Ensamble y análisis de SARS-CoV-2

Marcos E. R. Ramírez

Resumen pa

Aquí tendre que poner mi resumen, una vez que yo haya terminado el protocolo, sino seria todo un dolor de cabeza sin saber que voy a hacer

Mas vale que sobre y no que falte jajaja

Introducción

El Coronavirus del Síndrome Respiratorio Agudo Severo de tipo 2 [1] (SARS-CoV-2) es un virus valga la redundancia de la familia de los coronavirus que causan enfermedades respiratorias, tanto comunes como graves como menciona [2], esta enfermedad es conocida como COVID-19 [2] y para las personas de mayor edad y para las que están bajo condiciones médicas como enfermedades cardiovasculares, diabetes, enfermedades crónicas respiratorias o cáncer tienen más probabilidades de desarrollar una enfermedad grave [3].

El surgimiento de SARS-CoV-2 fue observado en casos de neumonía inexplicable en la ciudad de Wuhan China [4]. Se realizaron estudios hasta saber de dónde podría haber brotado el virus, hasta que en murciélagos se identificaron como los hospedadores de importantes virus zoonóticos como el virus de Nipah, el virus Hendra y SARS-CoV, incluyendo coronavirus con una considerable diversidad genética [4].

El surgimiento de la pandemia por SARS-CoV-2 es debido a la gran capacidad de virulencia. Esto es explicado por el descubrimiento de múltiples detalles estructurales y no estructurales semejantes al sitio único de escisión de FURIN, (SCoV2-PLpro), ORF3b y proteínas no estructurales, y así como cambios conformacionales en la estructura de la proteína spike durante el hospedaje de la célula [5]

Metodología y resultados

El propósito de este artículo es emplear técnicas bioinformáticas para el análisis de un organismo o virus en este caso SARS-CoV-2, el punto es poner en practica lo aprendido en clases y determinar como es estructural y funcionalmente.

Para ese punto primero será necesario entender desde la secuenciación cual método de secuenciación que se usó y sus características, con el propósito de saber que plataformas o recursos como softwares bioinformáticos se usarán y que indicaciones dar.

Una vez identificado cuales son las características de secuenciación se procede a hacer análisis de calidad de la secuenciación para identificar valga la redundancia, la calidad y decidir si son muestras aptas para su análisis o si sirven para el trabajo de investigación, ya que se ha dado la luz verde se filtran la secuencias y/o bases que tengan una muy baja calidad, con ello para minimizar los errores en el análisis.

Ya filtradas las secuencias será necesario hacer mapeos para determinar a que organismo se parece mas en su secuencia de nucleótidos, con ello en mente se procede a hacer el ensamble y determinar la estructura y función de dichos genes que provienen de su muestra.

Dichas herramientas usadas en este articulo serán mencionadas y descritas sus funciones.

Solo con fines prácticos cada una de estas herramientas se

presentará la página para que puedan revisar sus manuales de instalación y uso, pero no se explicara como se instalaran (más que solo se explicara porque se usaron dichas opciones), además cabe destacar que la computadora que se iba a usar para este proyecto no funciono ya que no se instalaron correctamente algunos programas, por lo que se optó por hacerlo con una vía remota.

Para empezar, se descargar la secuenciación del **SARS-CoV-2**, entonces se ingresó a la base de datos de **NCBI** en la base de datos de **SRA** con el **ID** de **SRP251618**. Por tanto, la primera muestra con el Accession: SRX10248073, es seleccionada para ver sus características. Se anexa apoyo visual en la (fig. 1 A).

Podemos ver las siguientes características. En la parte de show Abstract tiene una pequeña descripción, la cuales citada: **Isolate raw read files from the Washington SARS-CoV-2 outbreak**, así como en **Strategy** pues secuenciar el RNA, es con el propósito de saber los patrones de expresión [6,7]. Por otra parte, es de tener en cuenta que los reads son single, lo cual ayudara para pasos posteriores.

Descarga

La primera herramienta a disposición es

SRA Toolkit

Este es un software que incluye diversas herramientas y librerías para usar información en el "INSDC Sequence Read Archives (SRA)" [8]

Entonces, ya que dentro de la computadora que se operó no se pudo descargar el SRA, entonces se trabajó mediante una computadora remota, aplicando los siguientes comandos de SRA Toolkit (**comando 1**) y el run (fig. 1 B). Y saldrán los archivos como en la segunda parte del comando.

