Universidad Internacional de Valencia

Genómica y Filogenómica del SARS-CoV-2

De:
Planeta Formación y Universidades

Universidad Internacional de Valencia

ÍNDICE

1.-Abreviaturas

ORF : "*Open Reading Frame*" o Marco de Lectura Abierto

ARN : Ácido Ribonucleico

RdRp : "*RNA-dependent RNA polymerase*" o ARN polimerasa dependiente de ARN

kDa : Kilodalton

ARNi : ARN de interferencia

RTC : Complejo replicación-transcripción

OMS : Organización Mundial de la Salud

ICTV : "*International Committee on Taxonomy of Viruses*" o Comité Internacional en Taxonomía de Virus.

kb : Kilobase

nm : nanómetros

aa : aminoácido

RBD : "*Receptor-Binding Domain*" o Dominio de Unión al Receptor

ACE2 : "*Angiotensin-Converting Enzyme 2*" o Enzima Convertidora de Angiotensina 2.

TRMPRSS2 : *"Transmembrane Serine Protease 2*" o Serina proteasa transmembrana 2.

ARNm : ARN mensajero

ARNc : ARN complementario

TRS : "*Transcription Regulatory Sequence*" o Secuencia regulatoria de la transcripción.

RE : Retículo Endoplasmático.

ERGIC : "*Endoplasmic Reticulum-Golgi Intermediate Compartment*" o Compartimento Intermedio del RE y Golgi.

vRNP : "*viral Ribonucleoprotein Particles*" o Partículas virales de ribonucleoproteínas

IA : Inteligencia Artificial

pb : pares de bases.

WGS : "*Whole Genome Sequencing*" o Secuenciación de genoma completo

SNP : "*Single Nuleotide Polimorphism*" o variación/polimorfismo de un solo nucleótido

2.-Tablas

Tabla 1. Descripción de las múltiples proteínas que forman parte de los virus de la familia *Coronaviridae*.

Tabla 2. Descripción de las proteínas estructurales que conforman el virus del SARS-CoV-2.

Tabla 3. Listado de los 12 genomas descargados de NCBI para llevar a cabo el proceso de filogenómica.

3.- Gráficos y Figuras

4.-Resumen

La reciente pandemia mundial del COVID-19, provocada por el virus SARS-CoV-2, ha generado un gran número de contagios y muertes desde que empezó, 7 millones aproximadamente (1). Esta pandemia hizo que se aunaran todas las áreas de la biología, la medicina y las ciencias computacionales con el fin de encontrar soluciones efectivas frente al virus causante de esta enfermedad. La bioinformática y todas sus disciplinas han jugado un papel muy importante en esta búsqueda de soluciones. Con la gran cantidad de datos generados, ha sido preciso que se desarrollaran herramientas, softwares y pipelines, para limpiarlos, manejarlos e interpretarlos correctamente, para obtener resultados fiables y sólidos, que ayuden a una correcta toma de decisiones, y a adquirir mayores conocimientos sobre el agente causante de la enfermedad, monitoreando su expansión, y sus alteraciones o mutaciones, con el fin de prevenir nuevos brotes causados por nuevas variantes del virus.

En este trabajo, se ha llevado a cabo la comparación de tres pipelines utilizadas para la secuenciación, determinación de variantes, predicción y anotación de genomas, y determinación de linajes del SARS-CoV-2, fijando nuestro interés en la pipeline PipeCoV (2) . Por ello, realizamos una réplica a pequeña escala del trabajo llevado a cabo por su autor, comparándola con las pipelines nf-core/Viralrecon (3) y Vpipe(4), y posteriormente utilizando los ensamblajes obtenidos para realizar un protocolo de filogenómica utilizando otros 12 genomas pertenecientes a la familia *Coronaviridae*.

En los resultados de ensamblaje, se observó que, de cara al ensamblaje de genoma, PipeCoV es buena elección, puesto que disminuye el número de bases indeterminadas (N) con respecto a las otras dos pipelines, manteniendo una longitud de genoma similar a la de la referencia. Sin embargo, nf-core/Viralrecon, fue la única pipeline que obtuvo variantes fiables y respaldadas.

En cuanto a los resultados de la filogenómica, se obtuvieron resultados alineados parcialmente con respecto a la bibliografía y evidencia existente, situando como pariente más cercano evolutivamente del SARS-CoV-2 y los dos ensamblajes obtenidos, al SARS-CoV, otro Betacoronavirus de humanos, y al coronavirus de murciélago HKU9.

Palabras Clave o *Keywords*: Genómica, Filogenómica, SARS-CoV-2, COVID-19

5.-Introducción

En el año 2019, en Wuhan (China), apareció la enfermedad conocida como COVID-19, la cual, ha provocado una pandemia mundial, causando hasta la actualidad 7.100.000 muertes aproximadamente, según la OMS (1). El causante de esta enfermedad es el nuevo virus SARS-CoV-2, estrechamente relacionado en cuanto a su estructura con el virus causante del síndrome respiratorio agudo severo (SARS-CoV) causante de brotes de neumonía en los años 2002-2003 en China, y con el virus del síndrome respiratorio de Oriente Medio (MERS-CoV) en el año 2012 y que causó también varios brotes en años posteriores no solo en Arabia Saudita, sino también en otros países como Jordania o Corea del Sur (5–7).

El SARS-CoV-2 se trata de un virus zoonótico (se puede transmitir de animales a humanos), y al igual que los otros dos coronavirus que se han mencionado en el párrafo anterior, con los cual está relacionado, la enfermedad que produce, el COVID-19, se caracteriza por la generación de neumonía, en bastantes casos, graves y críticas, y que presenta una gran transmisión de persona a persona mediante gotas respiratorias de individuos infectados a individuos susceptibles (8,9).

Esta pandemia mundial, al tratarse de una emergencia sanitaria global, ha conseguido que todas las ramas de las ciencias (medicina, biología, biotecnología, bioquímica, ...) se implicaran de lleno en solventar el problema, experimentando así un rápido avance en los conocimientos sobre este virus, y un rápido desarrollo de vacunas frente al mismo. Un papel importante lo ha jugado la bioinformática y las distintas áreas de esta ciencia, como la genómica y/o la transcriptómica, entre otras, que han conseguido en un breve periodo de tiempo, dar a conocer las estructuras moleculares del virus, así como sus mecanismos de actuación.

A continuación, en los sucesivos apartados, se irán mencionando y describiendo las características más comunes de los virus de la familia *Coronaviridae*, y también las del virus SARS-CoV-2, así como los procesos y metodologías utilizados tanto para adquirir un mayor conocimiento sobre estos virus, como el parentesco y cercanía que comparten, puesto que se trata de una familia de virus con un gran potencial, con capacidad de provocar epidemias y pandemias globales, por lo que todo conocimiento

adquirido es útil para prevenir nuevas situaciones de pandemias como la ocurrida recientemente en el año 2019 con la enfermedad del COVID-19.

5.1.-Familia *Coronaviridae*.

En este apartado describiremos de la manera más brevemente posible las características moleculares y estructurales generales, y la clasificación de los virus pertenecientes a esta familia. Son numerosos los artículos que describen las características de los virus de la familia *Coronaviridae*, puesto que es una familia ampliamente estudiada desde hace tiempo, desde cuando fue descubierto en 1932 su primer miembro (10). Para llevar a cabo esta descripción nos centraremos en los artículos (11–16), y en el reporte sobre esta misma familia llevado a cabo por el Comité Internacional en Taxonomía de Virus o *International Committee on Taxonomy of Viruses* (ICTV) (17,18), principal organismo a nivel mundial encargado de la clasificación de todos los virus conocidos hasta el momento, siendo la autoridad más actualizada en lo que a virus se refiere. Destacar que mencionamos de antemano los artículos puesto que la información obtenida de ellos es muy similar, en muchos de los casos exactamente la misma, por lo que es difícil distinguir qué partes de la información corresponden a uno u otro, ya que esta información se contiene en todos estos artículos, sin prácticamente modificación.

Los virus de esta familia han sido asociados con la producción de diferentes enfermedades en distintos animales y también en humanos, causando una amplia variedad de estas enfermedades, entre ellas enfermedades respiratorias, gastrointestinales, neurológicas y/o multisistémicas, lo que señala su potencial como agentes patógenos de un gran abanico de enfermedades, unido a un gran rango de huéspedes que actúan como reservorios y transmisores del virus, y una gran cantidad de huéspedes en los que producir este tipo de enfermedades, puesto que distintos coronavirus infectan a distintas especies animales, desde aves, hasta mamíferos, incluido reptiles y anfibios.

Con respecto a su biología molecular, su principal característica es que son virus de ARN monocatenario lineal de sentido positivo, policistrónico, capado en su extremo 5', y poliadenilado en su extremo 3', con un tamaño de genoma que oscila entre 22 y 36 kb, en función del género al que pertenezcan dentro de la familia.

Atendiendo a la organización del genoma (Figura 1), estos virus presentan un numero variable de ORFs, de tal forma que casi dos tercios de su genoma, hacia el extremo 5', codifica para proteínas no estructurales utilizadas para procesos transcripción y replicación del genoma, así como para otros procesos relacionados con la interferencia y/o atenuación del sistema inmune del huésped al que producen al enfermedad, mientras que el tercio restante del genoma codifica para proteínas estructurales, siendo las principales las proteínas S, E, M y N, las cuales, aparecen en el mismo orden en todos los virus conocidos de esta familia, y algunas otras proteínas accesorias a estas principales, encargándose estas proteínas de funciones como la empaquetamiento del material genético (proteína N o proteína de la nucleocápside), dar forma y envoltura al virión (proteína E o proteína de envoltura, y proteína M o glicoproteína de membrana), y la unión a las células dianas del huésped (proteína S o proteína de espiga/espícula). En la Tabla 1 se puede encontrar un resumen de las principales proteínas tanto estructurales, como no estructurales, codificadas por lo coronavirus.

En cuanto a la estructura de los viriones generados, éstos presentan una morfología esférica, con un diámetro de entre 80 y 160 nm, con la proteína de espiga sobresaliendo de la envoltura a modo de proyección superficial, lo que le confiere un aspecto de corona, dando así nombre a la familia.

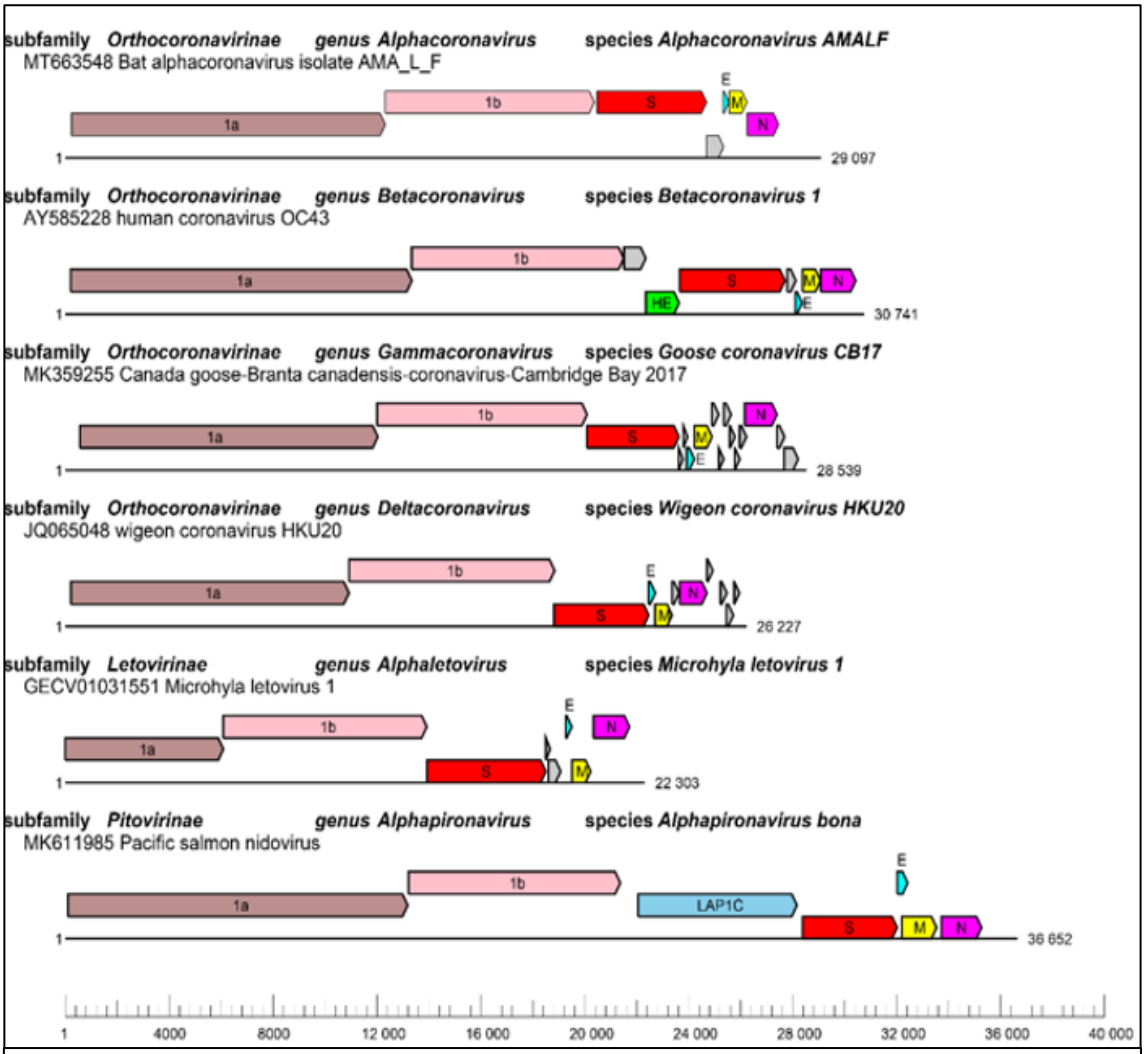

Figura 1.- Organización del genoma de diferentes virus pertenecientes a la familia *Coronaviridae*. Se puede observar como la longitud de los genomas oscila entre 22 y 36 kb, presentando tamaños de genomas bastantes similares dentro de cada subfamilia (91).

Cabe destacar que en este apartado no se explicará el ciclo biológico general de los virus pertenecientes a esta familia, ya que en apartados posteriores, se explicará el del SARS-CoV-2, siendo este similar al del resto de virus.

