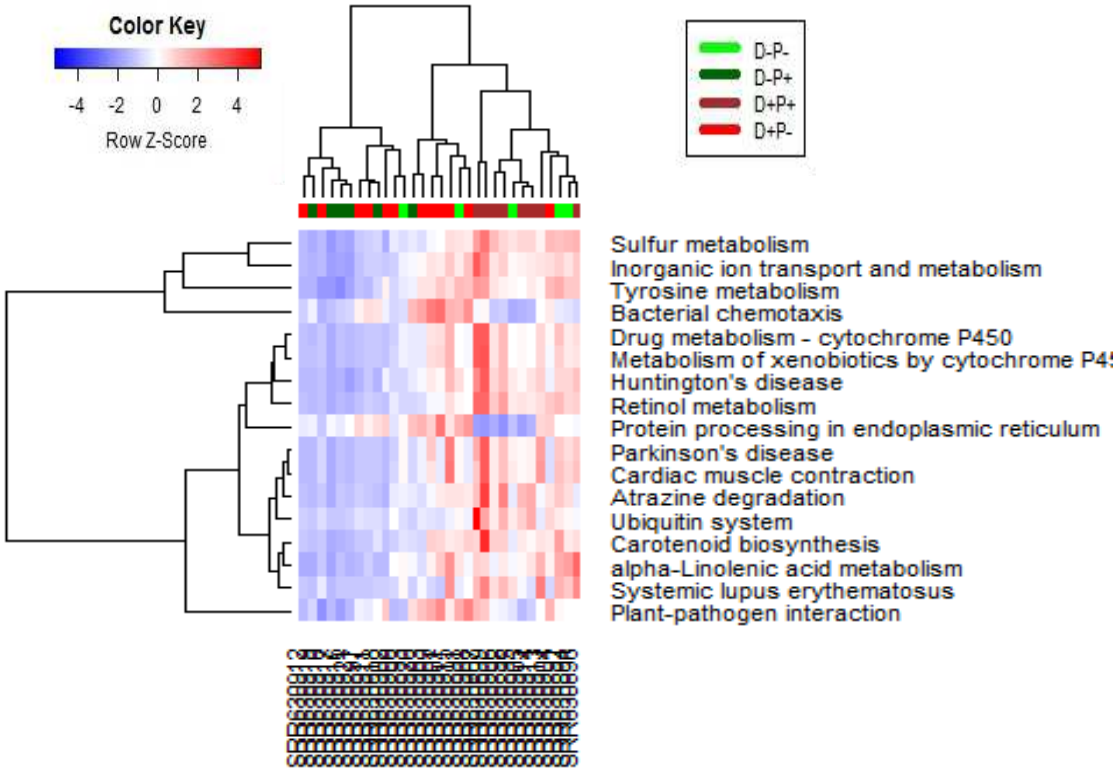

Figura 45. *Heatmap* de las *pathways* diferencialmente expresados en alguno de los cuatro grupos.

En la figura 45, se aprecia dos bloques de *pathways* claros, pero que no se conrresponden con grupos bien definidos. Sí cabe destacar que entre el bloque de *pathways* más *upreguladas* se encuentra el grupo de diabétidos con periodontitis mayoritariamente; mientras que el bloque de *pathways* más *downreguladas* se puede asociar más al grupo de periodontitis sin diabetes. Aunque en ambos bloques se encuentran muestras pertenecientes a otros grupos.

Efecto de la diabetes con periodontitis en los COGs

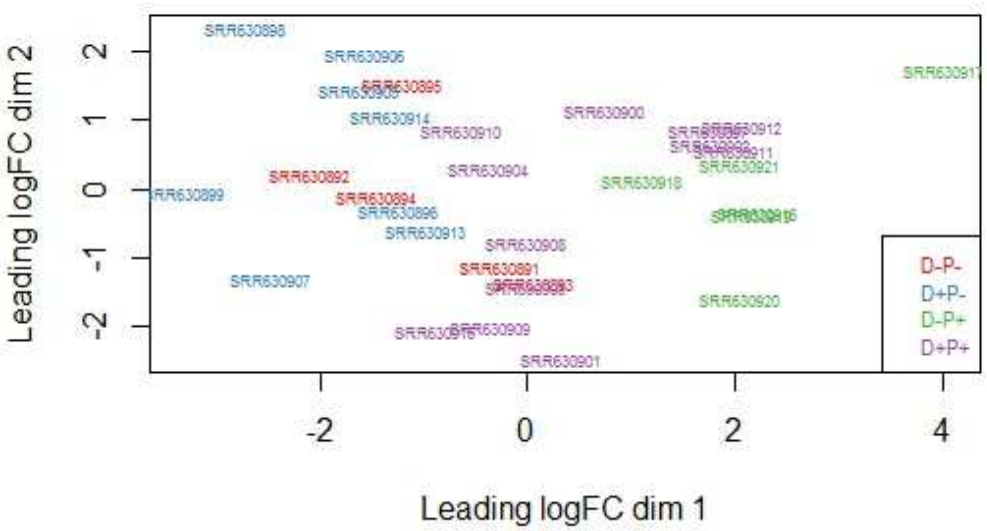

Figura 46. Gráfico MDS que muestra la ordenación de las muestras en las dos primeras dimensiones tras el filtrado y la normalización de las cuentas por el tamaño de las librerías y la transformación logarítmica de los datos. Las muestras pertenecientes al grupo D+P- aparecen en azul, las pertenecientes al grupo D-P- en rojo, el grupo D+P+ se representa en morado y el grupo D-P+ en verde.

En la figura 46 se puede apreciar que los grupos de interés se separan en torno a la primera dimensión, pero no hay una clara separación entre ellos.

