Universidad Nacional Autónoma de México
Facultad de Ciencias

Genómica computacional
Alineamiento genómico del SARS-CoV-2

Balmaseda Villarreal Diana, Loya Sanchez Gabriel, Tenorio Hernandez Luis
Facultad de Ciencias, Av. Universidad No. 3000, Universidad Nacional Autónoma de México, Circuito interior. Del. Coyoacán.

Introducción

El coronavirus 2 del síndrome respiratorio grave (SARS-CoV-2) es una cepa viral asociada al desarrollo de una enfermedad respiratoria denominada enfermedad por coronavirus o COVID-19. Actualmente, se conocen tres enfermedades relacionadas a coronavirus: SARS-CoV descrito en 2003, MERS-CoV descrita en 2012 y el reciente SARS-CoV-2 descrito en 2019 (Gorbalenya, 2020).

La subfamilia de coronavirus (*Othocoronaviridae)* es caracterizada por su relación con enfermedades del tracto respiratorio en aves y mamíferos. Causando desde diarreas, infecciones respiratorias o la muerte. Morfológicamente presentan una envoltura vírica que recubre la cápside donde se encuentra el RNA monocatenario positivo, el cual será finalmente utilizado para la replicación del virus (Cascella, 2020).

La importancia de su estudio reside en que cada cepa se ha descrito por su relación con importantes contagios a nivel mundial, algunos denominados epidemias y el actual COVID-19 con categoría de pandemia, siendo importante la transmisión directa de hospedero a hospedero.

Los datos con los que se cuenta provienen de una secuenciación metagenómica de RNA, obtenidos del lavado de fluido bronquioalveolar (BALF) en un paciente alojado en el Hospital Central de Wuhan. Los mismo se encuentran almacenados en la base de datos del Centro Nacional para la Información Biotecnológica (NCBI) (Fan ,2020).

Los datos metagenómicos son resultado de una secuenciación metatranscriptomica del RNA total obtenido del BALF, se debe considerar la calidad y cantidad obtenida, después se construye una librería con los fragmentos de mRNA cuidando la limpieza de rRNA de la muestra. Posteriormente, se sintetizan dos cadenas de cDNA por la transcriptasa reversa y la DNA polimerasa respectivamente. Finalmente, se unen los adaptadores a ambos extremos de las secuencias pareadas, ya sea por PCR o ligadura. (Peimbert, 2016).

El trabajo bioinformático consiste en la posterior filtración de las lecturas y la decisión si se ensambla *de novo* (con programas como MegaHit, SOAPdenovo, Trinity o

Velvet) o si se alinea con secuencias de referencia (con programas como Bowitie o BWA). Los resultados de alineación *de novo* son comparados con bases de datos para proteínas y nucleótidos no redundantes (usando BLAST) para la identificación de posibles agentes etiológicos en la muestra (Fan, 2020).

Finalmente cuando es reconocida una coincidencia importante con un largo par de bases se puede diseñar un primer para confirmación y detección de terminales del genoma mediante PCR. Así, para obtener la cobertura del genoma y se remapean todas las lecturas obtenidas anteriormente comparando con el genoma de mayor compatibilidad (Fan, 2020).

Las pruebas estadísticas son importantes para fundamentar los resultados obtenidos y así finalmente puedan almacenarse en la base SRA del NCBI.

Cabe destacar, que entre las nuevas tecnologías de secuenciación (NGS) encontraremos reportadas lecturas sencillas (single-end) o pareadas (pair-end), siendo estás últimas las más usadas. Siendo la principal diferencia que en las lecturas sencillas solo se lee un extremo de los fragmentos de DNA, mientras que en las pareadas se generan lecturas de ambos extremos del fragmento de DNA separados por una distancia conocida (Rodriguez, 2012).

Las secuencias pareadas pueden ser de dos tipos: mate-pairs, en las que los fragmentos de DNA pueden ser de 600 pb a 4kb; o pairs-end en donde los fragmentos suelen ser menores a 300 pb (Illumina).

Objetivo(s)

Obtener datos de secuenciación crudos referentes al SARS-CoV-2 almacenados en la plataforma NCBI.

Comparar los resultados obtenidos del alineamiento de lecturas con programas especializados.

Realizar el ensamble genómico del SARS-CoV-2 a partir de los datos genómicos.