Tabla 1.- Descripción de las múltiples proteínas que forman parte de los virus de la familia Coronaviridae. La definición de las funciones de las proteínas nsp1 hasta nsp16 (excepto nsp2) se ha parafraseado directamente de la "Tabla 17.1" del artículo (11). El resto de información, ha sido obtenida en conjunto entre este último artículo mencionado y los estudios (19–22)

Proteína		Funciones
nsp1	**ORF1a: nsp1-nsp10** **ORF1ab o ORF1b: nsp1-nsp10 + nsp12-nsp16**	"Antagonista del interferón (no presente en todos los coronavirus)"
nsp2		Control y regulación de funciones del huésped, así como necesaria para la replicación viral
nsp3		"Dominios de proteasa tipo papaína y varios otros dominios de interacción con proteínas. Pueden unir el genoma del ARN al complejo replicasa/transcriptasa."
nsp4		"Andamio transmembrana. Participa en la remodelación de la membrana."
nsp5		"Proteasa principal (M pro) (también llamada proteasa tipo 3C)"
nsp6		"Andamio transmembrana. Participa en la remodelación de la membrana."
nsp7		"Forma un gran complejo con nsp8."
nsp8		"Forma un gran complejo con nsp7. Este complejo puede actuar como un freno de procesividad para la RdRp."
nsp9		"Proteína de unión a ARN monocatenario"
nsp10		"Cofactor de unión al zinc para la 2′- O -metiltransferasa (nsp16)"
nsp12		"RdRp"
nsp13		"ARN 5′ trifosfatasa (síntesis de capuchón), ARN helicasa"
nsp14		"N7-metiltransferasa, exonucleasa Exo N 3′–5′ (proporciona una función de corrección para la RdRp del coronavirus)"
nsp15		"Endonucleasa Nendo U (corta el ARN monocatenario y bicatenario aguas abajo de los residuos de uridilato, produciendo 2-3 fosfatos cíclicos)."
nsp16		"2′- O -metiltransferasa (síntesis de caperuza)"
ORF3a		Proteinas variables entre subfamilias y especies. Funciones variadas: - Regulación de la respuesta inmunitaria del húesped: interferencia con el interferón tipo I (IFN I) de la respuesta inmune innata (IRI), IRI proinflamatoria, estrés del RE, apóptosis y autofagia - Actividad de canal iónico o/y viroporina, asociadas a patogenisis, morfogénesis, o proliferación viral. -Virulencia del virus
ORF3b		
ORF4a		
ORF6		
ORF7a		
ORF7b		
ORF8a		
ORF8b		
ORF10		
ORF9a		
ORF9b		

otras	
proteína M	Formación de viriones, y mediante su interacción con otras proteínas estructurales, pueden contribuir a la estabilidad de las mismas, e incluso afectar a la interacción del virus con la célula húesped
proteína N	Unión al RNA genómico para empaquetarlo en la nucleocápside. También funciones esenciales en la transcripción y replicación viral.
proteína E	Importante para el desarrollo y diseminación de nuevos viriones. También presenta actividad de canal iónico.
proteína S	Reconocimiento de receptores de la célula húesped y responsable de la fusión de la membrana viral con la membrana de la célula huésped
proteína HE	Proteína de membrana con actividad hemaglutinante y acetil-esterasa, presente únicamente en algunos *Betacoronavirus*. Puede actuar como segunda proteína de reconocimiento de los receptores de la célula huésped

Por último, en este apartado hablaremos sobre la taxonomía de la familia *Coronaviridae*. Esta familia pertenece al orden de los *Nidovirales*. Al mismo tiempo, si bien en el pasado se encontraba formada por las subfamilias *Coronavirinae* y *Torovirinae*, investigaciones recientes tras la pandemia mundial del COVID-19, han dejado fuera de la familia a la subfamilia *Torovirinae*, quedando la familia formada por tres nuevas subfamilias según (ICTV): *Orthocoronavirinae*, *Letovirinae*, y *Pitovirinae*. En el artículo (23), esta familia solo presentaba dos subfamilias (*Orthocoranavirinae* y *Letovirinae*), por lo que tras una búsqueda más intensiva en (ICTV) (24), se ve que la subfamilia *Pitovirinae* sí que está incluida como nueva subfamilia de *Coronaviridae* desde el 2021, por lo que, si bien este artículo mencionado es bastante actual, no utilizarían el único genoma disponible actualmente de esta subfamilia para su investigación.

La subfamilia más numerosa y de mayor interés, debido a que sus huéspedes son aves y mamíferos (humanos entre ellos), es la subfamilia *Orhocoronavirinae*, dividida a su vez en cuatro géneros: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* y *Deltacoronavirus*. En los dos primeros géneros mencionados, se encuentran siete especies diferentes que infectan y causan enfermedad en humanos, perteneciendo el SARS-CoV-2, virus objetivo de este trabajo, al género *Betacoronavirus*.

5.2.- SARS-CoV-2.

El presente apartado se centra en una descripción más precisa sobre el virus del SARS-CoV-2, objetivo del presente estudio, entrando en mayores detalles en diferentes áreas de conocimiento e investigación como su biología molecular, su estructura, su ciclo de vida, su clasificación taxonómica, …

Se irán desarrollando diferentes subapartados para poder abordar con mayor detalle, claridad y organización los diferentes conocimientos sobre este virus.

5.2.1.- Origen, estructura y clasificación taxonómica.

El SARS-CoV-2, al igual que otros coronavirus, es un virus zoonótico, surgido en animales que han actuado como reservorios y/u hospedadores intermedios, antes de ser transmitido a humanos. Actualmente, existe controversia sobre su origen, puesto que hay dos coronavirus muy similares en cuanto a identidad de secuencia y estructura genómica, el RaTG13, un coronavirus de murciélago, y un coronavirus de pangolín malayo.

El estudio de “Zhang C.” y colaboradores (25), muestra que secuencias ensambladas del coronavirus del pangolín cubren el 73 % del genoma del SARS-CoV-2, con una identidad de secuencia del 91%, y con un 92% de identidad con respecto a la proteína de espiga del SARS-CoV2, mientras que el estudio (26), precursor del anterior, muestra una identidad de aminoácidos del 100%, 98.2%, 96.7% y 90.4% con el virus SARS-CoV-2 en los genes E, M, N y S, respectivamente.

Otros estudios, como el llevado a cabo por Zhou P. y colaboradores (27), señalan que el coronavirus RaTG13, comparte una identidad de secuencia, a nivel de secuencia genómica completa, del 96.2% con el SARS-CoV-2, mientras que con respecto a la proteína de espiga (S) comparte una identidad de nucleótidos del 93.1%.

Los resultados de estos estudios, junto al conocimiento previo sobre los coronavirus y sus hospedadores, parecen indicar que el posible origen del SARS-CoV-2 se encuentra en coronavirus de murciélago, RaTG13, lo que indica que el murciélago podría ser el huésped más probable del SARS-CoV-2, puesto que los murciélagos son transmisores y hospedadores de diferentes especies de coronavirus, si bien, no es posible descartar

al pangolín también como huésped intermediario, puesto que los resultados de los estudios también son significativos, e indican una similitud también elevada. Por tanto, el origen del SARS-CoV-2 aún se encuentra por dilucidar, si bien, parece que todas las investigaciones indican al murciélago.

En cuanto a la estructura y/o morfología del virus, esta es esférica, y en cuanto a su tamaño, este puede oscilar desde los 50 nm hasta los 200nm de diámetro, si bien hay estudios que recogen que el tamaño del diámetro es máximo 140 nm (12,28) . Son 4 proteínas las encargadas de darle esta estructura al virus (“Figura 2”):

- La proteína S o proteína de espiga/espícula/pico
- La proteína N o proteína de la nucleocápside
- La proteína M o proteína de membrana
- La proteína E o proteína de envoltura

La Tabla 2 recoge información más detallada sobre estas proteínas. Como se puede observar en la “Figura 2”, este virus carece de otra proteína estructural accesoria presente en el grupo A de los *Betacoronavirus*, concretamente la proteína HE, o hemaglutinina esterasa (13,29), que se encuentra formando pequeñas estructuras en forma de espiga por debajo de la proteína S.

Tabla 2.- Descripción de las proteínas estructurales que conforman el virus del SARS-CoV-2.

Proteína	Estructura	Función	Bibliografía
S	• 150-200 kDa • 1250-1300 aa • Homotrímera • 3 dominios: extremo N-terminal extracelular, dominio transmembrana, y segmento C-terminal intracelular. • Organización: péptido señal + 2 subunidades: S1, responsable de unión al receptor, y S2, responsable de la fusión de membrana	• Interacción con célula huésped • Tras interacción, induce una reorganización estructural en su estructura y permite que el virus se fusione con la membrana celular.	(30,31)
N	• 43-50 kDA • 349-740 aa • 5 dominios: dominio N-terminal (NTD) intrínsecamente desordenado, dominio de unión a RNA (RBD), un enlazador central desordenado (LINK), dominio de dimerización, y dominio C-terminal desordenado previsto(CTD)	• Unión, ensamblaje y encapsulación del genoma de RNA viral. • Inhibidor viral de ARNi (interferencia de ARN) • Establecen un complejo de replicación-transcripción viral (RTC) y remodelan la membrana celular junto con nsp2-16	(22,31–33)
M	• 25-35 kDa • 218-263 aa • Dimero • 3 segmentos estructurales: 3 hélices transmembrana N-terminales, región visagra, y dominio C-terminal de lamina β en forma de sándwich orientado al interior.	• Impulsor del ensamblaje del virus y de la gemación de la membrana	(22,31,34)
E	• 74-109 aa • 8.4-109 kDa • 3 dominios: dominio N-terminal hidrófilo corto (NTD), dominio transmembrana hidrófobo, y dominio C-terminal hidrófilo largo (CTD) • En forma monomérica o pentamérica formando un haz de 5 α-hélices transmembrana de ~35 Å de longitud	• Ensamblaje y gemación viral • Activa el inflamasoma del hospedador, es decir, activa la respuesta inmune. • Afecta a la viabilidad de la célula hospedadora. • Actividad viroporina	(19,22,35,36)

En cuanto a su relación o taxonomía con respecto a otros virus de su misma familia, se encuentra en la subfamilia *Orthocoronavirinae*, perteneciendo al género *Betacoronavirus*, compuesto por 11 virus más. Dentro de este género, pertenece al subgénero *Sarbecovirus*, compuesto únicamente por el SARS-CoV-2 y el SARS-CoV (31,37).

5.2.2.- Organización genómica del SARS-CoV-2.

El SARS-CoV-2 presenta un tamaño de genoma cercano a las 30 kb, concretamente 29903 kb, actual tamaño del genoma de referencia (35,37–39), con 14 ORFs que codifican para 29 proteínas distintas según algunos estudios y la información obtenida manualmente de GISAID (35,40). Dentro de estos ORFs se encuentran 2 ORFs, el ORF1a y el ORF1ab, que abarcan dos tercios de la longitud del genoma y que codifican para las proteínas no estructurales (NSP1-NSP10, NSP12-NSP16), encargadas de procesos como la replicación y transcripción del virus, y el tercio del genoma restante se encuentra compuesto por 4 ORFs pertenecientes a las proteínas estructurales encargadas de la morfología y de la unión del virus a las células huésped (proteínas N,S,M y E), más el restos de ORFs restantes que se corresponden con proteínas accesorias relacionadas con funciones como la evasión o modificación del comportamiento del sistema inmune del hospedador (15,38,41). La organización del genoma del SARS-CoV-2 se puede observar en la “Figura 3”.

5.2.3.- Ciclo de vida del SARS-CoV-2.

El ciclo del SARS-CoV-2 comienza cuando este se une a la célula hospedadora, mediante la proteína de espiga (S), la cual, en su subunidad S1 contiene su RBD por el cual se une de manera específica a la enzima presente en la membrana plasmática de las células del epitelio respiratorio y otros tejidos (42,43). Tras esta unión, la proteína S se activa mediante la acción de proteasas de la célula huésped, principalmente la TMPRSS2 y la catepsina L, de tal forma que, si esta activación de la proteína S es llevada a cabo por la primera proteasa mencionada, la membrana del virus se fusiona directamente con la membrana plasmática de la célula huésped, mientras que en el caso de la segunda proteasa mencionada, esta activación ocurre en los endolisosomas de la misma manera, fusionándose la membrana del virus con la del lisosoma (42,44).

Tras la fusión de las membranas, se libera el RNA genómico al citoplasma celular, actuando como un mRNA celular, puesto que su caperuza en el extremo 5', y su cola poliadenilada en el extremo 3', hace que los ribosomas celulares lo reconozcan como un mRNA más (42,43). Seguidamente, comienza la traducción de los dos grandes ORFs principales, el ORF1a y el ORF1ab, dando lugar a las poliproteínas pp1a y pp1ab respectivamente, mediante un cambio en el marco de lectura ribosomal o "*ribosomal frameshifting*". Estas poliproteínas son posteriormente procesadas por las proteasas virales PLpro (*Papain-Like Protease, nsp3)* y 3CLpro (*Main/3-Chymotrypsion-Like Protease, nsp5*), provocando la liberación de las 16 proteínas no estructurales (nsp1-nsp10 y nsp12-nsp16), encargadas de formar el complejo de replicación-transcripción (RTC) (38,42).

Nsp3, nsp4 y nsp6 generan la remodelación de las membranas del retículo endoplasmático para la formación de vesículas, las cuales son utilizadas como orgánulos de replicación viral (45). El RTC, compuesto principalmente por la ARN polimerasa dependiente de ARN (RdRp), junto con otros cofactores como nsp7 y nsp8, y otros proteínas, como nsp13, nsp14, nsp10 y nsp16, se encargan de la síntesis de una cadena de ARNc de sentido negativo que se utilizará como molde tanto para la replicación de nuevos genomas de ARN de sentido positivo, como para la síntesis de ARN subgenómicos mediante mecanismos de transcripción discontinua, en los cuales la polimerasa salta entre secuencias TRS (38,42). Estos ARN subgenómicos se exportan al citoplasma celular donde son traducidos a proteínas estructurales (S, M, E, N) y accesorias, mientras que los nuevos ARN de sentido positivo generados, son encapsulados por la proteína N para formar la nucleocápside, o para servir como molde en nuevas rondas de replicación (42,45).

