**Taller Qiime 2: Beta Diversidad**

Universidad de los Andes, Bogotá - Colombia

*Presentad por*

Carlos Andrés Díaz - código:202010343
David León - código: 201615216
Cesar Patiño - código: 201924259

*Rutas absolutas a los datos de entrada para la resolución del taller*
*Taller7:*
`/hpcfs/home/ciencias/biologia/cursos/bcom4102/patiño_leon_diaz/Taller7`

**Tabla ASVs:**
/hpcfs/home/ciencias/biologia/cursos/bcom4102/Datasets/Talle6_Qiime2/Tabla_Dada2.qza
**Tabla taxonomía:**
/hpcfs/home/ciencias/biologia/cursos/bcom4102/Datasets/Talle6_Qiime2/TaxonomyTable.qza
**Árbol:**
/hpcfs/home/ciencias/biologia/cursos/bcom4102/Datasets/Talle6_Qiime2/insertion-tree.qza
**Metadata File:**
/hpcfs/home/ciencias/biologia/cursos/bcom4102/Datasets/Talle6_Qiime2/MFile_TallerQiime2.txt

**Manual de referencia para seguir este taller:**
**https://docs.qiime2.org/2021.2/plugins/**

**Esta parte del taller comprende 5 número de etapas, que se describen a continuación:**
En cada caso, los métodos a utilizar están indicados en *itálica*, y las métricas en **negrilla**

***Etapa 1: Generación de matrices asociadas a beta diversidad***

Para este punto usaremos dos métodos diferentes. Por un lado, el método *qiime diversity beta*, permitirá generar matrices de distancia con base en métricas no filogenéticas. Por otro lado, el método *beta-phylogenetic* permite usar métricas filogenéticas, como Unifrac, añadiendo un árbol filogenético de referencia.

**Indicación:** Corra el método *qiime diversity beta* con la métrica de **Bray Curtis**.

**Figura 1. Línea de comando empleada**

```
qiime diversity beta --i-table Tabla_Dada2.qza  --p-metric braycurtis --o-distance-matrix MatrixCurtis
```

**Indicación:** Para el uso de métricas filógenéticas debe usar el método *beta-phylogenetic*. Para este caso use la métrica **weighted_unifrac** (con el árbol indicado en las rutas absolutas al inicio del documento).

**Figura 2. Línea de comando empleada**

```
qiime diversity beta-phylogenetic --i-table Tabla_Dada2.qza --i-phylogeny
insertion-tree.qza --p-metric weighted_unifrac --o-distance-matrix
DistanceMatrixWeighted_unifrac
```

***Etapa 2: Generación de los PCoA***

Ya con las dos matrices de distancia creadas, podemos generar el análisis de componentes principales utilizando el método *pcoa*.

**Indicación:** Corra el método *pcoa* (poner solo los parámetros requeridos según el manual de referencia para este método)

**Figura 3.** Línea de comando empleada

```
qiime diversity pcoa --i-distance-matrix
DistanceMatrixWeighted_unifrac.qza --o-pcoa
PCoA_ResultsWeighted_unifrac

qiime diversity pcoa --i-distance-matrix MatrixCurtis.qza --o-pcoa
PCoA_ResultsCurtis
```

***Etapa 3: Generación del plot de ordenamiento interactivo***

El artefacto generado con el PCoA puede visualizarse interactivamente utilizando el plugin *emperor* y el método *plot*.

**Figura 4.** Línea de comando empleada

```
qiime emperor plot --i-pcoa PCoA_ResultsWeighted_unifrac.qza
--m-metadata-file MFile_TallerQiime2.txt --o-visualization
visualizationPCoAWeighted_unifrac

qiime emperor plot --i-pcoa PCoA_ResultsCurtis.qza --m-
metadata-file MFile_TallerQiime2.txt --o-visualization
visualizationPCoACurtis
```

**Indicación:** Genere un PcoA con el método *plot* . Realizarlo de manera independiente para las dos matrices de distancia generadas.

