

# Máster en Bioinformática

Secuenciación Genómica y Análisis De Variantes Para Medicina Personalizada y De Precisión

Curso académico 2024-25  
Edición Abril



**Universidad**  
Internacional  
de Valencia

Dra. Laura Gutiérrez Macías  
[laura.gutierrez.m@professor.universidadviu.com](mailto:laura.gutierrez.m@professor.universidadviu.com)

# Capítulo 6. Análisis de genomas de virus. El caso de SARS-CoV-2.

## 6.1.

- **Estructura de los virus. Aspectos generales.**
  - 6.1.1.- Bacteriófagos
  - 6.1.2.- Virus que infectan células eucariotas

## 6.2.

- **El genoma de SARS-CoV-2**

## 6.3.

- **Clasificación de SARS-CoV-2. Variantes de interés y linajes. Bases de datos.**
  - 6.3.1.- Clasificación de la OMS: variantes de preocupación, variantes de interés y variantes bajo vigilancia
  - 6.3.2.- Nextstrain: Monitorización de la evolución de SARS-CoV-2 a tiempo real.
  - 6.3.3.- GISAID: el repositorio de todos los genomas de SARS-CoV-2 secuenciados
  - 6.3.4.- PANGO: Clasificación de linajes y constelaciones de mutaciones.
  - 6.3.5.- Análisis de mutaciones de interés. Bases de datos de monitorización

## 6.4.

- **Análisis bioinformático de linajes de SARS-CoV-2**
  - 6.4.1.- Limpieza de lecturas crudas. Eliminación de genoma de hospedador.
  - 6.4.2.- Mapeo al genoma de referencia y determinación de linaje.
  - 6.4.3.- Ensamblaje *de novo* y anotación.

# 6.1. Estructura de los virus.

## Aspectos generales

Los virus constituyen el grupo de los organismos más simples que se conocen. De hecho, en términos biológicos, **los virus no son considerados organismos vivos**, por depender de una célula hospedadora y su maquinaria biológica para poder replicarse.

Algunos virus poseen genes que codifican su propio ADN o ARN polimerasa, pero algunos dependen de las enzimas del hospedador para replicarse y transcribirse. Todos los virus utilizan los ribosomas y la maquinaria de translación para la síntesis de los péptidos que construyen la envuelta de su progenie. Esto nos indica que los virus deben encajar con la maquinaria de su hospedador, lo que hace que sean **bastante específicos de organismos particulares**, y tipos de virus concretos no infectan un amplio espectro de especies.

Existen una multitud de tipos de virus, que podemos clasificar en dos grandes grupos, en función del tipo de célula al que infectan.

### 6.1.1. Bacteriófagos

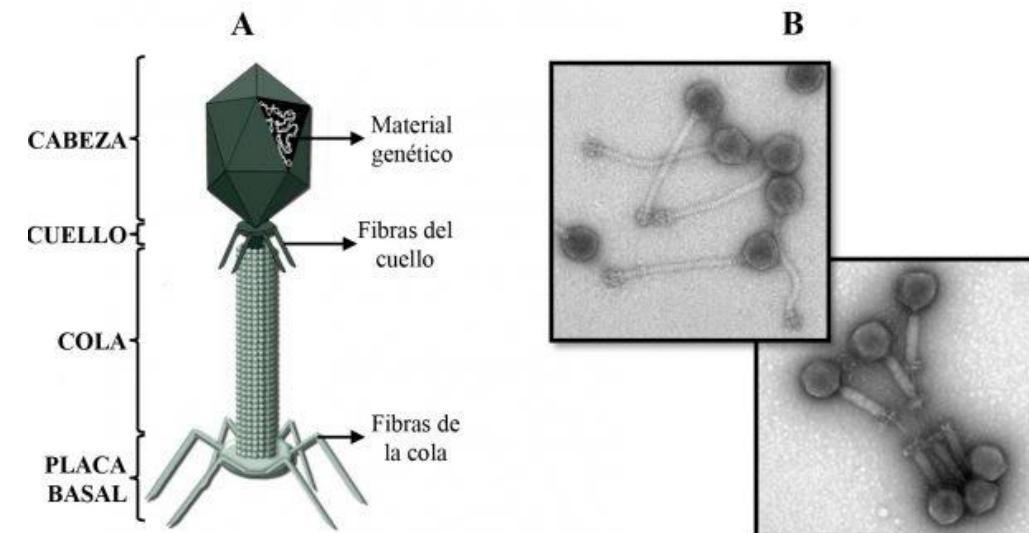
Virus que **infectan bacterias**.

Están compuestos por dos componentes básicos que son las proteínas (que forman la envuelta o cápside) y el ácido nucleico, contenido dentro de ella.

Este ácido nucleico puede ser tanto ADN como ARN y, a su vez, puede ser tanto de doble hélice como de hélice sencilla.

La cápside puede tener forma isosaédrica (fago MS2 que infecta *E. coli* o PM2 que infecta *P. aeruginosa*), filamentosa o con forma de hélice (M13 – *E. coli*) o con forma cabeza-cola, con una cabeza icosaédrica que contiene el ácido nucleico unida a una cola filamentosa que facilita la entrada del ácido nucleico en la célula infectada (fagos T4 y lambda).

Por lo tanto, existe una gran variedad de tipos de bacteriófagos, que podéis ver generalizada y resumida en la tabla siguiente.



+

Fago	Hospedador	Estructura cápside	Estructura del genoma	Tamaño genoma (kb)	Número de genes
Lambda	<i>Escherichia coli</i>	Cabeza-cola	ADN linear doble cadena	49,5	48
φX174	<i>E. coli</i>	Icosaédrico	ADN circular cadena sencilla	5,4	11
f6	<i>Pseudomonas phaseolicola</i>	Icosaédrico	ARN linear segmentado doble cadena	2,9; 4,0; 6,4	13
M13	<i>E. coli</i>	Filamentoso	ADN circular cadena sencilla	6,4	10
MS2	<i>E. coli</i>	Icosaédrico	ARN linear cadena sencilla	3,6	3
PM2	<i>Pseudomonas aeruginosa</i>	Icosaédrico	ADN linear cadena doble	10,0	21
SPO1	<i>Bacillus subtilis</i>	Cabeza-cola	ADN linear doble cadena	150	> 100
T2, T4, T6	<i>E. coli</i>	Cabeza-cola	ADN linear doble cadena	166	> 150
T7	<i>E. coli</i>	Cabeza-cola	ADN linear doble cadena	39,9	> 55

Una de las características generales de los fagos es su capacidad de **empaquetar un gran número de genes** en poco espacio, gracias a **contener genes solapantes**. Estos comparten secuencia genética, pudiendo estar un gen contenido dentro de otro, pero codificando para proteínas diferentes, ya que los transcritos comienzan en puntos de traducción diferentes y en marcos de lectura diferentes. Esta característica es común a muchos virus, que exhiben mayor complejidad cuantos más genes contienen.

Los bacteriófagos han tomado un especial interés porque **pueden arrastrar en su escisión genes de resistencia a antibióticos o virulencia, así como empaquetar elementos genéticos móviles como plásmidos de pequeño tamaño** (Colavecchio et al., 2017; Navarro y Muniesa, 2017; Prieto et al., 2016; Rodríguez-Rubio et al., 2020).

Por otra parte, están siendo ampliamente estudiados para combatir el problema de la resistencia a antibióticos ( Bragg et al., 2014; Brives y Pourraz, 2020; Moghadam et al., 2020; Principi et al., 2019).

Por otro lado, los bacteriófagos se clasifican en dos grupos según su ciclo de vida: **líticos** y **lisogénicos**.

Los fagos de tipo **lítico** exterminan a la bacteria hospedadora muy rápidamente tras la infección inicial, mientras que los fagos **lisogénicos** pueden permanecer quiescentes en la célula huésped por un tiempo sustancial, incluso varias generaciones.

Durante el ciclo lisogénico algunos fagos pueden integrarse en el genoma del hospedador, dando lugar a un profago. Esta integración se da mediante recombinación sitio-específica, en un lugar concreto del genoma del hospedador. Este profago integrado puede mantenerse durante generaciones y es replicado junto con el genoma bacteriano, pasando a las células hijas. Estímulos químicos o físicos hacen que este profago se induzca a una fase lítica, comenzando por una recombinación que lo escinde del genoma del hospedador, comienza su replicación y se sintetizan las proteínas de la envuelta. Finalmente, la célula es lisada y los nuevos fagos son liberados al medio, listos para infectar nuevas bacterias. Ejemplo de este proceso es el fago lambda de *E. coli*.

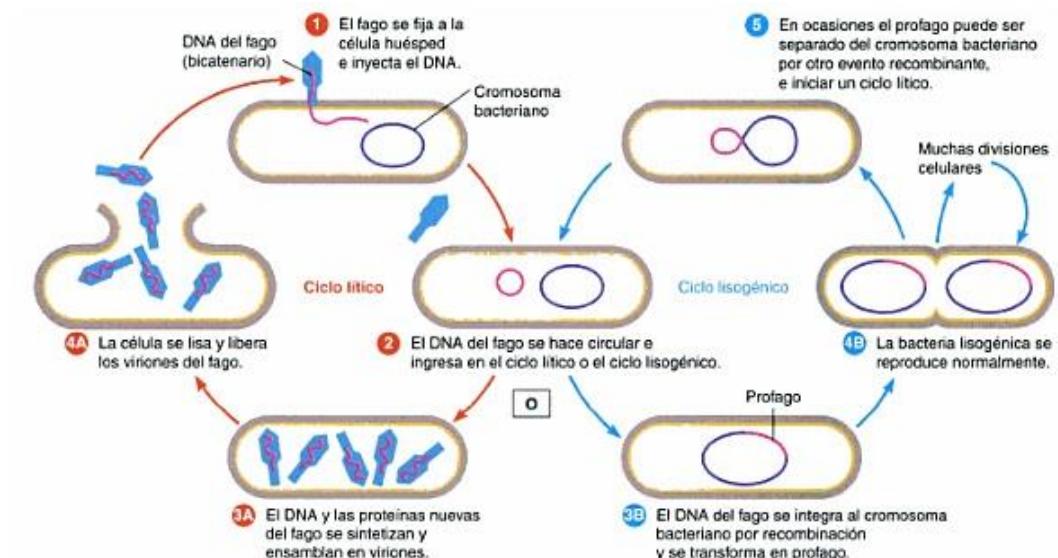


FIGURA Ciclo lisogénico del bacteriófago λ.

Para comprender las diferencias entre el ciclo lítico y lisogénico, este vídeo explica perfectamente los conceptos.

<https://www.youtube.com/watch?v=SY9zOhNN7Dw>

### 6.1.2. Virus que infectan células eucariotas.

La cápside de estos virus puede ser icosaédrica o filamentosa, pero no se han detectado estructuras de cabeza-cola, que permanecen exclusivas de bacteriófagos.

Una característica diferencial de los virus eucariotas, especialmente aquellos que infectan células animales, es que **la cápside puede estar rodeada de una membrana lipídica, formando un componente adicional de la estructura vírica**. Esta membrana deriva del hospedador cuando la nueva partícula vírica se libera y abandona la célula y puede estar modificada por la inserción de proteínas específicas del virus.