Los productos de la replicación y traducción, es decir, los ARN subgenómicos traducidos a proteínas estructurales, y los ARN de cadena positiva, son utilizados para el ensamblaje nuevos viriones de SARS-CoV-2, teniendo este proceso lugar principalmente en el ERGIC, donde las proteínas S, M y E, sintetizadas y procesadas en el RE, se insertan en la membrana del mismo (38,43,45).

El ARN de sentido positivo sintetizado, se asocia con la proteína N para la formación de complejos ribonucleoproteicos virales (vRNP) en el citosol, los cuales, adquieren una conformación *"beads-on-a-string"*, permitiendo empaquetar de manera eficiente el

extenso genoma del virus (45). La interacción de la proteína N y la M es un paso fundamental para el transporte de la nucleocápside al ERGIC y para la curvatura de la membrana que da lugar a la gemación de los viriones, proceso en el que también es fundamental la presencia de la proteína E que juega un papel importante en la morfología viral, la modulación de la curvatura y la permeabilidad de la membrana. La proteína M y E también son responsables de que retener y/o reclutar la proteína S en ERGIC para su incorporación a la envoltura de los nuevos viriones (38,45).

Por tanto, para que el ensamblaje de los viriones sea eficiente, se requiere que las proteínas M, N, S y E se expresen simultáneamente, de manera qué tras el ensamblaje de los nuevos viriones, estos avanzan hacia el lumen del ERGIC, donde adquieren su envoltura lipídica, y las glicoproteínas y proteínas de membrana virales mencionadas, necesarias para poder infectar otras células hospedadoras (43).

Por último, una vez formado completamente los viriones, estos son liberados. Hay estudios que señalan que su vía de liberación es la vía secretora clásica utilizada por otros virus, en la que pequeñas vesículas se fusionan con la membrana celular y liberan un solo virión al medio extracelular. Sin embargo, otros estudios más recientes, señalan el uso de una vía de exocítica lisosomal de liberación, mediante la formación de túneles que conectan vesículas con múltiples viriones con la membrana celular. Este tipo de liberación de los viriones, les permite evitar su destrucción celular y facilita la diseminación a nuevas células hospedadoras (43,45).

5.3.- Papel de la Bioinformática en la investigación, conocimiento, y desarrollo de soluciones frente al SARS-CoV-2.

Según el *National Human Genome Research Institute*, la bioinformática “es una subdisciplina científica que implica el uso de ciencias informáticas para recopilar, almacenar y analizar y diseminar datos e información biológicos, como secuencias de ADN y aminoácidos o anotaciones sobre esas secuencias” (46). Es por tanto un campo de las ciencias computacionales que se encarga de realizar análisis de secuencias de moléculas biológicas, permitiendo compararlas, establecer relaciones evolutivas y determinar la función de dichas secuencias. Esta disciplina de investigación abarca diferentes subáreas en función del campo de aplicación, recogiendo así múltiples

subdisciplinas: genómica, transcriptómica, metabolómica, filogenómica, proteómica, … englobando todas generalmente bajo el termino de ciencias ómicas.

La bioinformática y las ciencias ómicas han jugado un papel realmente importante en la pandemia del COVID-19, debido a que gracias a ellas, se han podido generar gran cantidad de datos de todo tipo (genomas, transcriptomas, proteomas, …) que han permitido llevar a cabo diferentes logros como la secuenciación y anotación del genoma del SARS-CoV-2, la secuencia de sus genes y proteínas, su relación con otros virus, …

Es por ello, que durante este apartado, se hará mención a distintas ciencias ómicas y los hitos que han conseguido en esta reciente pandemia global sufrida recientemente, poniendo de manifiesto su potencial, así como las perspectivas futuras de estas ciencias con respecto a la enfermedad del COVID-19 y su agente causante, el virus SARS-CoV-2.

La genómica ha jugado un papel crucial en el rápido control y conocimiento del SARS-CoV-2, llevando a cabo la secuenciación del genoma del SARS-CoV-2 a las pocas semanas de la detección de los primeros casos en la provincia china de Wuhan, permitiendo así el desarrollo de pruebas de diagnóstico, y dando comienzo a la vigilancia genómica a nivel mundial. El uso de manera específica de herramientas genómicas como el alineador BLAST (47), pipelines o softwares para anotación de genomas como VADR (48) o ensambladores de genomas como Velvet (49) o coronaSpades (50), han jugado un papel realmente importante en la consecución temprana del ensamblaje del SARS-CoV-2, el conocimiento de sus secuencias y estructuras y el posterior control de la aparición de variantes que pudieran causar alteraciones que influyeran en la patogenicidad y transmisibilidad del virus (51). La creación de esta gran cantidad de datos genómicos ha permitido la creación de bases de datos como GISAID (52), que recogen un gran número de secuencias de acceso libre y abierto, lo que ha permitido un seguimiento más actualizado y controlado a nivel global, tanto de la dispersión de este, como de su evolución y el surgimiento de nuevas variantes (53,54).

La vigilancia genómica llevada a cabo durante la pandemia, se ha extendido a otros entornos diferentes del entorno clínico, como por ejemplo en el entorno medioambiental, controlando diferentes medios, como las aguas residuales, donde se han desarrollado herramientas para la detección del virus y de variantes emergentes, siendo especialmente útil para anteponerse a posibles brotes, y controlar la circulación de

variantes de forma previa a la detección de las mismas en casos clínicos (55,56). Otro papel clave que ha jugado la genómica y la filogenómica ha sido en la identificación de recombinaciones y eventos evolutivos convergentes, permitiendo detectar variantes fenotípicamente similares, pero genéticamente de origen distinto, realizando así un análisis continuo de la diversidad genética del virus para comprender mejor sus mecanismos de adaptación viral o la capacidad de evasión del sistema inmune.

El papel más relevante que ha podido jugar la genómica ha sido la identificación y seguimiento de variantes, lo cual, mediante análisis comparativos y estructural de secuencias a través de alineamientos, ensamblajes y anotaciones de genomas, ha permitido detectar variaciones en regiones de gran importancia de la secuencia del virus, afectando a procesos vitales del mismo, como por ejemplo las mutaciones en el dominio RBD de la proteína S determinadas por “Durmaz V.” y sus colaboradores(57), que observaron que afectaban a la infectividad del virus y a la respuesta inmune del hospedador frente a él, y demostró que la variante Omicron presentaba una mayor afinidad de unión por el receptor ACE2.

Otra área de la bioinformática con un papel importante ha sido la proteómica y la bioinformática estructural, llevando a cabo funciones como el modelado estructural de las proteínas virales, clave para entender mejor los mecanismos del ciclo lítico del virus. Mediante técnicas como la cristalografía de rayos X o la predicción por IA, se ha conseguido la resolución de la estructuras de las proteínas del virus, de tal forma que se pudieron utilizar estas estructuras para la simulación de interacciones entre proteínas, o proteínas-ligandos, permitiendo identificar los sitios de mutación relevante y realizar el diseño de anticuerpos antivirales y neutralizantes, como por ejemplo en el estudio (57), donde la estructura de la proteína S de Ómicron mostró como las mutaciones estabilizaban la conformación activa para un mejor reconocimiento del receptor y una mejor evasión del sistema inmune. La bioinformática estructural, a través del modelado estructural, mediante herramientas como PyMOL(58) o AlphaFold (59) han conseguido logros como el modelado de epítopos y la predicción de regiones inmunogénicas claves para el diseño de vacunas y terapias basadas en anticuerpos, o la predicción y visualización de la estructura tridimensional de las proteínas que componen el virus, permitiendo la identificación de lugares susceptibles para el desarrollo de vacunas y fármacos (51,53,60).

Otra área de gran importancia en la pandemia ha sido la bioinformática farmacológica, muy similar, y estrechamente relacionada con la proteómica y la bioinformática estructural. Ha llevado a cabo procesos importantes para lograr controlar la pandemia, mediante el diseño y reposicionamiento de fármacos, a través de la bioinformática estructural y el cribado virtual de estos fármacos, así como la identificación de epítopos inmunogénicos y potenciales sitios de unión para moléculas inhibidoras de determinados procesos virales esenciales para la transmisibilidad y/o reproducción del virus. Los estudios de docking molecular y las simulaciones de dinámica molecular han conseguido agilizar el cribado de compuestos y la obtención de fármacos candidatos a antivirales (54). Softwares de docking como AutoDock (61) o SwissDock (62) han sido utilizados ampliamente para llevar a cabo estas simulaciones de interacción o docking entre compuestos y proteínas del virus, acelerando y facilitando la selección de candidatos a fármacos para ensayos preclínicos y clínicos posteriores.

Un papel también muy importante dentro de la bioinformática, lo juegan la ciencia o minería de datos, y el desarrollo y uso de la IA para diferentes funcionalidades relacionadas con la investigación y/o la medicina personalizada. En la pandemia de COVID-19, estas áreas de acción pertenecientes a una área tecnológica, se adaptaron y utilizaron para lograr un progreso más rápido en el control de la pandemia.

El estudio de “Weronika S.” y sus colaboradores (63) pone de manifiesto el papel importante de la minería de datos para identificar biomarcadores pronósticos y factores de riesgo asociados a la gravedad o mortalidad por contraer COVID19, marcando la utilidad de biomarcadores como MR-proADM, NLR y KL-6 para la predicción de casos clínicos adversos. Esta minería de datos también ha conseguido la identificación de factores de riesgo asociados a comorbilidades, como los estudios llevados a cabo por (64,65), que permitieron observar como la expresión diferencial del receptor ACE2 en pacientes con enfermedades cardiovasculares influye en la susceptibilidad y gravedad de la infección.

La IA ha sido también muy útil para el control de la pandemia, haciendo avances en todos los campos de la bioinformática y la medicina. Por ejemplo, se han llevado a cabo análisis de bases de datos hospitalarias, facilitando así la estratificación de pacientes y predicción de complicaciones mediante *Machine Learning* (54,65), se ha utilizado *Deep Learning* para el diagnóstico y/o pronóstico de la enfermedad producida por el SARS-

CoV-2 mediante la integración y análisis de datos de imágenes médicas, genómica y registros clínicos, es decir, se han conseguido modelos que detectan patrones en radiografías y tomografías, o que predicen la evolución clínica de los pacientes a partir de datos multiómicos de los mismos. También la IA ha sido utilizada para la optimización de ensayos clínicos, la identificación de combinaciones terapéuticas y la predicción de la respuesta a distintos tratamientos virales y vacunas (51,54).

La integración de todas estas herramientas de bioinformática e IA, han permitido que se puedan desarrollar y validar biomarcadores para lograr una medicina más personalizada, más enfocada al paciente o grupos de pacientes, permitiendo mejores resultados tanto en la prevención de la enfermedad, como principalmente en el tratamiento de la misma.

Si bien se han logrado gran cantidad de avances para controlar la pandemia, y mejorar el pronóstico y tratamiento de las personas que adquieran la infección por SARS-CoV-2, aún existen grandes desafíos como la estandarización de pipelines informáticos, lo que garantizaría que todos los datos se trataran por igual, y se obtuvieran resultados similares independientemente del personal investigador y clínico, o la necesidad de interoperabilidad entre las distintas bases de datos y la revisión de sus datos, para garantizar la presencia de datos de alta calidad para poder avanzar con mayor seguridad en los distintos procesos a llevar a cabo (65).

Toda la experiencia e información adquirida durante la pandemia, ayudará tanto a la prevención de nuevas pandemias, como el conocimiento de una respuesta más rápida y coordinada ante futuras posibles pandemias, principalmente mediante la integración de la bioinformática y todas sus ramas, con la IA y la medicina personalizada, mejorando así el desarrollo de soluciones frente a patógenos y enfermedades emergentes.

6.-Objetivos

1. Llevar a cabo el ensamblaje y análisis de datos genómicos del virus SARS-CoV-2 mediante el uso de las siguientes pipelines: PipeCoV (2), nf-core/Viralrecon (3), y Vpipe (4).
2. Evaluación y comparación de los resultados obtenidos con PipeCoV frente a los obtenidos por nf-core/Viralrecon y Vpipe.
3. Elaboración de árbol filogenético para establecer la relación evolutiva y/o filogenética entre las secuencias ensambladas con PipeCoV, y otras 12 secuencias de genomas completos pertenecientes a especies de la familia *Coronaviridae*.

7.- Materiales y Métodos.

Todos los procesos llevados a cabo en este trabajo se realizaron en una máquina virtual VirtualBox 7.1.6 de Oracle (66), donde se montó un sistema operativo Ubuntu 20.04, con 5 CPUs, 9.5 Gb de RAM y 2 Gb de SWAP.

Para llevar a cabo el desarrollo del trabajo, se utilizaron dos carreras de secuenciación Illumina, SRR33190466 con lecturas de secuenciación de 250 pb y siendo secuenciada bajo estrategia WGS, y SRR33436961 con lecturas de secuenciación de 150 pb y siendo secuenciada mediante secuenciación de amplicones, utilizando los primers Artic_V4-1 (67) con un tamaño de amplicón de 400 pb. En ambos casos, las carreras son de extremos emparejados o "*paired-end*".

Ambas carreras se descargaron de la base de datos NCBI (68). La carrera SRR33190466 forma parte del proyecto con identificador de acceso PRJNA732685, el cual, fue llevado a cabo por "Tennessee DOH Lab Services", en Estados Unidos, con fecha de registro en NCBI el 25/05/2021, mientras que la carrera SRR33436961 forma parte del proyecto PRJNA870735, con fecha de registro del 18/08/2022, llevado a cabo en Puerto Rico, con el fin de determinar la supervivencia de las distintas variantes del SARS-CoV-2 en la población puertorriqueña.

Para llevar a cabo el desarrollo de la filogenia, se obtuvieron varios genomas de referencia de NCBI pertenecientes a la familia de virus *Coronaviridae* con el fin de establecer su relación evolutiva tanto con el genoma de referencia del SARS-CoV-2, como con las secuencias consensos obtenidas en nuestros ensamblajes. Los genomas descargados se encuentran en la Tabla 3.

Tabla 3.- Relación de genomas utilizados para le proceso de filogenómica.