***Para analizar con respecto a las etapas 1-3 y entregar Curtis:***

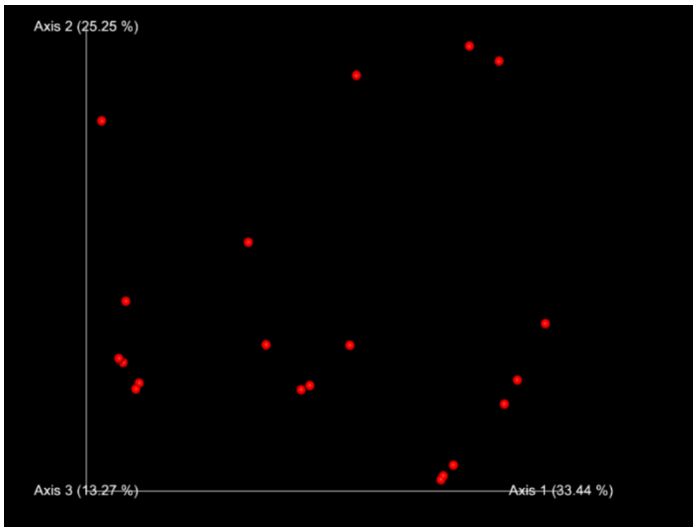


**Figura 5.** Análisis PcoA con el método Weighted Unifrac

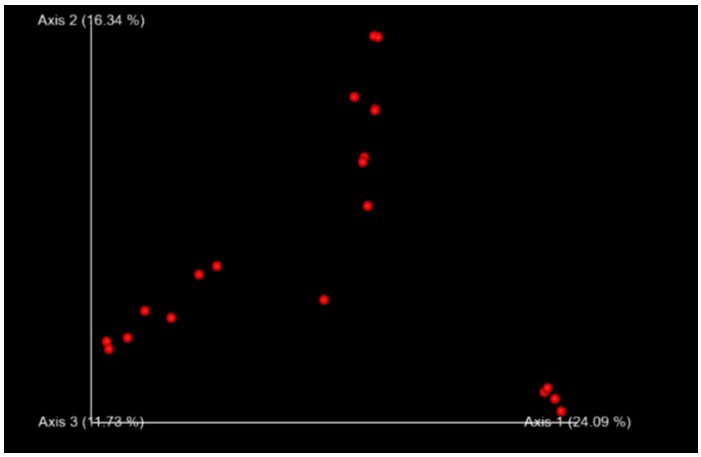


**Figura 6.** Análisis PcoA con el método BrayCurtis

¿Qué diferencia nota entre los dos PCoAs generados a partir de las dos matrices de distancia? Explique esto a qué puede deberse (piense en la base de las dos métricas). Adjunte las imágenes de los dos plots interactivos de emperor.

La diferencia se puede deber a que la métrica de Bray Curtis se basa en los datos de la abundancia / conteo de lecturas, en la que los valores son asignados desde 0 a 1 teniendo en cuenta las diferencias en abundancias de los microorganismos, mientas que la métrica Weighted_Unifrac tiene en cuenta la información de parentesco entre los miembros de una comunidad gracias a un cálculo de su filogenética (Lozupone, Lladser, Knights, Stombaugh & Knight, 2010). Por lo anterior, y teniendo en cuenta lo observado en las figuras 5 y 6, resulta evidente que el análisis mediante Unifrac logra explicar en un mayor porcentaje el porcentaje de varianza de las muestras, aquí es muy importante notar que en este análisis preliminar solamente se considera como única característica el origen de muestra y no las otras características asociadas (Ricotta & Podani, 2017).

Ahora escoja una de las gráficas e intente colorear las muestras por las categorías de su archivo de metadatos. Describa si encuentra patrones de agrupamiento por alguna categoría.