Los virus de eucariotas muestran una gran variedad de estructuras (Tabla 12). Nuevamente, pueden ser virus de ADN o ARN, cadena doble o sencilla, o incluso parte doble con regiones de cadena sencilla, lineares o circulares, segmentados o no segmentados.

Virus	Hospedador	Estructura del genoma	Tamaño genoma (kb)	Número de genes
Adenovirus	Mamíferos	ADN linear doble cadena	36,0	30
Hepatitis B	Mamíferos	ADN circular parcialmente doble cadena	3,2	4
Virus <i>Influenza</i>	Mamíferos	ARN linear segmentado cadena sencilla	22,0	12
Parvovirus	Mamíferos	ADN linear cadena sencilla	1,6	5
Poliovirus	Mamíferos	ADN linear cadena sencilla	7,6	8
Reovirus	Mamíferos	ARN linear segmentado doble cadena	22,5	22
Retrovirus	Mamíferos, pájaros	ARN linear cadena sencilla	6,0-9,0	3
SV40	Monos	ADN circular cadena doble	5,0	5
Virus mosaico del tabaco	Plantas	ARN linear cadena sencilla	6,4	6
Virus <i>Vaccinia</i> (viruela)	Mamíferos	ADN circular cadena doble	240	240

La mayor parte de estos virus siguen un ciclo de **infección lítico**, pero unos pocos pueden comportarse con ciclo lisogénico.

Algunos coexisten con las células hospedadoras durante largos periodos de tiempo (años), incluso integrándose en el genoma de la célula.

Si son virus ARN tienen un paso adicional de conversión a ADN, para integrarse y ser retroelementos virales. Gracias a una enzima llamada transcriptasa reversa, estos virus ARN se copian a ADN complementario, crean la hebra complementaria de ADN para generar la estructura de doble cadena y posteriormente son integrados en un lugar del genoma.

A diferencia del bacteriófago lambda, los retrovirus no tienen similitud genética con el ADN del hospedador. Esta integración es un prerequisito para la expresión de los genes del retrovirus que codifican para poliproteínas. Cada una de ellas es escindida en dos o más productos, incluyendo las proteínas de la envuelta viral y la transcriptasa reversa.

En el siguiente enlace, podrás encontrar un resumen de los tipos de virus que podemos encontrar.

<https://www.youtube.com/watch?v=jX3MhWWi6n4>

6.2.

El genoma de SARS-CoV-2

La enfermedad COVID-19 es una enfermedad infecciosa causada por el virus SARS-CoV-2. La primera datación de este virus se tiene en la provincia de Wuhan (China) en otoño-invierno de 2019 (estas dataciones están aún bajo estudio).

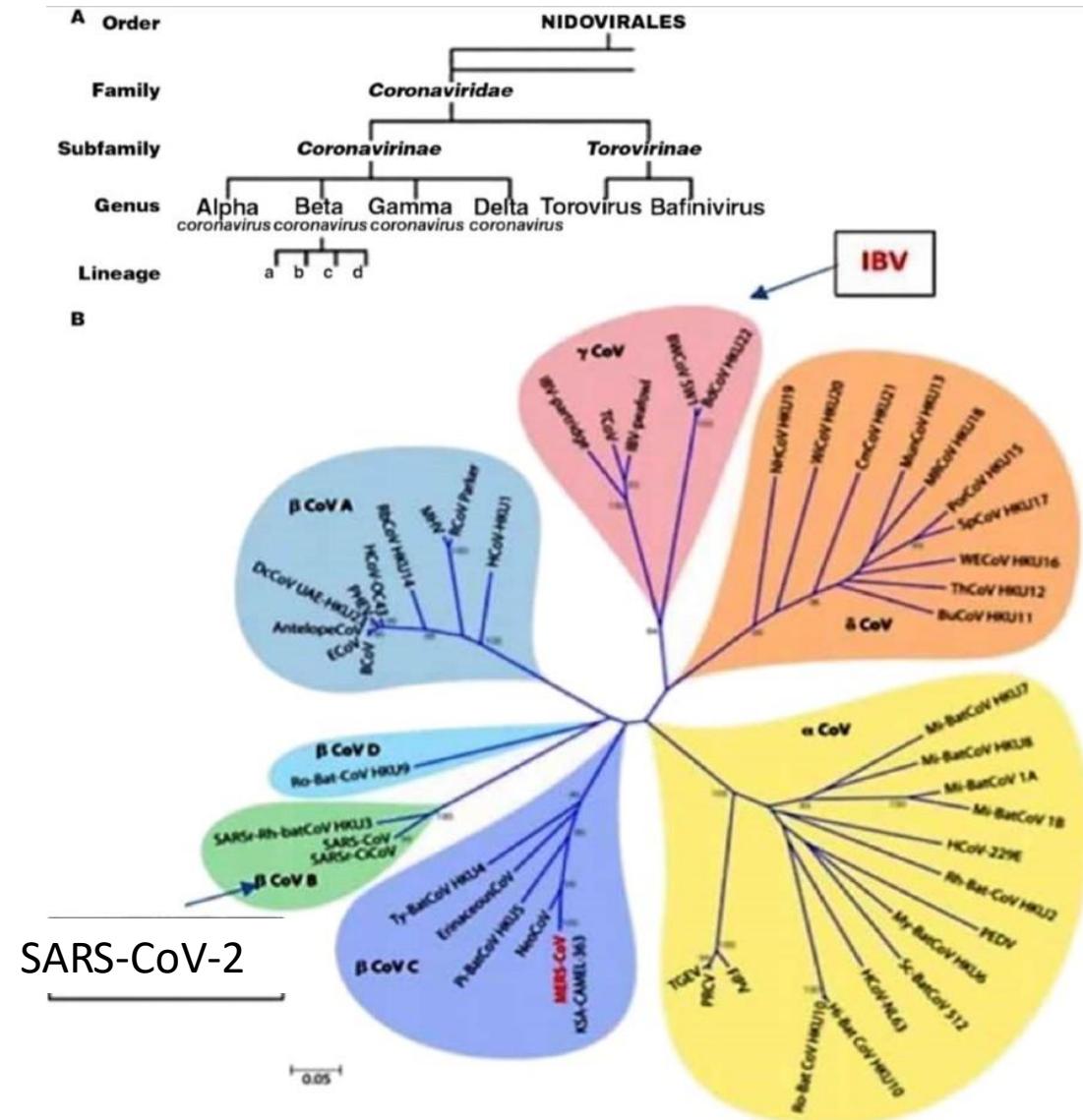
Al igual que otros coronavirus, SARS-CoV-2 afecta al sistema respiratorio, causando una enfermedad respiratoria con síntomas como tos, fiebre y, en casos más graves, disnea.

La mortalidad es alta en grupos de edad elevada ( $> 80$  años), así como en personas de distintas edades con factores de riesgo (inmunosupresión, obesidad, hipertensión, diabetes, etc.).

Además del síndrome respiratorio, hoy en día conocemos que COVID-19 puede dar un proceso inflamatorio, conduciendo a sepsis, daño cardiaco agudo, fallo cardiaco y disfunción multiorgánica en pacientes de alto riesgo; así como persistir con secuelas graves en algunos pacientes.

Los coronavirus (CoV) son un grupo de virus con envuelta, donde su material genético es una hebra de ARN de cadena única. Este grupo de virus pertenecen a la familia Coronaviridae del orden de los Nidovirales. Han sido clasificados en cuatro géneros que incluyen los alfa, beta, gamma y delta coronavirus. Entre ellos, alfa y beta infectan mamíferos, mientras que gamma infecta especies aviares y delta infectan ambos tipos, mamíferos y aves.

Los coronavirus como SARS-CoV, el coronavirus de la hepatitis murina (MHV), MERS-CoV, el coronavirus bovino (BCoV), el coronavirus de murciélagos HKU4 y los coronavirus humanos como OC43 y SARS-CoV-2 pertenecen al grupo de los beta coronavirus. Los más conocidos, que han dado lugar a epidemias o incluso pandemia, son SARS-CoV, MERS y SARS-CoV-2, todos ellos transmitidos a partir de un evento zoonótico.

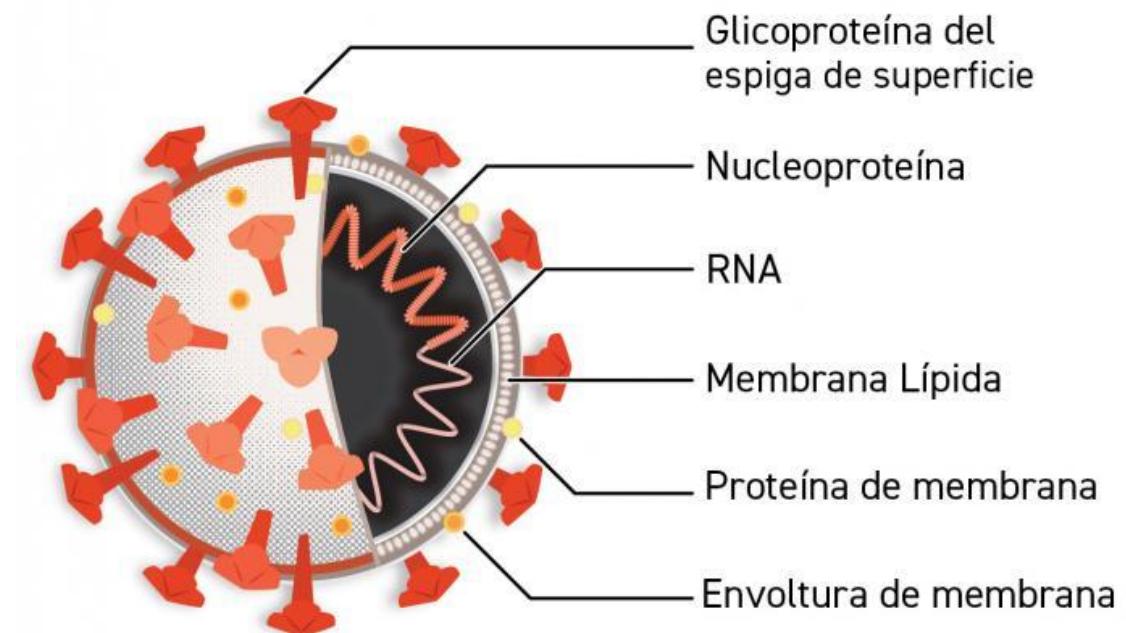


Los coronavirus tienen un genoma de ARN que comprende entre 26 a 32 kb de longitud.