Genomas para filogenómica.	
Human coronavirus 229E, complete genome 27,317 bp linear RNA AF304460.1 GI:12082738	Rousettus bat coronavirus HKU10 isolate 183A, complete genome 28,494 bp linear RNA JQ989270.1 GI:408796090
Human Coronavirus NL63, complete genome 27,553 bp linear RNA AY567487.2 GI:49035964	Bat coronavirus CDPHE15, complete sequence 28,035 bp linear RNA KF430219.1 GI:530341145
SARS coronavirus PC4-227, complete genome 29,728 bp linear RNA AY613950.1 GI:53854521	Human betacoronavirus 2c EMC/2012, complete genome 30,119 bp linear RNA JX869059.2 GI:409052551
Rousettus bat coronavirus HKU9, complete genome 29,114 bp linear RNA EF065513.1 GI:124389477	Duck coronavirus DK/GD/27/2014, complete genome 27,754 bp linear RNA KM454473.1 GI:817050609
Coronavirus SW1, complete genome 31,686 bp linear RNA EU111742.1 GI:159034149	MAG: Pacific salmon nidovirus isolate H14 36,652 bp linear RNA MK611985.1 GI:1726259608
Bulbul coronavirus HKU11 isolate HKU11-934, complete genome 26,487 bp linear RNA FJ376619.2 GI:212377306	Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome 29,903 bp linear RNA MN908947.3 GI:1798172431

7.1.- Ensamblaje del SARS-CoV-2.

Como se menciona en el apartado de objetivos, el ensamblaje del SARS-CoV-2 se ha llevado a cabo con 3 pipelines diferentes, con el fin de obtener una secuencia consenso como resultado de ensamblaje, y evaluar cuál de las tres podría ser más adecuada para llevar a cabo un análisis completo de datos de secuenciación, tanto del SARS-CoV-2 como de organismos.
Las 3 pipelines que se han utilizado son: PipeCoV (2), nf-core/ViralRecon (3), y Vpipe (4).

La instalación de PipeCoV se llevó a cabo tal como se indica en su repositorio de github (https://github.com/alvesrco/pipecov) . Se llevó a cabo un clonado de este repositorio en el equipo, y posteriormente se ejecutaron tanto los scripts de mejora de calidad “qc_docker.sh”, como el de mapeo y ensamblaje de lecturas “assembly_docker.sh”. En el primero de ellos, se ha realizado una pequeña modificación con el fin de que el software AdapterRemoval (69) incluido e implicado en este pipeline, detectara por sí

mismo los adaptadores de las lecturas, sin necesidad de tener que facilitarle un archivo que contuviera la información sobre los adaptadores utilizados en la secuenciación, puesto que no en todos los casos se dispone de información de dichos adaptadores, lo que dificultaría su versatilidad a la hora de la limpieza de lecturas. Se eliminó la opción "--adapter-list" del código fuente, de tal forma que AdapterRemoval identifique y localice por sí mismo los adaptadores, y los elimine de las lecturas. Sin embargo, si se quisiera mantener el código fuente, y tener que darle una lista de los múltiples adaptadores utilizados para la secuenciación, aún sin conocer la información sobre dichos adaptadores, lo más útil sería ejecutar AdapterRemoval de manera aislada con la opción de "--identify-adapters" de tal forma que obtendríamos la secuencia de dichos adaptadores y podríamos crear un archivo con dichas secuencias para poder pasárselo después a la pipeline para eliminar dichos adaptadores, si bien, esto no se realizó en este trabajo.

PipeCoV se ejecutó tal y como indica su creador, con los siguientes comandos:

1. ./qc_docker.sh -i illumina -1 ~/tfm/sarscov2/SRR33436961/SRR33436961_1.fastq.gz -2 ~/tfm/sarscov2/SRR33436961/SRR33436961_2.fastq.gz -q 20 -l 50 -t 2
2. ./assembly_docker.sh -i illumina -1 ~/tfm/sarscov2/assemblies/pipecov/output_quality/SRR33436961_good.pair1.truncated -2 ~/tfm/sarscov2/assemblies/pipecov/output_quality/SRR33436961_good.pair2.truncated -r ~/tfm/sarscov2/SARS_CoV2_GenRef_NC_045512.2.fasta -k 31 -m 2 -l 100 -c 10 -g 5 -o ~/tfm/sarscov2/assemblies/pipecov/output_assembly -t 2 -s SRR33436961

En este caso se muestra únicamente la ejecución de la pipeline para la carrera SRR33436961, si bien, para la otra carrera se ejecutó de la misma manera.

Tras la observación de los resultados, que serán expuestos en apartados posteriores a este, se observó una posible diferencia o "debilidad" de esta pipeline frente a las otras dos mencionadas, y es que, al contrario que lo leído en el artículo de su creador, no genera ningún archivo con extensión ".vcf", lo que indica que esta pipeline no tiene hasta el momento la posibilidad de llevar a cabo la llamada de variantes, si bien, habría que revisar si su creador se refiere en su artículo a la detección de linajes o a detección de variantes.

Al carecer de archivos ".vcf" que no arrojaran las variantes que es capaz de detectar esta pipeline, tuvimos que utilizar los 2 archivos ".bam" generados por dicha pipeline para llevar a cabo de manera manual la llamada de variantes con bcftools (70). Los dos archivos bam utilizados, son aquellos generados por samtools (71) en el mapeo de todas las lecturas frente al genoma de referencia, y aquel que recoge el mapeo de únicamente las lecturas que alinean con el genoma de referencia frente a un consenso que se genera tras alinear con minimap2 (72) el resultado del ensamble *de novo* contra el genoma de referencia.

Los comandos utilizados para llevar a cabo este proceso se encuentran recogidos en el script "get_vcf.vcf", script que se podrá encontrar en el repositorio de github del presente trabajo (https://github.com/Juanlu-bif/genomic-phylogenomic). En este caso, al tener dos secuencias de referencia (la secuencia de referencia original, y una secuencia consenso generada tras el ensamblaje *de novo* de las lecturas y su alineamiento con el genoma de referencia original), había que comparar y alinear ambas secuencias de referencia con minimap2 (72) para que las posiciones de las variantes en ambas secuencias no fueran distorsionadas o alteradas con respecto a su posición en el genoma de referencia original. En principio se iba a aceptar solo las variantes comunes a ambas secuencias como variantes o SNP que fueran confiables, pero ante la ausencia de dichas variantes, se decidió tomar en cuenta tanto las comunes, como las no comunes a ambas secuencias consenso, y filtrar dichas variantes en función de su DP (Profundidad de secuenciación) y su QUAL (Calidad), para lo cual utilizamos bcftools call y bcftools filter.

Tras ejecutar PipeCoV, se procedió a ejecutar las siguientes 2 pipelines, primero ViralRecon, y posteriormente Vpipe.

ViralRecon presenta dos posibles métodos para lanzar su ejecución en función de si el protocolo de secuenciación ha sido mediante amplicones, o mediante secuenciación de genoma completo o metagenoma.

En el caso de la carrera SRR33436961 se secuenció mediante amplicones, por lo que ViralRecon se tuvo que lanzar bajo un protocolo de amplicón, y para ello era necesario descargar el archivo ".bed" que contenía información sobre los primers utilizados para secuenciar dichos amplicones. Esta secuenciación se llevó a cabo utilizando el protocolo

ARTIC V_4.1, por lo que descargamos el archivo ".bed" con la información sobre dichos primers del repositorio de github de ARTIC (67). En lo referente a la otra carrera, la SRR3343190466, la metodología utilizada era la secuenciación de genoma completo, por lo que ViralRecon se lanzó estableciendo su parámetro "--protocol" en "metagenomic", como si fuera para metagenómica.

Lanzamiento de ViralRecon:

1. python3 fastq_dir_to_samplesheet.py -r1 SRR33436961_1.fastq.gz -r2 SRR33436961_2.fastq.gz ~/tfm/sarscov2/SRR33436961/ samplesheet.csv

2. nextflow run nf-core/viralrecon --input samplesheet.csv --outdir viralrecon_output --genome 'MN908947.3' -profile docker --platform illumina --protocol amplicon --primer_bed SARS-CoV-2-fit.primer.bed

Para la ejecución de Vpipe, primero se instaló dicha pipeline con su script de instalación rápida, pero bajo un entorno conda específico donde se instalaron todos sus requerimientos en lo referente a softwares secundarios necesarios, y tras su instalación, se organizaron las muestras o carreras en un subconjunto de directorios tal y como se indica en su repositorio de github (https://github.com/cbg-ethz/V-pipe), puesto que esta estructura era importante para que se llevara a cabo una correcta ejecución.

Tras organizar todo correctamente, se procedió a su lanzamiento siguiendo los siguientes 3 pasos:

1. Configuración del archivo “config.yml”:

```
general:
    virus_base_config: 'sars-cov-2'
    # e.g: 'hiv', 'sars-cov-2', or absent
    snv_caller: 'shorah'

input:
    samples_file: samples.tsv

output:
    datadir: output_vpipe

    trim_primers: true
    # see: config/README.md#amplicon-protocols
    snv: true
    local: true
    global: false
    visualization: false
    diversity: false
    QA: false
    upload: false
    dehumanized_raw_reads: false
```

2. Generación y preparación de la “samples.tsv” que recoge la información sobre las distintas carreras o muestras a secuenciar o analizar:

```
./vpipe --cores 2 –dryrun
```

3. Por último, ejecución de Vpipe:

```
./vpipe -p --cores 2 --conda-frontend conda
```

Tras finalizar todo el procedimiento de ejecución de las respectivas pipelines, se procedió a la comparación de los resultados. Para ello, para valorar los resultados de los análisis y los ensamblajes, se procedieron a comparar las siguientes características:

- Número de bases indeterminadas en las secuencias consenso obtenidas tras los ensamblajes (número de Ns)
- Longitud de la secuencia consenso obtenida.
- Número de variantes/SNP detectados
- Ejecución de QUAST (73) de todos los consensos frente al genoma de referencia.

Para llevar a cabo todas estas comparaciones, se utilizaron scripts desarrollados en Python. Los scripts se mencionan a continuación y sus resultados se expondrán posteriormente en el apartado 8, correspondiente a los resultados y discusión sobre los mismos:

- count_n_bases.py – script para determinar el número de bases indeterminadas de cada ensamblaje en formato fasta. Reporta un gráfico en el que muestra de forma comparativa el número de bases no determinadas de las secuencias consensos obtenidas por cada una de las pipelines.
- genome_comparison_long.py – script que permite determinar la longitud de los consensos. Reporta un gráfico comparativo de las distintas longitudes con respecto al genoma de referencia y también un mapa de calor que muestra el porcentaje del genoma de referencia cubierto por cada consenso.
- vcf_analysis.py – script que muestra un gráfico de barras, donde muestra las variantes detectadas por cada pipeline y por carrera de secuenciación (SRR).

Si bien es cierto, que el resultado de todos estos análisis, a excepción del determinar el número de variantes, se pueden realizar con QUAST (73), se ha preferido elaborar estos scripts para obtener una comparación más visual de los mismo. Igualmente, el reporte de QUAST (73) se encontrará en el repositorio de github para que pueda ser revisado más detenidamente.

7.2.- Filogenómica del SARS-CoV-2

Con el fin de determinar la cercanía evolutiva de las secuencias consensos obtenidas con PipeCoV al genoma de referencia, así como para determinar su cercanía a otros virus de la familia *Coronaviridae* (los recogidos en la "Tabla 3" del apartado 7), se ha llevado a cabo un protocolo de filogenómica, utilizando el método de máxima verosimilitud o *Maximum Likelihood* para llevar a cabo la inferencia filogenética. Con respecto a la elección de este método atendiendo a su precisión y/o fiabilidad, ésta ha sido aleatoria, puesto que hay multitud de publicaciones defendiendo todos los métodos de inferencia filogenética existentes y marcando los numerosos artefactos o errores que pueden introducir estos mismos métodos a la hora de realizar la inferencia. Es por ello, que para la elección del método se ha optado por tener en cuenta el tipo de datos del que partimos y la cantidad de datos de los que se disponen respecto a la capacidad computacional de la máquina virtual utilizada. Como al final, los datos no son numerosos, 14 genomas, y todos pertenecientes a una misma familia de virus, se optó por este método con el fin de acelerar el proceso de elaboración del árbol, puesto que métodos basados en inferencia bayesiana requieren un mayor tiempo de ejecución y elaboración. Además, utilicemos el método que utilicemos, siempre habrá errores introducidos, generándose artefactos en la filogenia, por lo que el árbol que se elabore no puede ser tomado como definitivo. Muchos de estos errores, así como muchos métodos a utilizar para llevar a cabo un mejor protocolo de filogenómica o biología evolutiva, y obtención de mejores resultados, son recogidos por "Lozano-Fernandez J." en su artículo (74).

Tras la elección del método se pasó a realizar los pasos necesarios para la elaboración del árbol filogenético. Para ello, se concatenaron todos los genomas en formato fasta, en un único fasta, junto con el genoma de referencia del SARS-CoV2 y las dos secuencias consensos.

Posteriormente, utilizamos mafft (75) para llevar a cabo un alineamiento múltiple de todos los genomas, y trimAl (76) para limpiar aquellas bases del alineamiento que sean todas gaps, y utilizando un modelo automatizado implementado en trimal (-automated1) para mejorar el alineamiento de cara a realizar la inferencia filogenética utilizando el método seleccionado (Máxima Verosimilitud).

Tras la limpieza del alineamiento, pasamos a ejecutar iqtree (77) para seleccionar el modelo de máxima verosimilitud al que se ajusta mejor el alineamiento mediante ModelFinder (78) , y posteriormente elaborar el árbol filogenético, que relaciona todas las secuencias.

Para visualizar el árbol, utilizamos la herramienta web iTOL (*Interactivate Tree Of Life*) (79). Adicionalmente, se llevó a cabo la representación de la matriz de distancias generada por iqtree (77) tras la elaboración del árbol filogenético, con el fin de visualizar de manera más intuitiva y visual la distancia evolutiva entre pares de secuencias (o genomas en este caso), y poder identificar mejor la relación y cercanía evolutiva entre los genomas utilizados.

8.- Resultados.

Tras finalizar las tareas descritas en el apartado 7 de este trabajo, se procedió a su interpretación. Hay que decir, que si bien, los resultados que arrojan son significativos, no se pueden calificar como concluyentes, principalmente de cara al ensamblaje de genomas, puesto que el tamaño muestral es demasiado pequeño debido a falta principalmente de recursos computacionales, motivo principal por el cual no se ha aumentado el tamaño muestral.

8.1.- Ensamblaje de genomas y análisis de variantes.