Efecto de la diabetes con periodontitis en los COGs

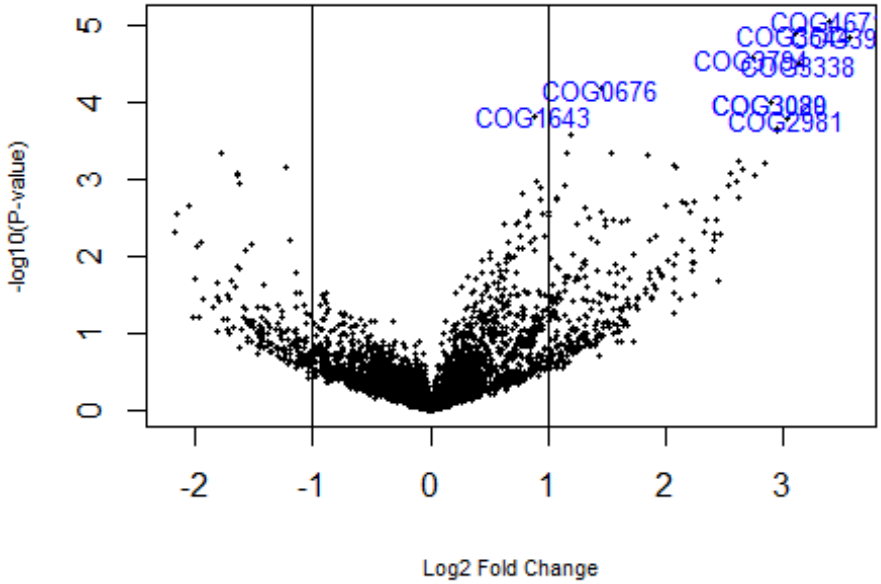

Figura 47. Gráfico volcano, donde se representan los COGs del contraste sobre el efecto de la diabetes con periodontitis. En abcisas se represantan los cambios de expresión en escala logarítmica y en ordenadas el menos logaritmo del p-valor.

Parece que la diabetes con periodontitis puede inducir más a una *upregulación* de ciertos COG, al igual que ocurría con los ortólogos KEGG.

Efecto de la diabetes sin periodontitis en los COGs

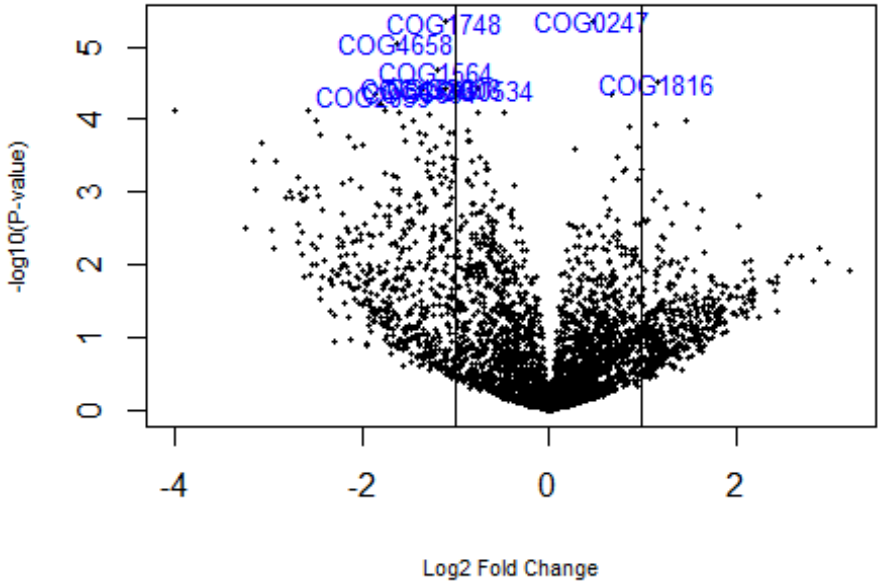

Figura 48. Gráfico volcano, donde se representan los COG del contraste sobre el efecto de la diabetes sin periodontitis. En abcisas se represantan los cambios de expresión en escala logarítmica y en ordenadas el menos logaritmo del p-valor.

La diabetes sin periodontitis puede tener más efecto en la *downregulación* de ciertos COGs.

Efecto de la periodontitis cuando no hay diabetes en los COGs

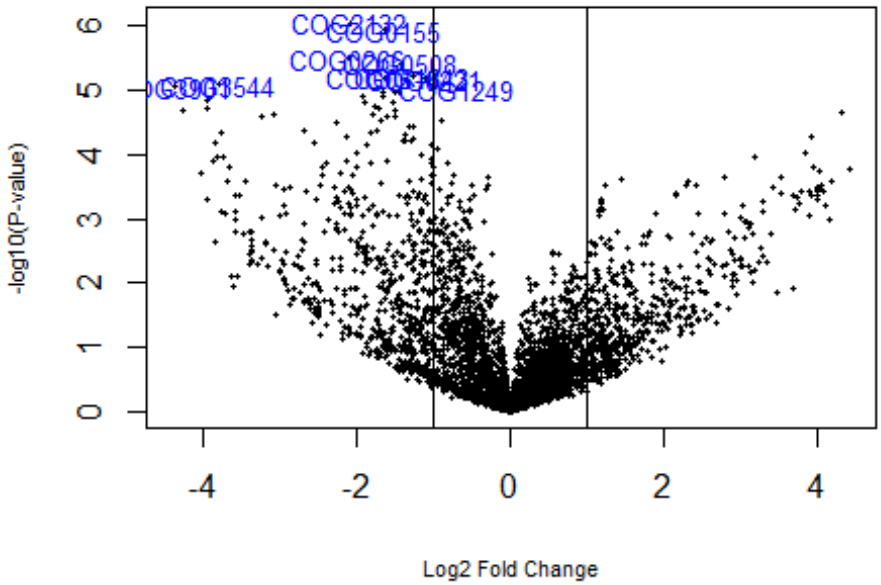

Figura 49. Gráfico volcano, donde se representan los COGs del contraste sobre el efecto de la periodontitis cuando no hay diabetes. En abcisas se represantan los cambios de expresión en escala logarítmica y en ordenadas el menos logaritmo del p-valor.