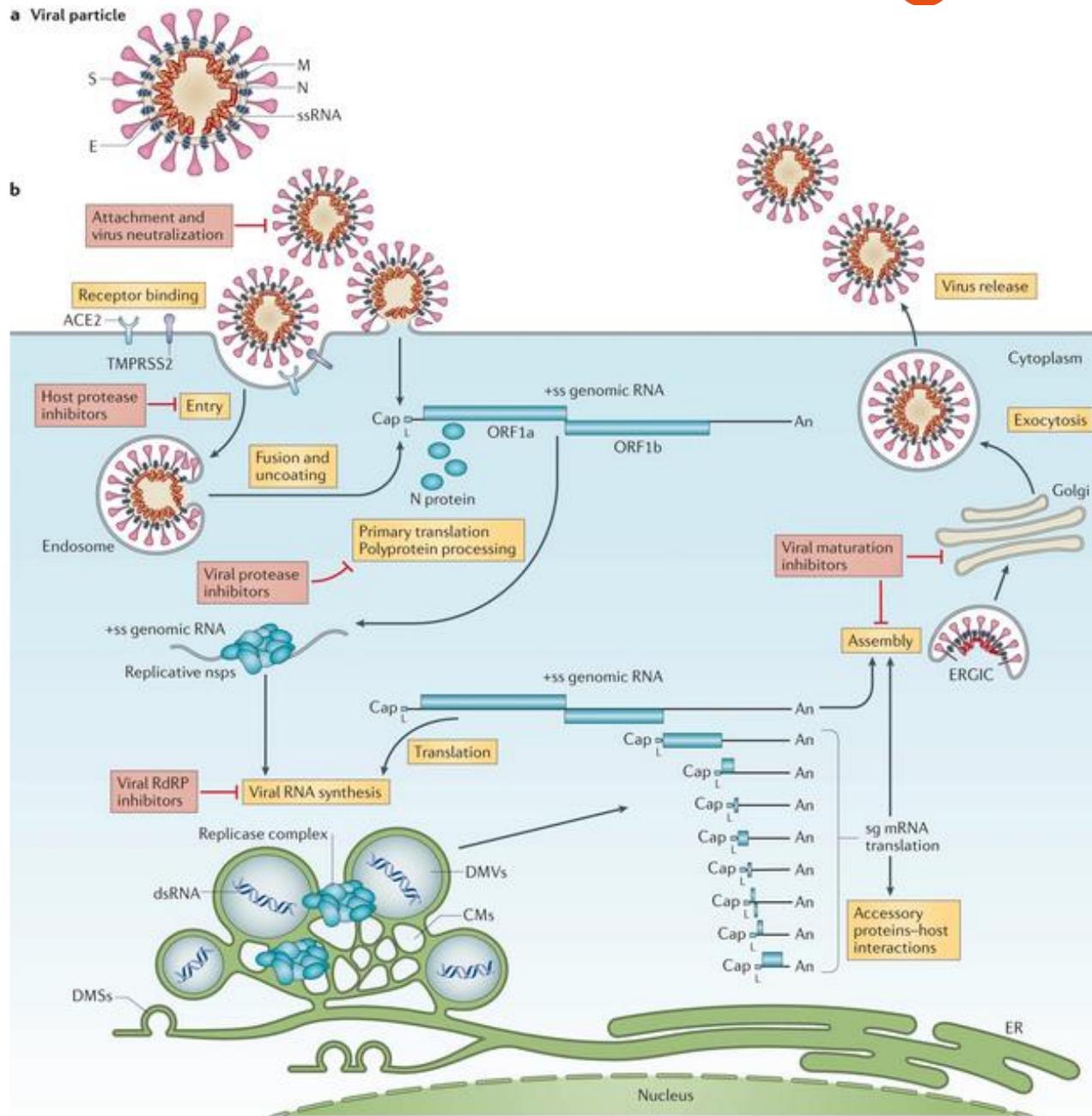
En el caso de SARS-CoV-2, comparte un 82 % de identidad en su secuencia con los genomas de SARS-CoV y MERS-CoV, y más de un 90 % de identidad en proteínas estructurales y con potencial enzimático. Este alto nivel de homología en su secuencia revela un mecanismo similar de patogénesis y está siendo clave en el desarrollo de tratamientos antivirales y, por supuesto, de la vacuna.

De manera principal, SARS-CoV-2 contiene cuatro proteínas estructurales que incluyen la espícula (spike, S), envuelta (envelope, E), la membrana (membrane, M) y la nucleocápside (nucleocapsid, N). Estas proteínas comparten alta homología con las correspondientes de los coronavirus asociados SARS-CoV y MERS-CoV.

## SARS-CoV-2



Los CoV dependen de sus proteínas de espícula (S) para unirse al receptor localizado en la superficie de las células huésped. El receptor de tipo angiotensina ACE-2 es la cerradura, donde la espícula (la llave) abre la puerta para entrar en la célula y provocar la infección. El dominio de unión a receptor (RBD) está compuesto por las subunidades S1 y S2.



El primer genoma disponible de SARS-CoV-2, originario de Wuhan, se encuentra disponible en la base de datos NCBI con el acceso NC\_045512.2 (acceso a fecha 12/12/2021).

Tiene una longitud de 29 903 pb y está compuesto por 13-15 marcos de lectura abiertos (ORF), doce de ellos funcionales. Existen once genes codificantes de doce proteínas totales que se expresan. La organización génica es muy similar a SARS-CoV y MERS-CoV, donde en sentido 5' a 3' contienen en primer lugar los genes codificantes de las proteínas no estructurales ORF1a y ORF1b, parcialmente solapantes.

Estas codifican para las poliproteínas pp1a y pp1ab, respectivamente, que se procesan mediante las proteinasas codificadas por el virus y producen 16 proteínas, que están bien conservadas en todos los CoV pertenecientes a la misma familia.

Posteriormente, se encuentra la codificación para las proteínas estructurales S, E, M y N, que son las dianas principales de las vacunas y tratamientos farmacológicos. Sus productos son vitales en la entrada del virus a la célula, la fusión y supervivencia en la célula hospedadora.

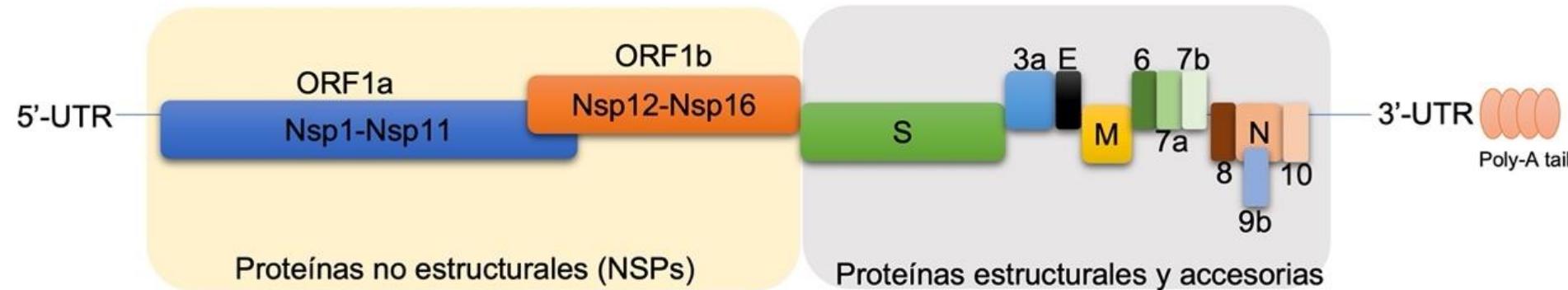
#### **Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome**

NCBI Reference Sequence: NC\_045512.2

[FASTA](#) [Graphics](#)

Go to: 

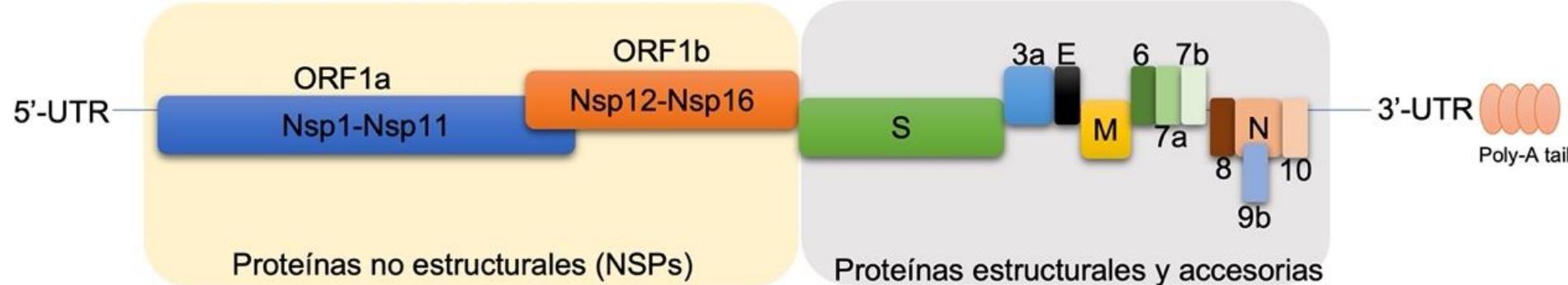
LOCUS	NC_045512	29903 bp ss-RNA	linear	VRL 18-JUL-2020
DEFINITION	Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.			
ACCESSION	NC_045512			
VERSION	NC_045512.2			
DBLINK	BioProject: <a href="#">PRJNA485481</a>			
KEYWORDS	RefSeq.			
SOURCE	Severe acute respiratory syndrome coronavirus 2			
ORGANISM	<a href="#">Severe acute respiratory syndrome coronavirus 2</a>			
	Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales; Cornidovirinae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus.			
REFERENCE	1 (bases 1 to 29903)			
AUTHORS	Wu,F., Zhao,S., Yu,B., Chen,Y.M., Wang,W., Song,Z.G., Hu,Y., Tao,Z.W., Tian,J.H., Pei,Y.Y., Yuan,M.L., Zhang,Y.L., Dai,F.H., Liu,Y., Wang,Q.M., Zheng,J.J., Xu,L., Holmes,E.C. and Zhang,Y.Z.			
TITLE	A new coronavirus associated with human respiratory disease in China			
JOURNAL	Nature 579 (7798), 265-269 (2020)			
PUBMED	<a href="#">32015508</a>			
REMARK	Erratum:[Nature. 2020 Apr;580(7803):E7. PMID: 32296181]			
REFERENCE	2 (bases 13476 to 13503)			
AUTHORS	Baranov,P.V., Henderson,C.M., Anderson,C.B., Gesteland,R.F., Atkins,J.F. and Howard,M.T.			
TITLE	Programmed ribosomal frameshifting in decoding the SARS-CoV genome			
JOURNAL	Virology 332 (2), 498-510 (2005)			
PUBMED	<a href="#">15680415</a>			
REFERENCE	3 (bases 29728 to 29768)			
AUTHORS	Robertson,M.P., Igel,H., Baertsch,R., Haussler,D., Ares,M. Jr. and Scott,W.C.			