Para determinar el resultado del ensamblaje de genomas y su análisis de variantes, nos fijamos principalmente en tres características. De cara al ensamblaje, nos centramos en la secuencia del genoma consenso obtenida tras la ejecución de cada pipeline, y evaluamos el número de bases indeterminadas que presentaba (N), y la longitud de la secuencia consenso. Con respecto al análisis de variantes, lo que se evaluó fue el número de variantes detectadas por cada pipeline, siempre tras realizar un filtrado de calidad de las mismas. Adicionalmente, se podría también incluir la determinación del taxón al que pertenece la secuencia consenso con Pangolin (80), pero en este caso, si bien sí que se llevó a cabo este proceso determinando en todos los casos, que ambas secuencias consensos pertenecían al linaje BA.2, concretamente a la variante Omicron (BA.2-like), no se tomó como un resultado a evaluar, debido a lo comentado con anterioridad, que el tamaño muestral era reducido y por tanto su análisis no fue considerado como determinante para la determinación de los resultados.

La evaluación realizada es la de la pipeline PipeCoV frente al resto, siguiendo pasos similares a los desarrollados por Oliveira y sus colaboradores (2) en su trabajo, comparando PipeCoV con el resto de pipelines.

Atendiendo al número de bases indeterminadas, se vio que PipeCoV es capaz de generar una secuencia consenso con un menor número de bases indeterminadas, hecho importante, puesto que podría significar una menor perdida de información sobre la secuencia. Cualquiera de las otras dos pipelines obtuvieron un mayor número de

bases indeterminadas en sus consensos, alcanzando en algunos casos entre el 1.60% y 1.65% de bases indeterminadas del total de bases de la secuencia, mientras que PipeCoV se mantiene su porcentaje más elevado en tan solo un 0.61% del total de bases del consenso. Los resultados obtenidos se muestran en la “Figura 4” y “Figura 5”. Por otro lado, se observó la longitud de las secuencias consenso y se compararon con la longitud del genoma de referencia. Se observó que las tres pipelines obtenían buenos resultados en cuanto a longitud, presentando una longitud muy similar a la del genoma de referencia, teniendo la longitud más baja PipeCoV, hecho que puede ser normal debido a la combinación de enfoques de ensamblaje *de novo* y con referencia, frente al enfoque únicamente de mapeo frente a genoma de referencia que realizan las otras dos pipelines, un método más seguro de cara a definir mejor la estructura concreta de los genes encontrados y del genoma, y no la estructura general que suele ser definida por el ensamblaje *de novo*, tal como explica brevemente en la introducción (2). Igualmente, en lo referente a la longitud del consenso obtenido, en ambos casos, PipeCoV, siendo el que genera un consenso de menor longitud, presenta una longitud equivalente al 99 % de la longitud del genoma de referencia, lo que quiere decir que tan solo presenta una diferencia de entre 200-300 pares de bases aproximadamente. Con respecto a las otras dos pipelines, su consenso presenta prácticamente la misma longitud de genoma que el genoma de referencia alcanzando prácticamente un 100% de longitud con respecto a longitud de la referencia. Los resultados se recogen en la “Figura 6”.

```
==================================
RESUMEN POR PIPELINE:
==================================
PIPECOV: 181 bases N totales
VIRALRECON: 821 bases N totales
VPIPE: 875 bases N totales

DETALLE POR MUESTRA:
----------------------------------

SRR33190466:
  pipecov: 1 N (0.00%)
  viralrecon: 331 N (1.11%)
  vpipe: 394 N (1.32%)

SRR33436961:
  pipecov: 180 N (0.61%)
  viralrecon: 490 N (1.64%)
  vpipe: 481 N (1.61%)
```

Figura 4.- Resumen de la ejecución del script “count_n_bases.py”

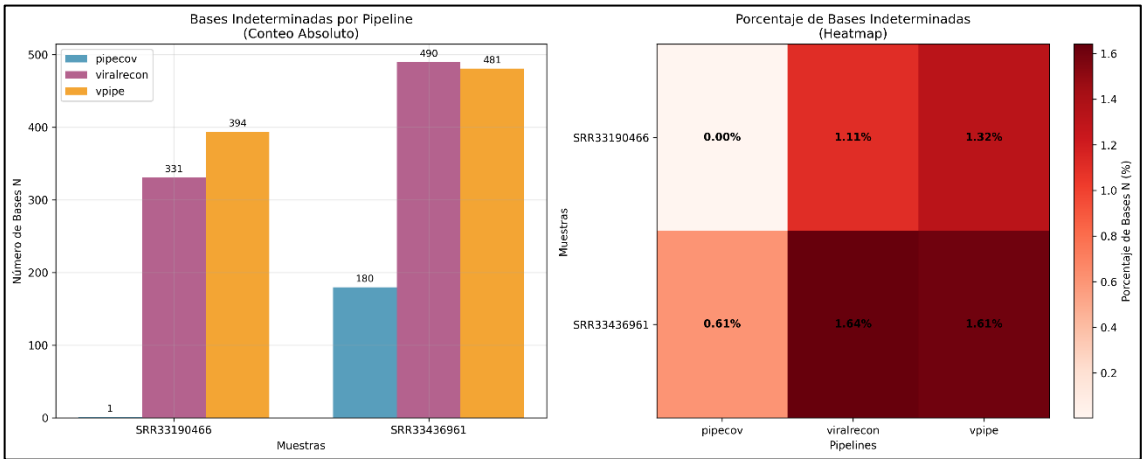

Figura 5.- A la izquierda, gráfico de barras agrupadas que muestra el número de bases indeterminadas de cada uno de los consensos obtenidos para cada muestra por las diferentes pipelines (PipeCoV, ViralRecon y V-pipe). A la derecha, mapa de calor que recoge el porcentaje de bases indeterminadas de cada secuencia consenso obtenida con respecto al total de bases de la misma. Dicho gráfico, se ha obtenido mediante la elaboración y ejecución del script “count_n_bases.py”.

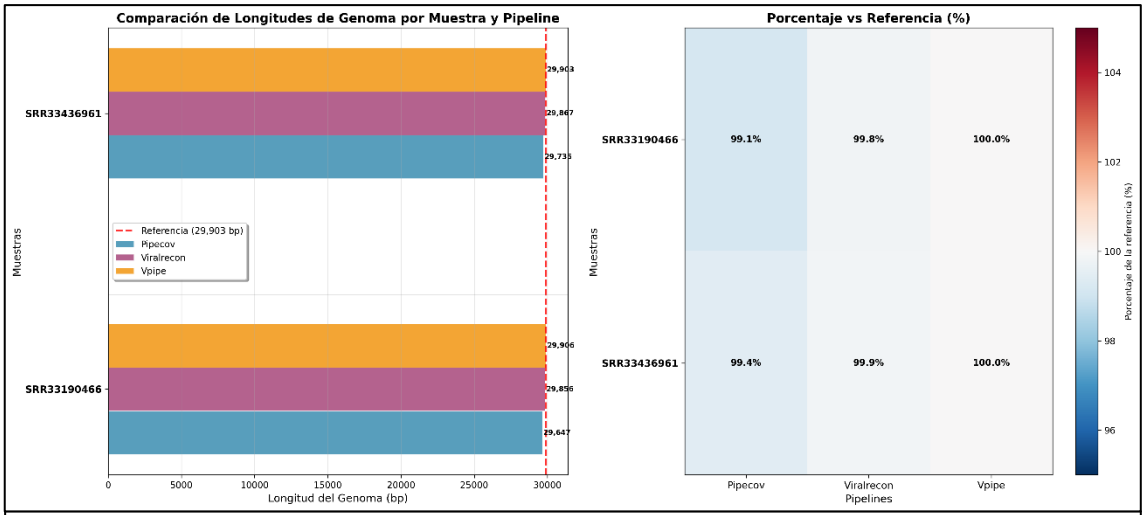

Figura 6.- A la izquierda un gráfico de barras que muestra la longitud de los consensos obtenidos por cada pipeline para cada carrera o muestra, frente a la longitud del genoma de referencia (línea roja). A la derecha, mapa que calor que indica el porcentaje de longitud de cada consenso con respecto a la longitud total de la secuencia de referencia. Ambas gráficas se han obtenido mediante la ejecución del script “genoma_comparison_long.py"

Finalmente, el último aspecto que se evaluó, fue la capacidad de las pipelines para detectar variantes. En este caso, PipeCoV es quizás la peor por un motivo, y es que esta pipeline no genera archivos que recojan dichas variantes, es decir, no realiza el proceso de llamada de variantes, por lo que lo tuvimos que hacer de forma manual/aislada con bcftools (70), tal como indicamos en el apartado 7.1 de este trabajo.

En el caso de PipeCoV, nos encontramos con que al generar dos archivos ".bam" distintos para detectar las variables, uno del alineamiento de todas las lecturas frente al genoma de referencia, y otro del alineamiento frente al consenso generado tras alinear el genoma de referencia con el genoma resultante del ensamblaje *de novo* con spades (81), se podría generar un mayor número de variantes, si bien, dichas variantes podrían ser artefactos, en especial aquellas localizadas sobre la secuencia consenso entre el ensamblaje *de novo* y el genoma de referencia, puesto que las posiciones de dicha secuencia consenso no coinciden con las posiciones de las bases de la referencia. Es por ello que en principio si iban a tomar únicamente en cuenta aquellas variantes comunes a ambos alineamientos, pero ante la observación de que no se producían variantes comunes, se decidió tomar en cuenta todas las variantes, pero con un filtrado previo de calidad. Tras este filtrado, nos encontramos que solo nos quedábamos con las variantes que pertenecían al alineamiento de todas las secuencias frente al genoma de referencia, lo que nos puede indicar que aquellas generas frente al consenso entre ensamblaje *de novo* y la referencia, se trataban de artefactos, puesto que presentaban baja calidad y baja profundidad de secuenciación.

Tras realizar los pasos descritos en el párrafo anterior, y determinar las variantes detectadas por PipeCoV, obtuvimos posteriormente las determinadas por ViralRecon y por V-pipe. Como resultado obtuvimos que PipeCoV y ViralRecon detectaban un número similar de variantes con respecto al genoma de referencia, mientras que V-pipe era más inestable en cuanto a la detección de SNPs, puesto que una de las carreras fue capaz de detectar 100 SNPs y en la otra 187 SNPs, mientras que las otras dos pipelines se mantuvieron entre 130 y 140 SNPs detectados. Estos resultados se muestran en la "Figura 8".

Ante esta inestabilidad de V-pipe en lo que a detección de SNPs se refiere, se decidió realizar un filtrado interno de todas las variantes, con un filtro de calidad (>100 para V-pipe y >20 para PipeCoV y ViralRecon) y de profundidad de secuenciación (>10 para ViralRecon y PipeCoV) por variante más estricto. Cuando realizamos dicho filtrado obtuvimos que todas las variantes detectadas por PipeCov y V-pipe desaparecieron, quedando únicamente las detectadas por ViralRecon, que mantenía los mismos números que aun sin el filtrado estricto. Dicho resultado se puede observar en la "Figura 7".

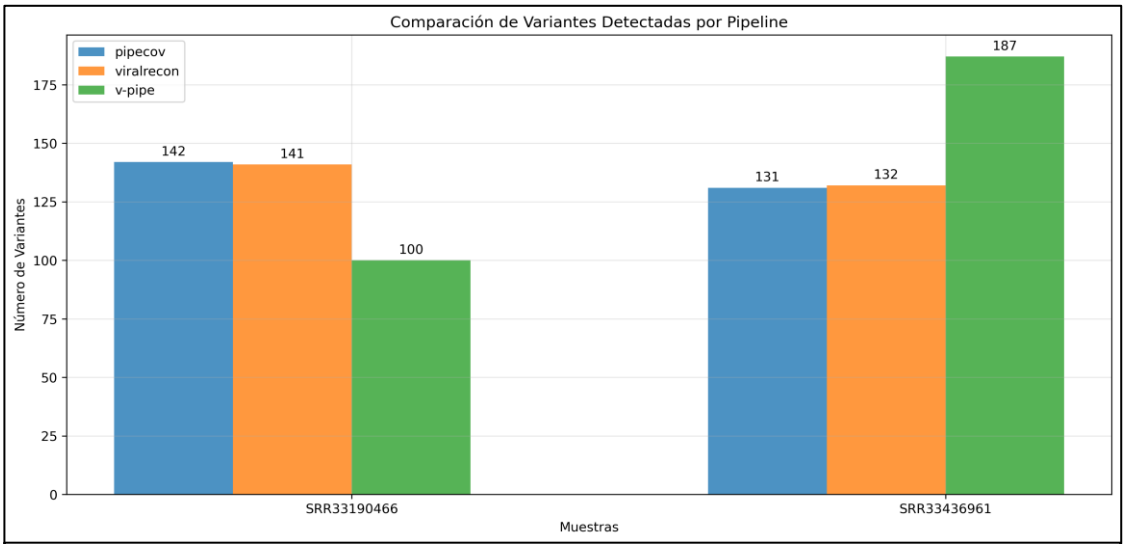

Figura 7.- Gráfico de barras que indica el número de variantes obtenidas por cada pipeline para cada carrera o muestra, de forma previa a realizar un filtrado estricto de dichas variantes. Este gráfico se ha obtenido mediante el script "vcf_analysis.py".

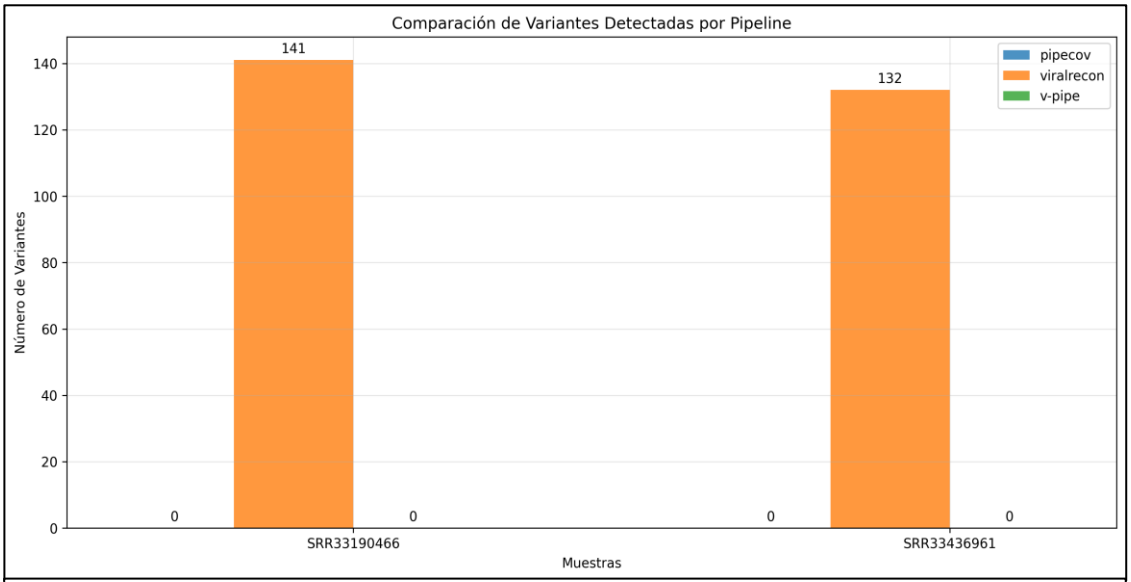

Figura 8.- Gráfico de barras que representa las variantes restantes al realizado de un filtrado estricto de SNPs mediante el script "filter_vcf.sh" y posterior representado con el script "vcf_analysis.py". Es posible observar como las variables detectadas por PipeCoV y V-pipe desaparecen por completo, mientras que las detectadas por ViralRecon se mantienen estable en número.