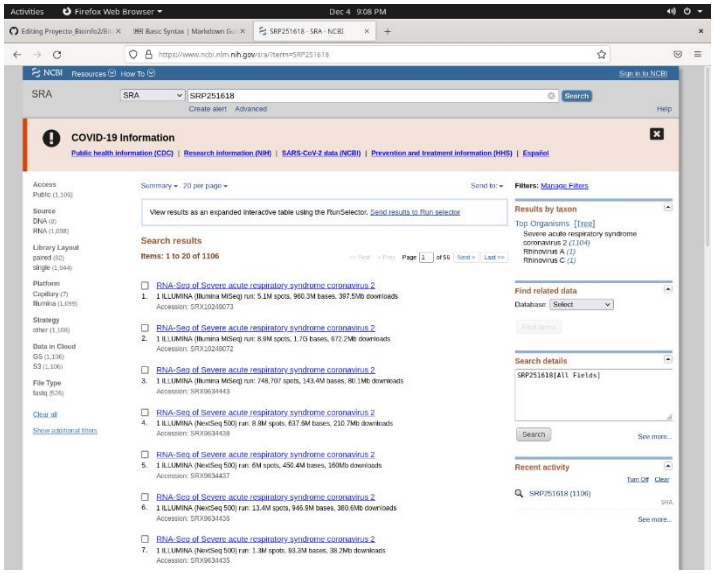
Control de calidad

FastQC

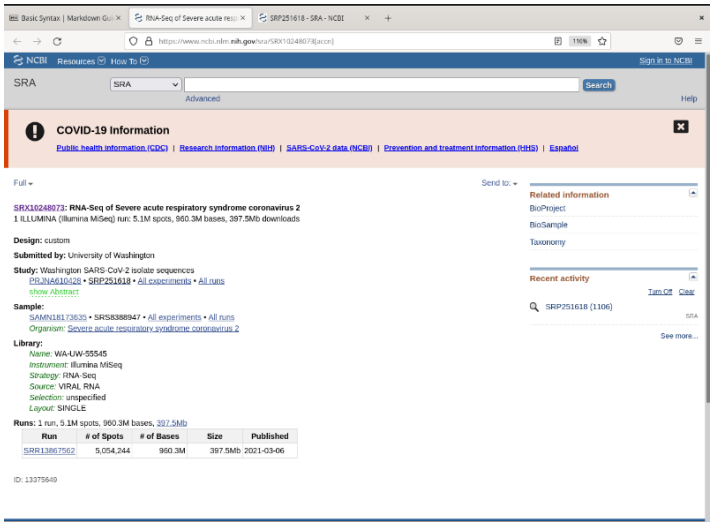
Este software [9] tiene la función de hacer un entorno gráfico la información del control de calidad de secuencias de alto rendimiento

Fig. 1: Ilustración de cómo se ven los pasos en la página

A.



B.



comando fastq-dump

```
$ fastq-dump SRR13867562
Read 5054244 spots for SRR13867562
Written 5054244 spots for SRR13867562

$ ls -lh
total 2.2G
-rw-rw-r-- 1 mramirez mramirez 2.2G Dec  5 01:32 SRR13867562.fastq
drwxrwxr-x 3 mramirez mramirez   28 Dec  5 01:23 ncbi
```

Para ello, con el siguiente comando (comando 2) y el **SRR13867562.fastq**, saldrán los dos archivos nuevos **SRR13867562_fastqc.html** y **SRR13867562_fastqc.zip** los cuales tendrán que abrirse el html para realizar el análisis de calidad.

En la (fig. 2) las características de la secuenciación se detallan como

Filename: Nombre del archivo fastq

File type: El tipo de archivo (el cual está bien ya que señala que tiene nombrada de forma correcta las bases)

Encoding: Indica que los valores de calidad ASCII están basadas en las tecnologías señaladas

Total sequences: La cantidad de secuencias procesadas es de poco más de 5 millones

Sequences flagged as por quality: El número de secuencias de baja calidad removidas es de 0

Sequences length: La longitud promedio de las secuencias es de 190 bases, ya que todas son de esa longitud

%GC: Y el contenido de GC consta de un porcentaje de casi la mitad (los detalles en la gráfica correspondiente)

Es observable que la calidad de las secuencias es buena, cerca del 100% de las bases representan calidades por encima de 28 de calidad ASCII (sanger / Ilumina 1.9) a excepción al final la base numero 190 llega a tener calidades pobres de hasta 11, pero aun en promedio sigue siendo de buena calidad ya que las bases con muy baja calidad (poco menos del 25% de ellas) tienen una pequeña influencia. (fig.3)

En la fig. 4 es más visible, ya que la distribución de las bases ronda en calidades mayores a 36, esto quiere decir que la tasa de error es menor a 0.2%

En la fig. 5 el resultado no es el más optimo al inicio y al final, ósea que hay diferencias mayores al 10% entre G y C en la posición 5, 150-154, 170-174 y 190. La misma situación entre A y T la diferencia es >10% en las posiciones 5-6, a mitad entre 150-154. Aunque esto será removido posteriormente

comando fastQC

```
$ fastqc SRR13867562.fastq
```

Fig. 2: Estadísticos basicos

Measure	Value
Filename	SRR13867562.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	5054244
Sequences flagged as poor quality	0
Sequence length	190
%GC	41

