

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/339617066>

Tutorial básico en español de Qiime2

Method · February 2020

DOI: 10.13140/RG.2.2.18061.69609

CITATIONS

2

READS

10,491

1 author:



[Diego A. Garza](#)

Centro de Investigación Científica de Yucatán

4 PUBLICATIONS 4 CITATIONS

SEE PROFILE

Tutorial básico en español de Qiime2

Biól. Diego Alberto Garza González

Unidad de Recursos Naturales, Centro de Investigación Científica de Yucatán, A. C. (CICY), Calle 43 No. 130 x 32 y 34,
Col. Chuburná de Hidalgo, 97205,
Mérida, Yucatán, México 25 febrero 2020

Contenidos

1. Buenas prácticas informáticas, de alguien que no es informático y comandos comunes para la Terminal
2. Creación de archivo *metadata*
3. Importación de secuencias *pair-end* crudas en formato CASAVA 1.8, *Demultiplexing* o demultiplexado, Número de *reads* y visualización de calidad de las secuencias
4. *Denoising* con DADA2 y generación de datos visuales después de control de calidad (Secuencias representativas, Tabla de frecuencia y estadísticas de filtrado de secuencias)
5. Asignación taxonómica de secuencias ITS y 16S
6. *Taxa collapse* y filtrado de cloroplastos y mitocondrias de secuencias 16S
7. *Barplots*, gráficos de abundancia relativa

Prefacio:

Realizo esta guía o pequeño tutorial como un bloc de notas para seguir una metodología y llevar cuenta de los comandos frecuentes y necesarios para llevar acabo la *pipeline* de Qiime2 ¹ (Se pronuncia *chaim*). Esto es independiente a los excelentes tutoriales que existen en la web oficial (<https://qiime2.org/>) en la sección de tutoriales, además del apoyo de la comunidad del foro. Esta intencionado a aquellos que se les dificulta navegar a través del lenguaje de línea de comando, más el lenguaje inglés.

Realizo este documento en formato libre y a mi gusto.

Asumo que lograron instalar Qiime2 versión 2019.10, y que están usando un sistema “tipo” Linux. Yo trabajo en mi MacBook Pro (13-inch, Mid 2012) con la modificación de 8G, si tengo la urgencia de correr un análisis utilizo nuestro equipo de supercómputo “Hobon”, pero todo el editado de comandos lo realizo en [Sublime Text](#) con licencia gratuita, así mismo trabajo en mi PC de escritorio con sistema Ubuntu.

Cualquier sugerencia es (tal vez) bienvenida.

~Diego

diego.garza@cicy.mx

¹ Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... & Bai, Y. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature biotechnology*, 37(8), 852-857.

Buenas prácticas informáticas, de alguien que no es informático

En informática hay un par de cosas importantes que pueden ser clave entre un **ERROR** y terminar a tiempo tu análisis. Me refiero a que no soy informático, porque no lo soy. Pero mi experiencia me ha llevado a solucionar y llevar a cabo estas recomendaciones.

- Mayúsculas y minúsculas cuentan, si algo se llama *secuencias.fasta* es otro archivo a *Secuencias.fasta* acostúmbrate a escribir en un solo tipo de formato.
 - Mayusculas_Cada_Palabra_Nueva,
 - todo_en_minsuculas,
 - TODO_EN_MAYUSCULAS
- Por buen uso de la lengua utilizamos acentos, sin embargo esto en informática puede ser confuso, sobre todo si compartimos nuestro trabajo con lenguajes de otras regiones. Sí, es poco elegante, pero evita usar acentos en carpetas o nombres “Secuenciacion_Diciembre” “secuencias_fungicas” “proyecto_yucatan”
- Habrás notado ya otra cosa, los guiones para separar palabras. Informáticamente es difícil que existan enunciados, cada palabra usualmente es una instrucción. Por lo que, carpetas o nombres si quieres separaciones agrega guiones, de preferencia guiones bajos.
 - Evita puntos como separadores, a veces estos indican también extensiones como archivo.txt, archivo.fasta, archivo.pdf
 - también evita guiones altos, sobre todo al inicio de las palabras “-secuencias-diciembre.fasta” algunos guiones significan comandos extras
 - Lo más adecuado es: “Secuencias_Febrero.fasta” está en la carpeta “Proyecto_Yucatan”

Creación de archivo *metadata*

Qiime2 nos pide un archivo con nuestros metadatos, es decir, toda la información relevante de nuestro estudio. Y también para realizar el *demultiplexing* o demultiplexado es necesario, ya que así relacionamos las corridas del secuenciador con su *barcode*.

El formato es bastante simple, un archivo de texto separado por tabulaciones, es decir en formato **.tsv**. Cada tabulación representa una columna. La información indispensable son las primeras dos columnas: id y barcode-sequence

El id representa la etiqueta con la que nombraremos ese *barcode*, es decir esa muestra.

El barcode-sequence como su nombre indica es el *barcode* al que asociaremos esta muestra

La elección del *software* para crearlo puede modificar el formato agregando caracteres. La manera más simple, a mi parecer, es crear una hoja de cálculo en Google (https://www.google.com/intl/es-419_mx/sheets/about/) aunque también se puede realizar en Excel de Microsoft.

En el caso de hojas de cálculo de Google es un formato como el siguiente,

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	id	barcode-sequence	Ubicación	Temporada	Tipo de Muestra	Obtencion	ng/uL								
2	Muestra1	AAAAAAA	Centro	Alta	Tratamiento	Humberto	76.4								
3	Muestra2	AAAAAAA	Sur	Baja	Control	Humberto	59.8								
4	Muestra3	AAAAAAC	Centro	Alta	Control	Abril	39.8								
5	Muestra4	AAAAAAC	Sur	Baja	Control	Abril	64.8								
6	Muestra5	AAAAAAG	Centro	Alta	Control	Humberto	63.6								
7	Muestra6	AAAAAAG	Sur	Baja	Tratamiento	Humberto	177								
8	Muestra7	AAAAAAT	Centro	Alta	Control	Diego	12.8								
9	Muestra8	AAAAAAT	Oeste	Baja	Tratamiento	Diego	12.8								
10	Muestra9	AAAAACC	Oriente	Alta	Control	Humberto	151								
11	Muestra10	AAAAACA	Norte	Alta	Tratamiento	Humberto	103								
12	Muestra11	AAAAACA	Centro	Alta	Control	Abril	115								
13	Muestra12	AAAAACC	Centro	Baja	Tratamiento	Abril	182								
14															
15															
16															
17															
18															
19															
20															
21															
22															
23															
24															
25															

disponible en este link: <https://docs.google.com/spreadsheets/d/1Y-urF6FSw7-w6H8A1x3vKD4q6L2aP4KFal-qEyrFSIg/edit?usp=sharing>