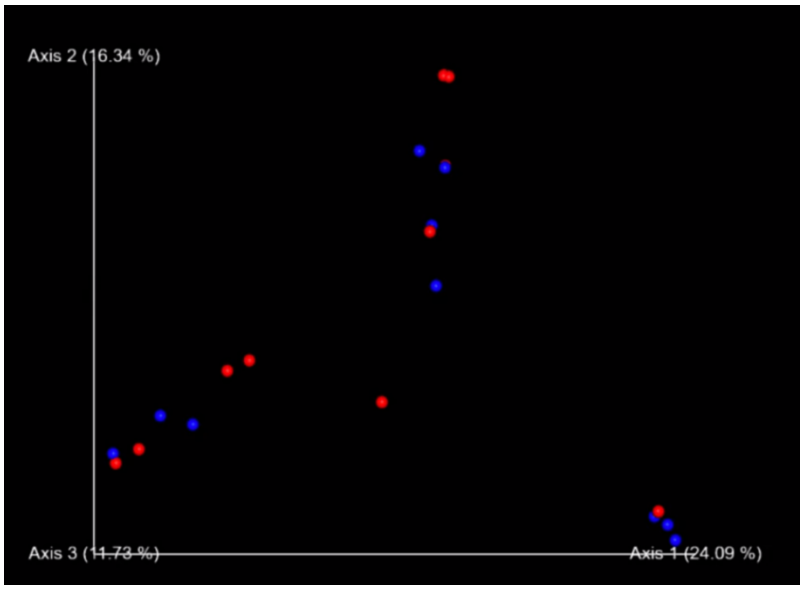


**Figura 7.** Análisis PcoA con el método BrayCurtis agrupando por la categoría estado de la enfermedad

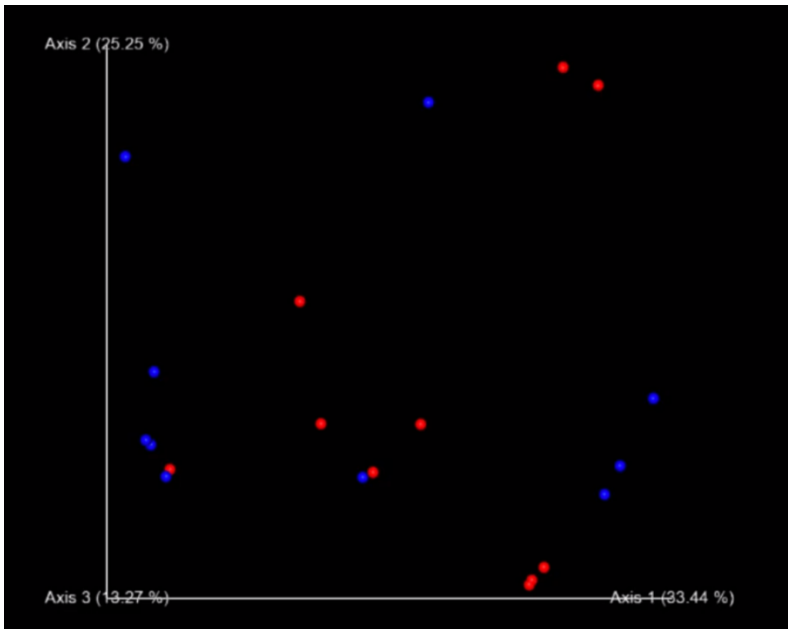


**Figura 8.** Análisis PcoA con el método Weighted Unifrac agrupando por la categoría estado de la enfermedad

Como se observa en las figuras 7 y 8 al utilizar la categoría de estado de la enfermedad resulta claro como ninguna de los métodos logra hacer una agrupación satisfactoria de los datos y por ende los agrupa correctamente, esto también se observa en la varianza explicada de los componentes principales, por lo que se recomienda usar otro tipo de métricas para ver si con la generación de otras matrices de distancia los datos se logran agrupar de una mejor forma.

***Etapa 4: Beta Group Significance***

El objetivo de este análisis es determinar si algunos grupos de muestras (grupos dados por los metadatos del estudio) son significativamente diferentes de otros. Para esto se pueden usar análisis estadísticos de permutaciones como PERMANOVA, Anosim, entre otros.

**Indicación:** Utilice el método *beta-group-significance* y no cambie los parámetros por defecto.

**Nota:** en el parámetro "--m-metadata-column" use la columna 'Disease estate' del Mapping file. En vista de que esta columna tiene un espacio, debe poner las comillas sencillas, de lo contrario reconocería la segunda palabra como parte del código y generaría un error.

***Para analizar con respecto a la etapa 4 y entregar***

¿Hay diferencias significativas entre los dos grupos de muestras analizados? Incluya imágenes de la salida.