Fig. 3: calidad de bases por secuencia

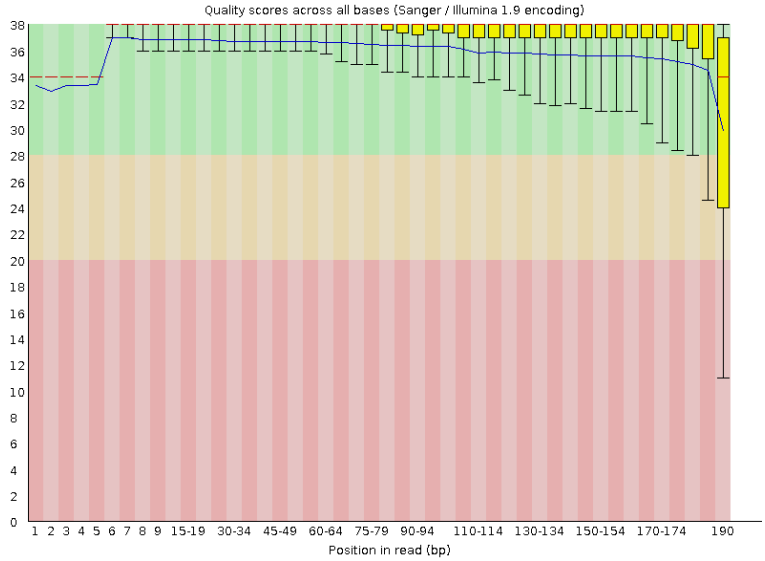


Fig. 4: Puntaje de calidad por secuencia

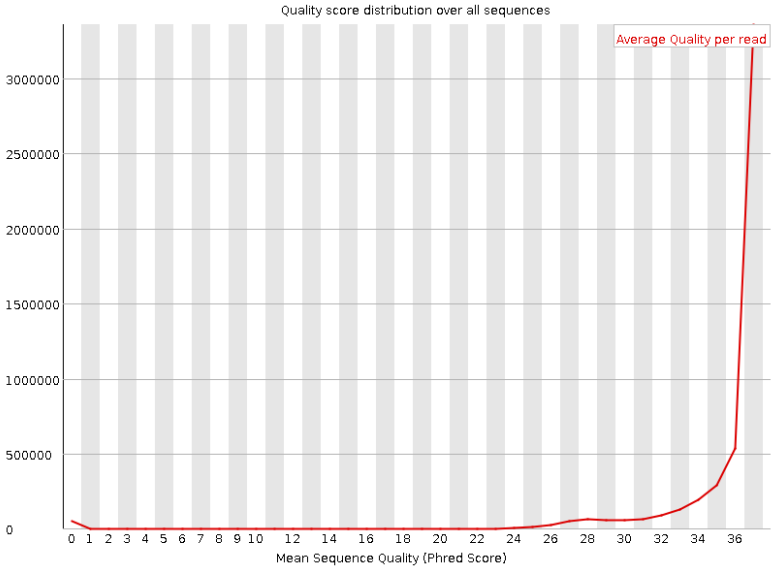


Fig. 5: Contenido de bases por secuencia

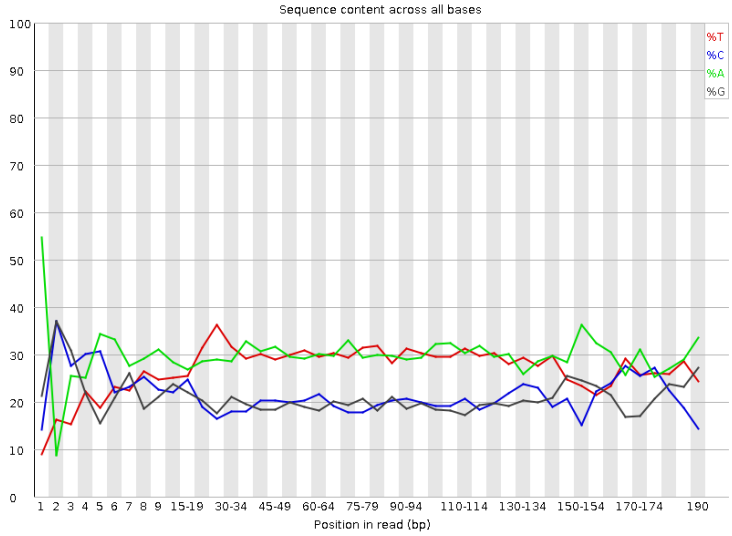


Fig. 6: Contenido GC por secuencia

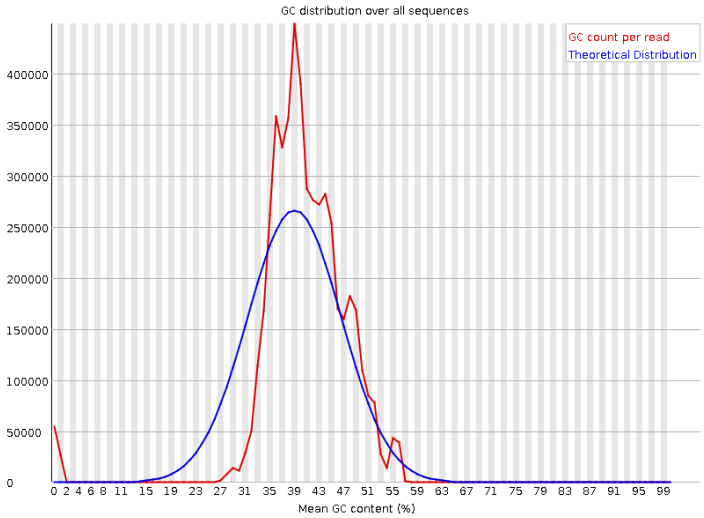
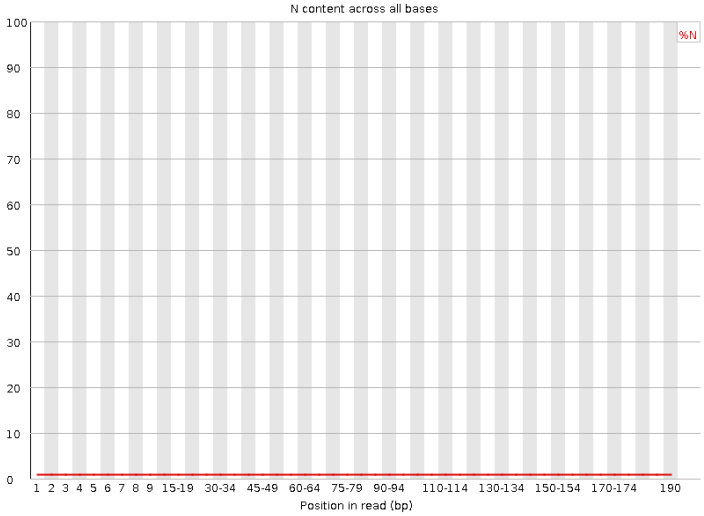


Fig. 7: Contenido de bases N



Es observable que la distribución en el contenido de GC está fuera del teórico, resultando en que mayoria (450,000 secuencias) tiene un porcentaje de GC entre 35-40% por tanto indica que no hay una buena distribución y hay muchas secuencias con tendencia a tener >35% (fig. 6)

Fig. 7 Indica que todas las bases fueron nombradas y no hubo ninguna que no se reconociera.

Como se indicó al inicio, no señalo el mínimo y el máximo en longitud de secuencias, la mayoría o me atrevería a decir que todas tienen longitud de 190 bases (fig. 8)

Los niveles de duplicación son muy altos, de hasta el 60% lo que podria significar que hay un bajo nivel de cobertura de la secuencia objetivo, y por tanto al eliminar los duplicados solo haria que nos quedemos con el 1.58% de secuencias (fig. 9)

Tener sobreexpresión de secuencias tampoco es bueno porque también indica que hay contaminación de organismos (tan solo 3 probablemente) y por otro que no es tan diverso o que también sean bastante signantes biológicamente. (fig. 10)

Finalmente, la cantidad de adaptadores universales Illumina es alta, de hasta el 40% en las posiciones 165-169 (fig. 11)

Filtrado de secuencias

Trimmomatic

Es un software que tiene útiles funciones de filtrado para secuencias Illumina paired-end y single-end. [10].

De acuerdo a las funciones que tiene trimmomatic, se traducen un poco para entender el siguiente comando:

Remover adaptadores (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10)

Remover las principales bases de baja calidad o bases N (debajo de calidad 3)

(LEADING:3)

Remover las bases finales de baja calidad o bases N (debajo de calidad 3) (TRAILING:3)