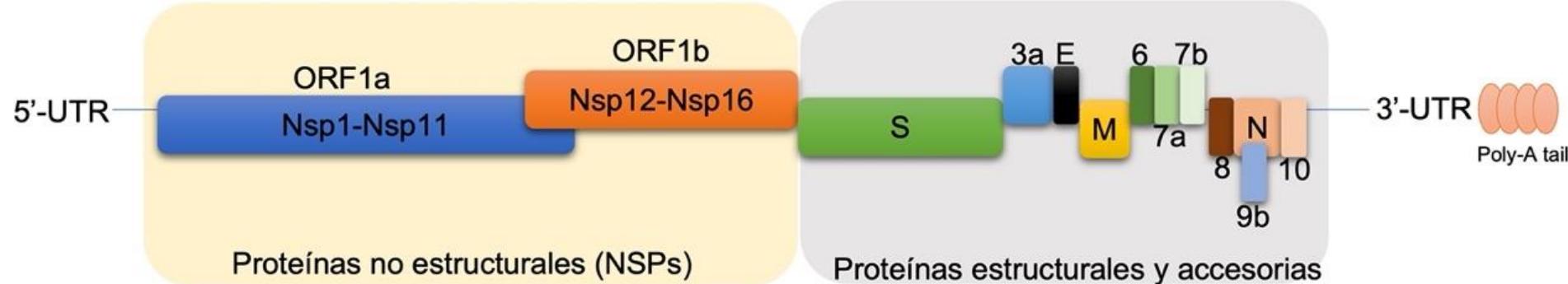
## 6.2. El genoma de SARS-CoV-2

Gen	Proteína	Localización*	ID NCBI	Descripción	Función propuesta
ORF1a	pp1a	266..13483	YP_009725295.1	Poliproteína	
	Nsp1	[1..180]	YP_009725297.1	Producto N-terminal de la poliproteína vírica	Media en la replicación del ARN y su procesado. Involucrada en la degradación de ARNm. Inhibición de la traducción.
	Nsp2	[181..818]	YP_009725298.1	Producto de replicasa esencial para corrección de la replicación viral	Modulación de la ruta de señalización de supervivencia de la célula hospedadora.
	Nsp3	[819..2763]	YP_009725299.1	Proteasa similar a la papalina.	Proteasa de tipo cisteína esencial para la replicación del virus.
	Nsp4	[2764..3263]	YP_009725300.1	Proteína transmembrana	Proteína de anclaje del complejo de replicación-transcripción.
	Nsp5	[3264...3569]	YP_009725301.1	Proteínaasas de tipo 3C	Relacionada con el procesamiento viral de las poliproteínas durante la replicación viral.
	Nsp6	[3570..3859]	YP_009725302.1	Dominio transmembrana	Juega un papel en la inducción inicial del <b>autofagoso</b> , desde el retículo <b>endoplasmático</b> de la célula hospedadora.
	Nsp7	[3860..3942]	YP_009725303.1	ARN polimerasa dependiente de ARN	Forma un complejo <b>hexadecamérico</b> junto con Nsp8. Adoptan la estructura de un cilindro hueco implicado en la replicación.
	Nsp8	[3943..4140]	YP_009725304.1	Replicasa ARN polimerasa <b>multienzimática</b> .	Forma un complejo <b>hexadecamérico</b> junto con Nsp7. Adoptan la estructura de un cilindro hueco implicado en la replicación.
	Nsp9	[4141..4253]	YP_009725305.1	Proteína viral de unión al ARN de cadena sencilla	Participa en la replicación viral actuando como una proteína de unión al ARN de cadena sencilla.
ORF1ab	Nsp10	[4254..4392]	YP_009725306.1	Proteína de tipo factor de crecimiento	Actúa en la transcripción viral, estimulando la función <b>exoribonucleasa</b> de Nsp14 y actividad <b>metiltransferasa</b> de Nsp16.
	Nsp11	[4393..4405]	YP_009725312.1	Proteína pequeña (13 aa) idéntica al primer segmento de Nsp12	Nsp16. Además, juega un papel esencial en la metilación de los ARNm. Desconocido
	Nsp12	[4393..5324]	YP_009725307.1	Poliproteína	Responsable de la replicación y transcripción del genoma del virus
	Nsp13	[5325..5925]	YP_009725308.1	ARN, dependiente de ARN (Pol/RdRp)	Dominio helicasa que se une a ATP. Dominio de unión a Zinc que se involucra en la replicación y transcripción.
	Nsp14	[5926..6452]	YP_009725309.1	Dominio <b>exoribonucleasa</b> de corrección de errores (ExoN/nsp14)	Actividad <b>exoribonucleasa</b> actuando en dirección 3' a 5' y con actividad N7-guanina <b>metiltransferasa</b> .
	Nsp15	[6453..6798]	YP_009725310.1	EndoRNAsa.	<b>Endoribonucleasa</b> , dependiente de Manganese ( $Mn^{2+}$ ).
	Nsp16	[6799..7096]	YP_009725311.1	Metiltransferasa 2'-O-ribosa	<b>Metiltransferasa</b> , que media en el proceso de metilación del ARNm.
S		21563..25384	YP_009724390.1	Proteína estructural	Proteína de superficie, espícula. Unión al receptor ACE2 de la superficie celular para promover la fusión del virus y la célula infectada.
ORF3a		25393..26220	YP_009724391.1		
E		26245..26472	YP_009724391.1	Proteína estructural	Proteína estructural de la envuelta. Formación de poros para el transporte de iones.
M		26523..27191	YP_009724393.1	Proteína estructural	Glicoproteína de membrana. Empaquetamiento del ARN.
ORF6		27202..27387	YP_009724394.1		
ORF7a		27394..27759	YP_009724395.1		
ORF7b		27756..27887	YP_009725318.1		
ORF8		27894..28259	YP_009724396.1		
N/ORF9		28274..29533	YP_009724397.2	Proteína estructural	Fosfoproteína nucleocápside
ORF10		29558..29674	YP_009725255.1		

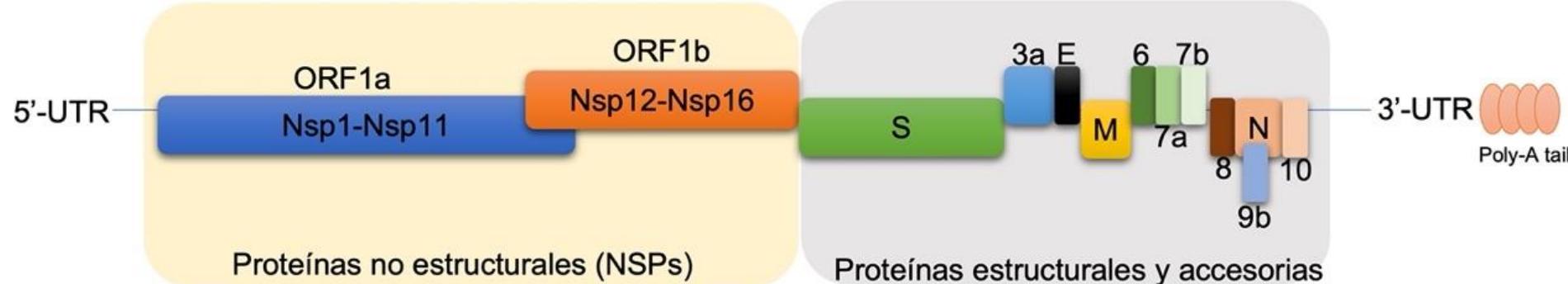
\*Las regiones indicadas entre corchetes se refieren a las regiones escindidas de la poliproteína correspondiente.



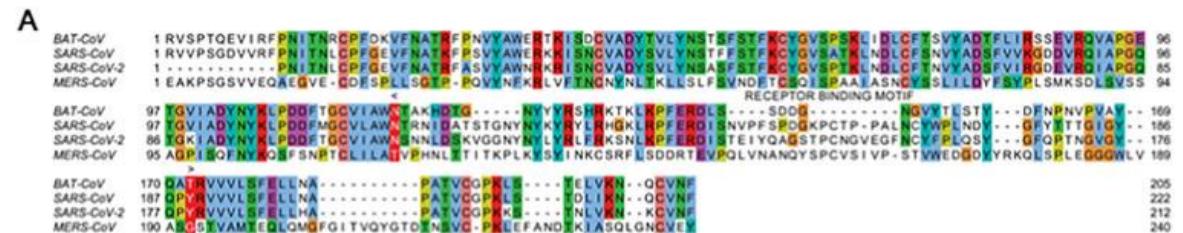
Además de las proteínas estructurales que forman la cápside, el genoma viral codifica para varias proteínas no estructurales, comúnmente denominadas *non-structural proteins* (NSP) que realizan diversas funciones en la replicación y ensamblaje del virus, así como en su patogénesis, modulando la transcripción temprana, función helicasa, inmunomodulación, activación génica y contrarrestar la respuesta antiviral. Las funciones de estas proteínas están especificadas en la Tabla 13.

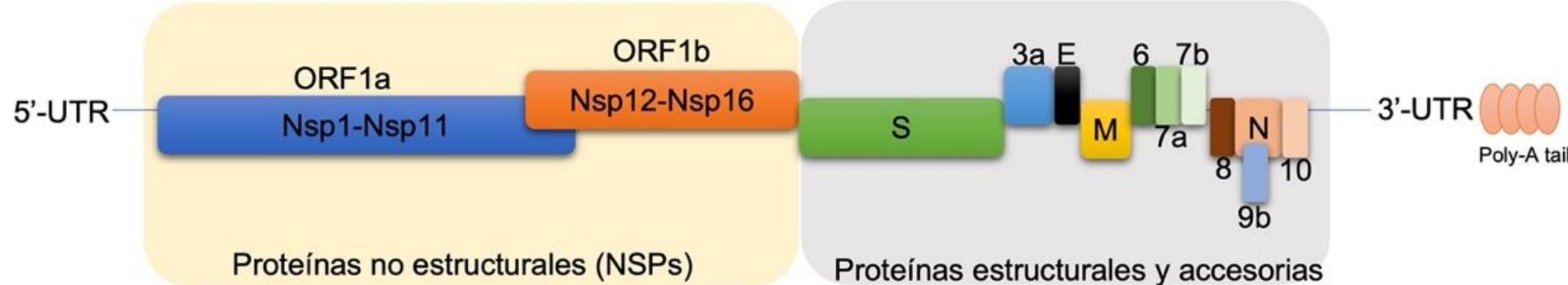


La **proteína estructural de la espícula** (*spike*, S) es una glicoproteína vital para la patogénesis, ya que es la proteína de unión al dominio RBD del receptor. En este punto es donde se inicia la infección, cuando el virión se introduce en la célula hospedadora. Se compone de 1273 aminoácidos y contiene tres subunidades: S1, S2 y S2'. El dominio S1 es el encargado en la unión de los viriones a la membrana celular, interaccionando con el receptor de tipo ACE2. En este proceso, la proteína S cambia de conformación. Conocemos que mutaciones en esta proteína provocan cambios conformacionales y, por tanto, una mejor o peor afinidad al receptor al que debe unirse. La subunidad S2 está involucrada en la fusión del virión con la membrana celular de la célula hospedadora. La región RBD de la proteína de la espícula es la región más variable y su comparación con otros coronavirus sugiere que comparten un perfil evolutivo similar (Figura 34A). En este caso, SARS-CoV-2 se muestra similar a los coronavirus procedentes de murciélagos HKU3 y SARS-CoV; mientras que MERS-CoV muestra divergencia, quizás por su origen no asociado a murciélagos. Los residuos identificados como críticos en esta región RBD para la unión al receptor ACE2 son Leu455, Phe486, Gln493, Ser494, Asn501 y Tyr505.



