

| Asignatura | Datos del alumno | Fecha |
|--|------------------------|-------|
| Secuenciación y Ómicas de Próxima Generación | Apellidos: Tomé Castro | 16/12 |
| | Nombre: Xosé Manuel | |



Este análisis de secuencias destaca un procesamiento de datos con lecturas forward y reverse similares en cantidad (mediana de 2083 lecturas). Algunas muestras, como BAQ1370.3, tienen 0 lecturas, posiblemente por baja calidad en los barcodes, mientras que otras, como BAQ4166.1.1, alcanzan 4201 lecturas. La calidad de las lecturas forward se mantiene alrededor de Q30-Q40, pero las reverse disminuyen a Q15 en las bases 100-150. Tras recortar adaptadores, crear submuestra (30%) y eliminar secuencias cortas (<100 lecturas), se lograron 1134 secuencias con una longitud promedio de

227.39 nucleótidos. La mayoría de las muestras superaron el filtro de calidad (mínimo 83.97%), aunque el emparejamiento entre forward y reverse varió, con valores tan bajos como 2% en algunas muestras. Finalmente, tras eliminar quimeras, se obtuvieron 1124 secuencias finales uniformes de 228 nucleótidos, demostrando un procesamiento eficiente.

Pregunta 1: ¿Qué profundidad de muestreo debemos seleccionar?

Al evaluar la **frecuencia por muestra**, observé que **64,377 secuencias totales** están distribuidas entre las 54 muestras en el que 3rd quartile (75% de las muestras) tienen **menos de 1,458 secuencias**. Hay muestras tienen una cobertura variable, con secuencias que oscilan entre 4 y 2,046. Con respecto a la **frecuencia por características**, observo que **64,377 secuencias totales** están distribuidas entre las **1,134 features** (OTUs). Cada feature representa una secuencia única (o un taxón). La mayoría de las **features** son poco frecuentes, con muchas apareciendo solo unas pocas veces, pero algunas son muy abundantes, con hasta 1,454 secuencias. La mayoría de las features tienen **frecuencias bajas** (por ejemplo, menos de 50 reads), lo que puede indicar una comunidad muy diversa, pero con muchos taxones poco representados, los que la feature mas abundante (1,454 reads) podría ser dominante en muchas muestras.

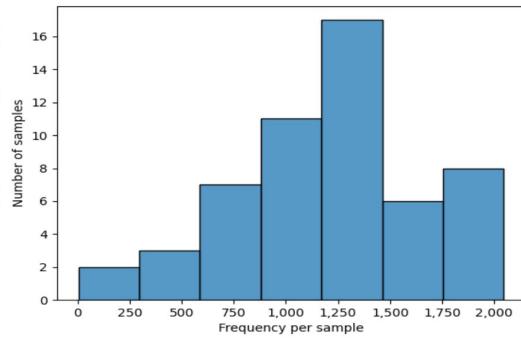
| Asignatura | Datos del alumno | Fecha |
|--|------------------------|-------|
| Secuenciación y Ómicas de Próxima Generación | Apellidos: Tomé Castro | 16/12 |
| | Nombre: Xosé Manuel | |