Metodología

Sección 1: Obtención de datos

Se accedió a los datos meta-transcriptómicos obtenidos del BALF de un paciente en Wuhan, China; que consisten en lecturas *paired-end* (150 pb) generadas en la plataforma MiniSeq (Illumina). Estos se encuentran en la base de datos *Sequence Read Archive (SRA)* de NCBI almacenados con referencia: PRJNA603194.

Se utilizó *SRA Toolkit (v. 2.10.5)* paras su descarga haciendo uso de los comandos *profetch* y *fastq-dump*. Al ser lecturas pareadas, se debía contar con dos archivos distintos de tipo fastq, mismos que se almacenaron con los nombres “SRR10971381_..fastq” y “SRR10971381_2.fastq”.

Sección 2: Análisis de calidad y limpieza.

El análisis de calidad se realizó con *FASTQC (v. 0.11.9)* en ambos archivos fastq.

Posteriormente, la limpieza se ejecutó en *ERNE (v. 1.4.4)* con los comandos *erne-filter* y la opción *--min.mean.phred-quality 10*; estableciendo que las lecturas con un phred score promedio menor a 10 fueran eliminadas. Adicionalmente se utilizó *Trimmomatic (v. 0.39)* mediante la plataforma usegalaxy.org, para limpiar las secuencias, se indicó un valor umbral de 10 en phred score para el filtrado de las lecturas.

Seccion 3: Alineamiento

Con base en los resultados obtenidos por las limpiezas, se optó por utilizar los datos crudos ya que la información perdida era significativa, debido a la mala calidad de las secuencias.

Dichos datos se alinearon con el genoma de referencia del SARS-CoV-2 almacenados en NCBI con la referencia MN908947.3. Para lo cual se utilizaron *BowTie2 (v. 2.4.1)* y *BWA (v 0.7.17)*.

En el caso de *BowTie2* se utilizó el alineamiento local, donde se elimina la restricción para alinear ambos extremos de las lecturas con el genoma de referencia. Se realizó de esta forma ya que las lecturas crudas aún contienen los adaptadores necesarios para la secuenciación, así como lecturas con phred scores bajos.

Para *BWA* se empleó *BWA-MEM* ya que este es el recomendado para lecturas largas y con errores, según sus desarrolladores, esto se logró con el comando *bwa mem BETA SRR10971381_1.fastq SRR10971381_2.fastq > bwa_mem_alineamientos.sam.*

Como resultado de ambos se obtuvo un archivo tipo SAM con todas las lecturas de los archivos fastq, con información propia del alineamiento (mapeos, flagstats, etc.). También se especificó en el caso de *BowTie2*, que las lecturas alineadas con el genoma de referencia se guardarán en archivos fastq separados cada uno correspondiente a las dos lecturas iniciales, los archivos fastq se lograron con awk en el caso de los archivos obtenidos con *BWA*.

Los archivos obtenidos de ambos softwares contenían sólo las lecturas alineadas. Para llevar a cabo dicha tarea se utilizó *SAMtools (v. 1.7)*. El archivo Bam o SAM obtenido se visualizó en *SeqMonk (v. 1.47.0)*.

Seccion 3: Ensamble

El ensamble del genoma se realizó *SPAdes (v.3.13.0)* usando como input ambos archivos fastq que contenían las lecturas alineadas.

A partir de los archivos fastq generados en *SPAdes* se visualizaron los contigs resultantes, haciendo uso de *Bandage.* El contig mejor representado se aisló en formato fasta y se validó su coincidencia con el genoma de referencia con *Web Blast* de NCBI.

Resultados

Los archivos fastq fueron obtenidos exitosamente, así como los análisis de calidad de los datos**(Fig.1)**.

