Taller 4: Qiime 2

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

Objetivo: Familiarizarse con el uso básico de softwares diseñados para el análisis de microbiomas o datos basados en amplicones.

Instrucciones:

Preparación:

1. Ingrese a la cuenta del cluster.
2. En su directorio, cree un nuevo directorio y nómbrelo Taller4_Qiime2_WF.
3. Copie todo el directorio 16S_CleanData a su directorio Taller4_Qiime2_WF. La siguiente es la ruta donde encuentran el archivo a copiar: `/hpcfs/home/ciencias/biologia/cursos/bcom4102/Datasets/16S_CleanData`
4. Inicie una sesión interactiva
5. Cargue el modulo de Qiime 2, disponible en el cluster.

Importar datos

1. En este paso, tomaremos el directorio que tiene **exclusivamente** las secuencias, para convertirlo al formato con el que trabaja qiime2 (.qza).

```
qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]'
--input-path 16S_CleanData --input-format
CasavaOneEightSingleLanePerSampleDirFmt --output-path demux-paired-
end.qza
```

2. Ahora crearemos el archivo de visualización, correspondiente a los datos importados

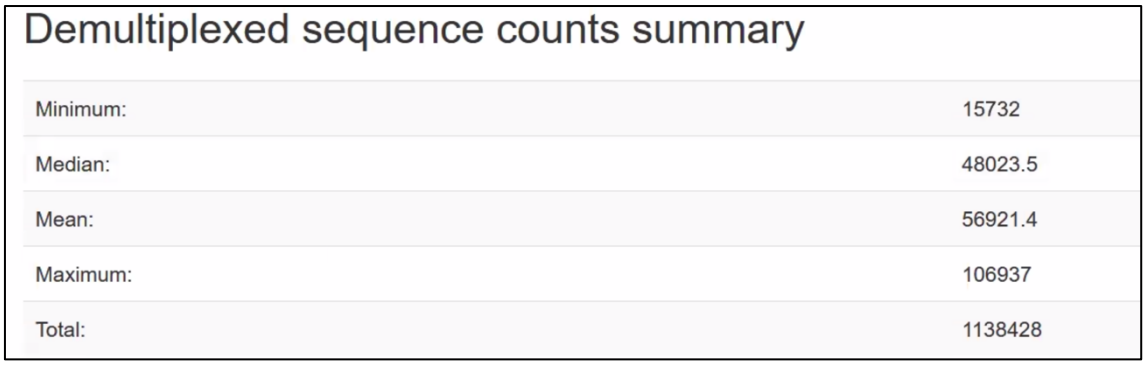
```
qiime demux summarize --i-data demux-paired-end.qza --o-visualization
demux-paired-end.qzv
```

3. Recuerde que el formato del mapping file ya es compatible con el sistema, por lo que no hay necesidad de importarlo.

Para entregar:

- Con la información dada por el summary, reporte las muestras con mayor y menor número de reads respectivamente, así como número de reads promedio.

  El mayor número de reads fue: 106937
  El menor número de reads fue: 15732
  El numero promedio de reads fue: 56921.5



Demultiplexed sequence counts summary

Minimum:	15732
Median:	48023.5
Mean:	56921.4
Maximum:	106937
Total:	1138428

- Revise el mapping file y describa que variables se pueden utilizar para posteriores análisis que impliquen clusterizar datos por categorías.

Revisando el archivo de mapping las variables que se utilizarían para hacer un clustering serían: Sample ID, Desease State y sample title

Generación de ASVs

1. En este paso se realiza simultáneamente la derreplicación, el denoising y la generación de las ASVs. Para este caso usaremos DADA2 en su modo de paired end:

```
qiime dada2 denoise-paired --i-demultiplexed-seq --p-trunc-len-f X --
p-trunc-len-r X --o-representative-sequences RepSeq_Dada2.qza --o-
table Tabla_Dada2.qza --o-denoising-stats Stats_Dada2.qza --p-n-
threads 4
```

Nota: usted deberá seleccionar los valores para --p-trunc-len-f y --p-trunc-len-R con base en la información disponible en la visualización de los datos importados.