Table summary

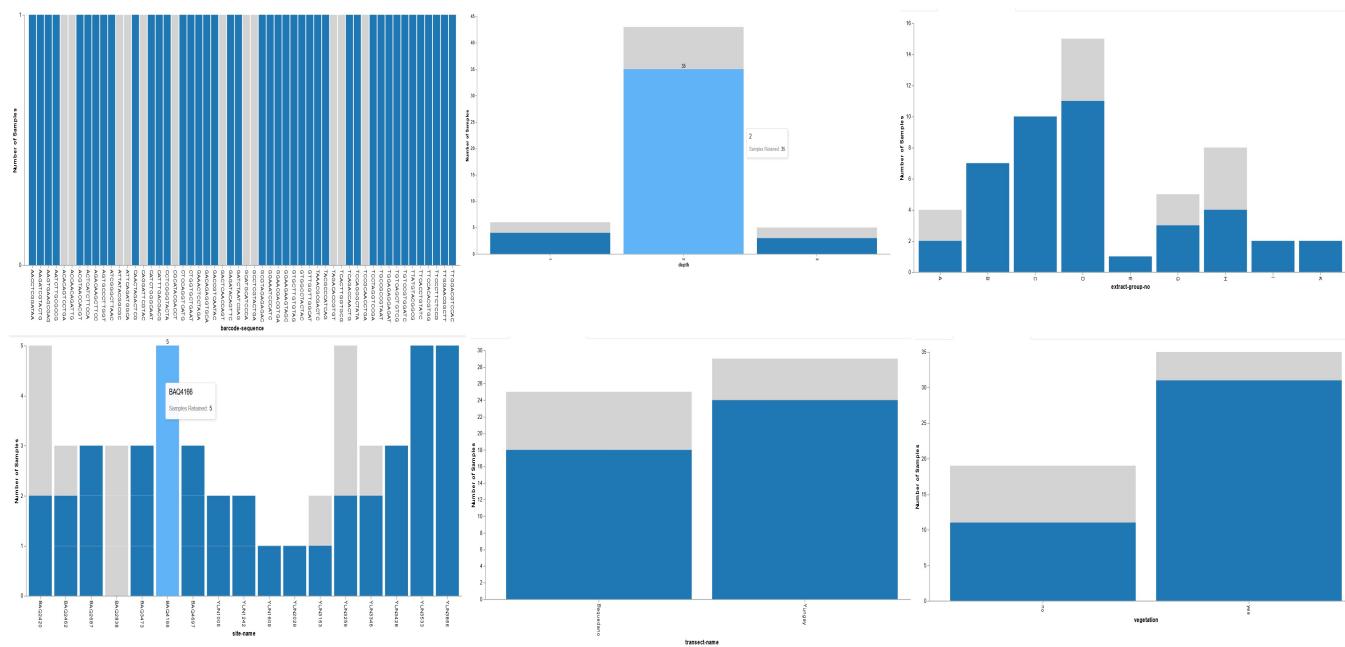
| Summary Statistic | Value |
|--------------------|--------|
| Number of samples | 54 |
| Number of features | 1,134 |
| Total frequency | 64,377 |

Frequency per sample

| | Frequency |
|-------------------|-----------|
| Minimum frequency | 4 |
| 1st quartile | 912.5 |
| Median frequency | 1,200.5 |
| 3rd quartile | 1,457.5 |
| Maximum frequency | 2,046 |
| Mean frequency | 1,192.2 |



Al probar con una profundidad de 1200, se eliminaron demasiadas secuencias, por lo que optar por valores cercanos al primer cuartil, como 900 o 887, resultó ser una mejor opción. Utilizando la herramienta interactiva, seleccioné una profundidad mínima para retener las muestras que cumplen con ese criterio. Por ejemplo, al elegir una profundidad de 900, 49 de 54 muestras (90.74%) cumplen con este umbral, lo que coincide con el hecho de que la mayoría de las muestras tienen más de Q1. A continuación, presento las gráficas de las variables categóricas y numéricas en el archivo de metadatos, con una **profundidad de 887**.