Adicionalmente a estas tres comprobaciones, se realizó una comparación de los distintos consensos frente al genoma de referencia con QUAST, donde se pueden ver todas estas características principales en las que hemos centrado los resultados, así como otras muchas características no incluidas en dichos resultados, y que pueden ser también de interés.

8.2.- Filogenómica del SARS-CoV-2.

Tras la elaboración del árbol filogenético (Figura 10) utilizando el método de máxima verosimilitud para realizar la inferencia genética de nuestros dos genomas consenso junto con otros 12 genomas pertenecientes también a la familiar *Coronaviridae*, se observó que como era de esperar, nuestras dos secuencias consenso (ambas pertenecientes al SARS-CoV-2) presentaban una distancia evolutiva mínima entre sí, y frente al genoma de referencia del SARS-CoV-2, lo que nos indica también que pertenecen al mismo virus. Además, dichas distancias estaban soportadas por un proceso de bootstrap del 100%, lo que indica, que en nuestro caso, como hemos indicado a iqtree que utilice un bootstrap de 1000, se ha obtenido el mismo árbol, relación y distancia evolutiva las mil veces que se ha elaborado el árbol con dichas secuencias, indicando así que dichas secuencias presentan esa distancia evolutiva con casi total seguridad. Para hacer una doble comprobación más exacta, se acudió a la matriz de distancias (Figura 9) generada tras la elaboración del árbol, donde se puede observar que las distancias en las secuencias consenso y el genoma de referencia del SARS-CoV-2 es de 0, lo que indica que son secuencias que pertenecen al mismo virus.
También se observó que estas tres secuencias (las 2 consenso y la referencia del SARS-CoV-2) forman un clado monofilético con las otras tres especies de coronavirus (EF065513.1, JX869059.2 y AY613950.1), perteneciendo todas estas secuencias al género *Betacoronavirus*, coincidiendo así con la información recogida en (ICTV) (17,18).
Hay que tener en cuenta, que todas estas especies no presentan entre sí, y con respecto a las secuencias consenso y al genoma de referencia del SARS-CoV-2, una distancia evolutiva superior a 1.21, lo que indica la proximidad de dichas secuencias, e indica la posible relación y/o descendencia del SARS-CoV-2 con uno de los coronavirus de murciélago, señalando esto a los murciélagos como uno de los huéspedes reservorio del SARS-CoV-2, por la similitud entre ambos virus.
Hay que señalar que todas las relaciones filogenéticas establecidas en este grupo monofilético están sostenidas por un bootstrap del 100%, lo que indica que estos genomas se encuentren así de próximos evolutivamente casi con total seguridad.

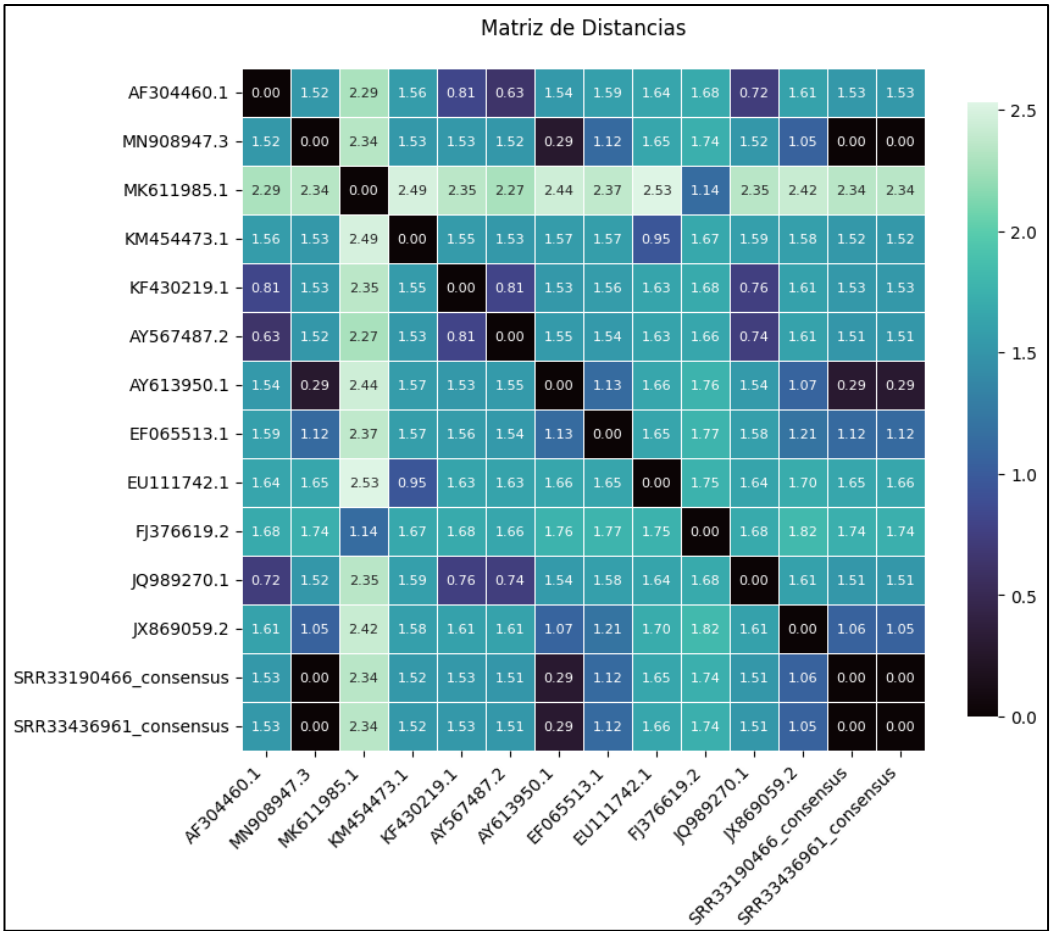

Figura 9.- Matriz de distancias que recoge la distancia evolutiva de cada especie de virus al resto de especies de virus. Los identificadores pueden ser consultados en la “Tabla 3” del apartado 7.

Observando el resto del árbol, vemos que los genomas con identificadores AY567487.2 y AF304460.1 (*Human coronavirus NL63*, y *Human Coronavirus 229E* respectivamente) conforman dos grupos externos del árbol, si bien, su distancia evolutiva al resto de genoma no es tan elevada. Esto no concuerda con los expuesto por (ICTV), y por lo que se observa en el árbol, vemos que la disposición de dichos grupos externos no se encuentra muy bien soportado por un bootstrap elevado, de hecho, ambas ramas o grupos presentan un Bootstrap por debajo del 65%, lo cual indica que se han generado distintos árboles con distintas topologías para dichos grupos, pero todos presentan una baja probabilidad de que las relaciones entre estas especies ocurran así en la realidad, mostrándose en el árbol la que mejor probabilidad tiene de que ocurra.

Si seguimos observando el árbol, vemos que los genomas JQ989270.1 y KF430219.1 constituyen clado entre ambos, si bien, dicho clado se encuentra soportado por un Bootstrap del 68 %, por lo que probablemente, dicho clado no se encuentre agrupado

y/o relacionado como se representa en el árbol. De hecho, atendiendo a lo recogido en (ICTV), que recoge toda la evidencia actual hasta el momento, este clado, debería conformar un único clado junto con los dos grupos externos, lo cual, unido a que el porcentaje de Bootstrap no es muy bueno, esta parte del árbol se podría revisar y afinar. Sin embargo, por las distancias recogidas en la matriz de distancias, es extraño que se hayan colocado como grupos externos, puesto que en ninguna de las distancias establecidas entre estas especies de virus, se supera una distancia evolutiva superior a 0.81.

Los genomas EU111742.1 y KM454473.1 constituyen un grupo o clado, el cual corresponde al género *Gammacoronavirus*, hecho también recogido en (ICTV). Lo que llama la atención, que en los resultados obtenidos el Bootstrap obtenido en el nodo donde se bifurcan *Betacoronavirus* y *Gammacoronavirus* se encuentre sostenido por un Bootstrap tan bajo, si bien, la matriz de distancia los sitúa bastante alejados con respecto a las especies del género *Betacoronavirus*.

Por último, los genomas FJ376619.2 y MK611985.1 conforman un mismo clado, si bien, según (ICTV) pertenecen a géneros distintos, perteneciendo el genoma FJ376619.2 al género *Deltacoronavirus* de la subfamilia *Orthocoronavirinae*, mientras que el genoma MK611985.1 pertenece a la familia *Alphapironavirus* (también perteneciente a la familia *Coronaviridae*).

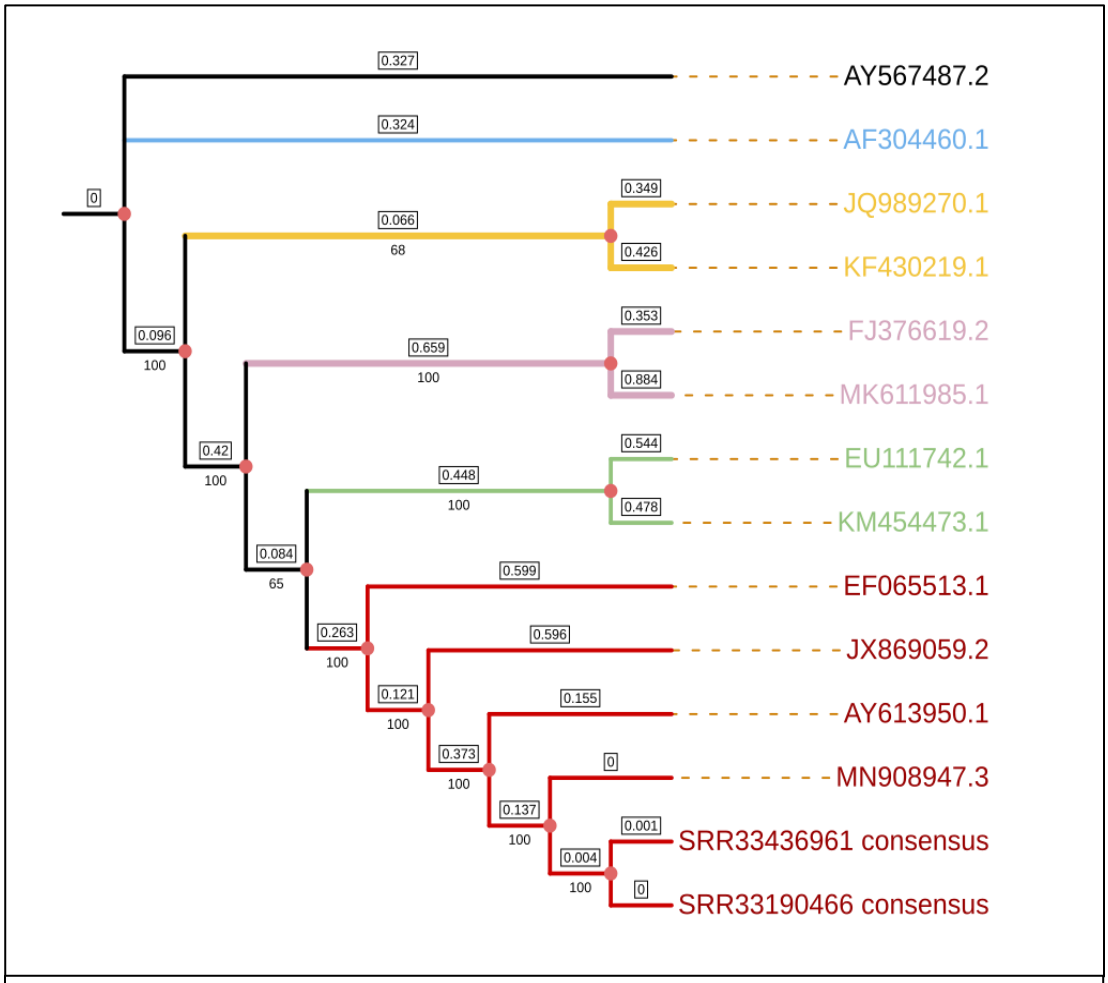

Figura 10.- Árbol filogenético obtenido tras llevar a cabo el proceso de filogenómica con las secuencias consenso, el genoma de referencia del SARS-CoV-2 y otras 11 secuencias de genomas pertenecientes a la familia *Coronaviridae*.

8.3.-Discusión.

La secuenciación de genomas, y la búsqueda y anotación de variantes, así como su relación con las distintas enfermedades o patologías, es un área en evolución constante. En el caso del SARS-CoV-2, la evolución ha sido aún más rápida, la generación de una pandemia mundial causada por este virus, ha conseguido que la bioinformática haya evolucionado de forma más rápida aún, puesto que la comunidad científica se ha volcado en el desarrollo de herramientas para el procesado de una mayor cantidad de datos y una obtención de resultados más rápida, para hacer frente a la pandemia.