La figura 49 también muestra un efecto de la periodontitis en la *downregulación* de ciertos COG.

Comparaciones múltiples

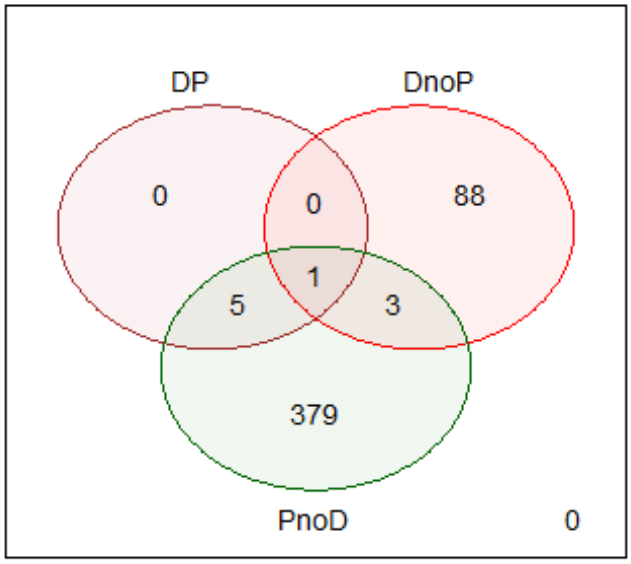

Figura 50. Diagrama de Venn, con el número de COG diferencialmente expresados en cada comparación y comunes a las tres comparaciones o a las comparaciones dos a dos.

La diabetes con periodontitis comparte un COG, de entre sus diferencialmente expresados, con el otro tipo de diabetes, que a su vez es compartido por la tercera condición. Por otro lado, comparte el resto de sus COGs diferencialmente expresados (5) con la periodontitis; pero esta condición, cuando se presenta por separado, presenta 379 COGs diferencialmente expresados en exclusiva y además comparte 3 con la diabetes sin periodontitis, que presenta adicionalmente 88 COGs diferencialmente expresados.

Perfiles de expresión de los COGs diferencialmente expresados

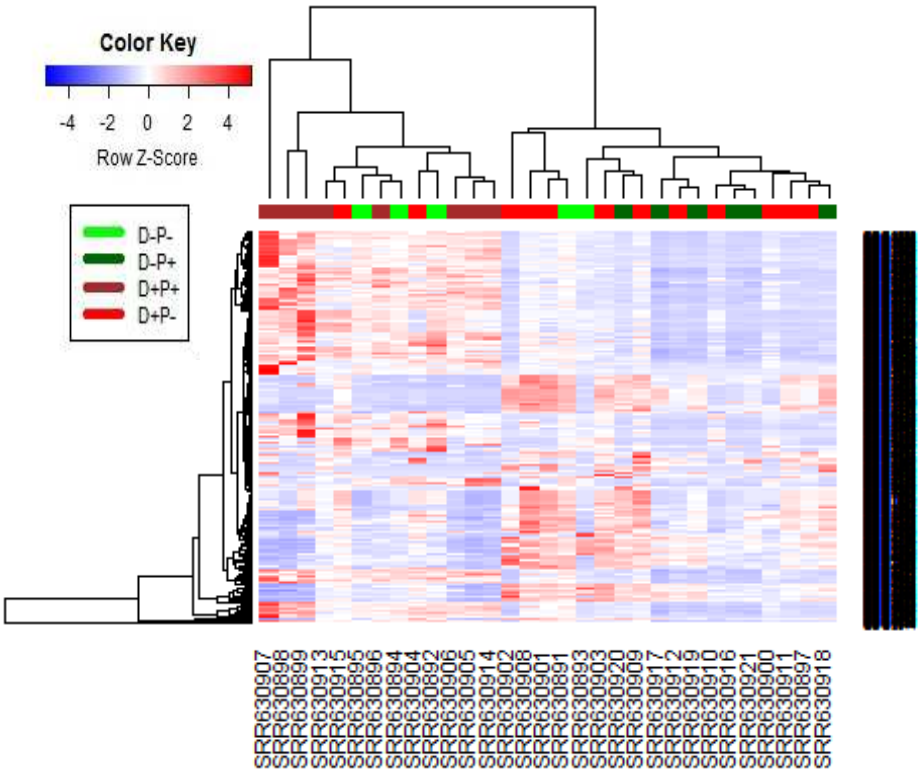

Figura 51. *Heatmap* de los COGs diferencialmente expresados en alguno de los cuatro grupos.

Por último, la figura 51 permite arrojar unas conclusiones similares a las comentadas para este mismo tipo de gráfico en el análisis de ortólogos KEGG.

4. Discusión

Con los análisis realizados, hemos visto que un ambiente diabético y periodontítico se caracteriza por presentar mayor abundancia de los filos *Synergistetes* y *Spirochaetes*, donde destaca el género *Treponema*. Estos filos, anteriormente ya habían sido asociados a la periodontitis, en particular *Synergistetes* se ha relacionado con la presencia de ciclopéptidos (cyclo (-leu-pro) y cyclo (-phe-pro)) de los que se ha sugerido que pueden tener un rol en la disbiosis producida en este entorno [33]. *Synergistetes* es un filo de bacterias anaerobias Gram negativas, con habilidad para fermentar aminoácidos [34]. Es frecuente encontrar miembros de este filo en sitios asociados a enfermedad, mientras que casi nunca se detectan en individuos sanos, sugiriendo que presentan un rol patogénico [35]. *Spirochaetes* también es un filo de bacterias Gram negativas, caracterizado por presentar células alargadas enrolladas helicoidalmente. El rasgo más característico de los miembros de este filo, es su posesión de flagelos especializados que les permiten desplazarse por movimientos giratorios

[36]. Especies de este género (*Treponema vincentii*) ya habían sido asociadas a problemas bucales como la gingivitis ulcerativa necrotizante y la propia periodontitis [37, 38].