**La proteína estructural de la espícula (spike, S)** es una glicoproteína vital para la patogénesis, ya que es la proteína de unión al dominio RBD del receptor. En este punto es donde se inicia la infección, cuando el virión se introduce en la célula hospedadora. Se compone de 1273 aminoácidos y contiene tres subunidades: S1, S2 y S2'. El dominio S1 es el encargado en la unión de los viriones a la membrana celular, interaccionando con el receptor de tipo ACE2. En este proceso, la proteína S cambia de conformación. Conocemos que mutaciones en esta proteína provocan cambios conformatacionales y, por tanto, una mejor o peor afinidad al receptor al que debe unirse. La subunidad S2 está involucrada en la fusión del virión con la membrana celular de la célula hospedadora. La región RBD de la proteína de la espícula es la región más variable y su comparación con otros coronavirus sugiere que comparten un perfil evolutivo similar (Figura 34A). En este caso, SARS-CoV-2 se muestra similar a los coronavirus procedentes de murciélagos HKU3 y SARS-CoV; mientras que MERS-CoV muestra divergencia, quizás por su origen no asociado a murciélagos. Los residuos identificados como críticos en esta región RBD para la unión al receptor ACE2 son Leu455, Phe486, Gln493, Ser494, Asn501 y Tyr505.

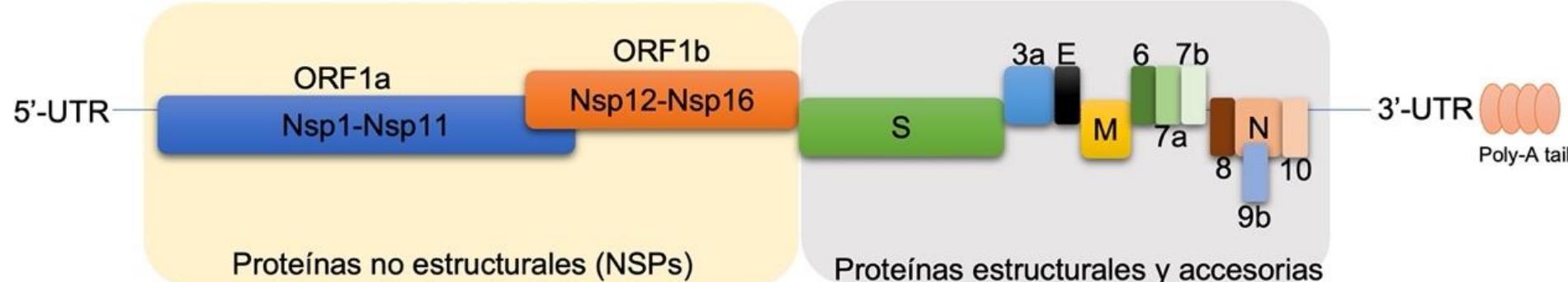




Las **proteínas de membrana** (*envelope*, E) son un grupo de pequeñas proteínas virales que asisten en el ensamblaje y liberación de los viriones. Es considerada una diana para fármacos ideal, por su pequeño tamaño (75 aa) y su papel fundamental en la morfogénesis y el ensamblaje del virus. Actúa formando unos poros lipídicos que actúan en el transporte de iones. La secuencia de la proteína de membrana entre los cuatro coronavirus principales está altamente conservado (Figura 34B).

**B**

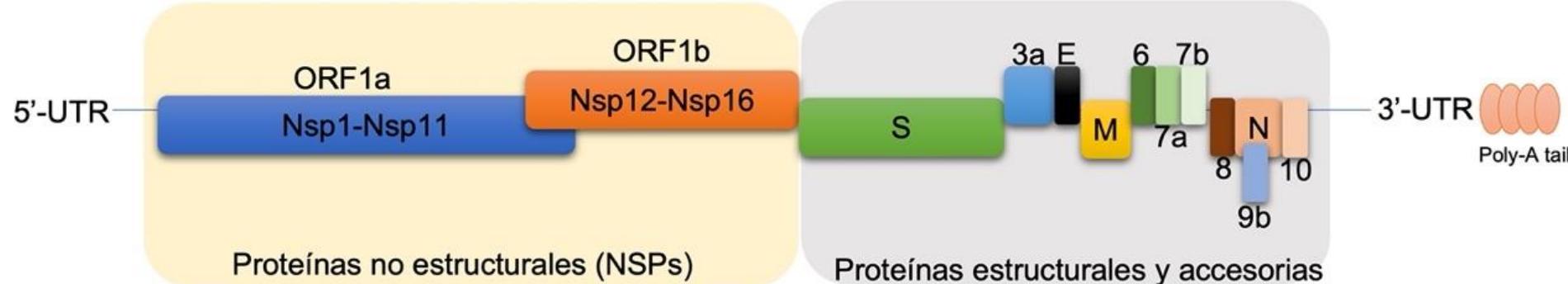
BAT-CoV	1	MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAII	TALRLCAYCCNIVNVSLSVPTVVVYSS	.....VKNLNSSEGVP	- 72
SARS-CoV	1	MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAII	TALRLCAYCCNIVNVSLSVPTVVVYSS	.....VKNLNSSEGVP	- 72
SARS-CoV-2	1	MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAII	TALRLCAYCCNIVNVSLSVPTVVVYSS	.....VKNLNSSEGVP	- 71
MERS-CoV	1	MLPFEVQERIGLFIIVNFFIFTVVCAIT	TALRLCVOCTMTGFNTLLVQALYLINYNTGRSVYVK	FQDSKPPPLPP	78
BAT-CoV	73	DLLV			76
SARS-CoV	73	DLLV			76
SARS-CoV-2	72	DLLV			75
MERS-CoV	79	DEWV			82



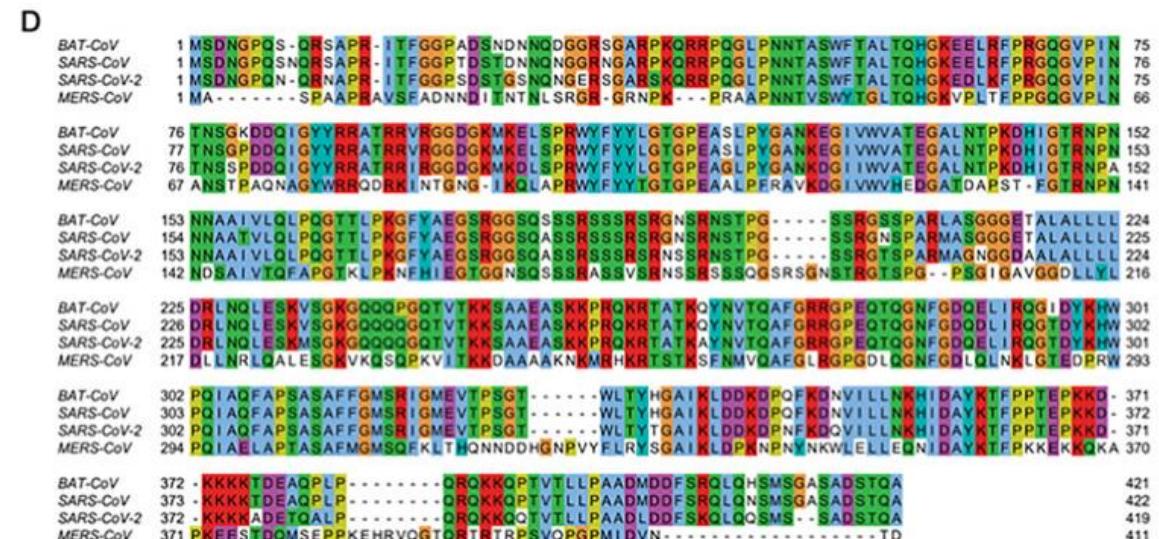
La **proteína de membrana** (*membrane*, M) tiene una longitud de 222 aa y juega un papel principal en el empaquetamiento del ARN, conteniendo tres dominios transmembrana (Figura 34C).

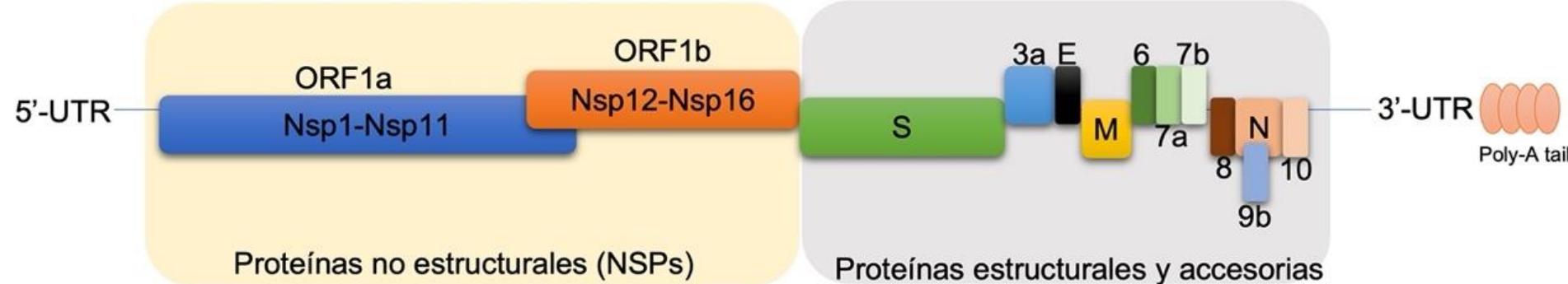
C

BAT-CoV	1 MAD - NGT ITVEELKQQLLEQWNLVIGFIFLAWIMLLQFAYSNRNRFLY I I KLVFLWL LWPVT LACFVLAAVYR IN NWVT	76
SARS-CoV	1 MAD - NGT ITVEELKQQLLEQWNLVIGFIFLAWIMLLQFAYSNRNRFLY I I KLVFLWL LWPVT LACFVLAAVYR IN NWVT	76
SARS-CoV-2	1 MAD SNGT ITVEELKQQLLEQWNLVIGFIFLFLTWICLLQFAYANRNRFLY I I KLIIFLWL LWPVT LACFVLAAVYR IN NWVT	77
MERS-CoV	1 MSN - MTQLTEADIIIAI KDWNFAWSLIFILLITIVLOYGYPSRSMTVYYVFKMFVLWL LWPSSMALS IFSAVYPIDLAS	76
BAT-CoV	77 GGIAIAAMACIVGLMWLSYFVASFRLFARTRSMMWSFNPE TNILLNVPLRG TILTRPLMESELVIGAVI IRGHLRMAGH	153
SARS-CoV	77 GGIAIAAMACIVGLMWLSYFVASFRLFARTRSMMWSFNPE TNILLNVPLRG TIVTRPLMESELVIGAVI IRGHLRMAGH	153
SARS-CoV-2	78 GGIAIAAMACLVGLMWLSYFIASFRLFARTRSMMWSFNPE TNILLNVPLHGTILTRPLLESELVIGAVI LRGHLRIAGH	154
MERS-CoV	77 QI ISGIVAAV SAMMWISYFVQSIIRLFMRITGSWWSFNPETNCLLNVPFGGT TVVPLVEDTSVTAVVINGHLKMGW	153
BAT-CoV	154 SLGRCDIKDLPKEITVATSRTLSYYKLGAASQRVGTDOSGFAAY NRYRIGNYKLNTDHSGSNDNIALLVO	221
SARS-CoV	154 SLGRCDIKDLPKEITVATSRTLSYYKLGAASQRVGTDOSGFAAY NRYRIGNYKLNTDHAGSNDNIALLVO	221
SARS-CoV-2	155 HLGRCDIKDLPKEITVATSRTLSYYKLGAASQRVAGDSGFAAYSPYRIGNYKLNTDHSSSDNIALLVO	222
MERS-CoV	154 HFGACDYDRLPNEVTVAKPNVLIAKMKVKROSYGTNSGVAIYHRYKAGNYR - SPPI TADIE ALLRA	219



La **proteína de nucleocápside (N)** se encarga del empaquetamiento del ARN viral en la nucleocápside. Media el ensamblaje viral interaccionando con el genoma del virus y la proteína M. Es también considerada una buena diana para los fármacos. Además, si inspeccionamos su conservación en distintos coronavirus, se observa que es una de las proteínas más conservadas entre ellos (aproximadamente 90 % de identidad en la secuencia) (Figura 34D). Basándose en esta alta similitud de secuencia se piensa que los anticuerpos frente a la proteína N de SARS-CoV deberían reconocer la proteína N de SARS-CoV-2.