Fig. 8: Distribución de la longitud de la secuencia

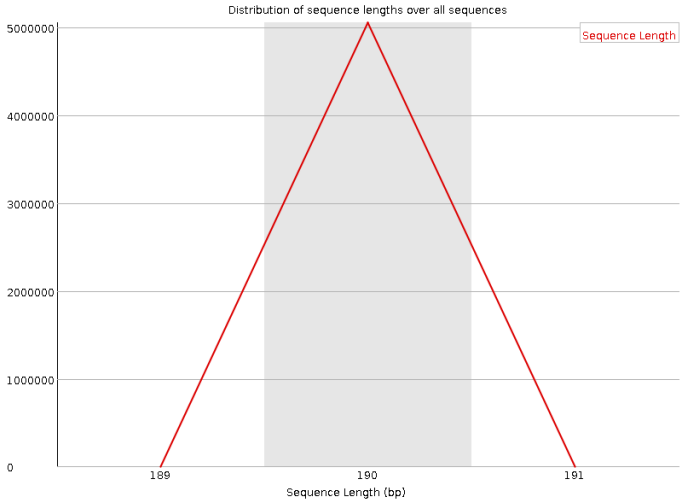
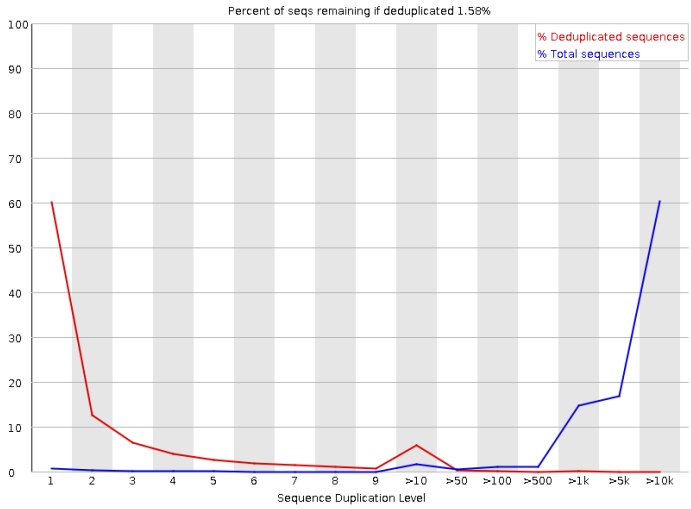


Fig. 9: Niveles de duplicación de las secuencias



Escanea la lectura (read) a razon de 4 en 4 bases y cortando cuando el promedio de calidad cae por debajo de 15 (SLIDINGWINDOW:4:15)

Quita lecturas que esten por debajo de 36 bases de longitud

(MINLEN:36)

Entonces el siguiente comando de acuerdo a la sintaxis del programa es para realizar el filtrado (**comando 3**)

Por lo que cada parte de la estructura indica

1.- La ruta que esta después de -jar es para llamar a trimmomatic, al cual se tuvo que buscar para llamarlo.

2.- Por otra parte, el SE es parte del protocolo de trimmomatic para hacer el recorte a fastq: con reads single end.

3.- En tanto -threads es el número de núcleos que usaran (se usaron 2).

4.- Se escoge el tipo de phred al que se va a basar el filtrado, por esta razón dentro de las características que se observaron en el html del fastqc, indicaba que la calidad es en referencia a Illumina 1.9 y por tanto esto lo hace a partir de phred 33.

5.- Lo siguiente se pone la ruta del fastq al que se le quiere limpiar (SRR13867562.fastq), y el siguiente es como debe de salir el resultado en. fastq (trimmSRR13867562.fastq), aqui se tuvo unos problemas con la ruta del archivo original SRR13867562.fastq, por lo que se decidió cambiar el fastq a la ruta actual a donde se estaba ejecutando el comando de java.

6.- Con la siguiente parte ILLUMINACLIP:TruSeq3-SE.fa:2:30:10 también se decidió copiar el archivo TruSeq3-SE.fa de la ruta original a donde estaba trabajando, por los mismos problemas. Además con las siguientes partes del comando LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 se añadieron por default como estaba en la descripción anterior.

Mapeo y cobertura

bwa

bwa es un software para realizar el mapeo de secuencias con baja convergencia sobre largas referencias genómicas semejantes al del ser humano. Este funciona bajo tres algoritmos BWA-backtrack, BWA-SW y BWA-MEM