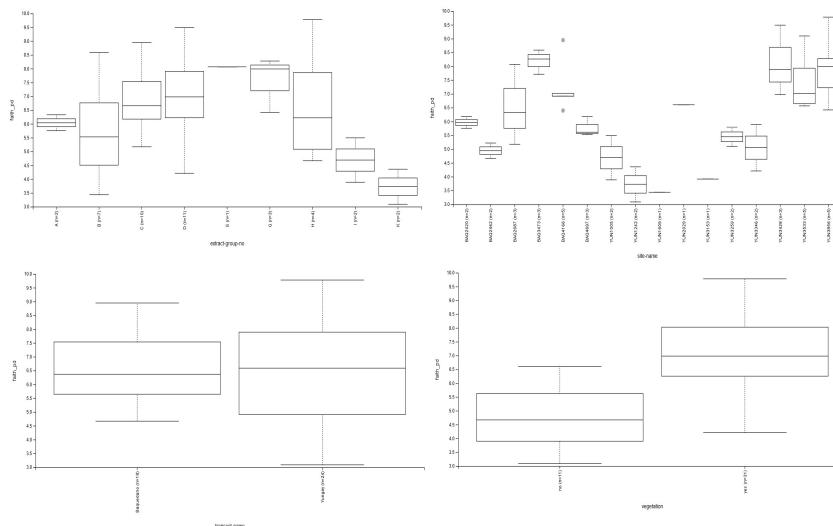


Pregunta 2: ¿Qué categorías de metadatos están asociadas con más fuerza con las diferencias en riqueza de la comunidad? ¿Y con igualdad?

El análisis de **alfa diversidad** se centró en cuatro métricas: evenness, faith, observed y shannon, evaluadas frente a todas las variables categóricas mostrando en la prueba de Kruskal-Wallis diferencias significativas en transect-name y vegetation, mientras que site-name, extract-

| Asignatura | Datos del alumno | Fecha |
|--|---|-------|
| Secuenciación y Ómicas de Próxima Generación | Apellidos: Tomé Castro Nombre: Xosé Manuel | 16/12 |

group-nº no mostró diferencias relevantes. Esto indica que ciertos hábitats o tipos de vegetación tienen comunidades microbianas más diversas, con una mayor cantidad de OTUs.

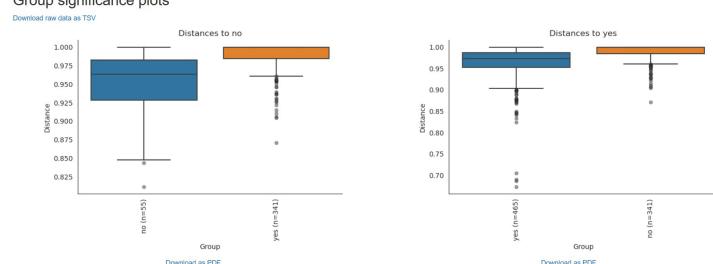


De las métricas, **faith** (mostrada a la izquierda) fue la más sensible en inferir las diferencias en **OTUs** (**p-value y q-value de 0.00008-0.7 y H 2.82-32**), seguida de shannon. En cambio, evenness y observed tuvieron resultados menos consistentes frente a todas las variables categóricas. Faith reveló que las diferencias en la riqueza de OTUs entre los grupos analizados no solo se debían al número de especies, sino a la presencia de grupos filogenéticamente distintos, con

variabilidad filogenética dentro de las comunidades microbianas.

En el análisis de **betadiversidad**, utilicé dos tipos de archivos: *emperor* y *qzv*. El índice de **Jaccard (gráfica izquierda)**, basado en la **similitud en términos de composición**, mostró el mejor desempeño en el análisis de betadiversidad. Este índice, que compara comunidades considerando únicamente la presencia o ausencia de especies, sugiere que las diferencias entre los grupos microbianos son principalmente estructurales (qué especies están presentes). Por lo tanto, las comunidades microbianas difieren más en la composición de

Group significance plots



especies que en la cantidad de cada una.