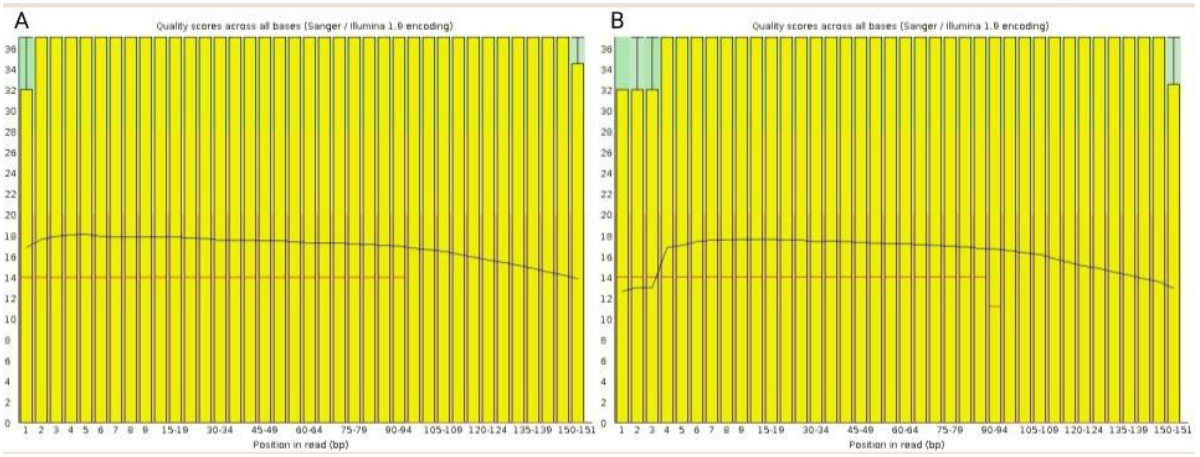

Figura 1. Distribución de phred scores a lo largo de las lecturas, resultado del análisis de calidad inicial con FastQC sobre las secuencias crudas para el archivo 1(**A**) y para el archivo 2 (**B**).

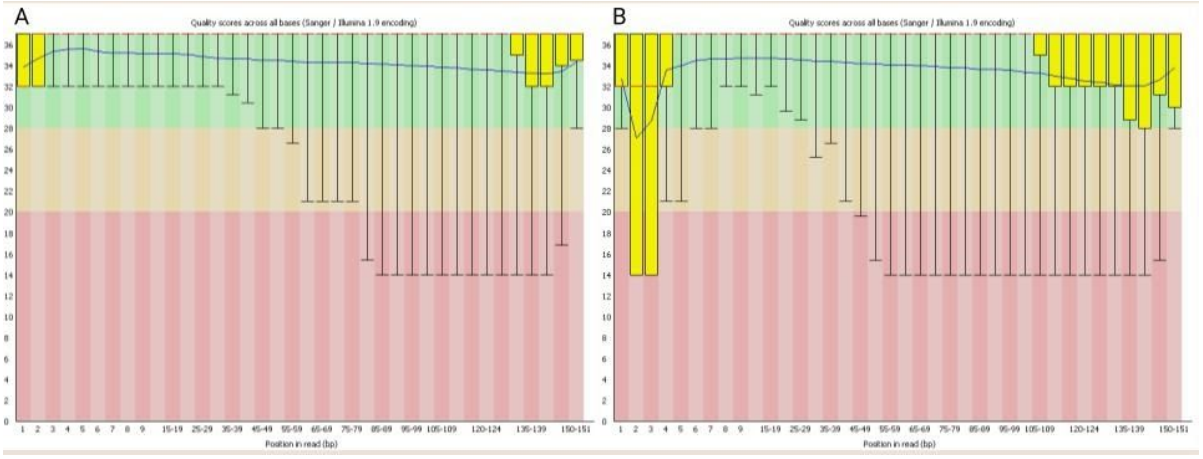

Figura 2. Distribución de phred scores a lo largo de las lecturas, resultado del análisis de calidad con FastQC sobre las secuencias después de ser filtradas en ERNE para el archivo 1(**A**) y para el archivo 2 (**B**).