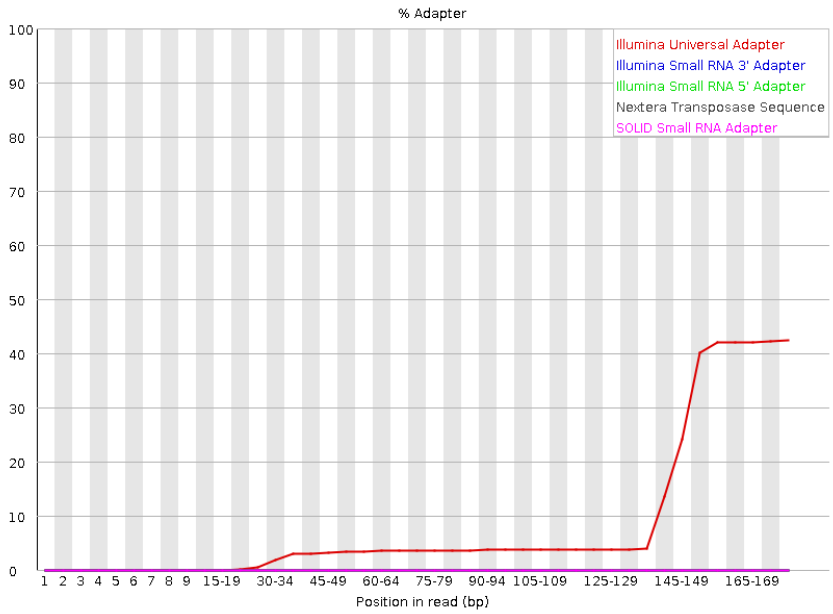
Para obtener la cobertura es necesario hacer el mapeo con bwa, para consequir el genoma de referencia, este caso será el genoma SARS-CoV-2 debe ser extraído de la base de datos NCBI, entonces en la base de datos de genomes, se escribe SARS-CoV-2 y en la referencia del genoma (reference genome) que aparece en la parte superior, se presiona, así en la siguiente página abrimos el apartado que dice FASTA y se descarga. Finalmente, el nombre del FASTA se cambió a SARS.fasta.

En una carpeta nueva, y dentro de ella seguimos el protocolo bwa index ref.fa [11] para hacer el índice, entonces con el **comando 4** podemos ver que salieron nuevos archivos, lo cuales serán necesarios para el mapeo. Con la siguiente sintaxis bwa mem ref.fa reads.fq > aln.sam, [12] ya que los reads considere por consenso que en la mayoria de ellos son de 70bp < reads < 190bp, ademas cabe destacar que provienen de la secuenciacion Illumina, tal como señala este repositorio. Entonces el comando resulta en (**comando 5**)

Fig. 10: Sobreexpresión de secuencias, en la bitácora se muestran el total de ellas

Sequence	Count	Percentage	Possible Source
AGGCAAACTGAGTTGGACGTGTGTTTTCTCGTTGAAACCAGGGACAAGGC	102026	2.0186203911010234	No Hit
CTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACAT	75126	1.4863944043857003	No Hit
GTCAAGCTGTCACGGCCAATGTTAATGCACTTTTATCTACTGATGGTAAC	73655	1.4572901506140186	No Hit
ACGCCTAAACGAACATGAAATTTCTTGTTTTCTTAGGAATCATCACAACT	73178	1.447852537392338	No Hit
AGGGGCTGAACATGTCAACAACTCATATGAGTGTGACATACCCATTGGTG	69383	1.3727671240248789	No Hit
AGCTAGTGGGGGACAACCAATCACTAATTGTGTTAAGATGTTGTGTACAC	69239	1.3699180332409753	No Hit

Fig. 11: Contenido de adaptadores



Comando Trimmomatic

```
$ java -jar /app/anaconda3/opt/trinity-2.9.1/trinity-plugins/Trimmomatic/trimmomatic.jar SE
-threads 2 -phred33 SRR13867562.fastq trimmSRR13867562.fastq ILLUMINACLIP:TruSeq3-
SE.fa:2:30:10 LEADING:3 TRAILING:

3 SLIDINGWINDOW:4:15 MINLEN:36

TrimmomaticSE: Started with arguments:

 -threads 2 -phred33 SRR13867562.fastq trimmSRR13867562.fastq ILLUMINACLIP:TruSeq3-
SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

Using Long Clipping Sequence: 'AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA'

Using Long Clipping Sequence: 'AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC'

ILLUMINACLIP: Using 0 prefix pairs, 2 forward/reverse sequences, 0 forward only sequences,
0 reverse only sequences

Input Reads: 5054244 Surviving: 4828483 (95.53%) Dropped: 225761 (4.47%)

TrimmomaticSE: Completed successfully


$ ls

SRR13867562.fastq  TruSeq3-SE.fa  trimmSRR13867562.fastq
```