La vegetación tuvo la mayor influencia en la estructura de las comunidades, con un rango alto de pseudo-F (2.3–8.42), lo que resalta su papel en la composición microbiana. Los resultados fueron altamente significativos (valores p y q de 0.001), confirmando la robustez del análisis. En el análisis de PCoA *unweighted* refleja diferencias en la presencia/ausencia de taxones, explicando menos variación (X: 8.82%, Y: 14.99%) y siendo útil para cambios generales en la composición. Por el contrario, el análisis *weighted* considera la abundancia relativa y explica mayor variación (X: 13.57%, Y: 28.44%), mostrando que las diferencias entre comunidades están influenciadas por la dominancia de ciertos taxones. Esto sugiere que gradientes

| Asignatura | Datos del alumno | Fecha |
|--|------------------------|-------|
| Secuenciación y Ómicas de Próxima Generación | Apellidos: Tomé Castro | 16/12 |
| | Nombre: Xosé Manuel | |

ambientales como el pH o la cobertura favorecen la proliferación de grupos específicos, reflejando adaptación funcional y estructuración diferencial en estas comunidades microbianas.

Pregunta 3: ¿Qué pasa si evaluamos algunas de las secuencias con BLAST? ¿Son las clasificaciones taxonómicas diferentes a las de qiime2? ¿A qué nivel taxonómico surgen las diferencias? fueron altamente significativos (0.001), lo que confirma la robustez de los resultados.

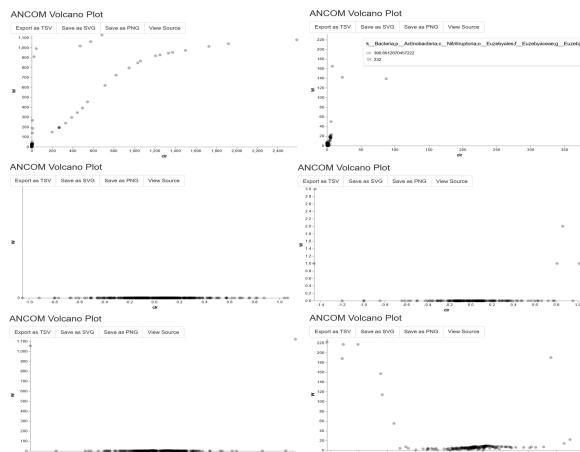
Usé el archivo taxonomy.qzv para visualizar los organismos hallados en la muestra junto a su filogenia. Después, seleccioné del archivo req.qzv el nombre asignado a los organismos con base en las tres secuencias identificadas y utilicé BLASTn con el parámetro **16S_ribosomal_RNA** para evaluar las coincidencias taxonómicas ya que contiene regiones altamente conservadas (idénticas en muchas especies) y están bien representadas en las bases de datos como BLASTn. Como se muestra en la tabla adjunta, las clasificaciones asignadas por QIIME2 y BLASTn presentan diferencias principalmente en el nivel de especie. QIIME2 es más conservador, asignando taxonomía solo cuando hay alta confianza estadística (e.g., dejando categorías como "género" o "familia" sin especificar). En contraste, BLASTn asigna taxonomía hasta el nivel más específico disponible, basado en similitudes directas con su base de datos. Esto generó discrepancias notables, especialmente en los niveles más bajos (género y especie), mientras que en niveles superiores (orden, familia) ambas herramientas fueron bastante parecidas.

| Id | Filogenia | Confianza | BLASTn Filogenia |
|----------------------------------|--|-----------------|--|
| 47db8912a5b600c9c2ecfb47ac4ee9a1 | k_Bacteria; p_Actinobacteria; c_Rubrobacteria; o_Rubrobacterales; f_Rubrobacteraceae; g_Rubrobacter; s_ | 0.9999999907808 | k_Bacteria; p_Actinobacteria; c_Rubrobacteria; o_Rubrobacterales; f_Rubrobacteraceae; g_Rubrobacter; s_Rubrobacter spartanus |
| 4987189e46ad46e93054a098eb5f74fb | k_Bacteria; p_Actinobacteria; c_Thermoleophilia; o_Solirubrobacterales; f_ g_; s_ | 0.9986925667971 | k_Bacteria; p_Actinobacteria; c_Thermoleophilia; o_Solirubrobacterales; f_ g_Solirubrobacter s_Solirubrobacter taibaiensis |
| 58bff7fe799e21d15a0d2891a44e60d2 | k_Bacteria; p_Actinobacteria; c_Actinobacteria; o_Acinomycetales; f_Glycomycetaceae; g_Glycomyces; s_harbinensis | 0.9986925667971 | k_Bacteria; p_Actinobacteria; c_Actinobacteria; o_Acinomycetales; f_Glycomycetaceae; g_Glycomyces; s_harbinensis |