Por otro lado, hay trabajos que señalan que la hiperglicemia modifica en ambiente de la cavidad oral, incrementando los niveles de glucosa, la permeabilidad vascular y los niveles de metalo-proteinasas de matriz, citoquinas y moléculas de adhesión. Esto hace que el entorno subgingival en diabéticos, sea rico en glucosa, pro-oxidante, rico en proteinas y anaeróbico [39]. Todo ello concuerda con la mayor presencia de los dos filos observados. Pues por ejemplo, hay trabajos que relacionan una quimiotripsina de *Treponema denticola* (TD-CTLP), como agente impulsor del desarrollo de la periodontitis y también se menciona que actúa activando otras enzimas, como las metaloproteinasas de matriz 8 y 9 [40].

Nuestros resultados sobre la riqueza de la comunidad, por otro lado, difieren de trabajos como Ganesan, et al., que afirman que en el contexto de diabetes y periodontitis, la riqueza de especies es menor a la presentada por no diabéticos con periodontitis [39]. En nuestro caso, se observa que en general en las distintas métricas, hay una disminución de la riqueza con la condición de diabetes y de periodontitis cuando se presentan por separado, mientras que al presentarse estas dos condiciones conjuntamente, no se aprecian cambios estadísticamente significativos en riqueza respecto al grupo control. Por lo que el efecto unido de éstas dos condiciones, es distinto al de cada una por separado. Esto puede ser debido a diferencias en la metodología usada y tampoco hay que olvidar que nuestros resultados deberían confirmarse en una cohorte mayor de individuos, pues una limitación del presente estudio es el bajo número de réplicas biológicas en cada grupo, lo que hace que, unido a la gran variabilidad de los datos, los análisis puedan no tener la suficiente potencia estadística. No obstante, este efecto diferencial cuando dos condiciones se presentan a la vez, ya ha sido documentado para otras circunstancias [39], por lo que no es de extrañar que pudiese pasar una situación similar en nuestro caso.

En lo referente a la diversidad beta, gracias a los gráficos de coordenadas principales y a los test estadísticos para las distintas métricas, hemos podido constatar que la diabetes y la periodontitis son las poblaciones con más distancia respecto al grupo control en nuestro estudio. Siendo la periodontitis el grupo más homogéneo de todos en general, pues los puntos pertenecientes a este grupo presentaban menos dispersión en el gráfico de coordenadas principales. Todo ello indicaría que estos dos grupos, compartirían menos características en común con el grupo control. Como se ha mencionado anteriormente la presencia de diabetes y periodontitis, tiene un impacto en el entorno en el que aparecen, en el caso de la diabetes aumentando su contenido en glucosa y creando un ambiente más pro-oxidante [39], mientras que en el caso de la periodontitis la hiperinflamación [41]. Estos factores pueden explicar estas diferencias en diversidad en nuestro estudio.

Por otra parte, de la inferencia de funciones, realizada gracias a PICRUST, se puede destacar que se presentaban más diferencias entre el grupo de diabéticos con periodontitis y la periodontitis en exclusiva, presentando el primero un patrón más marcado de *upregulación* de funciones y el segundo una mayor *downregulación* general, respectivamente. Entre medias de estos grupos más marcados, se observaba una transición gradual de patrones de expresión, formada por muestras pertenecientes a distintos grupos. Una examinación detallada de las tablas con los ortólogos KEGG diferencialmente *upregulados* en el grupo con diabetes y periodontitis y *downregulados* en el grupo con periodontitis, nos lleva a destacar ortólogos KEGG implicados en las rutas de procesado de la información genética para el grupo de diabéticos con periodontitis, entre los que se encuentran: la exodeoxiribonucleasa V gamma, la glutarredoxina 2 y la proteína ribosómica metiltiotransferasa S12. Curiosamente esta ruta presenta miembros *downregulados* en el grupo con periodontitis como: la propia glutarredoxina 2 o la S transferasa de glutatión putativa. Respecto a otros miembros *upregulados* en el grupo con diabetes y periodontitis, cabe destacar, la anhidrasa carbónica (perteneciente a la ruta de metabolismo energético, en concreto del metabolismo del nitrógeno), la proteína de transporte específica de tirosina (pertenecientes a la ruta de procesos celulares y señalización, en particular a transportadores acoplados a otros iones) y la proteína CysZ (perteneciente a la ruta del metabolismo de amino ácidos). Estos resultados a nivel funcional, en cuanto a la anhidrasa carbónica y el regulador NtrX de la respuesta de regulación de nitrógeno, también *upregulado*, respaldan los obtenidos en la taxonomía, pues existen trabajos que relacionan a especies del género *Treponema* con el ciclo del nitrógeno, eso sí en contextos mediambientales [42]. En el contexto de diabetes, por otro lado, autores como Qin et al., han destacado el enriquecimiento en el transporte de membrana de azúcares y en el transporte de aminoácidos de cadena ramificada en pacientes diabéticos [43], mientras que en nuestro estudio sí se aprecia una *upregulación* de transportadores, pero asociados a aminoácidos como tirosina o prolina y transportadores de manganeso y hierro. Esta divergencia puede darse al efecto conjunto de un ambiente diabético y periodontitico, que puede aportar características que lo diferencian de un ambiente condicionado solamente por la diabetes, además no hay que olvidar que en el caso del presente estudio, la inferencia de funciones se ha realizado sobre un metagenoma oral, mientras que en el caso del trabajo mencionado, el análisis de funciones se realizó sobre metagenoma intestinal. Cabe destacar que los hallazgos de Qin et at, para funciones *downreguladas* en un entorno intestinal diabético, sí se confirman en nuestras muestras diabéticas de la cavidad oral, pues se observa una reducción en la quemotaxis bacteriana (con *downregulación* de diversas proteínas de quemotaxis, como MotB), en el ensamblado flagelar (hay *downregulación* del factor de ensamblado flagelar FliW, por ejemplo) y en el metabolismo de cofactores y vitaminas (con reducción de glutamato formiminotransferasa entre otras); confirmándose que nuestra técnica de inferencia de funciones a través de PICRUST, es una buena aproximación para tener una idea general de lo que está pasando a nivel funcional bajo ciertas condiciones, aunque siempre hay que tener en cuenta que los

resultados solo sugieren lo que puede estar pasando, pues son inferencias y que los resultados deberían confirmarse con secuenciación *shotgun* [20].