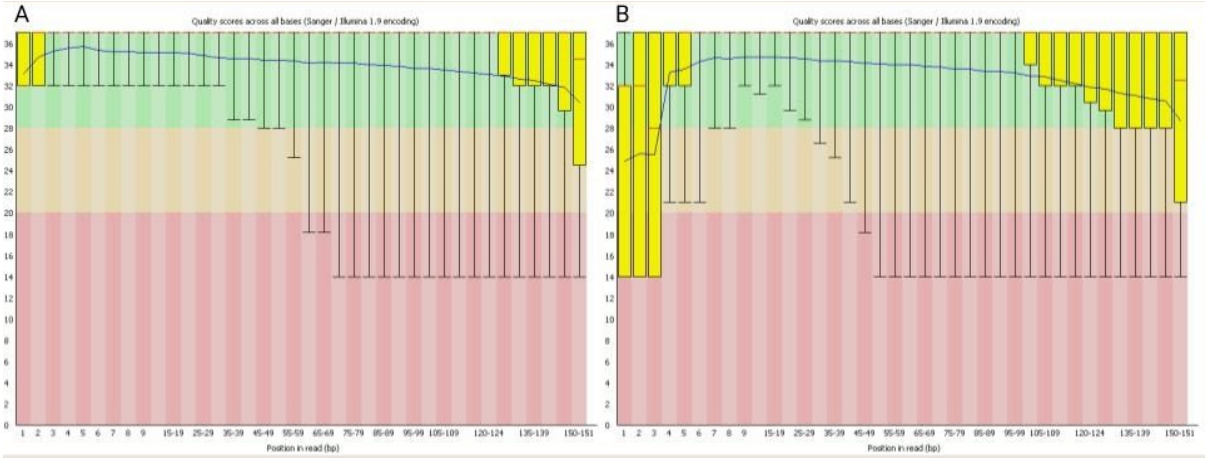

Figura 3. Distribución de phred scores a lo largo de las lecturas, resultado del análisis de calidad con FastQC sobre las secuencias después de ser filtradas con Trimmomatic para el archivo 1(**A**) y para el archivo 2 (**B**).

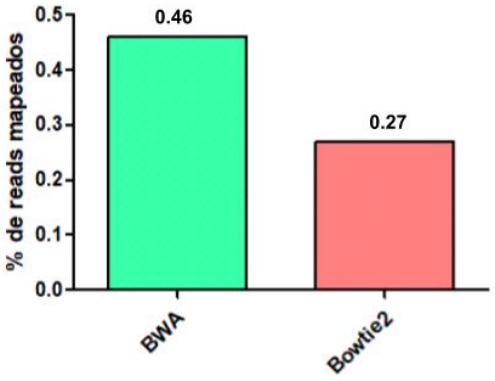

Figura 4. Comparación del porcentaje de reads mapeados usando los softwares BWA y Bowtie2. Total de lecturas: 56,624,252; lecturas alineadas con BWA: 260,042; lecturas alineadas con bowtie2: 154,529

Sin embargo, en los análisis de calidad posteriores a la limpieza de datos se observó una pérdida significativa de lecturas con ambos softwares (*Trimmomatic* y *ERNE*), razón por la cual se decidió trabajar con las lecturas sin limpiar. **(Fig. 2 y 3)**.

El alineamiento fue exitoso con *BowTie2* y *BWA*. Aunque, se determinaron diferencias en cuanto al porcentaje de lecturas alineadas; siendo en *BWA* donde se obtuvieron mayor cantidad de alineamientos **(Fig. 4)**.

En dichas lecturas, visualizadas en *SeqMonk* **(Fig. 5 y 6)**, se observó de forma global una distribución muy uniforme de las lecturas a lo largo del genoma del SARS-CoV-2, con diferencias sutiles sólo visibles a menor escala.

El ensamble del genoma se realizó en SPAdes con ayuda de un servidor del Instituto de Biotecnología, UNAM. La visualización de los contigs mostró gráficos similares a los obtenidos en *BowTie2* **(Fig. 8A)** y *BWA* **(Fig. 7A)**. El gráfico más destacado obtenido con los datos alineados en *BowTie2* es más pequeño que el generado por *BWA*, además se observa diferencia en los contigs generados, 10 y 35 respectivamente; lo cual se infere es reflejo de la menor cantidad de alineamientos generados por *BowTie2.*

De ambos ensambles, generados y obtenidos por cada programa, se extrajeron las secuencias representadas en los gráficos más destacados a un archivo fasta. Dichos archivos se alinearon con *Web BLAST* de NCBI para validar el alineamiento. En ambos casos los archivos se alinearon con el genoma del SARS-CoV-2 con un Querry Recovery del 99% **(Fig. 7B y 8B)**.

Conclusión

Los objetivos propuestos para este proyecto se cumplieron ya que fue posible trabajar con datos genómicos, alinearlos con dos programas distintos con fines de comparación y finalmente ensamblar un genoma con los datos alineados con ambos softwares.