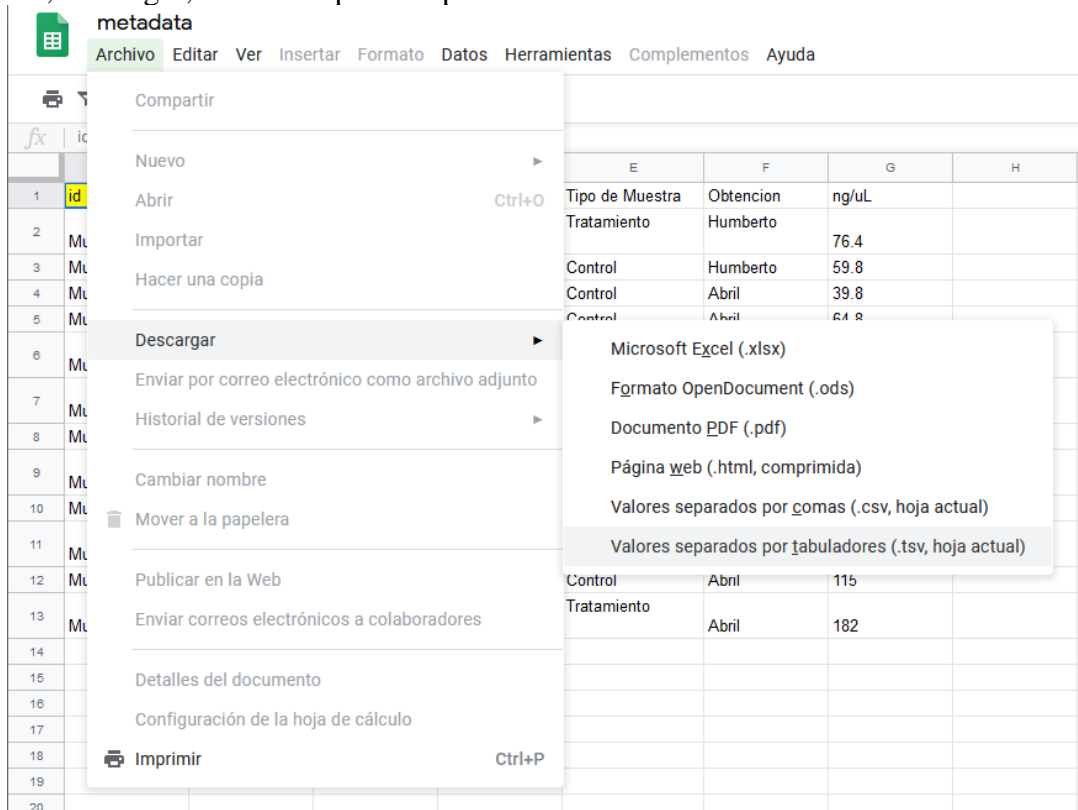
En mi caso puse Ubicación, Temporada, Tipo de muestra, quien tomó la muestra y su concentración.

Otros casos pueden ser: Enfermos vs Sanos, Temporada de secas vs lluvias, con antibiótico A o B, donde fue colectada la muestra, y cuantas cosas tengan que ver con nuestros datos.

Algo quizá importante de agregar es información metodológica de la extracción de ADN, por ejemplo como yo agregué en las últimas dos columnas, la concentración o calidad de nuestras muestras, ya que si una muestra

estaba comprometida de este modo podemos ver como se desempeña, o quien realizó dicha muestra, para saber si dicha persona pudo contaminar las muestras, etc.

Una vez llenada nuestra tabla, procedemos a descargarla de la siguiente manera:
Vamos a archivo, descargar, Valores separados por tabuladores



Guardamos el archivo, y ¡voilà! Ya tenemos nuestro archivo metadata.

Importación de secuencias *pair-end* crudas en formato CASAVA 1.8, *Demultiplexing* o demultiplexado, Número de *reads* y visualización de calidad de las secuencias

Para importar en este formato necesitamos tener a la mano los *barcodes* que nos proporciona la casa de secuenciación y nuestros archivos crudos (*raw sequences*) en formato FASTQ, es decir, **.fastq**

Los primeros pasos antes de siquiera activar qiime2 consisten en crear el entorno y los archivos necesarios para poder trabajar en qiime2.

Paso_1

Primeramente, crearemos una carpeta donde depositaremos dichas secuencias, esta carpeta debe contener únicamente las secuencias. Desde nuestra Terminal podemos dirigirnos al escritorio y crear la carpeta.

El comando **mkdir** creará un directorio, lo llamaremos “Carpeta_de_Trabajo”.

```
mkdir /data/home/diego.garza/Desktop/Carpeta_de_Trabajo
```

A continuación crearemos dentro de “Carpeta_de_Trabajo” otra carpeta llamada “Raw_sequences” y otra llamada Qiime2

```
mkdir /data/home/diego.garza/Desktop/Carpeta_de_Trabajo/Raw_sequences
```

```
mkdir /data/home/diego.garza/Desktop/Carpeta_de_Trabajo/Qiime2
```

Podemos depositar en esta carpeta recién creada nuestras secuencias.

Paso_2

Posteriormente debemos renombrar nuestros archivos para el formato CASAVA 1.8, en este formato la primera parte del nombre corresponde al “*id*” o identificador, la segunda al barcode, el tercero el número de línea (por default pongo L001), cuarto la dirección del *read* R1 *forward* o R2 *reverse*, quinto el número de set (por default pongo 001).

Por lo que, si mi archivo viene así:

```
6266-SP1-357wF-806R_R1.fastq
```

```
6266-SP1-357wF-806R_R2.fastq
```

Lo debo editar para que sea así:

```
Muestra1_AAATC_L001_R1_001.fastq
```

```
Muestra1_AAATC_L001_R2_001.fastq
```

Paso_3

Para poder introducirlos correctamente deben estar comprimidos en formato **.gz** por lo que debo aplicar el comando **gzip** a mis secuencias.

```
gzip Muestra1_AAATC_L001_R1_001.fastq
```

```
gzip Muestra1_AAATC_L001_R2_001.fastq
```

Los archivos salida tienen el siguiente formato:

```
Muestra1_AAATC_L001_R1_001.fastq.gz
```

```
Muestra1_AAATC_L001_R2_001.fastq.gz
```

Un atajo, en el caso de que tengamos 20 archivos, es utilizar el comodín * (asterisco).

```
gzip *.fastq
```

De este modo, todo en cuanto termine con **.fastq** se le aplicará la instrucción.

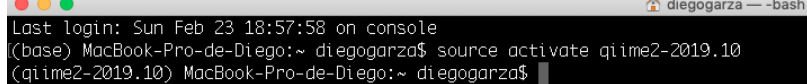
Paso_4

Lo siguiente es en nuestra Terminal activar Qiime2. Normalmente el comando es:

```
source activate qiime2-2019.10
```

Y este se debe activar en nuestra carpeta de instalación de Qiime2, donde está el archivo “qiime2-2019.10-py36-osx-conda.yml”

Un vez activo qiime lo podemos ver en la parte izquierda, como se ve en la imagen.



```
diegogarza — bash
Last login: Sun Feb 23 18:57:58 on console
(base) MacBook-Pro-de-Diego:~ diegogarza$ source activate qiime2-2019.10
(qiime2-2019.10) MacBook-Pro-de-Diego:~ diegogarza$
```

Paso_5

Sera movernos a la Carpeta de Trabajo y a la subcarpeta de Qiime2, para ello usamos el comando:

*Recuerden cambiar la dirección acorde a su equipo, en mi caso es /data/home/diego.garza pero varía de equipo a equipo

```
cd /data/home/diego.garza/Desktop/Carpeta_de_Trabajo/Qiime2
```

Paso_6

Aquí viene lo bueno, usaremos nuestro primer comando de Qiime2 para importar:

qiime2 tools import : representa la instrucción de importacion

--type el tipo de formato que estamos agregando

--input-path el directorio donde se encuentran nuestras secuencias modificadas como se explico anteriormente, importante, no debe existir otro archivo más que los **.fastq.gz** ya que de lo contrario nos marcará error

(Unrecognized file (/Users/ /Desktop/.. for CasavaOneEightSingleLanePerSampleDirFmt).