Por último, respecto al grupo con periodontitis, se aprecia que muchos de los ortólogos *upregulados* en el grupo con diabetes y periodontitis, aparecen *downregulados* en este grupo, como: la proteína CysZ, transportadores de manganeso y hierro o la anhidrasa carbónica. Las funciones *downreguladas* en el grupo coinciden por las mencionadas en otros trabajos para esta condición, llevados a cabo también en microbioma oral, como son el metabolismo lipídico, donde se presenta una reducción de miembros como la sintasa de ácidos grasos, el metabolismo de terpenoides (con *downregulación* de la fosforibosiltransferasa decaprenil-fosfato). Por otro lado, también se encuentra una *upregulación* de miembros necesarios para la síntesis de flagelos (como la proteína flagelar FlbB), de la biosíntesis de aminoacyl-tRNA (sinthetasa lysyl-tRNA), del metabolismo energético (dehydrogenasa NADH) y del de carbohidratos (pectinesterasa) [44]. Los análisis con COGs, están en la línea de lo comentado para las rutas KEGGs, por lo que su explicación sería similar a lo argumentado hasta el momento.

5. Conclusiones

Las conclusiones extraídas de los resultados son:

- La diabetes y la periodontitis, presentan una menor riqueza que los controles, en población china. Mientras que cuando se presentan en conjunción, este efecto se ve anulado, por lo que el efecto de una condición sobre la riqueza microbiológica, viene condicionado por su entorno.

- La igualdad no se ve afectada por la diabetes, la periodontitis o la suma de ambas en población china.

- En lo referente a la diversidad beta, la diabetes y la periodontitis comparten menos características en común con los controles.

- Se han identificado dos filos característicos asociados al grupo con diabetes y periodontitis: *Synergistetes* y *Spirochaetes*, y dentro de este último, se ha destacado el género *Treponema*.

- Las mayores diferencias a nivel génico, se han presentado entre los grupos con periodontitis, diabéticos y no diabéticos.

La falta de diferencias en igualdad entre grupos, puede deberse a que por falta de suficiente tamaño muestral, los test estadísticos no hayan presentado la suficiente potencia estadística o realmente puede que la

igualdad no presente un rol importante en la caracterización de los grupos en población china. Las diferencias detectadas a nivel funcional, pueden ser reflejo del mayor impacto de la condición de periodontitis a nivel bucal, mientras que, la diabetes es una condición más sistémica y se han aplicado mayor número de estudios en su análisis a nivel intestinal.

Recomendaciones de futuro

Las diferencias funcionales entre los efectos causados por la presencia de periodontitis, diabetes o la presencia de ambos, puede plantear que se dé una disbiosis a nivel funcional más allá de la que se observa a nivel de organismos microbianos (que no es demasiado fuerte). Habría que determinar si las diferencias funcionales se asociasen con la patofisiología de la diabetes de tipo 2 y con sus diferentes grados de severidad, pues no debemos olvidar que una de las limitaciones de este estudio es que no permite arrojar conclusiones de causalidad, que podrían ser determinadas con modelos murinos con ratones *germ-free*. Otras posibles acciones futuras podrían encaminarse hacia estudios de clasificación, que permitiesen discernir entre las formas más severas de diabetes de las que no lo son.

Métodos usados y limitaciones del trabajo

Para la inferencia de funciones, podría haberse probado a colapsar la tabla a nivel de género en lugar de especie, por ser un nivel en que se presenta más confianza a la hora de realizar la asignación taxonómica. En los análisis de expresión diferencial, no se decidió corregir por la longitud del gen, porque al comparar entre muestras, este efecto sería el mismo entre una muestra y otra (asumiendo que la longitud es la misma en todas las muestras). Pero tras seguir profundizando, parece que al tener más lecturas los genes con mayor longitud, la potencia del test es mayor y por eso se podrían encontrar más genes con expresión diferencial entre los genes más largos, un mayor tamaño muestral se asocia a un mayor probabilidad de rechazar la hipótesis nula de igualdad de medias para una misma diferencia de medias. Por último, otra objeción la metodología es que el análisis de las *pathways* KEGG, podría haberse hecho por un test hipergeométrico, pero con la información disponible no fue capaz de ejecutarse, por eso se eligió la estrategia descrita en el estudio.
Una limitación del trabajo, es que las conclusiones se limitan a la población china, no sabemos lo que ocurriría en otro contexto poblacional.

Logro de los objetivos

En general se puede afirmar que la planificación ha sido realista y ha facilitado su seguimiento, permitiendo el cumplimiento de los objetivos marcados al principio del estudio.

6. Glosario

16S rRNA: componente de la subunidad 30S de los ribosomas procariontes que se une a la secuencia de Shine-Dalgarno.

AL: Attachment Loss, término referido la pérdida de inserción de tejido conectivo. Es la manifestación clínica predominante y determinante de la enfermedad periodontal.

ANCOM: ANalysis Of Composition Of Microbiomes, metodología creada para comparar la abundancia de taxones individuales en dos o más poblaciones.

Anhidrasa carbónica: enzima de la familia de metaloenzimas que cataliza la conversión de dióxido de carbono y agua a bicarbonato y protones.