| Asignatura | Datos del alumno | Fecha |
|--|------------------------|-------|
| Secuenciación y Ómicas de Próxima Generación | Apellidos: Tomé Castro | 16/12 |
| | Nombre: Xosé Manuel | |

Pregunta: ¿Qué géneros son diferentes entre las muestras?

Utilicé dos archivos diferentes de ANCOM para evaluar las diferencias en la abundancia bacteriana entre los grupos. El primero, denominado ANCOM, contenía las etiquetas "extract-group-nº"(1º grafica), "transect-name" (2º grafica) y "vegetación" (3º grafica). El segundo archivo, 16-ANCOM (específicamente adaptado para datos **16S rRNA**), utilizaba un enfoque de análisis similar, pero con un orden taxonómico más detallado, de 6 niveles, es decir, hasta el orden taxonómico género. En ambos casos, al analizar la etiqueta "transect-name", no se obtuvieron resultados estadísticamente significativos, lo que me llevó a centrar la atención en otras etiquetas y en la estructura taxonómica de los datos.

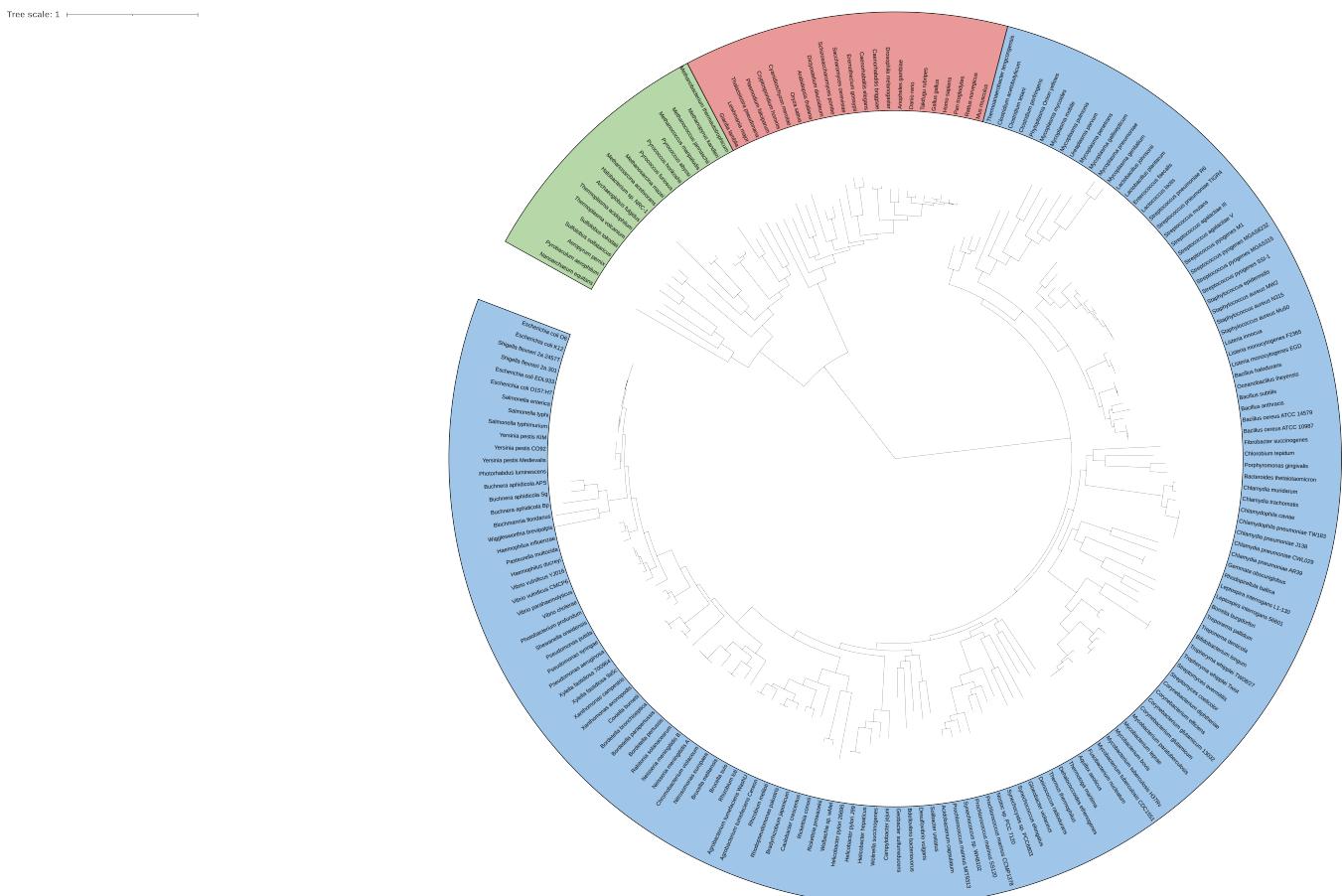


En el análisis ANCOM, la etiqueta **extract-group-nº** mostró diferencias en las abundancias bacterianas, con un patrón claro en géneros como *Acidimicrobia* y *Thermoleophilia* (grafica izquierda). En cuanto a **vegetación**, las diferencias más notables fueron en géneros como *Ralstonia* y *Arthrobacter*, que fueron más abundantes en el grupo "no" que en el grupo "yes". Sin embargo, ANCOM no detectó una gran complejidad taxonómica, con diferencias limitadas a unos pocos géneros. Por otro lado, el análisis **16-ANCOM** proporcionó un enfoque más detallado, revelando diferencias significativas en más características bacterianas. Por ejemplo, *Euzebya*