Este trabajo, con el fin de llevar a cabo una comparación a pequeña de algunas de estas herramientas elaboradas, evalúa y comparar los resultados de tres pipelines diferentes, siendo al menos dos de ellas, pipelines bien documentadas, con un amplio uso en estudios bioinformáticos relacionados con el SARS-CoV-2, como en el caso del ViralRecon y V-pipe, si bien, se centra en los resultados obtenidos por PipeCoV debido a que es una pipeline más sencilla e intuitiva de ejecutar, mientras que las otras dos requieren ciertos conocimientos en nextflow (82) y/o snakemake (83,84) para ser ejecutadas. Esto hace referencia a la reproducibilidad del trabajo, o de la ejecución del software o pipeline, y es que, la reproducibilidad de estas pipelines es esencial para que otros investigadores o grupos de investigación puedan utilizarlas, así como replicar el trabajo de este estudio. Para garantizar esta reproducibilidad, es importante tener en cuenta la documentación del trabajo realizado, haciendo referencia con esto a la documentación de las distintas versiones de las pipelines y softwares utilizados, así como a ciertos parámetros de procesamiento utilizados para el procesamiento, para que así se facilite la reproducibilidad futura del trabajo. Una buena manera de garantizar la reproducibilidad del trabajo es mediante Docker (85), empaquetando todas las dependencias y pasos del pipeline en contenedores, lo que a pesar de afectar en una mínima parte al rendimiento del análisis o trabajo, va a facilitar tanto la instalación como la reproducibilidad en diferentes sistema operativos (86), siendo este parte del motivo por lo que se centra más el trabajo en PipeCoV, ya que lo convierte en una pipeline más fácil e intuitiva de ejecutar en cualquier entorno, tal como hemos mencionado anteriormente.

Un hecho también importante o a tener en cuenta en la elaboración, uso y/o valoración de herramientas genómicas para ensamblaje de genomas es la determinación de los enfoques a utilizar, y es que, en un estudio reciente (87), se muestra que pipelines con enfoques híbridos (*de novo* y por mapeo, tal como PipeCoV), presentan mejores resultados de secuencias de ensamblaje que aquellos enfoques *de novo* o con mapeo frente a referencia, siempre que las lecturas a ensamblar sean similares o cercanas a la secuencia de referencia, observando un montaje más completo de la secuencia del genoma, y reduciendo los errores en aquellas regiones más variables.

Un paso muy importante a tener en cuenta en todas estas pipelines es el control de calidad antes de pasar al ensamblaje y al llamado de variantes. Si bien, todas ellas presentan softwares para llevar a cabo este control de calidad, la mayoría arrojan información sobre un determinado paso, mientras que la integración de herramientas

tales como MultiQC (88), únicamente integrado en ViralRecon, podría ofrecer información sobre los resultados de cada etapa de la pipeline, tanto en la calidad de las lecturas de secuenciación, del ensamblaje del genoma, la llamada de variantes, … puesto que crean un informe que recoge los resultados de todos los softwares ejecutados en la pipeline. Este es un fallo que muestra PipeCoV y Vpipe, puesto que poder acceder a toda la información de ejecución, te puede ayudar a afinar determinados parámetros para determinados pasos de la pipeline, con el fin de obtener resultados más fiables.

Por último, con respecto a la inferencia genética y/o evolutiva, el uso de un modelo u otro es igualmente válido, si bien hay modelos más precisos que otros, debido a que incorporan modelos y metodologías que se ajustan mejor a como funciona la evolución en un marco real. La metodología empleada en este caso como medida de fiabilidad de árboles filogenéticos, el bootstrap, es una medida muy extendida y válida. Sin embargo, una posible mejora a realizar sería el uso del software BEAST (89) para realizar la inferencia genética mediante inferencia bayesiana, tal como hizo Suchard M. y sus colaboradores (90) en su trabajo, integrando información adicional como modelos de reloj molecular y coalescencia, intervalos de credibilidad para tasas de mutación, y tiempos de divergencia. Incluso, sería interesante poder integrar una doble comparativa, mediante máxima verosimilitud e inferencia bayesiana, de tal forma que las relaciones soportadas por ambos enfoques, fueran consideradas como válidas.

9.- Conclusiones.

En este trabajo se han comparado 3 pipelines diferentes, elaboradas para el ensamblaje de genoma, búsqueda y anotación de variantes, anotación de genomas, y determinación de linajes. Concretamente, se han comparado las pipelines PipeCoV, ViralRecon y Vpipe, centrándonos principalmente en PipeCoV, puesto que es mucho más intuitivo y sencillo de cara a su ejecución que las otras dos pipelines, y por ello se ha llevado a cabo una réplica a pequeña escala del estudio llevado a cabo por "Oliveira R." y sus colaboradores (2). Como conclusión, tras la obtención de resultados, no podemos decir que PipeCoV presente mejores o peores resultados que las otras pipelines, puesto que en cuanto a la determinación de bases, presenta mejores resultados que las otras dos pipelines (presenta un menor número de bases indeterminadas en las secuencias consenso obtenidas del ensamblaje), y en cuanto a la longitud de la secuencia consenso obtenida del ensamblaje, es ligeramente inferior a la obtenida por las otras dos pipelines. Donde sí que presenta una desventaja frente a las otras dos pipelines, es que el proceso de búsqueda y anotación de variantes no lo tiene implementado, por lo que le puede restar puntos de cara a su utilidad en bioinformática clínica.

Con respecto a la filogenia obtenida, coincide parcialmente, en cuanto al género *Betacoronavirus* se refiere, con los datos recogidos por el (ICTV), siendo este el órgano más actualizado actualmente en cuanto a la taxonomía de virus se refiere, puesto que es el encargado de aprobar y recoger todos los cambios y actualizaciones. Los resultados obtenidos no son los más satisfactorio si vemos la clasificación recogida por el (ICTV), pero debido a limitaciones tanto logísticas, como la capacidad computacional y el tiempo de ejecución, como personales, como lo es la falta de experiencia en este tipo de investigaciones, sería necesario realizar una revisión más a fondo de los datos obtenidos, llevando a cabo algunos de los pasos recogidos por Lozano-Fernandez J. en su artículo (74), si bien, algunos de estos, como la revisión, limpieza, y mejora de los alineamientos, sí que ha sido llevada a cabo.

En resumen:

- La pipeline PipeCoV crea un consenso de mayor calidad al presentar menos bases indeterminadas, mientras que ViralRecon ha demostrado ser más fiable y robusta en lo que a la detección de variantes se refiere.

- PipeCoV presenta una desventaja frente al resto de pipelines, y es que no lleva integrada el proceso de llamada y anotación de variantes, siendo esto un punto a mejorar en dicha pipeline.
- La filogenia obtenida indica un posible origen en el murciélago, pero no se puede afirmar de forma definitiva, debido a que no se ha llevado a cabo con todas las especies pertenecientes a la familia *Coronaviridae*, por lo que son necesarios más datos y genomas.

Todo el proceso llevado a cabo en este trabajo, así como todos los archivos, script e información utilizados, y los resultados obtenidos, se encuentran recogidos en el repositorio de github:
https://github.com/Juanlu-bif/genomic-phylogenomic

10- Bibliografía.

1. COVID-19 deaths | WHO COVID-19 dashboard [Internet]. [cited 2025 Jun 18]. Available from: https://data.who.int/dashboards/covid19/deaths?n=c

2. Oliveira RRM, Negri TC, Nunes G, Medeiros I, Araújo G, de Oliveira Silva F, et al. PipeCoV: a pipeline for SARS-CoV-2 genome assembly, annotation and variant identification. PeerJ [Internet]. 2022 Apr 13 [cited 2025 Jul 3];10:e13300. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9013232/

3. Patel H, Monzón S, Varona S, Espinosa-Carrasco J, Garcia MU, bot nf core, et al. nf-core/viralrecon: nf-core/viralrecon v2.6.0 - Rhodium Raccoon. [cited 2025 Jul 3]; Available from: https://zenodo.org/records/7764938

4. Wagner DD, Marine RL, Ramos E, Ng TFF, Castro CJ, Okomo-Adhiambo M, et al. VPipe: an Automated Bioinformatics Platform for Assembly and Management of Viral Next-Generation Sequencing Data. Microbiol Spectr [Internet]. 2022 Apr 27 [cited 2025 Jul 3];10(2). Available from: https://pubmed.ncbi.nlm.nih.gov/35234489/

5. Fauci AS, Lane HC, Redfield RR. Covid-19 — Navigating the Uncharted. N Engl J Med [Internet]. 2020 Mar 26 [cited 2025 Jun 18];382(13):1268. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7121221/

6. De Wit E, Van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol [Internet]. 2016 Aug 1 [cited 2025 Jun 18];14(8):523. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7097822/

7. Zhong NS, Zheng BJ, Li YM, Poon LLM, Xie ZH, Chan KH, et al. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. Lancet [Internet]. 2003 Oct 25 [cited 2025 Jun 18];362(9393):1353. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7112415/

8. Garcia-Alamino JM. Epidemiological aspects, clinic and control mechanisms of SARS-CoV-2 pandemic: Situation in Spain. Enferm Clin. 2021 Feb 1;31:S4–11.

9. Dabanch J. EMERGENCIA DE SARS-COV-2. ASPECTOS BÁSICOS SOBRE SU ORIGEN, EPIDEMIOLOGÍA, ESTRUCTURA Y PATOGENIA PARA CLÍNICOS. Revista Médica Clínica Las Condes [Internet]. 2020 Jan 1 [cited 2025 Jun 19];32(1):14–9. Available from: https://www.elsevier.es/es-revista-revista-medica-clinica-las-condes-202-articulo-emergencia-de-sars-cov-2-aspectos-basicos-S0716864020300924

10. Hudson CB, Beaudette FR. Infection of the cloaca with the virus of infectious bronchitis. Science (1979) [Internet]. 1932 [cited 2025 Jun 26];76(1958):34. Available from: https://pubmed.ncbi.nlm.nih.gov/17732084/

11. Payne S. Family Coronaviridae. Viruses [Internet]. 2017 [cited 2025 Jun 25];149. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7149805/

12. Hidalgo P, Valdés M, González RA. Molecular biology of coronaviruses: an overview of virus-host interactions and pathogenesis. Bol Med Hosp Infant Mex [Internet]. 2021 [cited 2025 Jun 25];78(1):41–58. Available from: www.bmhim.com

13. Coronaviridae. Virus Taxonomy [Internet]. 2012 Jan 1 [cited 2025 Jun 26];806–28. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780123846846000689

14. Hartenian E, Nandakumar D, Lari A, Ly M, Tucker JM, Glaunsinger BA. The molecular virology of coronaviruses. J Biol Chem [Internet]. 2020 Sep 11 [cited 2025 Jun 26];295(37):12910. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7489918/

15. Das A, Ahmed R, Akhtar S, Begum K, Banu S. An overview of basic molecular biology of SARS-CoV-2 and current COVID-19 prevention strategies. Gene Rep [Internet]. 2021 Jun 1 [cited 2025 Jun 26];23:101122. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8012276/

16. Kasmi Y, Khataby K, Souiri A, Ennaji MM. Coronaviridae: 100,000 Years of Emergence and Reemergence. Emerging and Reemerging Viral Pathogens [Internet]. 2019 Jan 1 [cited 2025 Jun 26];127. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7149750/

17. Home | ICTV [Internet]. [cited 2025 Jun 25]. Available from: https://ictv.global/

18. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res. 2018 Jan 1;46(D1):D708–17.

19. Kakavandi S, Zare I, VaezJalali M, Dadashi M, Azarian M, Akbari A, et al. Structural and non-structural proteins in SARS-CoV-2: potential aspects to COVID-19 treatment or prevention of progression of related diseases. Cell Commun Signal [Internet]. 2023 Dec 1 [cited 2025 Jul 3];21(1):110. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10183699/

20. Colina SE, Serena MS, Echeverría MG, Metz GE. Clinical and molecular aspects of veterinary coronaviruses. Virus Res [Internet]. 2021 May 1 [cited 2025 Jul 3];297:198382. Available from: https://www.sciencedirect.com/science/article/pii/S0168170221000897

21. Fang P, Fang L, Zhang H, Xia S, Xiao S. Functions of Coronavirus Accessory Proteins: Overview of the State of the Art. Viruses 2021, Vol 13, Page 1139 [Internet]. 2021 Jun 13 [cited 2025 Jul 3];13(6):1139. Available from: https://www.mdpi.com/1999-4915/13/6/1139/htm

22. Darvishi M, Hajilou F. Clinics in Nursing Review on Virology of Coronaviridae. Clinics in Nursing. 2024;3(3).

23. Zmasek CM, Lefkowitz EJ, Niewiadomska A, Scheuermann RH. Genomic evolution of the Coronaviridae family. Virology [Internet]. 2022 May 1 [cited 2024 Jul 30];570:123. Available from: /pmc/articles/PMC8965632/

24. Taxon Details | ICTV [Internet]. [cited 2025 Jun 26]. Available from: https://ictv.global/taxonomy/taxondetails?taxnode_id=202113904&taxon_name=Pitovirinae#release_37

25. Zhang C, Zheng W, Huang X, Bell EW, Zhou X, Zhang Y. Protein structure and sequence re-analysis of 2019-nCoV genome refutes snakes as its intermediate host or the unique similarity between its spike protein insertions and HIV-1. J Proteome Res [Internet]. 2020 Apr 3 [cited 2025 Jun 27];19(4):1351. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7099673/

26. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou JJ, et al. Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. bioRxiv [Internet]. 2020 Feb 20 [cited 2025 Jun 27];2020.02.17.951335. Available from: https://www.biorxiv.org/content/10.1101/2020.02.17.951335v1

27. Zhou P, Yang X Lou, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature [Internet]. 2020 Mar 12 [cited 2025 Jun 27];579(7798):270. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7095418/

28. Fernández-Pérez GC, Oñate Miranda M, Fernández-Rodríguez P, Velasco Casares M, Corral de la Calle M, Franco López, et al. SARS-CoV-2: what it is, how it acts, and how it manifests in imaging studies. Radiologia. 2021 Mar 1;63(2):115–26.