ASV: Amplicon Sequence Variant, nuevo método para analizar datos de secuenciación de alto rendimiento de genes marcadores, que controla los errores de manera que las variantes de secuencia de amplicón se puedan resolver exactamente, a nivel de las diferencias en un solo nucleótido sobre la región del gen secuenciado.

CCA: Constrained Correspondence Analysis, técnica de ordenación restringida multivariante que extrae los gradientes principales entre combinaciones de variables explicativas en un conjunto de datos.

Clasificador Naive Bayes: en teoría de la probabilidad y minería de datos, es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales.

COG: Clusters of Orthologous Groups, base de datos generada por comparación de proteínas predichas y conocidas en todos los genomes microbianos secuanciados completamente para inferir sets de ortólogos.

Deshidrogenasa NADH: gran complejo multienzimático que cataliza la transferencia de electrones del NADH al coenzima Q en la cadena respiratoria.

Diabetes tipo 2: trastorno metabólico que se caracteriza por hiperglucemia en el contexto de resistencia a la insulina y falta relativa de insulina. Representa alrededor del 90 % de los casos de diabetes.

Diagrama Venn: esquemas usado en la teoría de conjuntos. Muestra colecciones (conjuntos) de cosas (elementos) por medio de líneas cerradas; el conjunto de elementos que pertenecen simultáneamente a otros dos es la intersección de ambos (en la intersección las regiones encerradas por sus líneas límite se superponen).

Disbiosis: desbalance del equilibrio microbiano de la microbiota normal, debido a cambios cuantitativos o cualitativos de su composición, su funcionamiento o actividades metabólicas, o bien, a cambios en su distribución.

Distancia Jaccard: mide el grado de similitud entre dos conjuntos, toma valores entre 0 y 1, correspondiente este último a la igualdad total entre ambos conjuntos.

Distancia Bray-Curtis: estadístico usado para cuantificar la diferencia de composición entre dos sitios diferentes, en función de los recuentos en cada sitio. Toma valores entre 0 y 1, donde 0 significa que los dos sitios tienen la misma composición (que comparten todas las especies), y 1 significa que los dos sitios no comparten ninguna especie.

Distancia UniFrac: distancia métrica utilizada para comparar comunidades biológicas. Incorpora información sobre la relación relativa de los miembros de la comunidad mediante la incorporación de distancias filogenéticas entre los organismos observados en el cálculo. Posee dos variantes: *weighted* (cuantitativa) y *unweighted* (cualitativa), donde la primera representa la abundancia de organismos observados, mientras que la segunda solo considera su presencia o ausencia.

Diversidad alfa: riqueza de especies de una comunidad particular a la que consideramos homogénea.

Diversidad beta: grado de cambio en la composición de la comunidad o grado de diferenciación de la comunidad, en relación a un ambiente de gradiente complejo o un patrón de ambientes.

Diversidad Faith: índice para medir diversidad filogenética. Es el análogo filogenético a la riqueza de taxones y recoge el número de unidades del árbol que se encuentran en una muestra.

Downregulación: proceso por el que una célula disminuye la cantidad de un componente celular en respuesta a un estímulo externo.

EMPEROR: herramienta de nueva generación interactiva para el análisis, visualización y la mejor comprensión de datasets de ecología microbiana de alto rendimiento.

ENA: European Nucleotide Archive, es un repositorio que proporciona acceso libre e inrestringido a secuencias anotadas de AND y ARN.

Exodeosiribonucleasa V gama: es una exonucleasa que destruye el ADN.

Filo: categoría en taxonomía situada entre el reino y la clase.

FlbB: proteína motora flagelar localizada alrededor del cuerpo basal flagelar.

Gingivitis ulcerativa necrotizante: enfermedad gingival inflamatoria, dolorosa y rápidamente destructiva, de etiología compleja.

Glutamato formiminotransferasa: enzima que interviene en el metabolismo del ácido fólico.

Glutarredoxina 2: enzima redox caracterizada por usar glutatión como cofactor.

Glutatión S transferasa putativa: enzima que cataliza reacciones transferencia de un grupo funcional dependientes de glutatión sobre substratos hidrófobos.

Gráfico volcano: en estadística, es un tipo de diagrama de dispersión que se usa para identificar rápidamente los cambios en grandes conjuntos de datos compuestos de datos duplicados.

Gram negativas: bacterias que no se tiñen de azul oscuro por la tinción de Gram. Esta característica está ligada a la estructura de la envoltura celular, que presenta doble membrana celular, una externa y la otra citoplasmática.

Heatmap: representación gráfica de los datos donde los valores individuales contenidos en una matriz se representan como colores.

Hemoglobina glicosilada: heteroproteína de la sangre que se forma por la unión de la hemoglobina (Hb) con glúcidos unidos a cadenas carbonadas con funciones ácidas en el carbono 3 y el 4. Su medición es una prueba de laboratorio usada en la diabetes para saber si el control que realiza el paciente sobre la enfermedad ha sido bueno durante los últimos meses.

Igualdad de Pielou: medida independiente del número de especies, su valor va entre 0 y 1, donde 1 corresponde a situaciones donde todas las especies son igualmente abundantes.

KEGG: Kyoto Encyclopedia of Genes and Genomes, es una base de datos para entender funciones de alto nivel y utilidades de los sistemas biológicos, desde información a nivel molecular.

Leu: leucina, uno de los veinte aminoácidos que utilizan las células para sintetizar proteínas.

Máquina virtual: software que simula un sistema de ordenador. Se puede instalar en el sistema operativo de elección y puede ejecutar programas como si fuese un ordenador real.

Metagenoma: conjunto de genes microbianos presentes en un entorno o ecosistema determinado.

Metaloproteinasa de matriz: enzima que descompone colágeno de los espacios entre las células de los tejidos. Es relevante en la participación

de procesos como curación de heridas, la angiogénesis y la metástasis de células tumorales.