Cabe también destacar algunas observaciones: En primer lugar la importancia de la generación de datos de buena calidad desde que salen del secuenciador, datos de buena calidad siempre facilitarán el trabajo. También es muy importante documentarse acerca de los softwares existentes

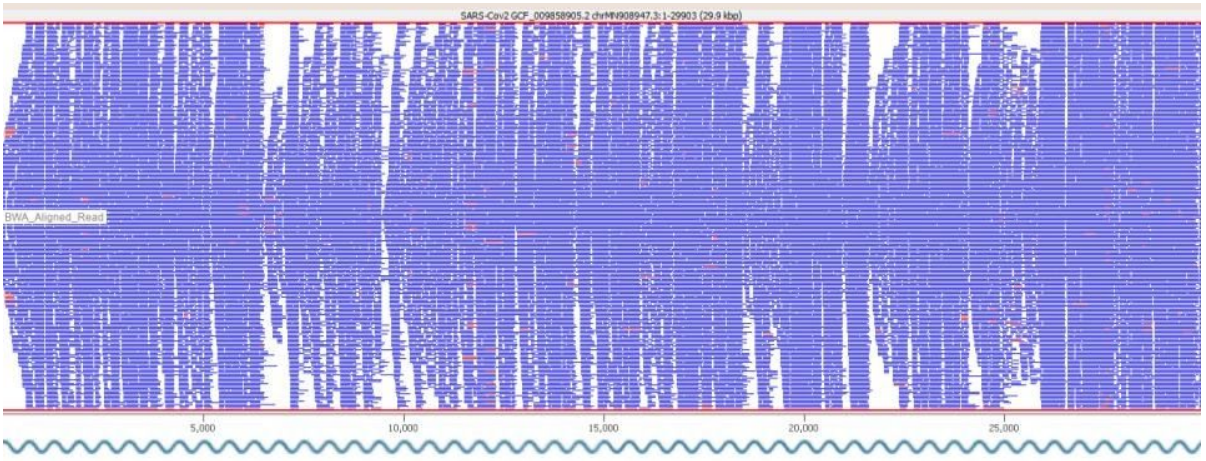

Figura 5. Visualización en el software SeqMonk de lecturas mapeadas al genoma de referencia obtenidas con el software BWA. En la parte posterior se muestra una escala que abarca las 29.9 kbs del genoma del SARS-CoV2. Las barras azules representan pares de reads que mapearon en esa ubicación del genoma, las barras rojas representan reads que alinearon con el genoma pero su par no alineó o lo hizo a una distancia mayor a 1kb.

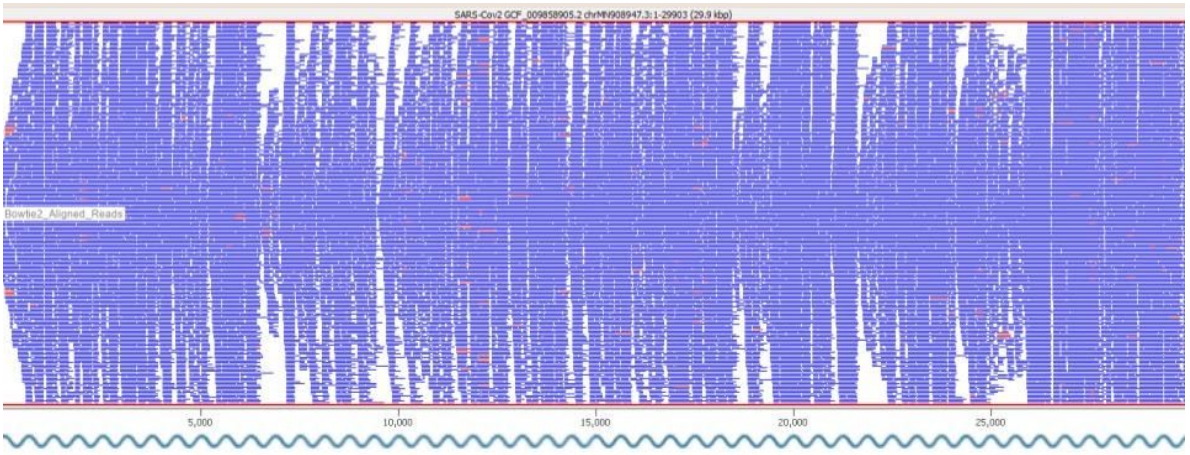

Figura 6. Visualización en el software SeqMonk de lecturas mapeadas al genoma de referencia obtenidas con el software bowtie2. En la parte posterior se muestra una escala que abarca las 29.9 kbs del genoma del SARS-CoV2. Las barras azules representan pares de reads que mapearon en esa ubicación del genoma, las barras rojas representan reads que alinearon con el genoma pero su par no alineó o lo hizo a una distancia mayor a 1kb.