F: 140
R :130

1. Ahora convertiremos la tabla y las estadísticas a archivos visualizables.

```
qiime feature-table summarize --i-table Tabla_Dada2.qza --m-sample-
metadata-file MFile_TallerQiime2.txt --o-visualization Tabla_Dada2.qzv

qiime metadata tabulate --m-input-file Stats_Dada2.qza --o-
visualization Stats_Dada2.qzv
```

Para entregar:

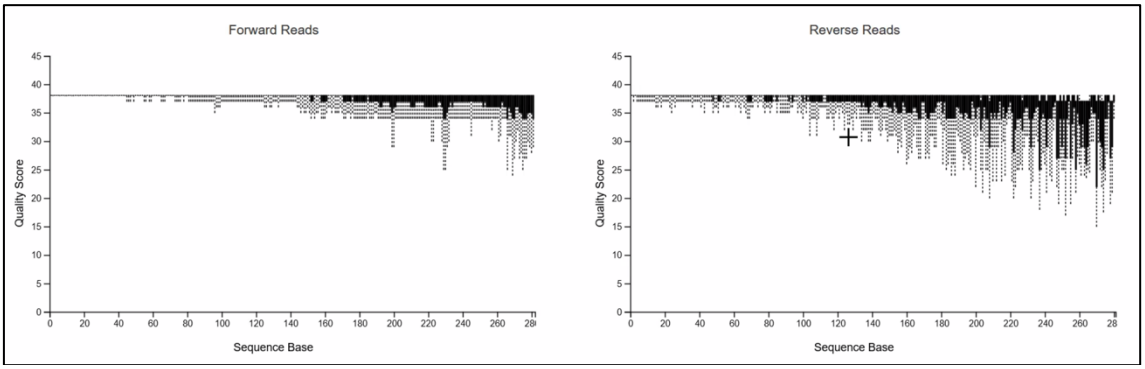
- Explique la utilidad de los flags: --p-trunc-len-f y --p-trunc-len-R. Averigüe por qué se consideran como requeridos para DADA2.

Es importante entender que DADA2 (ASV métodos conocidos actualmente) está utilizando observaciones repetidas de la verdadera secuencia biológica para distinguir las secuencias reales de los errores R y f. Por lo tanto, para que se detecten variantes de secuencia, debe haber al menos 2 lecturas sin errores de esa secuencia en el conjunto de datos, y la sensibilidad de estos métodos a variantes raras está restringida por la fracción de lecturas sin errores en los datos.

El parámetro --p-trunc-len solo debe usarse para recortar secuencias de referencia si las secuencias de consulta se recortan a esta misma longitud o menos. Las secuencias de extremos emparejados que se unen correctamente suelen tener una longitud

variable. Las lecturas de un solo extremo que no están truncadas en una longitud específica también pueden tener una longitud variable. Para la clasificación de lecturas de extremos emparejados y lecturas de un solo extremo sin recortar, recomendamos entrenar un clasificador en secuencias que se han extraído en los sitios de cebadores apropiados, pero que no están recortadas. (Callahan, 2021)

- Proporcione los valores que utilizó para el proceso y la razón por la cual los escogió.



La razón por la cual escogimos los valores 140 para --p-trunc-len-f, y 130 para --p-trunc-len- es por la disminución en la calidad de las lecturas.

- Visualice el archivo de estadísticas de DADA2 y describa que filtro se hace en cada uno de los pasos del proceso.

sample-id #q2:types	input numeric	filtered numeric	denoised numeric	merged numeric	non-chimeric numeric
SRR12614450	79769	79769	79733	15	15
SRR12614451	28083	28083	28050	2	2
SRR12614452	40707	40707	40684	6	5
SRR12614453	38811	38811	38789	0	0
SRR12614454	37470	37470	37454	4	4
SRR12614455	42582	42582	42410	4	4
SRR12614456	106937	106937	106873	10	10
SRR12614457	84492	84492	84442	16	16
SRR12614458	56463	56463	56442	4	4
SRR12614459	80986	80986	80915	24	24
SRR12614506	43488	43488	43423	53	53
SRR12614507	49132	49132	49111	3	3
SRR12614508	15732	15732	15709	2	2
SRR12614509	46203	46203	46111	19	19
SRR12614510	46915	46915	46888	13	12
SRR12614511	33488	33488	33322	61	61
SRR12614512	82576	82576	82507	22	22
SRR12614513	83065	83065	82985	20	20
SRR12614514	74957	74957	74702	0	0
SRR12614515	66572	66572	66514	109	109

Paso 1: Inspeccionar los perfiles calidad de las lecturas

Se debe empezar visualizando los perfiles de calidad de las lecturas forward

Paso 2: Filtrado y trimming

Asignar los nombres para los archivos filtrados fastq.gz

Paso 3: Derreplicación

El paso de derreplicación combina todas las lecturas de secuenciación identicas en unas "secuencias únicas" con su correspondiente "abundancia" que corresponde al numero de

lecturas con esa única secuencia. El paso de dereplicación reduce sustancialmente el tiempo de computo debido a que elimina las comparaciones redundantes.

Paso 4: **Merge de lecturas**

En este paso se hace merge de las las lecturas "foward" y "reverse" para obtener las secuencias completas sin ruido. El proceso de merge se realiza alineando las lecturas forward sin ruido con el complemento directo de su correspondiente lectura reverse sin ruido y así construir las secuencias denominadas "contig" unidas. Por defecto, la union de las secuencias solo es producida si se sobreponen al menos 12 bases entre las lecturas forward y reverse, y si son identicas la una a la otra en la región que se sobreponen.