Microbioma humano: conjunto de genes de los microorganismos presentes en nuestro organismo.

Microbiota: población microbiana existente en un organismo.

MOTHUR: paquete de software de código abierto para procesamiento de datos bioinformáticos usado en el análisis de DNA de microorganismos no cultivados.

PBS: *2h postpandrial blood sugar*, test que mide la glucosa en sangre exactamente 2 horas tras haber comida. Para este tiempo, el azúcar en sangre ha disminuido en gente sana pero sigue elevado en gente con diabetes.

Pectinesterasa: enzima hidrolasa que hidroliza enlaces estermetílicos de la pectina de alto metoxilo, y junto con la poligalacturonasa, puede hidrolizar las sustancias pécticas produciendo metanol y pectato.

Periodontitis: enfermedad que puede cursar con gingivitis, para proseguir con una pérdida de inserción colágena, recesión gingival y pérdida del soporte óseo del diente.

PERMANOVA: *Permutational Multivariate Analysis of Variance*, es un test estadístico multivariante no paramétrico usado para comparar grupos de objetos y testar la hipótesis nula de que los centroides y dispersión de los grupos, definida como media de espacio, son equivalentes para todos los grupos.

Phred+33: Puntuación de calidad de una base, es un valor entero que representa la probabilidad estimada de un error, es decir, que la base es incorrecta. Se representan con caracteres ASCII, como se trata te phred+33, se representan con ASCII_BASE 33.

PICRUST: *Phylogenetic Investigation of Communities by Reconstruction of Unobserved States*, es un paquete de software bioinformático que permite la inferencia de un perfil functional de una comunidad microbiana basandose en la encuesta de un gen marcador a lo largo de una o más muestras.

Pirosecuenciación: tecnología que permite determinar una secuencia de ADN mediante luminiscencia. Este método de secuenciación se basa en la detección de la liberación de pirofosfato cuando se incorporan los nucleótidos.

Primer: es una cadena corta de ARN o ADN (de alrededor de 18-22 bases) que sirve como punto de partida para la síntesis de ADN.

Pro: prolina, uno de los aminoácidos que usan las células para sintetizar las proteínas de los seres vivos.

Proteina CysZ: transportador de sulfato dependiente de protón que media la obtención de sulfato. Proporciona sulfuro para la ruta de síntesis de la cisteína.

Python: lenguaje de programación orientado a objetos, de código abierto e interpretado, que no necesita compilar el código fuente para poder ejecutarse.

Quimiotaxis: reacción por la que bacterias u otras células de organismos uni o pluricelulares dirigen sus movimientos según la concentración de un estímulo químico en su medio ambiente.

Quimotripsina: enzima proteinasa que puede realizar proteólisis, con una cadena polipeptídica de 245 residuos y cinco enlaces disulfuro. Es una enzima digestiva que degrada las proteínas de los alimentos en el intestino.

Recesión gingival: proceso en el que las encías se retraen, dejando al descubierto la raíz del diente y otras partes que se encontraban ocultas bajo el tejido. Es una patología de origen multicausal.

Regulador NtrX: proteína reguladora de asimilación de nitrógeno.

Riqueza de especies: es el número de especies diferentes representado en una comunidad ecológica, paisaje o región.

Secuenciación *shotgun*: técnica de laboratorio para determinar la secuencia del ADN del genoma de un organismo. Se fragmenta el genoma en una colección de pequeños fragmentos de ADN que se ordenan de forma individual. Un programa de ordenador busca coincidencias en las secuencias de ADN y las utiliza para colocar los fragmentos individuales en el orden correcto para reconstruir el genoma.

Terpenoides: terpenos (compuestos orgánicos derivados del ácido mevalónico) modificados químicamente, por oxidación o reorganización del esqueleto hidrocarbonado (como la vitamina A o retinol, que contiene un átomo de oxígeno).

Tiometiltransferasa S12: enzima que produce metabolitos metilados.

Tirosina: uno de los veinte aminoácidos que forman las proteínas. No es esencial en los mamíferos ya que su síntesis se produce a partir de la hidroxilación de otro aminoácido: la fenilalanina.

Upregulación: proceso por el que una célula aumenta la cantidad de un componente celular, como ARN o proteína, en respuesta a un estímulo externo.