--input-format el formato que estamos utilizando, en este caso el CASAVA1.8

--output-path es el archivo de salida, para nuestro caso dejaremos el nombre estándar para identificar adecuadamente nuestro archivo. Si quieren hacerlo identificativo pueden agregar abrevaciones antes del nombre, por ejemplo, Yucatán_Fecal_demux-paired-end.qza

Podemos modificar los parámetros, según la información relevante, si quieren usar otro directorio, por ejemplo.

```
qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]' --input-path
/data/home/diego.garza/Desktop/Carpeta_de_Trabajo/Raw_sequences/ --input-format CasavaOneEightSingleLanePerSampleDirFmt
--output-path /data/home/diego.garza/Desktop/Carpeta_de_Trabajo/Qiime2/demux-paired-end.qza
```

Este es el comando en una sola línea, sin embargo, agregando el carácter \ es posible dar órdenes en forma de lista de la siguiente manera:

```
qiime tools import \
--type 'SampleData[PairedEndSequencesWithQuality]' \
--input-path /data/home/diego.garza/Desktop/Carpeta_de_Trabajo/Raw_sequences/ \
--input-format CasavaOneEightSingleLanePerSampleDirFmt \
--output-path /data/home/diego.garza/Desktop/Carpeta_de_Trabajo/Qiime2/demux-paired-end.qza
```

De este modo es más visual los parámetros que vamos agregando.

Otro aspecto para resumir la escritura es que, si no ponemos una carpeta de salida qiime2 guardará el archivo en nuestra carpeta actual que en este caso es **/Qiime2**

```
qiime tools import \
--type 'SampleData[PairedEndSequencesWithQuality]' \
--input-path /data/home/diego.garza/Desktop/Carpeta_de_Trabajo/Raw_sequences/ \
--input-format CasavaOneEightSingleLanePerSampleDirFmt \
--output-path demux-paired-end.qza
```

Una vez copiado y pegado el comando, esperamos a que Qiime2 importe, puede tardar algunos segundos dependiendo la cantidad de muestras.

Si aparece verde como en la imagen, ¡Listo! Ya tenemos nuestras secuencias importadas y demultiplexadas en qiime2

```
(qiime2-2019.10) MacBook-Pro-de-Diego:~ diegogarza$ qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]' --input-path /Users/diegogarza/Desktop/SECUENCIAS/Secuencias+Barcode/ITS+Barcode/ --input-format CasavaOneEightSingleLanePerSampleDirFmt --output-path demux-paired-end.qza
Imported /Users/diegogarza/Desktop/SECUENCIAS/Secuencias+Barcode/ITS+Barcode/ as CasavaOneEightSingleLanePerSampleDirFmt to demux-paired-end.qza
(qiime2-2019.10) MacBook-Pro-de-Diego:~ diegogarza$
```

Paso_7

Sin embargo, no podemos visualizar nada. Esto debido a que el archivo que se genera es tipo *data artifact* en formato **.qza** qiime2 tiene un formato para poder visualizar estos archivos *data artifact*, en un formato *data visual* **.qzv** si quieres más información sobre los conceptos, [aquí](#).

Vamos a generar el archivo mediante el comando:

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

De nuevo, debe aparecer el comando verde. Este archivo lo podemos encontrar en la carpeta, y podemos visualizarlo ya sea utilizando el comando (que funciona sin internet) o en línea en la página

<https://view.qiime2.org/>

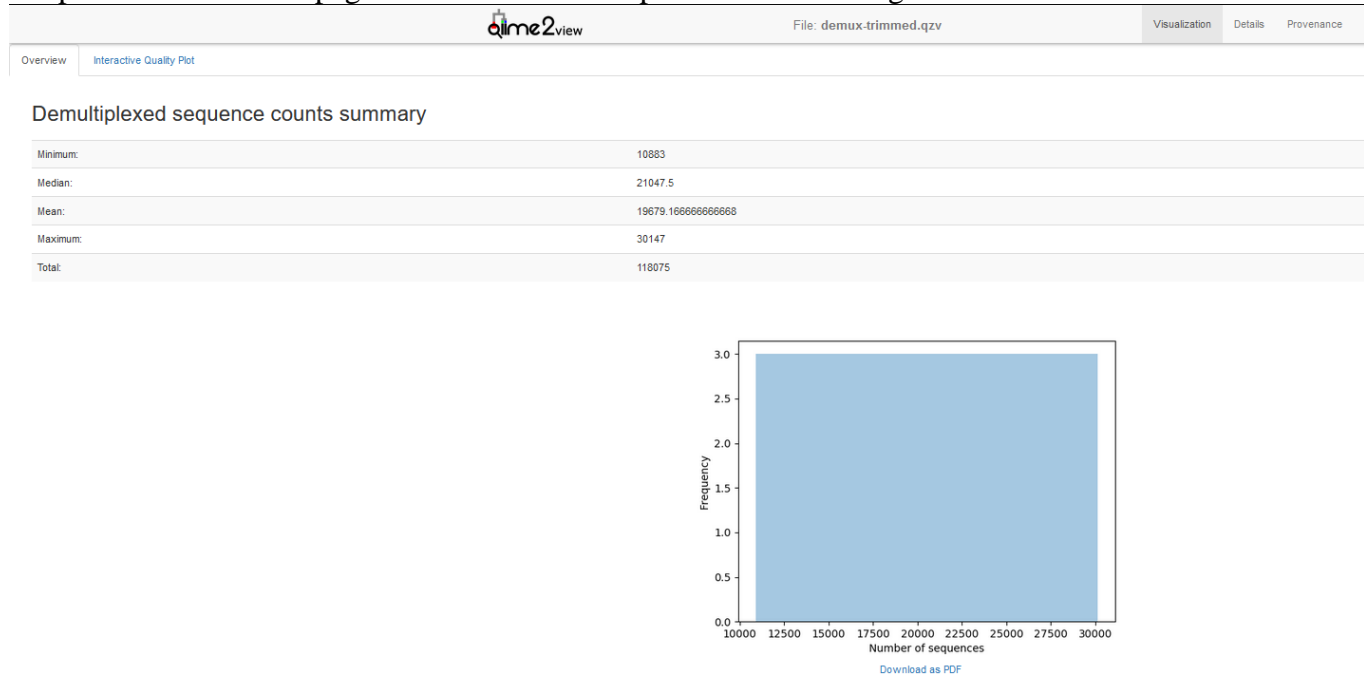
Noten como ahora es más sencillo compartir los datos con colegas simplemente enviando el archivo **.qzv** y que ellos lo visualicen en la página web sin necesidad de tener qiime2 instalado.