**Figura 9.** Línea de comando empleada

```
qiime diversity beta-group-significance --i-distance-matrix DistanceMatrixWeighted_unifrac.qza --m-metadata-file MFile_TallerQiime2.txt --m-metadata-column 'Desease estate' --o-visualization BetaSignificanceMatrixWeighted_unifrac

qiime diversity beta-group-significance --i-distance-matrix MatrixCurtis.qza --m-metadata-file MFile_TallerQiime2.txt --m-metadata-column 'Desease estate' --o-visualization BetaSignificanceMatrixCurtis
```

Overview

| | PERMANOVA results |
|---|---|
| method name | PERMANOVA |
| test statistic name | pseudo-F |
| sample size | 20 |
| number of groups | 2 |
| test statistic | 0.725425 |
| p-value | 0.753 |
| number of permutations | 999 |

**Figura 10.** Resultados PERMANOVA usando la matriz generada por BrayCurtis.

Al observar el valor p generado por el análisis de PERMANOVA para esta métrica, es claro que es demasiado alto (0.753) por lo que no existe una diferencia significativa entre los dos tipos de muestras analizados tomando como base la distancia generada por la métrica BrayCurtis.

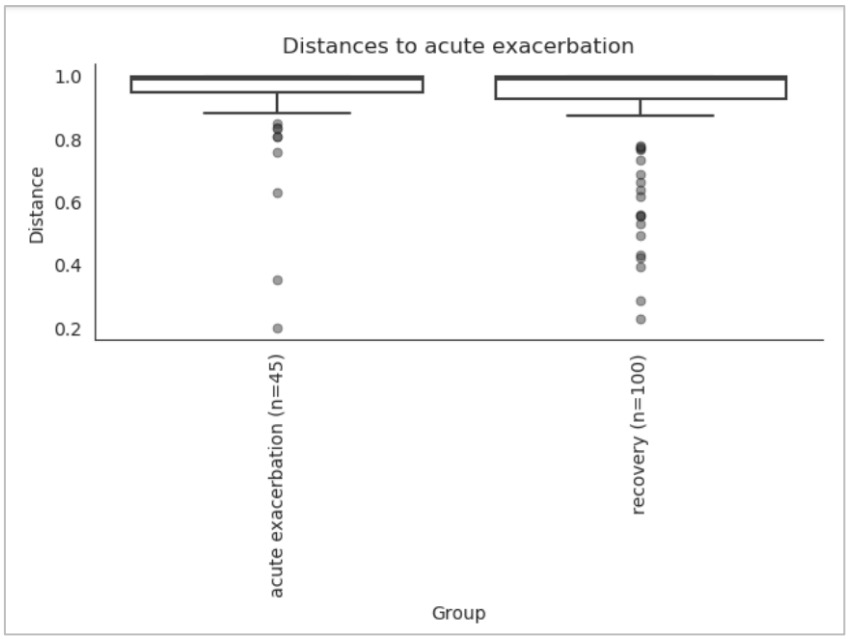


**Figura 11.** Diagrama de cajas y bigotes por las distancias generadas por BrayCurtis.

De la figura 11 es claro como de acuerdo con la métrica establecida por BrayCurtis, el rango intercuartílico comprendido por las muestras provenientes de distintas localizaciones no permite una diferenciación clara, resulta interesante además observar como existen muestras con valores de distancias muy pequeñas las cuales generan valores atípicos con respecto a las demás.

Overview

| | PERMANOVA results |
|---|---|
| method name | PERMANOVA |
| test statistic name | pseudo-F |
| sample size | 20 |
| number of groups | 2 |
| test statistic | 0.696363 |
| p-value | 0.646 |
| number of permutations | 999 |

**Figura 12.** Resultados PERMANOVA usando la matriz generada por Weighted Unifrac.

De la misma forma, al observar el valor p generado por el análisis de PERMANOVA para esta métrica, es claro que es demasiado alto (0.6963) por lo que no existe una diferencia significativa entre los dos tipos de muestras analizados tomando como base la distancia generada por la métrica **Weighted Unifrac.**