29. Made Artika I, Dewantari AK, Wiyatno A. Molecular biology of coronaviruses: current knowledge. 2020 [cited 2024 Feb 24]; Available from: https://doi.org/10.1016/j.heliyon.2020.e04743

30. Huang Y, Yang C, Xu X feng, Xu W, Liu S wen. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin [Internet]. 2020 Sep 1 [cited 2025 Jul 2];41(9):1141–9. Available from: https://www.nature.com/articles/s41401-020-0485-4

31. Asghari A, Naseri M, Safari H, Saboory E, Parsamanesh N. The Novel Insight of SARS-CoV-2 Molecular Biology and Pathogenesis and Therapeutic Options. DNA Cell Biol [Internet]. 2020 Oct 1 [cited 2025 Jun 27];39(10):1741–53. Available from: /doi/pdf/10.1089/dna.2020.5703?download=true

32. Cubuk J, Alston JJ, Incicco JJ, Singh S, Stuchell-Brereton MD, Ward MD, et al. The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. Nature Communications 2021 12:1 [Internet]. 2021 Mar 29 [cited 2025 Jul 2];12(1):1–17. Available from: https://www.nature.com/articles/s41467-021-21953-3

33. Wu W, Cheng Y, Zhou H, Sun C, Zhang S. The SARS-CoV-2 nucleocapsid protein: its role in the viral life cycle, structure and functions, and use as a potential target in the development of vaccines and diagnostics. Virology Journal 2023 20:1 [Internet]. 2023 Jan 10 [cited 2025 Jul 2];20(1):1–16. Available from: https://virologyj.biomedcentral.com/articles/10.1186/s12985-023-01968-6

34. Zhang Z, Nomura N, Muramoto Y, Ekimoto T, Uemura T, Liu K, et al. Structure of SARS-CoV-2 membrane protein essential for virus assembly. Nature Communications 2022 13:1 [Internet]. 2022 Aug 5 [cited 2025 Jul 2];13(1):1–12. Available from: https://www.nature.com/articles/s41467-022-32019-3

35. Yang H, Rao Z. Structural biology of SARS-CoV-2 and implications for therapeutic development. Nature Reviews Microbiology 2021 19:11 [Internet]. 2021 Sep 17 [cited 2025 Jun 29];19(11):685–700. Available from: https://www.nature.com/articles/s41579-021-00630-8

36. Zhou S, Lv P, Li M, Chen Z, Xin H, Reilly S, et al. SARS-CoV-2 E protein: Pathogenesis and potential therapeutic development. Biomedicine & Pharmacotherapy [Internet]. 2023 Mar 1 [cited 2025 Jul 2];159:114242. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9832061/

37. Helmy YA, Fawzy M, Elaswad A, Sobieh A, Kenney SP, Shehata AA. The COVID-19 Pandemic: A Comprehensive Review of Taxonomy, Genetics, Epidemiology, Diagnosis, Treatment, and Control. J Clin Med [Internet]. 2020 Apr 1 [cited 2025 Jun 27];9(4):1225. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7230578/

38. Alsobaie S. Understanding the molecular biology of SARS-CoV-2 and the COVID-19 pandemic: A review. Infect Drug Resist [Internet]. 2021 [cited 2025 Jun 29];14:2259–68. Available from: https://www.tandfonline.com/doi/pdf/10.2147/IDR.S306441

39. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. Author Correction: A new coronavirus associated with human respiratory disease in China (Nature, (2020), 579, 7798, (265-269), 10.1038/s41586-020-2008-3). Nature. 2020 Apr 16;580(7803):E7.

40. GISAID - hCoV-19 Reference Sequences [Internet]. [cited 2025 Jun 29]. Available from: https://gisaid.org/wiv04/

41. López-Ayllón BD, de Lucas-Rius A, Mendoza-García L, García-García T, Fernández-Rodríguez R, Suárez-Cárdenas JM, et al. SARS-CoV-2 accessory proteins involvement in inflammatory and profibrotic processes through IL11 signaling. Front Immunol [Internet]. 2023 Jul 20 [cited 2025 Jun 29];14:1220306. Available from: https://coronavirus.jhu.edu/map.html

42. Mingaleeva RN, Nigmatulina NA, Sharafetdinova LM, Romozanova AM, Gabdoulkhakova AG, Filina Y V., et al. Biology of the SARS-CoV-2 Coronavirus. Biochemistry (Mosc) [Internet]. 2023 Dec 1 [cited 2025 Jun 27];87(12–13):1662. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9839213/

43. Eduardo Oliva Marín J. SARS-CoV-2: origen, estructura, replicación y patogénesis Artículo de Revisión. [cited 2025 Jun 30]; Available from: https://doi.org/10.5377/alerta.v3i2.9619

44. Jackson CB, Farzan M, Chen B, Choe H. Mechanisms of SARS-CoV-2 entry into cells. Nat Rev Mol Cell Biol [Internet]. 2022 Jan 1 [cited 2025 Jun 30];23(1):3–20. Available from: https://www.nature.com/articles/s41580-021-00418-x

45. Pizzato M, Baraldi C, Boscato Sopetto G, Finozzi D, Gentile C, Gentile MD, et al. SARS-CoV-2 and the Host Cell: A Tale of Interactions. Frontiers in Virology [Internet]. 2021 Jan 12 [cited 2025 Jun 30];1:815388. Available from: www.frontiersin.org

46. Bioinformática [Internet]. [cited 2025 Jun 30]. Available from: https://www.genome.gov/es/genetics-glossary/Bioinformatica

47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol [Internet]. 1990 [cited 2025 Jul 1];215(3):403–10. Available from: https://pubmed.ncbi.nlm.nih.gov/2231712/

48. Schäffer AA, Hatcher EL, Yankie L, Shonkwiler L, Brister JR, Karsch-Mizrachi I, et al. VADR: Validation and annotation of virus sequence submissions to GenBank. BMC Bioinformatics [Internet]. 2020 May 24 [cited 2024 Sep 1];21(1):1–23. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3537-3

49. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res [Internet]. 2008 May [cited 2025 Jul 1];18(5):821. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2336801/

50. Meleshko D, Hajirasouliha I, Korobeynikov A. coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. Bioinformatics [Internet]. 2021 Dec 22 [cited 2025 Jul 1];38(1):1–8. Available from: https://dx.doi.org/10.1093/bioinformatics/btab597

51. How did bioinformatics allow for swift development of the SARS-CoV-2 vaccine? | Scientia News [Internet]. [cited 2025 Jul 1]. Available from: https://www.scientianews.org/articles/how-did-bioinformatics-allow-for-swift-development-of-the-sars-cov-2-vaccine%3F

52. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, et al. GISAID's Role in Pandemic Response. China CDC Weekly, 2021, Vol 3, Issue 49, Pages: 1049-1051 [Internet]. 2021 Dec 3 [cited 2025 Jul 1];3(49):1049–51. Available from: https://weekly.chinacdc.cn/en/article/doi/10.46234/ccdcw2021.255

53. Hufsky F, Lamkiewicz K, Almeida A, Aouacheria A, Arighi C, Bateman A, et al. Computational strategies to combat COVID-19: Useful tools to accelerate SARS-CoV-2 and coronavirus research. Brief Bioinform [Internet]. 2021 Mar 1 [cited 2025 Jul 1];22(2):642–63. Available from: https://pubmed.ncbi.nlm.nih.gov/33147627/

54. Ma L, Li H, Lan J, Hao X, Liu H, Wang X, et al. Comprehensive analyses of bioinformatics applications in the fight against COVID-19 pandemic. Comput Biol Chem [Internet]. 2021 Dec 1 [cited 2025 Jul 1];95. Available from: https://pubmed.ncbi.nlm.nih.gov/34773807/

55. Jahn K, Dreifuss D, Topolsky I, Kull A, Ganesanandamoorthy P, Fernandez-Cassi X, et al. Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using COJAC. [cited 2025 Jul 1]; Available from: https://doi.org/10.1038/s41564-022-01185-x

56. Liu Y, Sapoval N, Gallego-García P, Tomás L, Posada D, Treangen TJ, et al. Crykey: Rapid identification of SARS-CoV-2 cryptic mutations in wastewater. Nat Commun [Internet]. 2024 Dec 1 [cited 2025 Jul 1];15(1):1–13. Available from: https://www.nature.com/articles/s41467-024-48334-w

57. Durmaz V, Köchl K, Krassnigg A, Parigger L, Hetmann M, Singh A, et al. Structural bioinformatics analysis of SARS-CoV-2 variants reveals higher hACE2 receptor binding affinity for Omicron B.1.1.529 spike RBD compared to wild type reference. Sci Rep [Internet]. 2022 Dec 1 [cited 2025 Jul 1];12(1):1–13. Available from: https://www.nature.com/articles/s41598-022-18507-y

58. Schrödinger LLC. The PyMOL Molecular Graphics System, Version~1.8. 2015 Nov.

59. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021 596:7873 [Internet]. 2021 Jul 15 [cited 2025 Jul 10];596(7873):583–9. Available from: https://www.nature.com/articles/s41586-021-03819-2

60. Chatterjee R, Ghosh M, Sahoo S, Padhi S, Misra N, Raina V, et al. Next-generation bioinformatics approaches and resources for coronavirus vaccine discovery and development—a perspective review. Vaccines (Basel) [Internet]. 2021 Aug 1 [cited 2025 Jul 1];9(8). Available from: https://pubmed.ncbi.nlm.nih.gov/34451937/

61. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem [Internet]. 2010 Jan 30 [cited 2025 Jul 1];31(2):455–61. Available from: https://pubmed.ncbi.nlm.nih.gov/19499576/

62. Bugnon M, Röhrig UF, Goullieux M, Perez MAS, Daina A, Michielin O, et al. SwissDock 2024: major enhancements for small-molecule docking with Attracting Cavities and AutoDock Vina. Nucleic Acids Res [Internet]. 2024 Jul 5 [cited 2025 Jul 1];52(W1):W324–32. Available from: https://dx.doi.org/10.1093/nar/gkae300

63. Weronika S, Lesniak S, Maschio D, Neria F, Rey-Delgado B, Cuerda VM, et al. Novel Biomarkers for SARS-CoV-2 Infection: A Systematic Review and Meta-Analysis. Journal of Personalized Medicine 2025, Vol 15, Page 225 [Internet].

2025 Jun 1 [cited 2025 Jul 1];15(6):225. Available from: https://www.mdpi.com/2075-4426/15/6/225/htm

64. Babadaei MMN, Hasan A, Bloukh SH, Edis Z, Sharifi M, Kachooei E, et al. The expression level of angiotensin-converting enzyme 2 determines the severity of COVID-19: lung and heart tissue as targets. J Biomol Struct Dyn [Internet]. 2021 [cited 2025 Jul 1];1–7. Available from: https://www.tandfonline.com/doi/abs/10.1080/07391102.2020.1767211

65. Fagyas M, Bánhegyi V, Úri K, Enyedi A, Lizanecz E, Mányiné IS, et al. Changes in the SARS-CoV-2 cellular receptor ACE2 levels in cardiovascular patients: a potential biomarker for the stratification of COVID-19 patients. Geroscience [Internet]. 2021 Oct 1 [cited 2025 Jul 1];43(5):2289–304. Available from: https://pubmed.ncbi.nlm.nih.gov/34674152/

66. Oracle VirtualBox [Internet]. [cited 2025 Jul 4]. Available from: https://www.virtualbox.org/

67. primer-schemes/nCoV-2019/V4.1 at master · artic-network/primer-schemes [Internet]. [cited 2025 Jul 4]. Available from: https://github.com/artic-network/primer-schemes/tree/master/nCoV-2019/V4.1

68. Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Connor R, et al. Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Res [Internet]. 2024 Jan 6 [cited 2025 Jul 4];53(D1):D20. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11701734/

69. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016 Feb 12;9(1).

70. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics [Internet]. 2011 Nov [cited 2025 Jul 4];27(21):2987–93. Available from: https://pubmed.ncbi.nlm.nih.gov/21903627/

71. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience [Internet]. 2021 Jan 29 [cited 2025 Jul 4];10(2):1–4. Available from: https://dx.doi.org/10.1093/gigascience/giab008

72. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics [Internet]. 2018 Sep 15 [cited 2025 Jul 4];34(18):3094–100. Available from: https://dx.doi.org/10.1093/bioinformatics/bty191

73. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics [Internet]. 2013 Apr 15 [cited 2025 Jul 4];29(8):1072–5. Available from: https://dx.doi.org/10.1093/bioinformatics/btt086

74. Lozano-Fernandez J. A Practical Guide to Design and Assess a Phylogenomic Study. Genome Biol Evol [Internet]. 2022 Sep 1 [cited 2025 Jul 4];14(9):evac129. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC9452790/

75. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol [Internet]. 2013 Apr [cited 2025 Jul 4];30(4):772. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3603318/

76. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics [Internet]. 2009 Aug [cited 2025 Jul 4];25(15):1972. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2712344/

77. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol [Internet]. 2014 Jan 1 [cited 2025 Jul 4];32(1):268. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC4271533/

78. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. Nat Methods [Internet]. 2017 May 30 [cited 2025 Jul 9];14(6):587. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC5453245/

79. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res [Internet]. 2024 Jul 5 [cited 2025 Jul 4];52(W1):W78–82. Available from: https://dx.doi.org/10.1093/nar/gkae268

80. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol [Internet]. 2021 Dec 16 [cited 2025 Jul 4];7(2). Available from: https://dx.doi.org/10.1093/ve/veab064

81. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology [Internet]. 2012 May 1 [cited 2025 Jul 4];19(5):455. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3342519/

82. DI Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol [Internet]. 2017 Apr 11 [cited 2025 Jul 4];35(4):316–9. Available from: https://www.nature.com/articles/nbt.3820

83. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Research 2021 10:33 [Internet]. 2021 Jan 18 [cited 2025 Jul 4];10:33. Available from: https://f1000research.com/articles/10-33

84. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics [Internet]. 2012 Oct 1 [cited 2025 Jul 4];28(19):2520–2. Available from: https://dx.doi.org/10.1093/bioinformatics/bts480

85. MerkelDirk. Docker. Linux Journal [Internet]. 2014 Mar 1 [cited 2025 Jul 4]; Available from: https://dl.acm.org/doi/10.5555/2600239.2600241

86. Di Tommaso P, Palumbo E, Chatzou M, Prieto P, Heuer ML, Notredame C. The impact of Docker containers on the performance of genomic pipelines. PeerJ [Internet]. 2015 [cited 2025 Jul 4];2015(9). Available from: https://pubmed.ncbi.nlm.nih.gov/26421241/

87. Zsichla L, Zeeb M, Fazekas D, Áy É, Müller D, Metzner KJ, et al. Comparative Evaluation of Open-Source Bioinformatics Pipelines for Full-Length Viral Genome Assembly. Viruses [Internet]. 2024 Dec 1 [cited 2025 Jul 4];16(12):1824. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11680378/

88. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics [Internet]. 2016 Oct 1 [cited 2025 Jul 4];32(19):3047–8. Available from: https://dx.doi.org/10.1093/bioinformatics/btw354

89. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol [Internet]. 2019 [cited 2025 Jul 4];15(4):e1006650. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006650

90. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol [Internet]. 2018 Jan 1 [cited 2025 Jul 4];4(1). Available from: https://dx.doi.org/10.1093/ve/vey016