El comando funciona para cualquier archivo **.qzv** pero no para archivos tipo **.qza**

El comando para ver es el siguiente:

```
qiime tools view demux-paired-end.qzv
```


Después nos abrirá una página web de nuestro explorador como la siguiente:



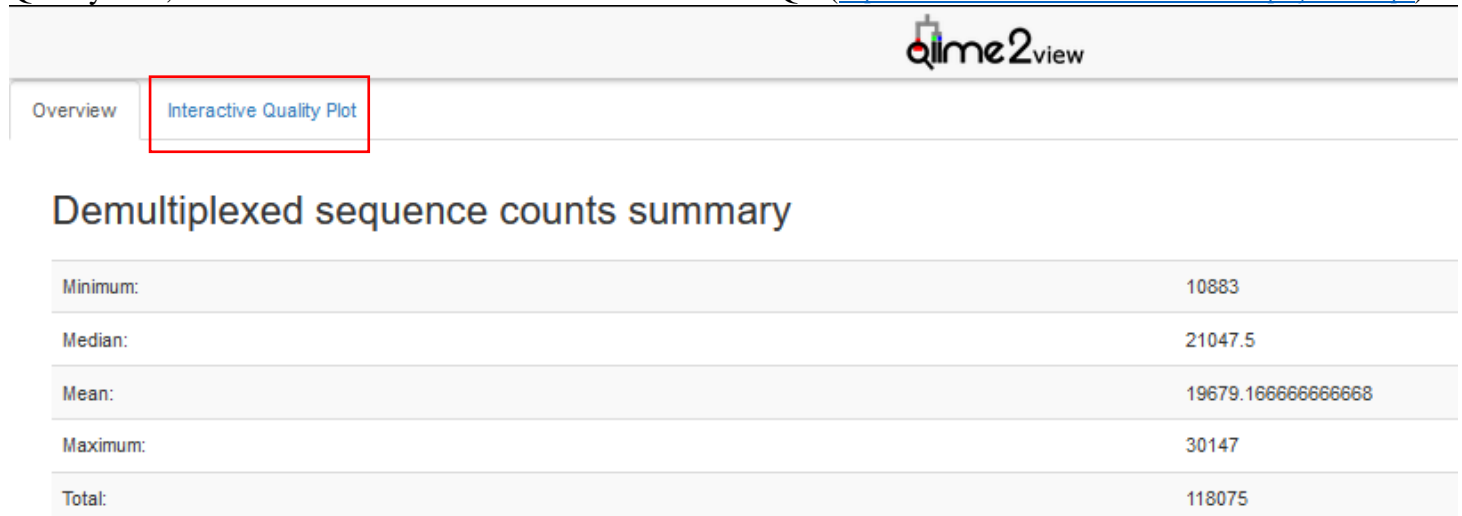
Per-sample sequence counts

Total Samples: 6

Sample name	Sequence count
T12	30147
T10	23670
SP7-8	22856
SP4	19239
SP2	11280
SP6	10883

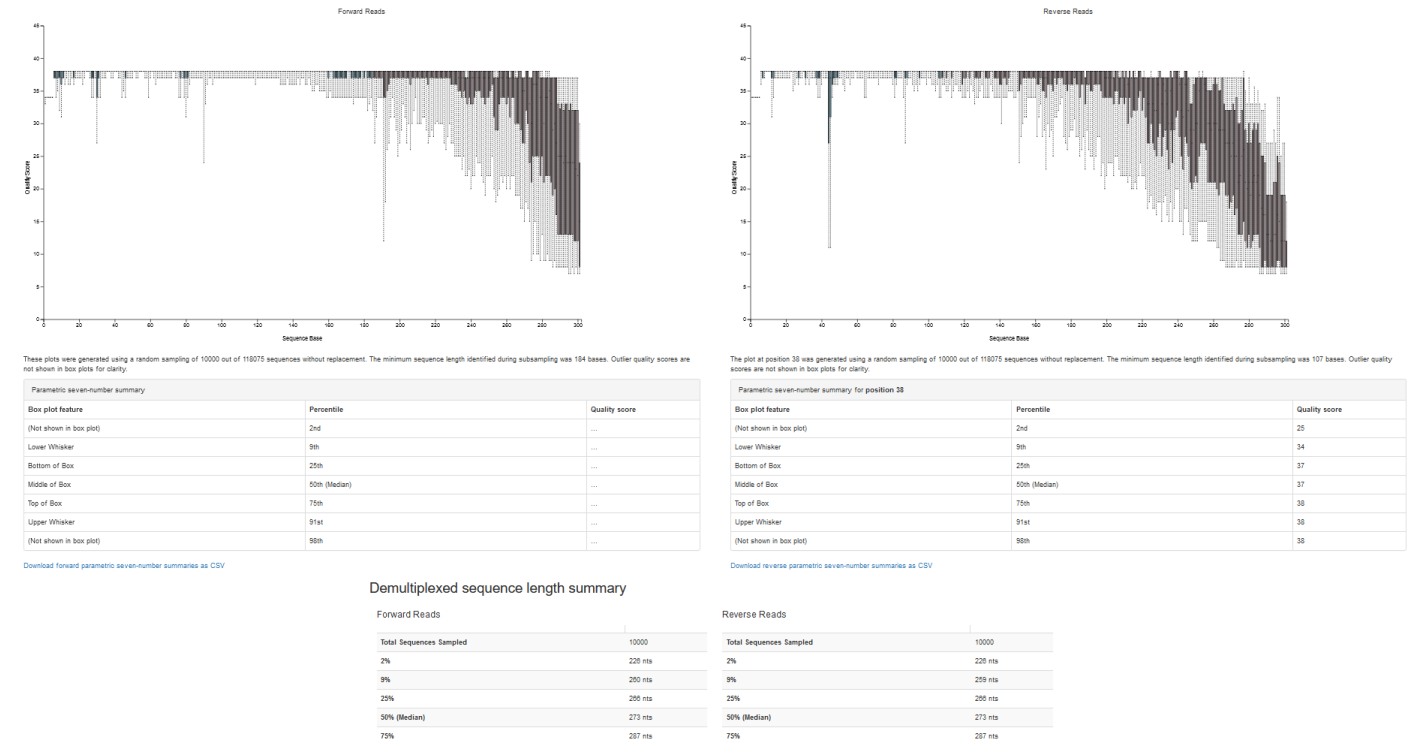
Aquí podemos ver nuestro número de *reads* totales y por muestra. Otra información relevante como el promedio, máximo y mínimo.

Algo que es de gran importancia para realizar la limpieza de nuestros datos es ver en la pestaña Interactive Quality Plot, la cual nos muestra archivos similares a FASTQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)



Al abrir esta pestaña nos aparecerá la siguiente ventana:

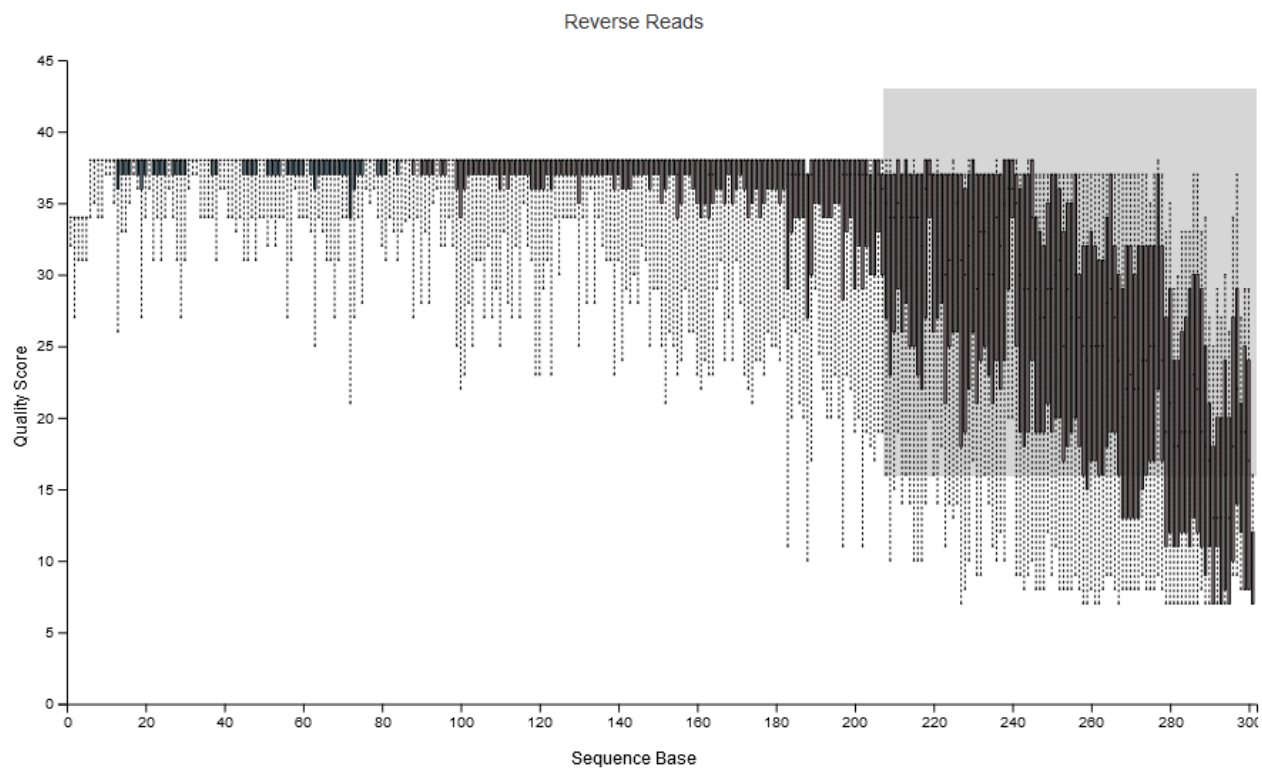
Click and drag on plot to zoom in. Double click to zoom back out to full size. Hover over a box to see the parametric seven-number summary of the quality scores at the corresponding position.