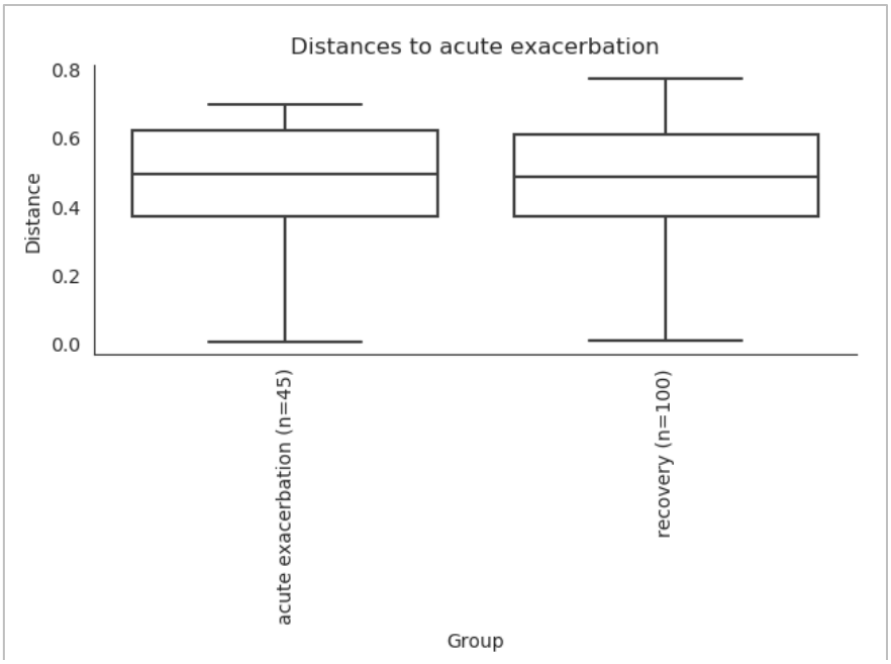


**Figura 13.** Diagrama de cajas y bigotes por las distancias generadas por Weighted Unifrac.

De la figura 13 es claro como de acuerdo con la métrica establecida por Weighted Unifrac, el rango intercuartílico comprendido por las muestras provenientes de distintas localizaciones no permite una diferenciación clara, siendo el sitio de recolecta aquel que presenta mayor variación, resulta interesante ver como con esta métrica desaparecen los valores atípicos generados con la métrica de BrayCurtis por lo que probablemente se deba al hecho de su estrecha relación filogenética.

***Etapa 5: Generación de Biplots***

Los biplots son otra forma de ordenamiento la cual permite proyectar algunos features del estudio en el mismo plot de ordenamiento. Esto sirve para saber si hay hay features que estén explicando la distribución, o el agrupamiento de algunas muestras.

Indicaciones: El primer paso es convertir su matriz de ASVs que tiene las frecuencias de cada feature en una matriz de abundancia relativa, ya que este será el input para el siguiente método.
Para esto use el método *relative-frequency* del plugin *feature-table*.
**Nota**: La matriz de frecuencia que vamos a usar en este caso es una matriz que está colapsada por taxonomía a nivel de género. La encuentran en el siguiente path: /hpcfs/home/ciencias/biologia/cursos/bcom4102/Datasets/Talle6_Qiime2/CollapseTable_L5.qza

**Figura 14.** Línea de comando empleada

```
qiime feature-table relative-frequency --i-table CollapseTable_L5.qza
--o-relative-frequency-table CollapseTable_L5_relativeFrequency
```

El segundo paso es generar el PCoA pero en vista de que la idea es incluir los features, el mismo método que se usó anteriormente no es adecuado. Para esto, use el método *pcoa-biplot*.

**Figura 15.** Línea de comando empleada

```
qiime diversity pcoa-biplot --i-pcoa PCoA_ResultsWeighted_unifrac.qza
--i-features CollapseTable_L5_relativeFrequency.qza
--o-biplot PCoAFeaturesResults_Weighted_unifrac
```

**Nota:** Uno de los inputs para dicho método es un PCoA ya construido, ya que se tomará de base para añadir los features. Use el PCoA basado en Unifrac que realizó anteriormente.

Finalmente se puede volver a usar *emperor* para visualizar el plot de ordenamiento pero en este caso se usa el método *biplot*.

**Figura 16.** Línea de comando empleada

```
qiime emperor biplot --i-biplot
PCoAFeaturesResults_Weighted_unifrac.qza --m-sample-metadata-
file MFile_TallerQiime2.txt --o-visualization
```

¿Qué features fueron añadidos al PCoA? Puede adjuntar imágenes.