En el siguiente enlace podéis encontrar la estructura de cada una de las proteínas de SARS-CoV-2 descritas.

<https://www.ncbi.nlm.nih.gov/Structure/SARS-CoV-2.html>

En el siguiente vídeo se encuentra una explicación del genoma de SARS-CoV-2.

<https://www.youtube.com/watch?v=t0lurlmjwu4>

## 6.3.

Clasificación de SARS-CoV-2.  
Variantes de interés y linajes. Bases  
de datos.

Basáandonos en la secuencia nucleotídica del virus, se han desarrollado varios métodos de tipificación y clasificación de los distintos linajes que han ido apareciendo desde la irrupción de SARS-CoV-2.

- Clasificación OMS: Variantes de preocupación (VOC), variantes de interés (VOI) y variantes bajo vigilancia (VUM)
- Nextstrain: Monitorización a tiempo real, clasificación filogenética
- GISAID: el repositorio mundial de todos los genomas de SARS-CoV-2 secuenciados
- PANGO: Clasificación de linajes y constelaciones de mutaciones
- El análisis de mutaciones de interés. Bases de datos de monitorización.

### 6.3.1. Clasificación de la OMS: variantes de preocupación, variantes de interés y variantes bajo vigilancia

Todos los virus cambian con el paso del tiempo, en cada uno de los pasos de replicación, debido fundamentalmente a errores en este proceso. La mayoría de estos cambios tienen escaso o nulo efecto sobre las propiedades del virus. Sin embargo, algunos de ellos pueden influir sobre algunas de ellas, como por ejemplo su facilidad de propagación, la gravedad de la enfermedad asociada o la eficacia de las vacunas, los fármacos para el tratamiento, los medios de diagnóstico u otras medidas de salud pública y social.

La OMS, en colaboración con redes de expertos, autoridades nacionales, instituciones e investigadores, ha estado vigilando y evaluando la evolución del SARS-CoV-2 desde enero de 2020. La aparición de variantes que suponían un mayor riesgo para la salud pública mundial, especialmente a finales de 2020, hizo que se empezaran a utilizar las categorías específicas de variante de interés (VOI) y variante de preocupación (VOC), con el fin de priorizar el seguimiento y la investigación a escala mundial y, en última instancia, orientar la respuesta a la pandemia de COVID-19.

### 6.3.1. Clasificación de la OMS: variantes de preocupación, variantes de interés y variantes bajo vigilancia

El seguimiento de estos cambios tiene importancia, para que en caso de que se detecten mutaciones significativas, se pueda informar de manera rápida y precisa a los distintos países, para implementar las medidas de contención adecuadas y evitar su propagación. Las medidas recomendadas actualmente por la OMS siguen funcionando, independientemente de la variante circulante, demostrando en países con amplia transmisión de variantes preocupantes que las medidas sociales y de salud pública, como las de prevención y control de la infección, reducen eficazmente el número de casos, hospitalizaciones y muertes por COVID-19.

Los sistemas de nomenclatura establecidos para nombrar y rastrear los linajes genéticos del SARS-CoV-2 por GISAID, Nextstrain y PANGO, que veremos en apartados posteriores, se utilizan en círculos científicos y en la investigación científica. Sin embargo, a nivel de comunicación con la población y para el debate no científico, se ha desarrollado un sistema de nomenclatura basado en el alfabeto griego, con el fin de facilitar su identificación y eliminar los estigmas que puedan derivar de la utilización del lugar de detección de una variante para designarla.

### 6.3.1. Clasificación de la OMS: variantes de preocupación, variantes de interés y variantes bajo vigilancia

Semanalmente la OMS publica una actualización de la clasificación de SARS-CoV-2, la distribución geográfica de las variantes preocupantes y los resúmenes de sus características fenotípicas (transmisibilidad, gravedad de la enfermedad, riesgo de reinfección e impactos en el diagnóstico y la eficacia de la vacuna) basada en los estudios más recientes publicados.

Para una actualización a tiempo real se puede consultar:

<https://www.who.int/es/activities/tracking-SARS-CoV-2-variants/tracking-SARS-CoV-2-variants>

### 6.3.1. Clasificación de la OMS: variantes de preocupación, variantes de interés y variantes bajo vigilancia

#### Variante preocupante (variant of concern, VOC)

Se trata de una variante que ha mostrado uno o varios cambios significativos en los siguientes criterios:

- Aumento de la transmisibilidad o cambio perjudicial en la epidemiología de la COVID-19.
- O aumento de la virulencia o cambio en la presentación clínica de la enfermedad.
- O disminución de la eficacia de las medidas sociales y de salud pública o de los medios de diagnóstico, las vacunas y los tratamientos disponibles.

### 6.3. Clasificación de SARS-CoV-2

#### Currently circulating variants of concern (VOCs):

WHO label	Pango lineage*	GISAID clade	Nextstrain clade	Additional amino acid changes monitored <sup>o</sup>	Earliest documented samples	Date of designation
Omicron*	B.1.1.529	GR/484A	21K, 21L, 21M, 22A, 22B, 22C, 22D	+S:R346K +S:L452X +S:F486V	Multiple countries, Nov-2021	VUM: 24-Nov-2021 VOC: 26-Nov-2021

\* Includes BA.1, BA.2, BA.3, BA.4, BA.5 and descendant lineages. It also includes BA.1/BA.2 circulating recombinant forms such as XE. WHO emphasizes that these descendant lineages should be monitored as distinct lineages by public health authorities and comparative assessments of their virus characteristics should be undertaken.

#### Previously circulating VOCs:

WHO label	Pango lineage*	GISAID clade	Nextstrain clade	Earliest documented samples	Date of designation
Alpha	B.1.1.7	GRY	20I (V1)	United Kingdom, Sep-2020	VOC: 18-Dec-2020 Previous VOC: 09-Mar-2022
Beta	B.1.351	GH/501Y.V2	20H (V2)	South Africa, May-2020	VOC: 18-Dec-2020 Previous VOC: 09-Mar-2022
Gamma	P.1	GR/501Y.V3	20J (V3)	Brazil, Nov-2020	VOC: 11-Jan-2021 Previous VOC: 09-Mar-2022
Delta	B.1.617.2	G/478K.V1	21A, 21I, 21J	India, Oct-2020	VOI: 4-Apr-2021 VOC: 11-May-2021 Previous VOC: 7-Jun-2022

## 6.3. Clasificación de SARS-CoV-2

### Variantes preocupantes (VOC) actualmente en circulación, a 15 de marzo de 2023

Nota: Para reflejar mejor el panorama actual de variantes, dominado por linajes descendientes de la variante ómicron, la OMS actualizó su sistema de seguimiento y las definiciones de trabajo de VOC y VOI el 15 de marzo de 2023.



#### XBB.1.5 Updated Risk Assessment, 24 February 2023

XBB.1.5 is a descendant lineage of XBB, which is a recombinant of two BA.2 descendant lineages. Previous risk assessments can be found here.<sup>1,2</sup>

From 22 October 2022 to 21 February 2023, 45 193 sequences of the Omicron XBB.1.5 variant have been made available from 74 countries. Most of these sequences are from the United States of America (72.2%). The other countries include the United Kingdom (7.3%), Canada (5.0%), Germany (2.7%), Austria (1.8%), Denmark (1.1%), and France (1.0%).

Based on its genetic characteristics and available growth rate estimates, XBB.1.5 is likely to further contribute to increases in case incidence globally. There is high-strength of evidence for increased risk of transmission and moderate-strength of evidence for immune escape. The number of cases associated with XBB.1.5 is still low in many countries, and from reports by several countries, no early signals of changes or increases in severity have been observed. At this time, because there is limited data currently available globally, a full assessment of the severity of XBB.1.5 cannot yet be confidently assessed. Taken together, available information does not suggest that XBB.1.5 has additional public health risk relative to the other currently circulating Omicron descendant lineages.

[https://www.who.int/docs/default-source/coronavirus/22022024xbb.1.5ra.pdf?sfvrsn=7a92619e\\_3](https://www.who.int/docs/default-source/coronavirus/22022024xbb.1.5ra.pdf?sfvrsn=7a92619e_3)

## Declaración sobre la actualización de las definiciones de trabajo y del sistema de seguimiento de las variantes preocupantes y las variantes de interés del SARS-CoV-2

16 de marzo de 2023 | Declaración

La OMS ha actualizado su sistema de seguimiento y las definiciones de trabajo de las variantes del SARS-CoV-2, el virus causante de la COVID-19, para reflejar mejor la situación actual de las variantes a nivel mundial, evaluar de forma independiente los sublinajes de la variante ómicron en circulación y clasificar con mayor claridad las nuevas variantes cuando sea necesario.

El SARS-CoV-2 sigue evolucionando. Desde el inicio de la pandemia de COVID-19, la OMS ha designado múltiples variantes preocupantes y variantes de interés en función de la evaluación de su potencial para propagarse, sustituir las variantes anteriores y provocar nuevas oleadas de infecciones con un mayor alcance, así como de la necesidad de ajustar las medidas de salud pública.

Basándose en comparaciones de la reactividad cruzada antigenética utilizando sueros animales, estudios de replicación en modelos experimentales de las vías respiratorias humanas y pruebas de estudios clínicos y epidemiológicos en humanos, existe consenso entre los expertos del Grupo Consultivo Técnico sobre la Evolución del Virus SARS-CoV-2 en que, en comparación con las variantes anteriores, la variante ómicron representa la variante preocupante más divergente vista hasta la fecha. Desde su aparición, los virus ómicron han seguido evolucionando en los planos genético y antigenético y han dado lugar a una gama cada vez mayor de sublinajes, todos los cuales se han caracterizado hasta ahora por su capacidad de evadir la inmunidad existente de la población y por infectar preferentemente las vías respiratorias superiores (frente a las inferiores), en comparación con las variantes preocupantes anteriores a la ómicron.