y elegir el indicado para la tarea que deseamos realizar, dada la gran cantidad de softwares que existen. Algunos podrían funcionar mejor que otros por la naturaleza de nuestros datos o para los intereses particulares del proyecto. En este proyecto se encontró que BWA resultó más eficaz, al menos en términos de cantidad de lecturas alineadas, en comparación con Bowtie2. Finalmente también dentro de los mismo programas es importante conocer las posibilidades que estos nos ofrecen para poder tomar la mayor ventaja posible de las herramientas bioinformáticas a la hora de realizar un proyecto.

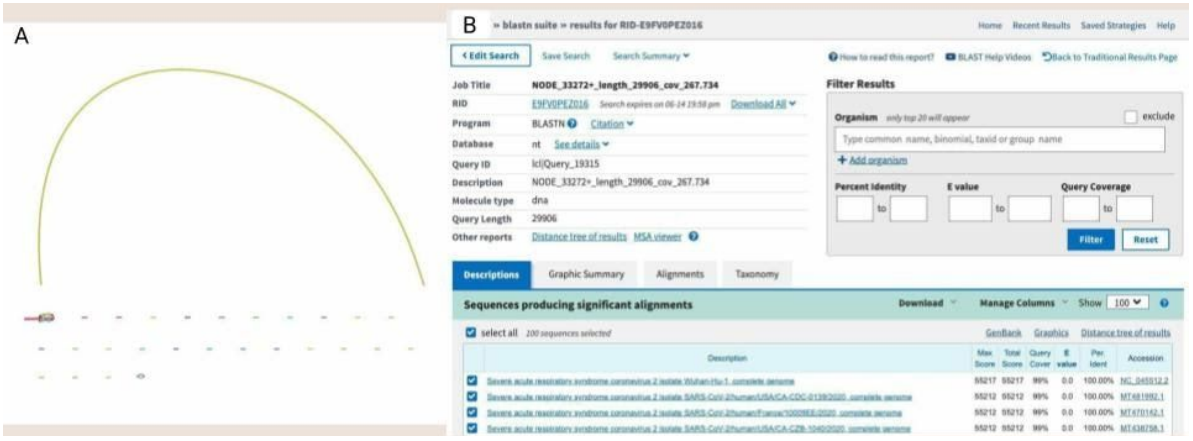

Figura 7. (**A**)Visualización en el software Bandage de los contigs generados con SPAdes a partir de lecturas mapeadas al genoma de referencia obtenidas con el software BWA . (**B**)Validación del ensamble representado por el gráfico más conspicuo en A, al hacer un alineamiento local en BLAST