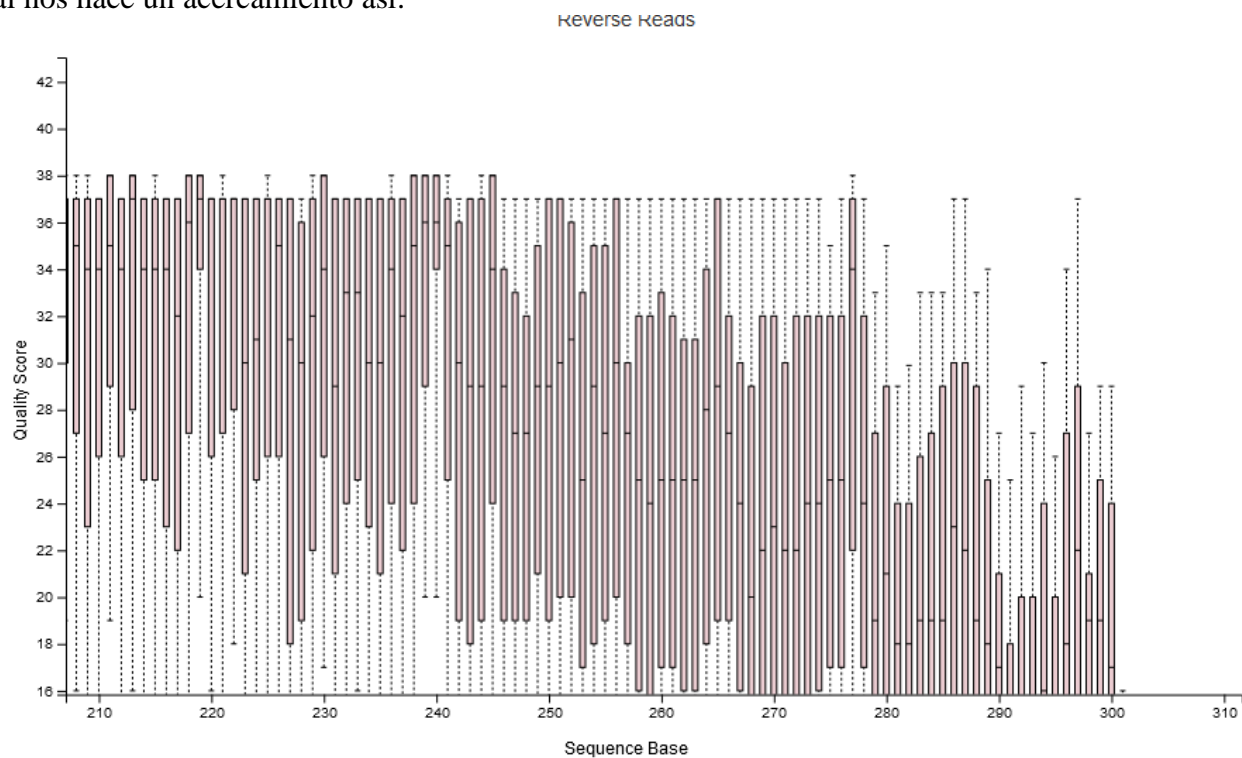
Ya que la intención de este tutorial no es explicarte sobre *cómo se debe* limpiar las secuencias, no me detendré en explicar más allá de lo necesario.

En general la gente utiliza las gráficas para visualizar la caída de la calidad, aunque esto es un tanto cualitativo y “a ojo”. Si tu forma de trabajar es interpretando solo la gráfica y tu investigador jefe tiene renuencia al respecto, continua a ese modo. Pero considera que de nuevo el criterio ¡es cualitativo!

Una herramienta útil para ver más a profundidad la gráfica es el *zoom* dando clic izquierdo y arrastrando el cuadro en la sección que es de nuestro interés, esto es como si seleccionáramos múltiples archivos dentro de la gráfica, aparecerá un recuadro para que seleccionemos el área de interés.



Lo cual nos hace un acercamiento así:



Para salir del *zoom* hay que dar dos clics izquierdos dentro del gráfico.

Qiime 2 nos ofrece un criterio más cuantitativo. Esto es usando la tabla interactiva debajo de la gráfica. Al mover el cursos podemos ver que los parámetros cambian para la posición en la que seleccionamos.

Parametric seven-number summary for position 251		
Box plot feature	Percentile	Quality score
(Not shown in box plot)	2nd	7
Lower Whisker	9th	11
Bottom of Box	25th	20
Middle of Box	50th (Median)	30
Top of Box	75th	37
Upper Whisker	91st	37
(Not shown in box plot)	98th	37

[Download reverse parametric seven-number summaries as CSV](#)

Aquí podemos ver que, en la posición 251 el 50% o la media, de las secuencias tienen una calidad *phred* de 30, por lo que son admisibles para mí gusto. Así podemos avanzar el cursor hasta ver en que posición cae la media por debajo del *score* que nosotros consideremos, a mi gusto es 20

Parametric seven-number summary for position 281		
Box plot feature	Percentile	Quality score
(Not shown in box plot)	2nd	7
Lower Whisker	9th	7
Bottom of Box	25th	11
Middle of Box	50th (Median)	18
Top of Box	75th	24
Upper Whisker	91st	29
(Not shown in box plot)	98th	37

Por ejemplo, aquí ya sabemos que la calidad cae en la posición 281, por lo que la óptima sería la posición 280.

Este sistema, aunque útil, depende de que tan cuidadosos somos con el cursor, por lo que buscar una forma más precisa de hacerlo sería descargando los parámetros de calidad en formato CVS, esto se encuentra en el enlace debajo de la tabla interactiva. Guardamos el archivo y lo podemos abrir en Excel.

These plots were generated using a random sampling of 10000 out of 118075 sequences without replacement. The minimum sequence length identified during subsampling was 184 bases. Outlier quality scores are not shown in box plots for clarity.

Parametric seven-number summary		
Box plot feature	Percentile	Quality score
(Not shown in box plot)	2nd	...
Lower Whisker	9th	...
Bottom of Box	25th	...
Middle of Box	50th (Median)	...
Top of Box	75th	...
Upper Whisker	91st	...
(Not shown in box plot)	98th	...