Comando bwa index

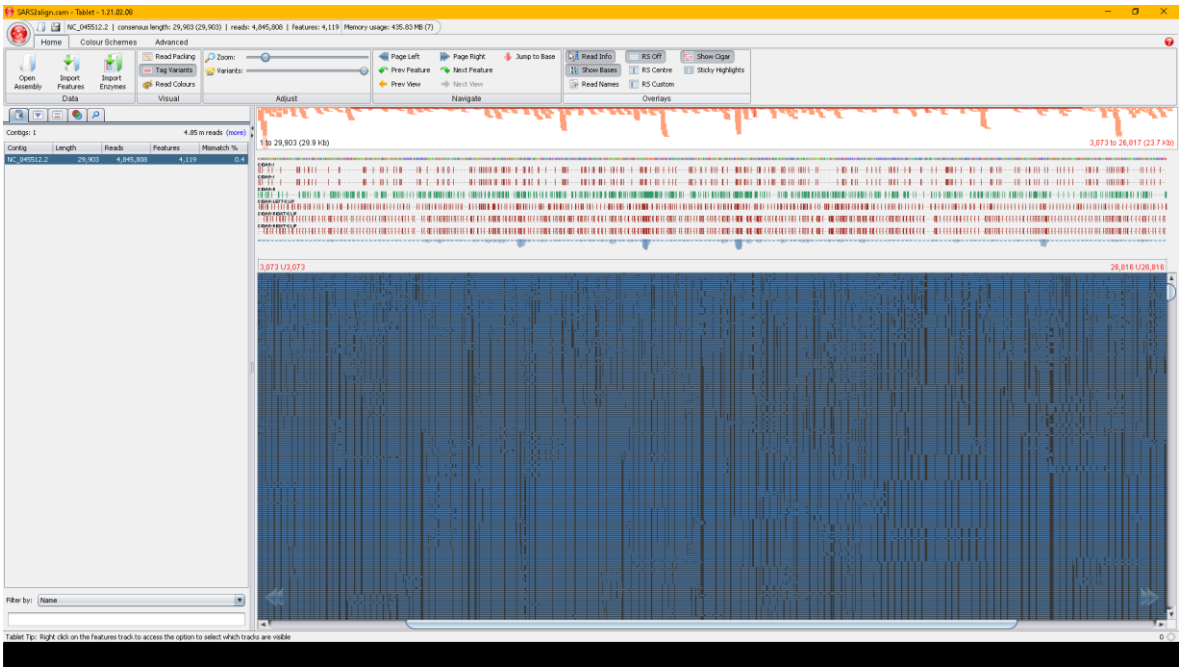
```
(base) [ SARS]$ bwa index SARS.fasta

(base) [ SARS]$ ls
SARS.fasta  SARS.fasta.amb  SARS.fasta.ann  SARS.fasta.bwt  SARS.fasta.pac  SARS.fasta.sa
```

Comando bwa mem

```
(base) [ SARS]$ bwa mem SARS.fasta ../SRR13867562.fastq >SARS2align.sam
(base) [ SARS]$ ls
SARS.fasta  SARS.fasta.amb  SARS.fasta.ann  SARS.fasta.bwt  SARS.fasta.pac  SARS.fasta.sa
SARS2align.sam
```

Fig. 12: Tablet, mapeo con el genoma de referencia



Tablet

Es un visualizador grafico de ensambles y alineamientos NGS de alto rendimiento [13]

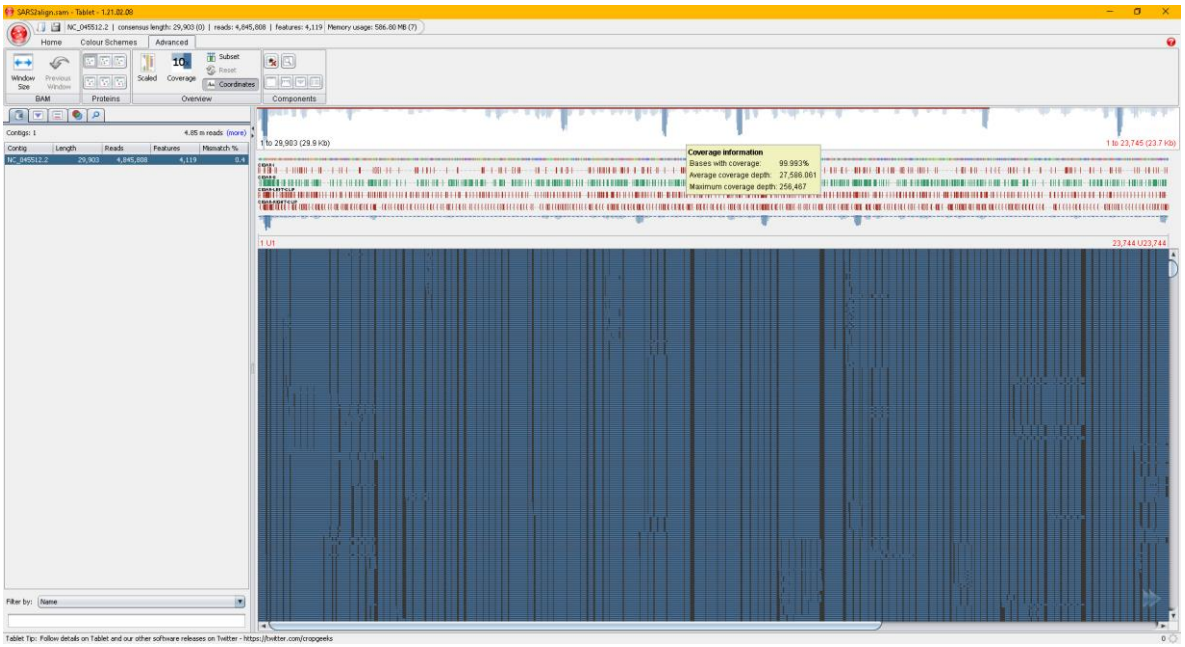
Entonces para obtener la cobertura, con Tablet en la parte que dice home > Open assembly y en el primer recuadro se coloca el archivo .sam y en el segundo recuadro se pone .fasta, con el ensamble terminado (fig. 12), para obtener la cobertura se dirigió a advanced y en donde dice coverage se seleccionó y también Coordinates para finalmente ver el porcentaje de cobertura