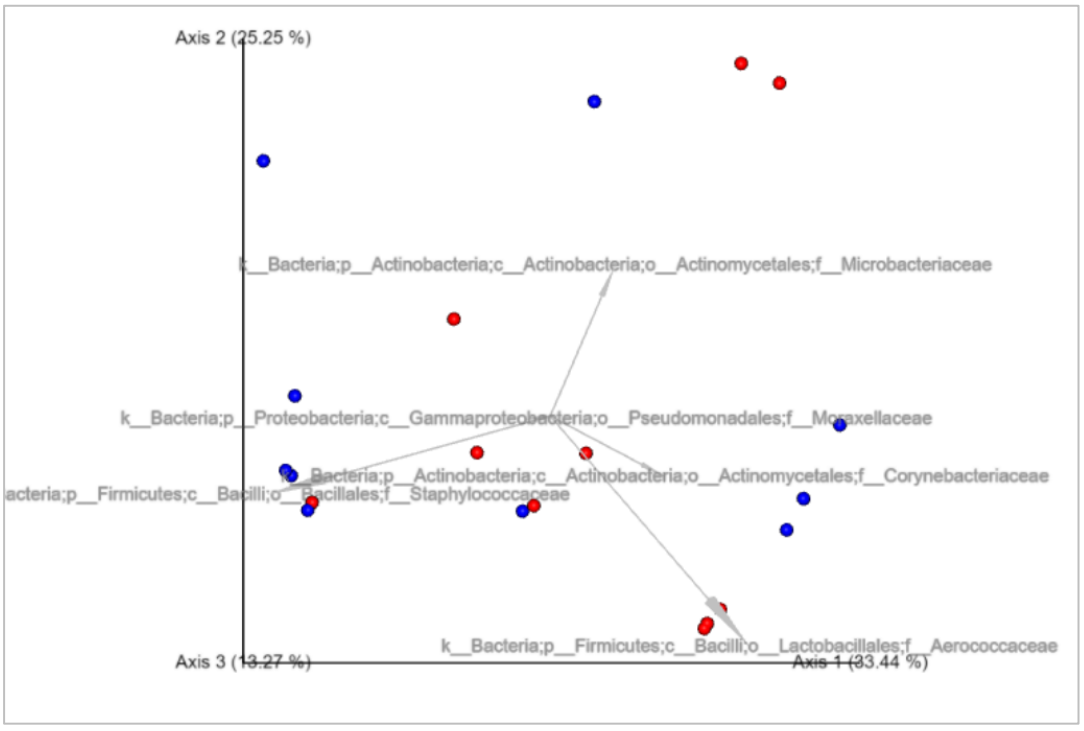


**Figura 16.** Análisis de componentes principales utilizando el método biplot.

Al observar la figura anterior resulta evidente que ahora existen nuevas características que describen en mayor forma las diferentes bacterias presentes en las muestras pero todavía no permite realizar una discriminación exitosa que permita realizar algún tipo de análisis acerca de las distintas muestras