[Download forward parametric seven-number summaries as CSV](#)

Al abrir el archivo podemos ver una tabla que contiene lo siguiente:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	count	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
2	2%	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000
3	9%	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34
4	25%	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34
5	50%	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34
6	75%	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34
7	91%	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34
8	98%	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34	34

La fila 1 que va de 0 hasta 300 (esto es porque mi secuenciación fue 2x300 MiSeq), esto representa esa posición de nucleótido.

La fila 2, representa el muestreo de 10,000 secuencias de todas nuestras muestras. Qiime2 toma los datos de calidad a partir de un muestreo equitativo de las muestras, es para tomar en consideración que no todas las muestras han llegado a esa longitud y por tanto se debe tener preocupación a la hora de interpretar.

Las filas 3 hasta la 9 representan los percentiles y sus calidades respectivas

Una forma útil de organizar sería [inmovilizando](#) la primera columna, aplicando un formato condicional a todos los datos de calidad resaltando en rojo todo aquello menor a 20.

The screenshot shows the Excel interface with the 'Conditional Formatting' menu open. The 'Highlight Cells Rules' option is selected, and a submenu is visible with options like 'Greater Than...', 'Less Than...', 'Between...', 'Equal To...', 'Text that Contains...', 'A Date Occurring...', and 'Duplicate Values...'. The spreadsheet data is visible in the background, showing the same data as the previous table.

Como a mi me interesa saber la media, elimino las filas 3,4 y 5. Y *voilà* tenemos la posición donde comienza a caer la calidad, con más precisión. Desde luego esto depende de nuestros objetivos y que tan conservadores deseamos ser.

A	JO	JP	JQ	JR	JS	JT	JU	JV	JW	JX	JY	JZ	KA	KB
	273	274	275	276	277	278	279	280	281	282	283	284	285	286
count	9980	9980	9980	9980	9980	9980	9980	9980	9980	9980	9980	9980	9980	9980
50%	24	25	25	34	24	19	21	18	18	19	19	19	23	22
75%	32	32	32	37	32	27	29	24	24	26	27	29	30	30
91%	37	35	37	38	37	33	35	29	29.89	33	33	33	37	37
98%	37	37	37	38	37	37	37	37	37	37	37	37	37	37

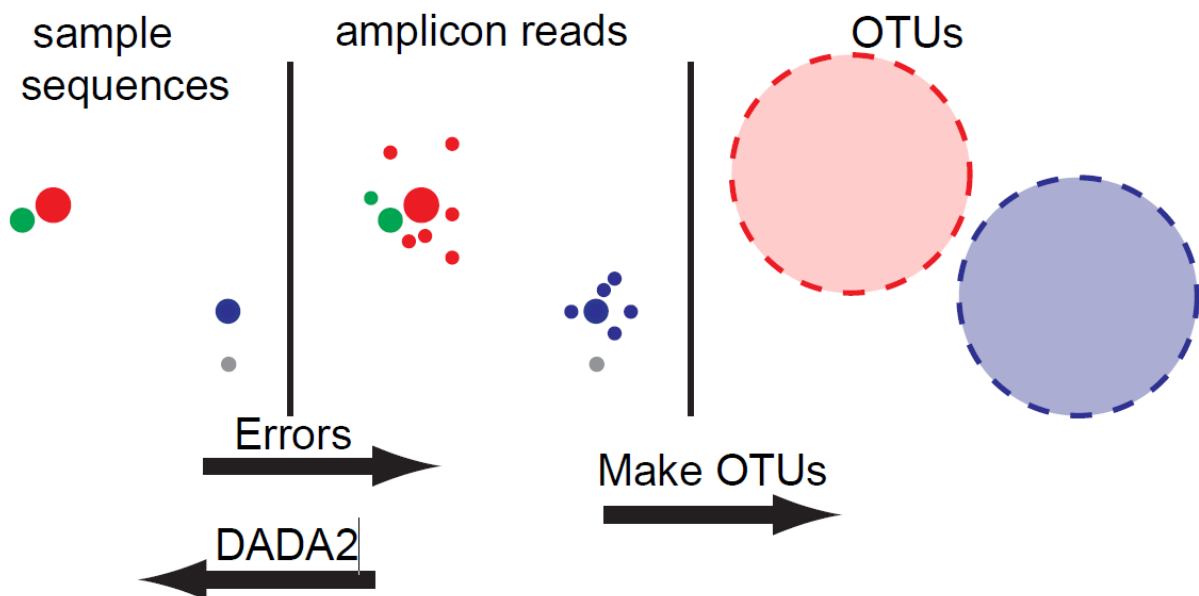
Denoising con DADA2

Aquí si me voy a detener a explicar, ya que me encontrado con dificultad de entender este proceso, más aún si no se domina el tema o se es nuevo en esta área y si existe [renuencia](#) al cambio en tu laboratorio. A partir de aquí meteremos las manos sobre nuestras secuencias, por lo que es importante comprender que *** estamos haciendo.

El *denoising* es un tema enorme sobre manejo de nuestros datos, traducido literalmente significa “quitar ruido” lo cual es una definición bastante apropiada de lo que es el proceso. El proceso consiste básicamente en: filtrado, dereplicado (*dereplication*), identificación y remoción de quimeras y la unión de *reads* pareados (merging pair-end reads). Así es, no hay agrupación o *clustering*

Aunque en quime2 existe el modo “tradicional” de limpieza de secuencias, estos nuevos algoritmos de control de calidad, como [DADA2](#), son la nueva generación dedicada a Illumina aprovechando al máximo la profundidad que esta plataforma nos provee, ya que los previos estaban adecuados a la secuenciación 454 y sus limitantes. De hecho, la forma de hacer OTU's (*Operational Taxonomic Units*) estaba considerada a partir de los errores típicos de secuenciación, llegándose al consenso general de usar 3% de disimilitud entre secuencias para definir “especie”, lo cual es considerado como una forma muy arbitraria. Actualmente los OTU's son motivo de [humor](#) y discusión (dar clic en [renuencia](#)) por los bioinformáticos. Si quieres informarte más adecuadamente al respecto, hay un [artículo](#) de la revista *Nature* que vale la pena revisar, yo me limitaré a explicar un poco más sencillo utilizando un gráfico del artículo de DADA2, en mi lenguaje coloquial.

Schematic of OTU and DADA2 approaches towards amplicon sequencing errors.



Tomada de Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods*, 13(7), 581.

Afortunadamente, la gráfica es auto explicativa. El color representa una secuencia libre de errores y el tamaño del círculo su abundancia. Dependiendo la resolución pueden representar cepas distintas, especies distintas, o algo que sencillamente, son variación biológica. Al movernos hacia las secuenciación, el proceso introduce errores, los cuales introducen variación no-biológica. Hacer OTU's nos permite limpiar estos errores no-biológicos, pero perdemos aquella fineza removiendo también la variación-biológica (se pierde el verde y gris).

DADA2 **intenta** quitar el ruido (*denoise*) de la secuenciación y volver al motivo original de nuestras muestras manteniendo la variación biológica.

Volviendo a DADA2 y qiime2, el comando requiere de una serie de parámetros, básicamente si queremos cortar de 3' a 5' n número de nucleótidos. Y truncado, que básicamente es *cortar los nucleótidos después de cierta posición*. A mi me costo mucho trabajo entender estos conceptos, pero una vez entendidos son muy sencillos.