Más del 98% de las secuencias genéticas públicamente disponibles desde febrero de 2022 corresponden a virus ómicron, y estos virus constituyen el fondo genético a partir del cual es más probable que surjan nuevas variantes del SARS-CoV-2, aunque sigue siendo posible la aparición de variantes derivadas de variantes preocupantes anteriormente en circulación o de variantes completamente nuevas. En el sistema anterior, todos los sublinajes ómicron se clasificaban en la categoría de variantes preocupantes ómicron, lo que no ofrecía la granularidad necesaria para comparar los nuevos linajes descendientes con fenotipos alterados con los linajes parentales de ómicron (BA.1, BA.2, BA.4/BA.5). Por tanto, a partir del 15 de marzo de 2023, los sublinajes ómicron serán clasificados de manera independiente en el sistema OMS de seguimiento de variantes como variantes bajo vigilancia, variantes de interés o variantes preocupantes.

La OMS también está actualizando las definiciones de trabajo de variante preocupante y variante de interés. La principal actualización consiste en hacer más específica la definición de variante preocupante para incluir los principales pasos evolutivos del SARS-CoV-2 que requieren importantes intervenciones de salud pública. Las definiciones actualizadas se pueden consultar en el [sitio web de seguimiento de variantes de la OMS](#).

Además, a partir de ahora, la OMS seguirá nombrando las variantes preocupantes con letras del alfabeto griego, pero ya no lo hará con las variantes de interés.

Teniendo en cuenta estos cambios, las variantes alfa, beta, gamma y delta, así como el linaje parental de la variante ómicron (B.1.1.529), se consideran variantes preocupantes anteriormente en circulación. La OMS ha clasificado ahora la variante XBB.1.5 como una variante de interés.

La OMS también seguirá publicando periódicamente evaluaciones de riesgos con respecto tanto a las variantes de interés como a las variantes preocupantes ([véase la última evaluación de riesgos respecto de la variante XBB.1.5](#)).

La OMS hace hincapié en que estos cambios no implican que la circulación de los virus ómicron ya no suponga una amenaza para la salud pública. Más bien, los cambios se han introducido para identificar mejor las amenazas adicionales o nuevas además de las que plantean los virus ómicron actualmente en circulación.

### 6.3.1. Clasificación de la OMS: variantes de preocupación, variantes de interés y variantes bajo vigilancia

#### Variante de interés (variant of interest, VOI)

Son variantes del SARS-CoV-2 que:

- Presentan cambios en el genoma que, según se ha demostrado o se prevé, afectan a características del virus como su transmisibilidad, la gravedad de la enfermedad que causa y su capacidad para escapar a la acción del sistema inmunitario, ser detectado por medios diagnósticos o ser atacado por medicamentos.
- Y además, según se ha comprobado, dan lugar a una transmisión significativa en medio extrahospitalario o causan varios conglomerados de COVID-19 en distintos países, con una prevalencia relativa creciente y ocasionando números cada vez mayores de casos con el tiempo, o bien que presentan, aparentemente, otras características que indiquen que pueden entrañar un nuevo riesgo para la salud pública mundial.

Linaje Pango	Clado Nextstrain	Características genéticas	Primeras muestras documentadas	Fecha de designación y evaluaciones del riesgo
				11-01-2023
XBB.1.5	23A	Variante recombinante de los sublinajes BA.2.10.1 y BA.2.75, es decir, BJ.1 y BM.1.1.1, con un punto de rotura en S1	21-10-2022	XBB.1.5 Evaluación rápida del riesgo, 11 de enero de 2023
		XBB.1 + S:F486P (perfil genético de la espícula similar al de XBB.1.9.1)		XBB.1.5 Actualización de la evaluación del riesgo, 24 de febrero de 2023
				XBB.1.5 Actualización de la evaluación del riesgo, 27 de junio de 2023
XBB.1.16	23B	Variante recombinante de los sublinajes BA.2.10.1 y BA.2.75, es decir, BJ.1 y BM.1.1.1  XBB.1 + S:E180V, S:K478R and S:F486P	09-01-2023	17-04-2023  <u>Evaluación inicial de riesgos de la variante XBB.1.16 – 17 de abril de 2023</u>  <u>XBB.1.16 Actualización de la evaluación del riesgo, 5 de junio de 2023</u>
EG.5	No asignado	XBB.1.9.2 + S:F456L  Incluye EG.5.1 (23F): EG.5 + S:Q52H	17-02-2023	09-08-2023  <u>Evaluación inicial de riesgos de la variante EG.5, 9 de agosto de 2023</u>

## 6.3. Clasificación de SARS-CoV-2

### 6.3.1. Clasificación de la OMS: variantes de preocupación, variantes de interés y variantes bajo vigilancia

#### Variante bajo vigilancia (variant under monitoring, VUM)

Cualquier variante del SARS-CoV-2 que presente modificaciones en el genoma que, según se sospeche, puedan afectar a las características del virus y parezcan indicar que la variante puede entrañar riesgos en el futuro, a pesar de que no se disponga de pruebas claras de los cambios que pueda causar en el fenotipo o en las características epidemiológicas del virus y sea necesario mantener el seguimiento y continuar estudiándola hasta que no se disponga de más información.

#### Variantes bajo vigilancia (VUM) actualmente en circulación, a 17 de agosto de 2023

Linaje Pango	Clado Nextstrain	Características genéticas	Primeras muestras documentadas	Fecha de designación
BA.2.75	22D	BA.2 + S:K147E, S:W152R, S:F157L, S:I210V, S:G257S, S:D339H, S:G446S, S:N460K, S:Q493R reversión	31-12-2021	06-07-2022
CH.1.1	23C	BA.2.75 + S:L452R, S:F486S	27-07-2022	08-02-2023
XBB*	22F	BA.2+ S:V83A, S:Y144-, S:H146Q, S:Q183E, S:V213E, S:G252V, S:G339H, S:R346T, S:L368I, S:V445P, S:G446S, S:N460K, S:F486S, S:F490S	19-08-2022	12-10-2022
XBB.1.9.1	23D	Variante recombinante de los sublinajes BA.2.10.1 y BA.2.75, es decir, BJ.1 y BM.1.1.1 XBB.1 + S:F486P (perfil genético de la espícula similar al de XBB.1.5)	05-12-2022	30-03-2023
XBB.1.9.2#	23D	Variante recombinante de los sublinajes BA.2.10.1 y BA.2.75, es decir, BJ.1 y BM.1.1.1 XBB.1 + S:F486P, S:Q613H	05-12-2022	26-04-2023
XBB.2.3	23E	Variante recombinante de los sublinajes BA.2.10.1 y BA.2.75, es decir, BJ.1 y BM.1.1.1 XBB + S:D253G, S:F486P, S:P521S	09-12-2022	17-05-2023
		BA.2.86 (solo hay tres secuencias disponibles. Se incluye como variante bajo vigilancia debido al gran número de mutaciones detectadas)	No asignado	Mutaciones de un posible ancestro BS.2 24-07-2023 17-08-2023

\*Se excluyen los sublinajes XBB enumerados aquí como variantes de interés (VOI) y variantes bajo vigilancia (VUM)

#Se excluyen los sublinajes XBB.1.9.2 enumerados aquí como variantes de interés (VOI) y variantes bajo vigilancia (VUM)

### 6.3.1. Clasificación de la OMS: variantes de preocupación, variantes de interés y variantes bajo vigilancia

En el siguiente enlace se encuentra un vídeo sobre las variantes de interés.

<https://www.youtube.com/watch?v=hxkoePuNUJw>

Un vídeo sobre las bases de datos que veremos en apartados posteriores.

<https://www.youtube.com/watch?v=nuecT8RMCyM>

Un vídeo explicando la importancia de la epidemiología genómica.

<https://www.youtube.com/watch?v=njS0Tw4uQUo>

### 6.3.2. Nextstrain. Monitorización de la evolución de SARS-CoV-2 a tiempo real.

Nextstrain (<https://nextstrain.org/>) es un proyecto de código abierto que fue diseñado para aprovechar el potencial científico y de salud pública de los datos de genomas de distintos patógenos. Se basa en la actualización continua de los genomas que están disponibles de manera pública, ofreciendo una visualización y análisis potente. Su objetivo es facilitar la comprensión epidemiológica y mejorar la respuesta frente a brotes epidémicos. Contienen conjuntos de datos para los patógenos de tuberculosis, dengue, ébola, enterovirus, gripe, virus de Nilo y COVID-19, entre otros.

Todo su trabajo se base en una premisa: durante el curso de una infección y una epidemia, los patógenos acumulan de manera natural mutaciones aleatorias en sus genomas. Esta es una consecuencia inevitable de la replicación del genoma, que es propenso a sufrir errores. Distintos genomas acumulan distintas mutaciones, por lo que estas pueden utilizarse como marcador de transmisión, donde los genomas estrechamente relacionados indican infecciones estrechamente relacionadas. La reconstrucción del árbol filogenético, relacionando los distintos genomas, nos permite estudiar fenómenos epidemiológicos importantes como la dispersión espacial, los tiempos de introducción y la tasa de crecimiento epidémico.

Si queremos que este análisis de secuencias del genoma del patógeno sirva para informar sobre las intervenciones de salud pública, debe realizarse rápidamente y los resultados deben difundirse ampliamente. Las prácticas actuales de publicación científica dificultan la rápida difusión de resultados epidemiológicamente relevantes. Por ello, esta web implementa un sistema en línea abierto con protocolos bioinformáticos sólidos para extraer datos de todos los grupos de investigación, mejorando la capacidad de realizar inferencias epidemiológicamente relevantes.

### 6.3.2. Monitorización de la evolución de SARS-CoV-2 a tiempo real. Nextstrain

Por lo tanto, esta web tiene como objetivo proporcionar una instantánea en tiempo real de la evolución de las poblaciones de patógenos y proporcionar visualizaciones de datos interactivas a virólogos, epidemiólogos, funcionarios de salud pública y científicos. A través de visualizaciones de datos interactivas, su objetivo es permitir la exploración de conjuntos de datos en actualización continua, proporcionando una nueva herramienta de vigilancia a las comunidades científicas y de salud pública.

La nomenclatura de los linajes de SARS-CoV-2 es variable en función de la base de datos utilizada. Como se puede observar en el árbol filogenético de la Figura, está basado en un código que determina el año (19, 20 o 21), seguido de una letra que determina el linaje exacto. En algunos casos se especifica su correlación con la nomenclatura proporcionada por la OMS, determinada por letras griegas (Alpha-Omicron, en la fecha actual). Para más información sobre el sistema de nomenclatura de Nextstrain, puedes consultar: <https://nextstrain.org/blog/2021-01-06-updated-SARS-CoV-2-clade-naming>.

En el siguiente enlace se encuentra un vídeo sobre el uso de Nextstrain.

<https://www.youtube.com/watch?v=aubtBo-dAhw>



DOCS HELP LOGIN

# Genomic epidemiology of SARS-CoV-2 with subsampling focused globally over the past 6 months

Built with nextstrain/ncov. Maintained by the Nextstrain team. Enabled by data from [GISAID](#).