mostró una abundancia constante entre grupos en el **extract-group-nº**, mientras que en el análisis de **vegetación**, *Ralstonia* fue más prominente en el grupo "no" y *DA101* en el grupo "yes". El enfoque de 16-ANCOM permitió identificar patrones microbianos más específicos y destacó una mayor variabilidad en las abundancias bacterianas, ofreciendo una resolución más alta para distinguir entre las comunidades microbianas en función de la vegetación. En los análisis ANCOM y 16-ANCOM, se identificaron diferencias significativas en la abundancia de varios géneros bacterianos entre las muestras, especialmente en relación con las variables **extract-group-nº** y **vegetación**. En **extract-group-nº**, géneros como *Ralstonia* y *Arthrobacter* mostraron diferencias claras en su abundancia entre los grupos A, B y C. En **vegetación**, *Ralstonia* y *Arthrobacter* fueron más abundantes en el grupo "no", mientras que *DA101* (familia *Chthoniobacteraceae*, orden *Chthoniobacterales*, clase *Verrucomicrobia*) destacó en el grupo "yes". Estos resultados indican que tanto las condiciones de vegetación como el factor **extract-group-nº** influyen de manera significativa en la composición y abundancia de estos géneros bacterianos en las muestras.

| Asignatura | Datos del alumno | Fecha |
|--|---|-------|
| Secuenciación y Ómicas de Próxima Generación | Apellidos: Tomé Castro Nombre: Xosé Manuel | 16/12 |

ANEXO



qiime tools export \

--input-path rooted-tree.qza \

--output-path exported-tree

Los colores son: Azul: Bacteria, Naranja: Eukaryota y Verde: Archaea