Paso_1

Antes de comenzar con la línea de comando, ¿tus secuencias tienen insertos los primers? De ser así necesitas saber la longitud del primer (de cuantos nucleótidos es) para removerlos. ¿Ya sabes en que posición cae la calidad de tus secuencias? A este punto como explico hay tres formas de hacerlo, mediante la gráfica a “ojo”, mediante la tabla interactiva, o mediante el Excel.

Para mi caso si tengo los primers por lo que:

Forward primer tamaño: 22

Reverse primer: 20

Forward caída de calidad: 280

Reverse caída de la calidad: 277

Paso_2

Vamos a modificar los parámetros del comando según nuestros datos

```
qiime dada2 denoise-paired\  
--i-demultiplexed-seqs demux-paired-end.qza\  
--p-trunc-len-f 280\  
--p-trunc-len-r 277\  
--p-trim-left-f 22\  
--p-trim-left-r 20\  
--o-table PE-table.qza\  
--o-representative-sequences PE-rep-seqs.qza\  
--o-denoising-stats PE-stats.qza
```

Lo mismo, si tenemos líneas verdes el resultado fue exitoso.

Paso_3

Ahora vamos a crear nuestros archivos visuales para poder ver los resultados

```
qiime metadata tabulate\  
--m-input-file PE-stats.qza\  
--o-visualization PE-stats.qzv  
qiime feature-table summarize\  
--i-table PE-table.qza\  
--o-visualization PE-table.qzv \  
--m-sample-metadata-file metadata_16S.tsv  
qiime feature-table tabulate-seqs\  
--i-data PE-rep-seqs.qza\  
--o-visualization PE-rep-seqs.qzv
```


El resultado después de ejecutar DADA2 son 3 archivos.

Un archivo artefacto.qza (PE-stats.qza) que describe en números como fueron sufriendo cambios, en cada etapa de tratamiento las secuencias. Así como la proporción relativa a la entrada inicial de secuencias. Esta tabla es importante interpretarla, ya que, si perdemos todas las secuencias en un proceso, o demasiadas, es en esa etapa donde debemos buscar resolución.

Search:

sample-id	input	filtered	percentage of input passed filter	denoised	non-chimeric	percentage of input non-chimeric
#q2-type#	numeric	numeric	numeric	numeric	numeric	numeric
SP2	11280	10716	95	10665	10665	94.55
SP4	19239	18143	94.3	18065	18006	93.59
SP6	10883	8237	75.69	7945	7942	72.98
SP7-8	22856	21679	94.85	21319	21312	93.24
T10	23670	20975	88.61	20521	19718	83.3
T12	30147	26729	88.66	26156	25789	85.54

Showing 1 to 6 of 6 entries

Previous 1 Next

Un artefacto `FeatureData[Sequence]` (PE-rep-seqs.qza) que son nuestras secuencias representativas, es de tipo , o lo que se podría considerar “OTU’s al 100%”, es decir aquellas secuencias únicas. Estas secuencias son los “Sequence Count” para mi caso son 138, también nos brinda información del tamaño que tienen, min, max, promedio, etc. Importante es que, podemos ver directamente la secuencia resaltada en azul. Esto es debido a que esta enlazada con BLAST, si damos clic en alguna nos abrirá una página de BLAST con la secuencia en cuestión. Otro dato importante es que se le asigna un id “Feature id” por lo que cada una de estas secuencias representativas ya esta “mapeada”. También podemos descargar estas secuencias en formato FASTA, sin embargo, no contienen datos de abundancia, por lo que sólo son útiles para asignación taxonómica, no para índices de diversidad.

QIIME2view

File: PE-rep-seqs.qzv

Visualization

Details

Provenance

Sequence Length Statistics

Download sequence-length statistics as a TSV

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
138	280	435	401.74	155	30.83

Seven-Number Summary of Sequence Lengths

Download seven-number summary as a TSV

Percentile:	2%	9%	25%	50%	75%	91%	98%
Length* (nts):	280	398	401	401	424	424	427

*Values rounded down to nearest whole number.

Sequence Table

To BLAST a sequence against the NCBI nt database, click the sequence and then click the View report button on the resulting page.

Download your sequences as a raw FASTA file

Click on a Column header to sort the table.

Feature ID	Sequence Length	Sequence
05e97bfc76fcafe8c55897f969ca84be	401	AATTTTCGCGCAATGGCGAAAGCCTGACGGAGCAATGCCGCGTGGAGGTAGAAGGCCACGGGTGCTGAACCTCTTTTCCGCGAGAAGAAACAATGACGCTATCTGGGGAATAAGCATCGGCTAACTCTGTGCCAGCAGCGCGGTAAATACAG
bf384a0d7a5dec8e6933a16bff71d996	401	AATTTTCGCGCAATGGCGAAAGCCTTACGGAGCAATGCCGCGTGGAGGTAGAAGGCCACGGGTGCTGAACCTCTTTTCCGCGAGAAGAAACAATGACGCTATCTGGGGAATAAGCATCGGCTAACTCTGTGCCAGCAGCGCGGTAAATACAG

Lo siguiente es un archivo tipo `FeatureTable[Frequency]` (PE-table.qza) la cual tiene la relación de todos los Feature id y su abundancia, esto está en la pestaña Feature Detail. Este archivo, es el que nos servirá más adelante para los índices de diversidad. Otro aspecto importante es el Interactive Sample Detail ya que nos indica los *reads* por muestra. Así podemos saber cual es la abundancia mínima de nuestro estudio.

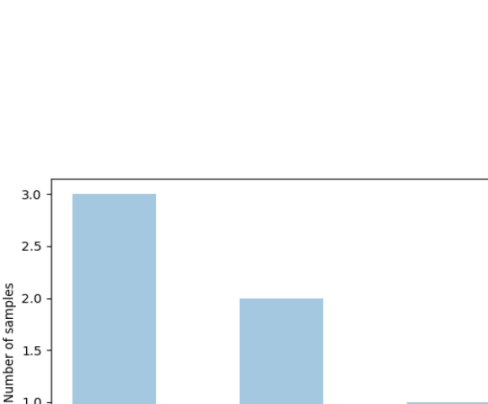
Table summary

Metric	Sample
Number of samples	6
Number of features	138
Total frequency	204,984

Frequency per sample

	Frequency
Minimum frequency	22,762.0
1st quartile	27,249.0
Median frequency	32,587.5
3rd quartile	38,037.0
Maximum frequency	51,724.0
Mean frequency	34,164.0

Frequency per sample detail ([csv](#) | [html](#))



Estos archivos pesan menos de 1 megabyte, y son capaces de darnos toda la información de nuestras secuencias. ¡A esto se le llama portabilidad! En lugar de mandar un archivo de 2-5gb a un colega, se le pueden enviar estos sin mayor problema.

Asignación taxonómica de secuencias ITS y 16S

Aquí también es importante mencionar un par de cosas. La base de datos que elijas va a determinar en gran medida que tantos éxitos tengas en asignar un “nombre” a un taxón. Hay varias bases populares, sin embargo parece que [SILVA](#) es la más actualizada para el caso de bacterias (16s) y para el caso de hongos la base de datos [UNITE](#) (ITS), estas son independientes de Qiime2, así que se deben citar por separado.

Para asignar taxonómicamente necesitas un clasificador (`classifier.qza`) la manera más sencilla de obtener uno, es que lo descargues de algún sitio. Qiime2 tiene una página de [recursos](#) donde se encuentran algunos, aunque, no están siempre actualizados. Para obtener la versión más nueva puedes entrenar tu clasificador de acuerdo a tus primers y secuencias objetivo siguiendo el [tutorial](#) dedicado a ello en qiime2.