1. ***C = coverage***
2. ***L = read length (bp)***
3. ***N = number of reads***
4. ***G = haploid genome length (bp)***

Entonces con G considere la longitud del fasta

```
# grep -v ">" SARS.fasta | wc -c
30332
```

asi como o ***L = 190bp*** tales que pueden ser menores o iguales a 190bp y ***N*** lo saque de tablet ***4,845,808***. Por tanto ***C = (L x N) / G*** talque ***C = (190bp x 4,845,808)/30,332bp = 30,354x***



Asi mediante la ecuación de Lander-Waterman [14] tambien calculamos la cobertura, (**cálculos 1**)

Discusión y Conclusiones

Explico un poco finalmente en que resulto y otras cosas que me invente jeje

Bien, mañana va a tocar hacer la introducción, contando el comienzo del covid, que es el covid y la situación de porque se secuencia, además de contar porque escogi este virus (para hacer un ensamble pequeño)

Fig. 13: Tablet, cobertura

Bibliografía

1.- Instituto Nacional del Cáncer (2 de julio de 2020) *SARS-CoV-2* Definición de SARS-CoV-2 https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/sars-cov-2

2.- Secretaría de Salud (12 de marzo de 2020) *COVID-19 PREGUNTAS FRECUENTES* Gobierno de México https://www.gob.mx/salud/documentos/covid-19-preguntas-frecuentes

3.- World Health Organization (2021) *Coronavirus disease*

(COVID-19) World Health Organization https://www.who.int/health-topics/coronavirus#tab=tab_1

4.- WHO-China. (2021). WHO-convened Global Study of Origins of SARS-CoV-2: China Part. https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part

5.- Kumar A, Prasoon P, Kumari C, Pareek V, Faiq MA, Narayan RK, Kulandhasamy M, Kant K. SARS-CoV-2-specific virulence factors in COVID-19. J Med Virol. 2021 Mar;93(3):1343-1350. doi: 10.1002/jmv.26615. Epub 2020 Nov 1. PMID: 33085084.

6.- Mackenzie R. J., (6 de abril de 2018) *RNA-seq: conceptos básicos, aplicaciones y protocolo* News Curier http://www.news-courier.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461

7. National Human Genome Research Institute (27 sept 2019) *Transcriptoma* National Human Genome Research Institute https://www.genome.gov/es/about-genomics/fact-sheets/Transcriptoma

8.- Klymenko A. *The NCBI SRA (Sequence Read Archive)* Github https://github.com/ncbi/sra-tools

9.- Simon Andrews *FastQC* Babraham Bioinformatics https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

10.- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170. http://www.usadellab.org/cms/?page=trimmomatic

11.- Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-1760. [PMID: 19451168] (no es la cita que es para el algoritmo que se usó, pero sirve para tenerlo en cuenta) http://bio-bwa.sourceforge.net/bwa.shtml#3

12.- Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-1760. [PMID: 19451168] (Leer el README.md) https://github.com/lh3/bwa

13.- Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD and Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics 14(2), 193-202. https://ics.hutton.ac.uk/tablet/

14.- The Sequencing Center (sin fecha) *What is sequencing coverage?* The Sequencing Center https://thesequencingcenter.com/knowledge-base/coverage/

x.- Poojary, M., Jolly, B., Scaria, V., & Shantaraman, A. (2020). Computational Protocol for Assembly and Analysis of SARS-nCoV-2 Genomes. Research Reports, 4, 1–14. DOI: 10.9777/rr.2020.10001 https://www.researchgate.net/publication/341592466

y.- Alla Mikheenko, Andrey Prjibelski, Vladislav Saveliev, Dmitry Antipov, Alexey Gurevich, Versatile genome assembly evaluation with QUAST-LG, Bioinformatics (2018) 34 (13): i142-i150. doi: 10.1093/bioinformatics/bty266 First published online: June 27, 2018