**Figura 17.** Análisis de componentes principales biplot con feature "all"

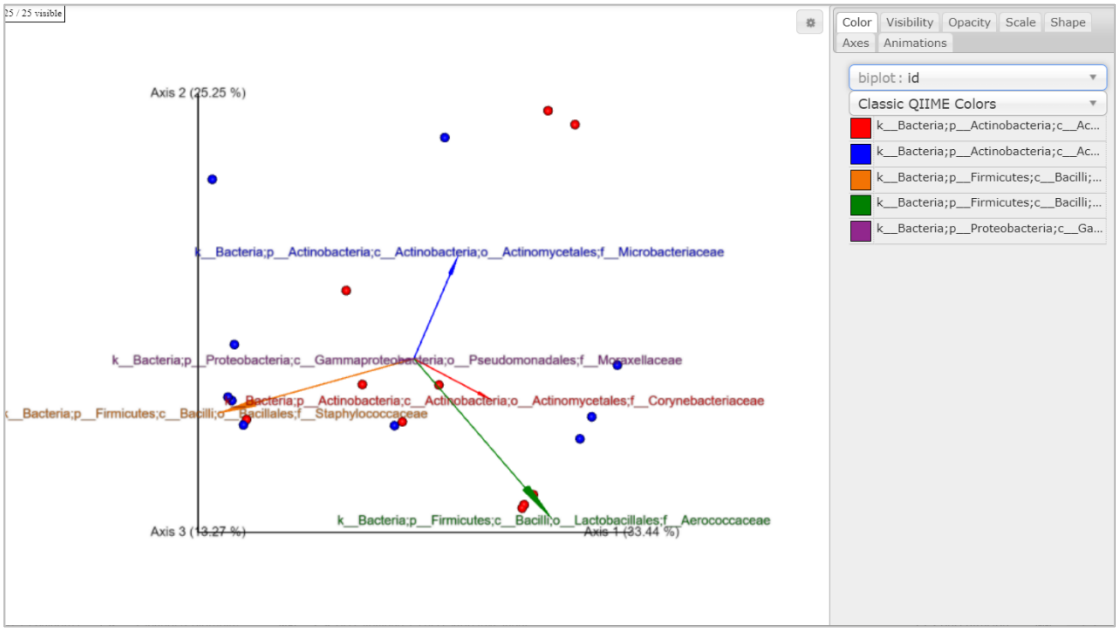


**Figura 18.** Análisis de componentes principales biplot con feature "id"

Al observar cuidadosamente las figuras 17 y 18 resultan evidentes las dos nuevas características que se generan, las cuales permiten caracterizar de una mejor forma la diversa afiliación taxonómica que tienen las muestras, siendo el filo Firmicute uno de los que mas se encuentra en las muestras denotadas de color rojo mientras que las muestras de color azul se encuentran asociaddas al filo Actinobacteria por otra parte, los otros dos vectores no tiene una capacidad de resolución muy alta pues tienen una mezcla de muestras y por ende es necesario un análisis más detallado para describir en mejor forma la comunidad.

Describa las relaciones o asociaciones entre determinados features y algunas muestras. Intente explicar esto con base en los met

**Abundancia Diferencial: LEfSe**

LEfSe es un análisis estadístico que intenta discriminar los features que están diferencialmente abundantes en determinado grupo de muestras. Esto da indicios muy claros del papel de ciertos taxa en ambientes o condiciones específicas.

En este caso usaremos un set de datos diferente que pueden encontrar en: /hpcfs/home/ciencias/biologia/cursos/bcom4102/Datasets/Talle6_Qiime2/Lefse/hmp_small_aerobiosis.txt
Dichos datos son resultado de un estudio en el que se evalua la diversidad bacteriana en diferentes zonas del cuerpo humano, relacionando estos datos con la cantidad de oxigeno disponible en cada zona.

Para cargar lefse use los siguientes comandos:

source ~/anaconda3/bin/activate
conda activate Lefse

El primer paso en LEfSe es formatear la tabla para que pueda ser leída por el programa:

format_input.py hmp_aerobiosis_small.txt hmp_aerobiosis_small.in -c 1 -s 2 -u 3 -o 1000000

Posteriomente se corre el análisis:

run_lefse.py hmp_small_aerobiosis.in hmp_small_aerobiosis.res

```
Number of significantly discriminative features: 51 (131 )
before internal wilcoxon
Number of discriminative features with abs LDA score > 2.0: 51
```

Y finalmente se genera el plot:

plot_res.py hmp_small_aerobiosis.res hmp_small_aerobiosis.png

***Para entregar***
Revise la tabla que usó de input para LEfSe “hmp_small_aerobiosis.txt” y comparela con una de las tablas de ASVs generadas por qiime2 (/hpcfs/home/ciencias/biologia/cursos/bcom4102/Datasets/Talle6_Qiime2/Lefse/feature-table.txt). Que cambios principales nota en el formato.

El cambio principal que encontramos en los fotmatos es que, los datos en la tabla **hmp_small_aerobiosis.txt** investiga de manera más profunda la asociación filogenética de las muestras, incluso presentando información sobre la disponibilidad de oxígeno o la parte del cuerpo de la cual procede. Por otra parte, la tabla **feature-table.txt** nos muestra la información de una manera más superficial

Describa que hace cada una de las opciones del comando que formatea la tabla. (El manual más claro lo pueden desplegar en la misma linea de comandos)