Paso 5: Remoción de quimeras

El método core dada corrige los errores de sustitución y sus valores, pero las quimeras permanecen. Afortunadamente, la precisión de las variantes de secuencia después de la eliminación de ruido hace que la identificación de quimeras sea más simple de lo que es cuando se trata de OTU difusas. Las secuencias quiméricas se identifican si se pueden reconstruir exactamente combinando un segmento izquierdo y un segmento derecho de dos secuencias "parentales" más abundantes

- Indique el número de features totales que obtuvo después del proceso.

El número de features fue: 22, (Estaki et al., 2020)

Metric	Sample
Number of samples	20
Number of features	22
Total frequency	385

Generación de árbol filogenético

1. Con el fin de realizar métricas de diversidad filogenéticas, es necesario construir primero el árbol filogenético. En este caso se construirá un árbol tipo insertion placement.

```
qiime fragment-insertion sepp --i-representative-sequences
RepSeq_Dada2.qza --p-threads 4 --o-tree insertion-tree.qza --o-
placements insertion-placements.qza
```

2. Ya con el árbol, se puede filtrar, y volver a visualizar la tabla de ASVS.

```
qiime fragment-insertion filter-features --i-table Tabla_Dada2.qza --
i-tree insertion-tree.qza --o-filtered-table TreeFiltered_Table.qza --
o-removed-table TreeRemoved_table.qza

qiime feature-table summarize --i-table t TreeFiltered_Table.qza --m-
sample-metadata-file MFile_TallerQiime2.txt --o-visualization
TreeFiltered_Table.qzv
```

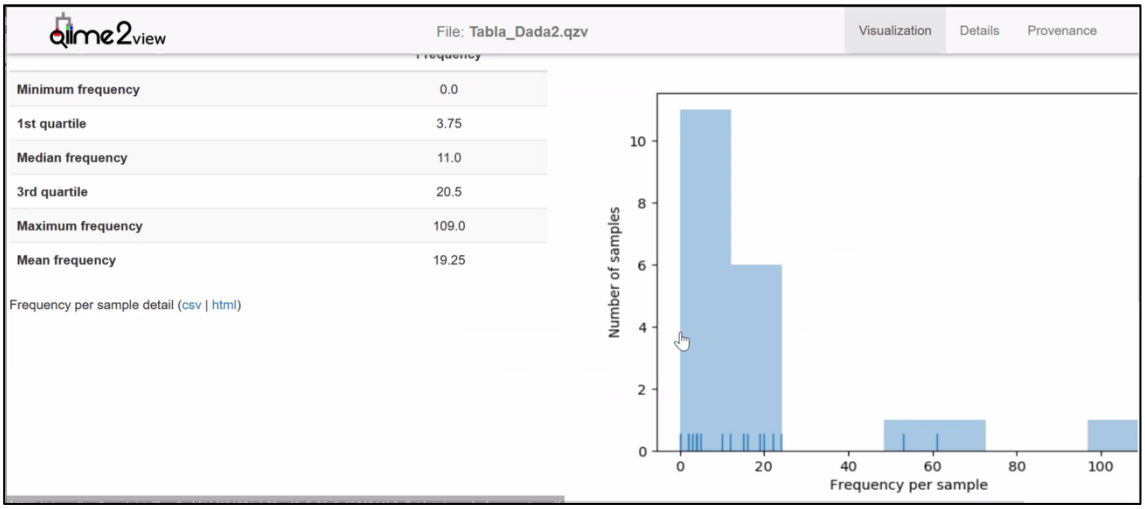
Para entregar:

- Describa en que consiste el método de insertion tree/insertion placement, y resalte sus diferencias con el método tradicional de construcción de árboles.