Para el caso de SILVA [versión 138](#), 16 diciembre 2019, lo pueden descargar del foro de qiime2, [aquí](#) Para el caso de UNITE [versión 8.2](#) dynamic, 20 febrero 2020, lo pueden descargar de aquí, pero no lo consideren un link permanente ya que es de mi almacenamiento personal y en cualquier momento puede cambiar.

Paso_1

Simplemente es seguir este script, reemplazando el `--i-classifier` por el directorio+nombre de nuestro clasificador y nuestras `rep-seqs.qza` de del paso del *denoising*. Esto aplica tanto para ITS como par 16s. Este proceso demora bastante, dependiendo de las características de tu computadora, si no tienes un equipo mayor a 8G de [RAM](#) no te recomendaría proceder este paso en el equipo actual. Para mejorar la eficiencia cierra todos los programas activos y deja el equipo trabajar.

```
qiime feature-classifier classify-sklearn\  
--i-classifier /data/home/diego.garza/Beebread/SECUENCIAS/taxonomy_reference/silva-132-99-515-806-nb-  
classifier.qza\  
--i-reads PE-rep-seqs.qza\  
--o-classification taxonomy_SILVA_138_PE.qza
```

Por default el intervalo de confianza para el clasificador es 0.70, es decir 70% si deseas cambiarlo, agrega la opción `--p-confidence` puedes modificar la línea de comando siguiente según el intervalo que quieras.

```
qiime feature-classifier classify-sklearn\  
--i-classifier /data/home/diego.garza/Beebread/SECUENCIAS/taxonomy_reference/silva-132-99-515-806-nb-  
classifier.qza\  
--i-reads PE-rep-seqs.qza\  
--p-confidence 0.7 \  
--o-classification taxonomy_SILVA_138_PE.qza
```

El resultado es un artefacto `FeatureData[Taxonomy]` adivinaron, en formato **.qza**

Paso_2

Vamos a hacer los visuales de nuestra clasificación, ¿ya van viendo un patrón entre un proceso, un data y un visual? Básicamente es eso, cada proceso te pide un input, sale un **.qza** y para verlo lo haces **.qzv**

Para este caso vamos a tabularlo con:

```
qiime metadata tabulate \  
--m-input-file taxonomy_SILVA_138_PE.qza \  
--o-visualization taxonomy_SILVA_138_PE.qzv
```

El resultado es algo como lo siguiente:

qiime2view

File: taxonomy_SILVA_138_PE.qzv

VisualizationDetailsProvenance

Download metadata TSV file

This file won't necessarily reflect dynamic sorting or filtering options based on the interactive table below.

Feature ID	Taxon	Confidence
008ab874b013135e99fbfa1d9a251d3f	d__Bacteria; p__Cyanobacteria; c__Cyanobacteria; o__Chloroplast; f__Chloroplast; g__Chloroplast	0.9999999968401893
00d85da877742cc47a14ae0b8062dd2f	d__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus	0.9646314693228784
0262437a711c496a1aa63b5f0e705079	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Burkholderiales; f__Neisseriaceae; g__Snodgrassella	0.9994097494198872
05400db6177781f32f04c3ba70fe2abb	d__Bacteria; p__Cyanobacteria; c__Cyanobacteria; o__Chloroplast; f__Chloroplast; g__Chloroplast	0.999999992096348
05743ba087bac927e9625cb25a633a0e	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__Mitochondria; g__Mitochondria	0.999999999971527

En esta tabla encontramos el feature ID el cual podemos buscar por ejemplo en nuestra FeatureData[Sequence] (rep-seqs.qza) para hacer BLAST y corroborar la asignación o buscar en nuestra FeatureTable[Frequency] (table.qza) para ver su abundancia.

Taxa collapse y filtrado de cloroplastos y mitocondrias de secuencias 16S

Ahora que sabemos que bichos tenemos en nuestra muestra, podemos hacer un par de cosas interesantes. Por ejemplo, filtrar nuestros datos si queremos eliminar la presencia de un taxon. También podemos hacer mapas de calor o abundancias relativas

Comenzaremos por filtrar nuestros datos. Vamos a filtrar nuestros datos de abundancia

`FeatureTable[Frequency]` (`table.qza`). Para ello Utilizamos el comando `qiime taxa filter-table` donde las opciones importantes son `--p-exclude` y `--p-include`, para `exclude` o `excluir` pones los nombres de los taxones como nos aparece en nuestra tabla `FeatureData[Taxonomy]` que queremos fuera, por ejemplo cloroplastos y mitocondrias: `mitochondria,chloroplast`. Como se ve cada taxon se separa por un carácter de coma ,

La opción `--p-include` nos permite seleccionar solo aquellos que entren en el criterio expuesto. Para este ejemplo no lo usaremos, pero si estuvieras interesado, por ejemplo, solo en lactobacilos puedes utilizar esta opción.

```
qiime taxa filter-table\  
--i-table PE-table.qza\  
--i-taxonomy taxonomy_SILVA_138_PE.qza\  
--p-exclude mitochondria,chloroplast\  
--o-filtered-table table-no-mitochondria-no-chloroplast_PE.qza
```

Ahora para nuestras secuencias representativas `FeatureData[Sequence]` (`rep-seqs.qza`) hacemos lo mismo, esto tiene su utilidad porque nos provee del FASTA el cual podemos usar para corroborar nuestra taxonomía.

```
qiime taxa filter-seqs\  
--i-sequences PE-rep-seqs.qza\  
--i-taxonomy taxonomy_SILVA_138_PE.qza\  
--p-exclude mitochondria,chloroplast\  
--o-filtered-sequences sequences-no-mitochondria-no-chloroplast_PE.qza
```

El siguiente punto es crear una nueva una `FeatureTable[Frequency]` (`table.qza`) pero anotada según taxonomía, esto es especialmente útil para sustituir cuando nos pidan la (`table.qza`), tener en lugar de los feature ID la asignación que acabamos de clasificar. Aquí es importante seleccionar el parámetro `--p-level` ya que indica el nivel taxonómico al cual vamos a crear esta nueva tabla. Yo lo haré a nivel 7 que es especie. Vamos además a usar nuestra tabla recién filtrada.

```
qiime taxa collapse \  
--i-table table-no-mitochondria-no-chloroplast_PE.qza \  
--i-taxonomy taxonomy_SILVA_138_PE.qza \  
--p-level 7 \  
--o-collapsed-table table-no-mitochondria-no-chloroplast_PE_lvl_spp.qza
```

Barplots, gráficos de abundancia relativa

Finalmente, vamos a hacer gráficos de abundancias relativas. Ciertamente hay una enorme discusión sobre su precisión, ya que, son **relativos**. Es decir, se hacen según el número de copias que tienes de 16S, esto no es exactamente una medida directa de abundancia, ya que algunos grupos tiene varias copias del 16S, además de que son porcentuales así que se afectan en una magnitud distinta a la real. Para mas información puedes leer este [artículo](#). La manera más precisa sería completar nuestro estudio con un PCR en tiempo real, por ejemplo, como en este [artículo](#), sin embargo no siempre está al alcance (\$\$\$).

De cualquier forma, es una herramienta bastante común y nos puede mostrar algunos patrones

```
qiime taxa barplot\  
--i-table table-no-mitochondria-no-chloroplast_PE.qza\  
--i-taxonomy taxonomy_SILVA_138_PE_clean.qza\  
--m-metadata-file metadata_16S.tsv\  
--o-visualization taxa-bar-plots_SILVA_138_PE_clean.qzv
```