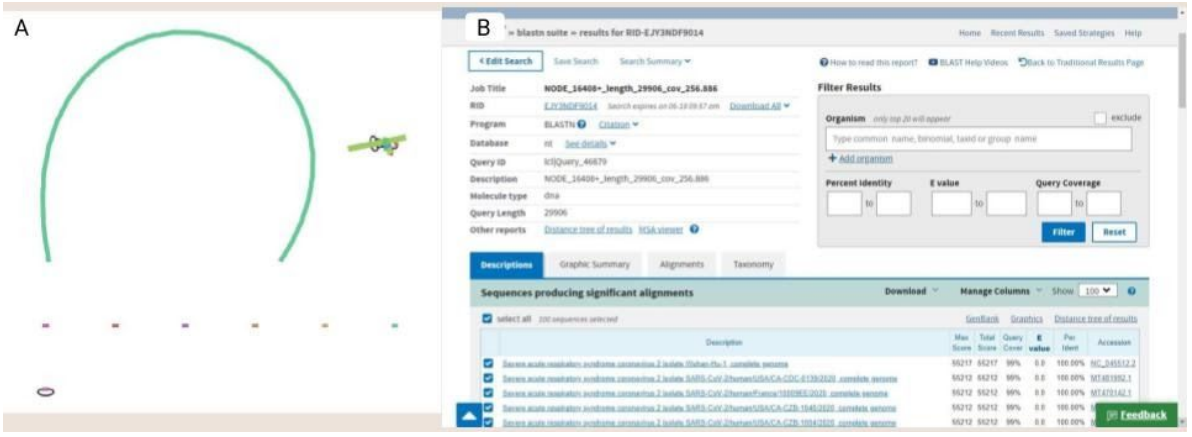

Figura 8. (**A**)Visualización en el software Bandage de los contigs generados con SPAdes a partir de lecturas mapeadas al genoma de referencia obtenidas con el software bowtie2. (**B**)Validación del ensamble representado por el gráfico más conspicuo en A, al hacer un alineamiento local en BLAST

Referencias

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed
- Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner.Journal of Computational Biology.May 2012.455-477.http://doi.org/10.1089/cmb.2012.0021
- Bankevich, A., Nurk, S., Antipov, D., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its

Applications to Single-Cell Sequencing. J Comput Biol. 2012 May;19(5):455-77. https://doi.org/10.1089/cmb.2012.0021

- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.
- Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Di Napoli, R. (2020). Features, Evaluation and Treatment Coronavirus (COVID-19). In StatPearls. StatPearls Publishing.
- Carmona Muñoz, R. (2012). Aplicaciones de las nuevas tecnologías de secuenciación. *¿CÓMO FUNCIONA?*, (Vol 5.), 138-139. Retrieved from http://www.encuentros.uma.es/encuentros138_9/secuenciacion.pdf
- Del Fabbro C., Scalabrin S., Morgante M., Giorgi F.M.
- PlosOne, December 23, 2013, Vol. 8(12):e85024, doi:10.1371/journal.pone.0085024
- Grabherr, M. G. et al. (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29, 644–652 . https://doi.org/10.1038/nbt.1883
- Gorbalenya, A.E., Baker, S.C., Baric, R.S. et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 5, 536–544 (2020). https://doi.org/10.1038/s41564-020-0695-z
- ICTV Taxonomy history: Severe acute respiratory syndrome-related coronavirus https://talk.ictvonline.org/taxonomy/p/taxonomy-history?taxnode_id=20181868
- Fan, W., Su, Z., Yan-Mei, C., Wen, W., & et, a. (2020). A new coronavirus associated with human respiratory disease in China. Retrieved 20 May 2020, from https://www.nature.com/articles/s41586-020-2008-3.pdf
- FastQC: a quality control tool for high throughput sequence data (2010)
- Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.
- Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676 . http://doi.org/10.1093/bioinformatics/btv033
- Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, Epub. PMID:20080505

- Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]
- Nurk S. et al. (2013) Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. In: Deng M., Jiang R., Sun F., Zhang X. (eds) Research in Computational Molecular Biology. RECOMB 2013. Lecture Notes in Computer Science, vol 7821. Springer, Berlin, Heidelberg
- Paired-End vs. Single-Read Sequencing Technology. (2020). Retrieved 14 June 2020, from https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html
- Peimbert M, Alcaraz L D. A Hitchhiker's Guide to Metatranscriptomic sequencing [M]// Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing. Springer International Publishing, 2016.
- Rodríguez-Santiago, B., & Armengol, L. (2012). Tecnologías de secuenciación de nueva generación en diagnóstico genético pre- y postnatal. *Diagnóstico Prenatal*, (Vol. 23. Núm. 2.), 56-66. Retrieved from https://www.elsevier.es/es-revista-diagnostico-prenatal-327-articulo-tecnologias-secuenciacion-nueva-generacion-diagnostico-S2173412712000273
- SeqMonk: A tool to visualise and analyse high throughput mapped sequence data (2007).
- NCBI SRA Toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software)
- Sutton, T.D.S., Clooney, A.G., Ryan, F.J. et al. (2019). Choice of assembly software has a critical impact on virome characterisation. Microbiome 7, 12 . https://doi.org/10.1186/s40168-019-0626-5
- Wick R.R., Schultz M.B., Zobel J. & Holt K.E. (2015). Bandage: interactive visualization of de novo genome assemblies. Bioinformatics, 31(20), 3350-3352. https://doi.org/10.1093/bioinformatics/btv383