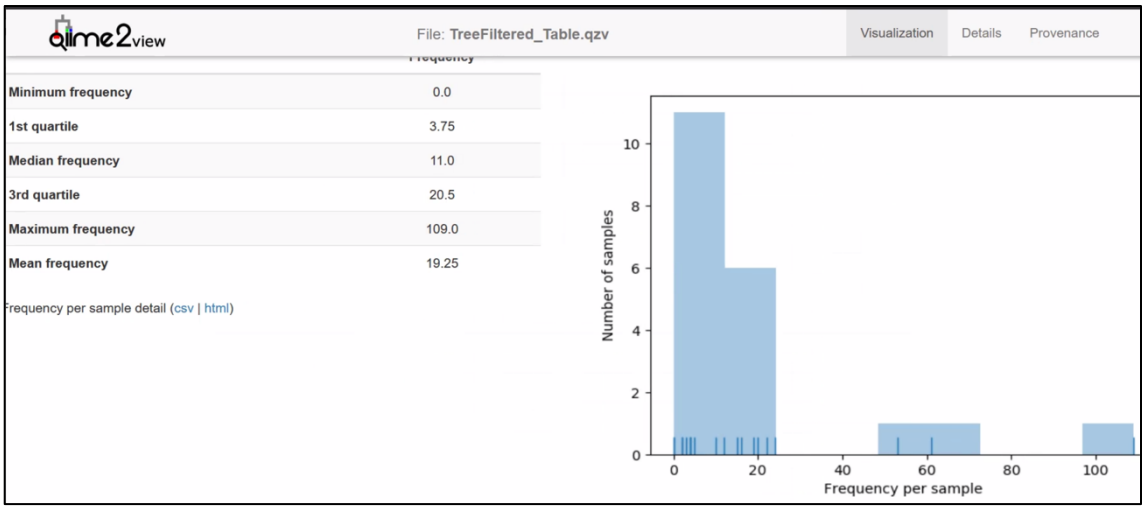
Hasta la fecha, muchos modelos evolutivos en métodos de inferencia filogenética solo han tenido en cuenta los eventos de sustitución, no las inserciones y eliminaciones que pueden tener lugar en la información genética del organismo analizado. Como resultado, los métodos de inferencia de árboles no solo usan menos información de secuencia de la que podrían, sino que también han resultado de difícil integración al modelado filogenético en los métodos de alineación de secuencias (como perfiles y modelos ocultos de Markov) que inherentemente requieren un modelo de inserción. y eventos de eliminación. Por lo tanto, un objetivo importante en el campo ha sido desarrollar modelos evolutivos que permitan el manejo de inserciones y deleciones a lo largo del tiempo con suficiente precisión para aumentar el poder de las búsquedas de homologías de secuencias basadas en perfiles. (Hall, 2004)

Una de las principales características del método tradicional es el uso de marcadores filogenéticos con fines informativos para verificar la veracidad de las referencias en los arboles, dichos marcadores se establecen manualmente. (Reddy, 2011); por otra parte, al analizar el método insertion tree/insertion placement se observa la implementación de algoritmos que ayudan a establecer las referencias de manera mecánica, ya que el proceso de selección de referencias manual se ha vuelto insostenible debido a la gran cantidad de datos. (Czech, Barbera & Stamatakis, 2018). En otras palabras, insertion placement ataca el problema de encontrar la posición óptima para una nueva referencia existente. La colocación es todo lo que se necesita y, en términos de precisión, es tan buena como, y quizás incluso mejor, que la reconstrucción de *novo*. Además, la ubicación puede ser más escalable que la reconstrucción de *novo* cuando se trata de árboles muy grandes. (Balaban, Sarmashghi & Mirarab, 2019)

- Compare la tabla inicial de ASVs con la nueva tabla filtrada y discuta que sucede a nivel de conteos y frecuencia de ASVs.



Al realzar la comparación de los datos de ASV obtenidos inicialmente y los ASV que se obtienen luego de realizar el árbol filogenético, es evidente que las frecuencias por muestra vs el número de muestras se mantienen exactamente igual, tal y como se muestra en las siguientes dos figuras:



Por tanto, ambos análisis son similares debido a que los fragmentos fueron insertados tanto para realizar el análisis del árbol como para los demás procesos ejecutados con QIMME2

Bibliografía

Janssen, S., McDonald, D., Gonzalez, A., Navas-Molina, J., Jiang, L., & Xu, Z. et al. (2018). Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information. *Msystems*, *3*(3). doi: 10.1128/msystems.00021-18

Hall, B. (2004). Comparison of the Accuracies of Several Phylogenetic Methods Using Protein and DNA Sequences. *Molecular Biology And Evolution*, *22*(3), 792-802. doi: 10.1093/molbev/msi066

Estaki, M., Jiang, L., Bokulich, N., McDonald, D., González, A., & Kosciolek, T. et al. (2020). QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data. *Current Protocols In Bioinformatics*, *70*(1). doi: 10.1002/cpbi.100

Callahan, B. (2021). DADA2 Pipeline Tutorial (1.16). Retrieved 20 February 2021, from https://benjjneb.github.io/dada2/tutorial.html

Czech, L., Barbera, P., & Stamatakis, A. (2018). Methods for automatic reference trees and multilevel phylogenetic placement. *Bioinformatics*, *35*(7), 1151-1158. doi: 10.1093/bioinformatics/bty767

Reddy, Niranjan. (2011). Basics for the Construction of Phylogenetic Trees. Webmedcentral BIOLOGY. 1. WMC002563.

Balaban, M., Sarmashghi S., Mirarab S.(2019). APPLES: Scalable Distance-based Phylogenetic Placement with or without Alignments
bioRxiv 475566; doi: https://doi.org/10.1101/475566