**Opciones de formato que formatea la tabla**

INPUT_FILE archivo de entrada, en nuestro caso **hmp_aerobiosis_small.txt**
OUTPUT_FILE archivo de salida, hmp_aerobiosis_small.in // este es el archivo que usamos posteriormente para el análisis.
**Argumentos opcionales:**
**--output_table** tabla de salida en .txt
**-f {c,r}** revisa si la información está presentada en filas o columnas (por defecto está en filas)
**-c [1..n_feats]** define que feature usa como clase (por defecto 1)
**-s [1..n_feats]** define que feature usa como subclase (por defecto -1 significando que no haya subclase)
**-o** define el valor de la normalización (por defecto es -1.0 significando que no haya normalización)
**-u [1..n_feats]** define que feature usar como sujeto (por defecto es -1 significando que no haya sujeto)
**-m {f,s}** Establece que hacer con valores faltantes (por defecto elimina los features con valores faltantes, por otra parte, s elimina las muestras donde hayan valores faltantes)
**-n int** establece la cardinalidad mínima de cada una de las subclases (las subclases con cardinalidades bajas son agrupadas y si la cardinalidad permanece baja no se realizan comparaciones con ellas
**-biom_c** BIOM_CLASS define que features usar como clase para archivos biom
**-biom_s** BIOM_SUBCLASS define que features usar como subclase para archivos biom

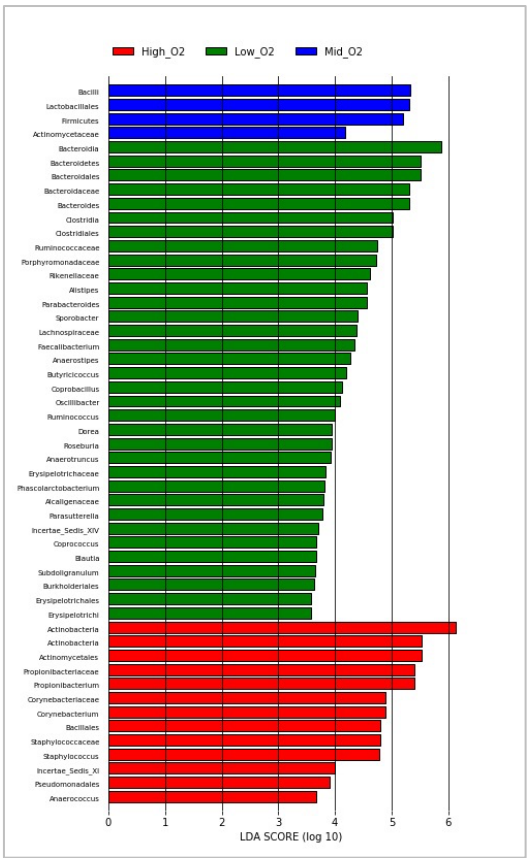


**Figura 19.** Análisis discriminante de la muestra

En la figura 19 se observa claramente como aquellos organismos con un LDA score más alto son aquellas que contribuyen a separarlos de las otras dos características (consumo de oxígeno alto e intermedio) existe un número más alto de bacterias que crecen en condiciones limitadas de oxígeno que aquellas facultativas o que aquellas aeróbicas (Segata y colaboradores 2011). Concretamente aquellos organismos pertenecientes a los Actinobacterias permiten caracterizar de una forma adecuada a los microorganismos que crecen en condiciones con tensión alta de oxígeno mientras que los organismos pertenecientes a los Bacteroidetes permiten un mejor análisis discriminatorio para aquellos que crecen a condiciones de baja tensión de oxígeno.

Por otro lado, resulta evidente que existe un mayor número de filos asociadas a condiciones bajas de oxígenos lo cual probablemente está relacionado con el nicho que ocupan en los distintos lugares de donde fueron obtenidas las muestras como por ejemplo el estómago, resulta interesante también observar cómo los ambientes con una tensión de oxígeno media posee el menor número de filo asociados a este tipo de condiciones por lo que se supondría que los sitios en el cuerpo de donde se obtienen las muestras son lugares con ambientes con condiciones de oxígeno bien marcadas limitando así la presencia de estos organismos facultativos.

**Referencias**


- Lozupone, C., Hamady, M., Kelley, S., & Knight, R. (2007). Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That

Structure Microbial Communities. Applied And Environmental Microbiology, 73(5), 1576-1585. doi: 10.1128/aem.01996-06

- Lozupone, C., Lladser, M., Knights, D., Stombaugh, J., & Knight, R. (2010). UniFrac: an effective distance metric for microbial community comparison. The ISME Journal, 5(2), 169-172. doi: 10.1038/ismej.2010.133

- Ricotta, C., & Podani, J. (2017). On some properties of the Bray-Curtis dissimilarity and their ecological meaning. Ecological Complexity, 31, 201-205. doi: 10.1016/j.ecocom.2017.07.003

- Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. Genome Biology, 12(6), R60. doi: 10.1186/gb-2011-12-6-r60