7. Bibliografía

1. Metzker, M.L., *Sequencing technologies—the next generation.* Nature reviews genetics, 2010. **11**(1): p. 31.
2. Wellen, K.E. and G.S. Hotamisligil, *Inflammation, stress, and diabetes.* J Clin Invest, 2005. **115**(5): p. 1111-9.
3. Riserus, U., W.C. Willett, and F.B. Hu, *Dietary fats and prevention of type 2 diabetes.* Prog Lipid Res, 2009. **48**(1): p. 44-51.
4. Mathers, C.D. and D. Loncar, *Projections of global mortality and burden of disease from 2002 to 2030.* PLoS Med, 2006. **3**(11): p. e442.
5. Haffner, S.M., *Epidemiology of type 2 diabetes: risk factors.* Diabetes Care, 1998. **21 Suppl 3**: p. C3-6.
6. Pihlstrom, B.L., B.S. Michalowicz, and N.W. Johnson, *Periodontal diseases.* Lancet, 2005. **366**(9499): p. 1809-20.
7. Karthik, S.J., et al., *Predictors for Gingival Index in Middle-Aged Asian Indians with Type 2 Diabetes from South India: A Cross-Sectional Observational Study.* ScientificWorldJournal, 2018. **2018**: p. 9081572.
8. Navarro Sánchez, A.B., R. Faria Almeida, and A. Bascones Martínez, *Relación entre diabetes mellitus y enfermedad periodontal.* Avances en Periodoncia e Implantología Oral, 2002. **14**: p. 9-19.
9. Garrett, W.S., *Cancer and the microbiota.* Science, 2015. **348**(6230): p. 80-6.
10. Hsiao, E.Y., et al., *Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders.* Cell, 2013. **155**(7): p. 1451-63.
11. Baquero, F. and C. Nombela, *The microbiome as a human organ.* Clinical Microbiology and Infection. **18**: p. 2-4.
12. Patel, J.B., *16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory.* Mol Diagn, 2001. **6**(4): p. 313-21.
13. Vijay-Kumar, M., et al., *Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5.* Science, 2010. **328**(5975): p. 228-31.
14. Petersen, C. and J.L. Round, *Defining dysbiosis and its influence on host immunity and disease.* Cell Microbiol, 2014. **16**(7): p. 1024-33.
15. Suárez-Moya, A., *Microbioma y secuenciación masiva.* Revista Española de Quimioterapia, 2017.
16. Ozsolak, F. and P.M. Milos, *RNA sequencing: advances, challenges and opportunities.* Nat Rev Genet, 2011. **12**(2): p. 87-98.
17. Rondon, M.R., et al., *Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms.* Appl Environ Microbiol, 2000. **66**(6): p. 2541-7.
18. S., A., *FastQC: a quality control tool for high throughput sequence data. .* 2010.
19. Caporaso, J.G., et al., *QIIME allows analysis of high-throughput community sequencing data.* Nat Methods, 2010. **7**(5): p. 335-6.
20. Langille, M.G.I., et al., *Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.* Nature biotechnology, 2013. **31**(9): p. 814-821.
21. Afgan, E., et al., *The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update.* Nucleic Acids Res, 2016. **44**(W1): p. W3-W10.
22. Team, R.D.C., *R: A language and environment for statistical computing.* 2008, R Foundation for Statistical Computing: Vienna, Austria.

23. Zhou, M., et al., *Investigation of the effect of type 2 diabetes mellitus on subgingival plaque microbiota by high-throughput 16S rDNA pyrosequencing.* PLoS One, 2013. **8**(4): p. e61516.
24. Callahan, B.J., et al., *DADA2: High-resolution sample inference from Illumina amplicon data.* Nat Methods, 2016. **13**(7): p. 581-3.
25. Lozupone, C., et al., *UniFrac: an effective distance metric for microbial community comparison.* ISME J, 2011. **5**(2): p. 169-72.
26. Vazquez-Baeza, Y., et al., *EMPeror: a tool for visualizing high-throughput microbial community data.* Gigascience, 2013. **2**(1): p. 16.
27. Anderson, M.J., *A new method for non-parametric multivariate analysis of variance.* Austral Ecology, 2001. **26**(1): p. 32-46.
28. McGarigal, K., S. Cushman, and S.G. Stafford, *Multivariate statistics for wildlife and ecology research*. 2000, New York: Springer. xiii, 283 p.
29. Mandal, S., et al., *Analysis of composition of microbiomes: a novel method for studying microbial composition.* Microb Ecol Health Dis, 2015. **26**: p. 27663.
30. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.
31. Law, C.W., et al., *RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR [version 2; referees: 3 approved].* F1000Research, 2016. **5**.
32. Whittaker, R.H., *Vegetation of the Siskiyou Mountains, Oregon and California. .* Ecological Monographs, 1960. **30**: p. 280-338.
33. Marchesan, J.T., et al., *Association of Synergistetes and Cyclodipeptides with Periodontitis.* J Dent Res, 2015. **94**(10): p. 1425-31.
34. Bhandari, V. and R.S. Gupta, *Molecular signatures for the phylum Synergistetes and some of its subclades.* Antonie van Leeuwenhoek, 2012. **102**(4): p. 517-540.
35. Vartoukian, S.R., R.M. Palmer, and W.G. Wade, *Diversity and morphology of members of the phylum "synergistetes" in periodontal health and disease.* Appl Environ Microbiol, 2009. **75**(11): p. 3777-86.
36. Ryan, K.J. and C.G. Ray, *Sherris Medical Microbiology*. 4th ed. ed. 2004: McGraw Hill.
37. Loesche, W.J., et al., *The bacteriology of acute necrotizing ulcerative gingivitis.* J Periodontol, 1982. **53**(4): p. 223-30.
38. Riviere, G.R., et al., *Identification of spirochetes related to Treponema pallidum in necrotizing ulcerative gingivitis and chronic periodontitis.* N Engl J Med, 1991. **325**(8): p. 539-43.
39. Ganesan, S.M., et al., *A tale of two risks: smoking, diabetes and the subgingival microbiome.* ISME J, 2017. **11**(9): p. 2075-2089.
40. Nieminen, M.T., et al., *Treponema denticola chymotrypsin-like proteinase may contribute to orodigestive carcinogenesis through immunomodulation.* British Journal Of Cancer, 2017. **118**: p. 428.
41. Kilian, M., et al., *The oral microbiome – an update for oral healthcare professionals.* Bdj, 2016. **221**: p. 657.
42. van de Water, J.A., et al., *Spirochaetes dominate the microbial community associated with the red coral Corallium rubrum on a broad geographic scale.* Sci Rep, 2016. **6**: p. 27277.
43. Qin, J., et al., *A metagenome-wide association study of gut microbiota in type 2 diabetes.* Nature, 2012. **490**(7418): p. 55-60.
44. Duran-Pinedo, A.E., et al., *Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis.* The Isme Journal, 2014. **8**: p. 1659.

8. Anexos

- Scripts con las instrucciones de la línea de comandos en QIIME2.
- Script en R Markdown para el gráfico de ordenación.
- Scripts en R Markdown para el análisis de ortólogos KEGG, *pathways* KEGG y COG diferencialmente expresados.
- Tablas de ASVs, de secuencias representativas y de taxones.
- Tablas con los ortólogos, *pathways* KEGG y COGs diferencialmente expresados en los efectos de interés.