Showing 3715 of 3715 genomes sampled between Dec 2019 and Aug 2023.

## Dataset

- ncov
- gisaid
- global
- 6m

## Date Range

2019-12-16 2023-08-18

## Color By

- Clade

## Filter Data

Type filter query here...

## Tree Options

### Layout

RECTANGULAR

RADIAL

UNROOTED

CLOCK

SCATTER

### Branch Length

TIME DIVERGENCE

Show confidence intervals

### Branch Labels

- clade

Show all labels

### Tip Labels

## Phylogeny

### Clade

- 20H (Beta)
- 20I (Alpha)
- 20J (Gamma)
- 21A (Delta)
- 21I (Delta)
- 21J (Delta)
- 21B (Kappa)
- 21D (Eta)
- 21H (Mu)
- 21K (BA.1)
- 21L (BA.2)
- 22A (BA.4)
- 22B (BA.5)
- 22C (BA.2.12.1)
- 22D (BA.2.75)
- 22E (BQ.1)
- 22F (XBB)
- 23A (XBB.1.5)
- 23B (XBB.1.16)
- 23C (CH.1.1)
- 23D (XBB.1.9)
- 23E (XBB.2.3)
- 23F (EG.5.1)
- 19A
- 19B
- 20A
- 20E
- 20C
- 20G
- 20B
- 20D



ZOOM TO SELECTED

RESET LAYOUT

2020 2021 2022 Date

20,000 18,000 16,000 14,000 12,000 10,000 8,000 6,000 4,000 2,000 0

## Diversity

1.2

1.0

0.8

0.6

0.4

0.2

0.0

ENTROPY EVENTS AA NT

## Geography

mapbox

Leaflet | © Mapbox © OpenStreetMap Improve this map

+ -

1

### 6.3.3. GISAID. El repositorio de todos los genomas de SARS-CoV-2 secuenciados

La iniciativa GISAID (<https://www.gisaid.org/>) promueve el intercambio rápido de datos de todos los virus Influenza y SARS-CoV-2. Esto incluye la secuencia genética y los datos clínicos y epidemiológicos asociados con virus humanos, y datos geográficos y específicos de especies asociados con virus aviares y otros virus animales, para ayudar a los investigadores a comprender cómo evolucionan y se propagan los virus durante epidemias y pandemias. En la actualidad (12 diciembre 2021), GISAID alberga más de seis millones de secuencias de SARS-CoV-2, que sirven de fuente de información para otras bases de datos como Nextstrain, anteriormente citada.

Una vez que se accede con registro a la base de datos, cada investigador puede remitir la secuencia nucleotídica del genoma de SARS-CoV-2 secuenciado en su laboratorio, junto con datos epidemiológicos básicos como la edad, sexo, condición y localización del paciente del que ha sido extraído el virus. Asimismo, se pueden depositar secuencias de SARS-CoV-2 aisladas de animales. Existe una gran cantidad de información en el apartado siguiente: <https://www.epicov.org/epi3/frontend#326ae6>

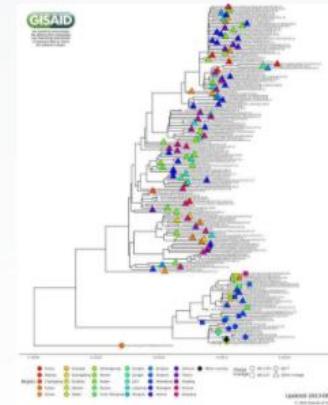
Adicionalmente, se pueden consultar todas las secuencias depositadas junto con estos datos epidemiológicos, seleccionando por localización, linaje, fecha, etc.; así como la dinámica de aparición y transmisión de variantes de interés.

## In Focus

### Since lifting the zero-COVID policy, data from China continues to resemble globally seen patterns

The genomic surveillance of data from China with preliminary [phylogenetic analyses](#) over recent months demonstrated a pattern of variant introduction and emergence risk which is substantially similar to that observed globally. Therefore, the daily updates will no longer be shown in the In Focus section. Future GISAID analyses for China will be viewable on the dedicated page accessible [here](#).

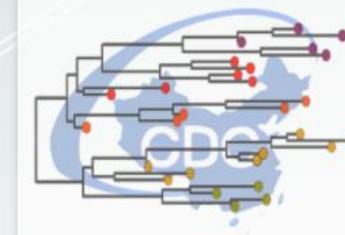
Global public health relies on timely genomic surveillance efforts in all countries and regions of the world to detect new evolutionary trends early on.



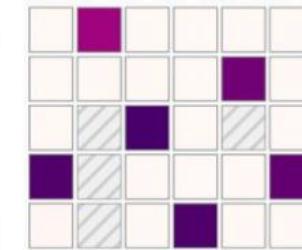
Representative genome sequences from China in the global context



## 中国新冠疫情



## Lineage comparison



## hCoV-19 data sharing via GISAID

**15,895,364**

genome sequence submissions

Submission Tracker	Phyldynamics	Tracking Variants	Frequency Dashboards
hCoV-19 Global	hCoV-19	hCoV-19 Variants	hCoV-19
hCoV-19 USA	hMpxV	hMpxV Variants	hMpxV
hMpxV	RSV	Influenza Subtypes	Influenza
RSV		RSV Subtypes	RSV

## GISAID Resources

## Data Acknowledgement Locator



Example ID

Register for Access Credentials

### 6.3.4. PANGO. Clasificación de linajes y constelaciones de mutaciones.

El sistema de clasificación más extendido para la monitorización de linajes de SARS-CoV-2 es el sistema PANGO, basado en un sistema de asignación filogenética y una nomenclatura basada en la combinación correlativa de letras y números.

La siguiente página web documenta todos los linajes actuales de SARS-CoV-2, con parámetros relacionados con su aparición y propagación, así como varias herramientas de software que los investigadores pueden utilizar para realizar análisis sobre la secuencia del virus

<https://cov-lineages.org/>

La explicación del sistema de clasificación de linajes está detallada, junto con las últimas novedades en él, en la web:

<https://www.pango.network/>

### 6.3.4. Clasificación de linajes y constelaciones de mutaciones: PANGO

Este sistema fue propuesto por primera vez en abril 2020 (Rambaut et al., 2020). El sistema es dinámico y flexible, pudiéndose adaptar a la naturaleza cambiante de esta pandemia y crecer conforme crezcan las evidencias genómicas de SARS-CoV-2. Cada linaje de PANGO define un grupo de secuencias del genoma de SARS-CoV-2 y se crea de acuerdo a dos principios:

1. Los linajes PANGO significan grupos o clústeres de infecciones que comparten un ancestro común. Si pensamos en la pandemia como un árbol que va ramificándose en cada transmisión, los linajes PANGO representan las ramas individuales dentro de ese árbol.
2. Los linajes PANGO están destinados a resaltar eventos epidemiológicamente relevantes, como la aparición del virus en una nueva ubicación, un rápido aumento de casos o la evolución del virus hacia fenotipos distintos.

### 6.3.4. Clasificación de linajes y constelaciones de mutaciones: PANGO

Los criterios actuales para la creación de nuevos linajes PANGO constan de una serie de criterios, definidos en el siguiente enlace y que incluyen unos estándares mínimos de tamaño de linaje, calidad del genoma, distinción genética e importancia epidemiológica. Estos criterios son cambiantes en el tiempo, para adaptarse a las necesidades y circunstancias cambiantes de esta pandemia.

<https://www.pango.network/the-pango-nomenclature-system/statement-of-nomenclature-rules/>

Este sistema jerárquico de nomenclatura se refleja en la forma de nombrar cada linaje. Cada uno de ellos recibe un código alfanumérico único que incluye información parcial, pero no completa, sobre la historia filogenética del mismo. Esta nomenclatura es un compromiso entre la comprensión humana y la legibilidad para los sistemas informáticos.

Los linajes pango actuales se muestran en el apartado: [https://cov-lineages.org/lineage\\_list.html](https://cov-lineages.org/lineage_list.html)

## Lineage List

This is a list of active lineages which have been seen in the last year. A complete list of lineages can be found on [the PANGO designation GitHub](#) and a JSON file containing the equivalent data from this page for the full set of lineages can be downloaded in JSON format [here](#).

All Fields

Search for lineage...

Lineage	Most common countries	Earliest date	# designated	# assigned	Description	WHO Name
A.5	Spain 62.0%, United Kingdom 14.0%, Uruguay 5.0%, Peru 2.0%, Portugal 2.0%	2020-03-01	439	525	A lineage with a lot of representation from Spanish-speaking countries. A Spanish/ South-American lineage, but now with sequences from an outbreak in Scotland. Also now includes what was previously A.10.	
B	United States of America 40.0%, United Kingdom 15.0%, China 7.0%, Mexico 6.0%, Germany 3.0%	2019-12-24	4001	10236	One of the two original haplotypes of the pandemic (and first to be discovered)	
B.1	United States of America 46.0%, Turkey 11.0%, United Kingdom 6.0%, Canada 4.0%, France 3.0%	2020-01-01	46228	118911	A large European lineage the origin of which roughly corresponds to the Northern Italian outbreak early in 2020.	
B.1.1	United Kingdom 26.0%, United States of America 16.0%, Japan 6.0%, Russia 6.0%, Turkey 5.0%	2020-02-03	22790	51416	European lineage with 3 clear SNPs `28881GA` `28882GA` `28883GC`	

### 6.3.4. Clasificación de linajes y constelaciones de mutaciones: PANGO

La clasificación PANGO propone un sistema de **constelaciones de mutaciones**, donde se entiende por constelación un conjunto de mutaciones que son significativas y que han podido emerger de manera independiente, en procesos evolutivos independientes, varias veces.

La definición y detalles de este sistema se encuentra en: <https://github.com/cov-lineages/constellations>

Las constelaciones definidas se encuentran en el siguiente enlace (Figura 38) y comprenden aquellos linajes de preocupación y regiones RBD de interacción con el receptor ACE2 de interés por sus mutaciones.

<https://cov-lineages.org/constellations.html>

# Constellations

A constellation is a collection of mutations which are functionally meaningful, but which may arise independently a number of times.

More information about constellations as well as their definitions can be found in the [Constellations github repo](#).

There we include files that define:

- lineages of concern - for these lineages, the defining mutations have been extracted so that individual sequences can be classified deterministically
- genetically interesting regions e.g. RBD sites which interact with ACE2

## Summary of Currently Defined Constellations

Label	Description	Sources	Type	Variant	Tags	Sites	Rules
Omicron (Unassigned)	B.I.1529 lineage defining mutations		variant	Pango_lineages: B.I.1529 WHO_label: Omicron mrca_lineage: None representative_genome: lineage_name: B.I.1529 incompatible_lineage_calls:	B.I.1529	nuc:C3037T, orflab:T3255I, orflab:P3395H, orflab:P4715L, orflab:I5967V, spike:G142D, spike:G339D, spike:S373P, spike:S375F, spike:N440K, spike:S477N, spike:T478K, spike:E484A, spike:Q498R, spike:N501Y, spike:Y505H, spike:D614G, spike:H655Y, spike:N679K, spike:P681H, spike:N764K, spike:D796Y, spike:Q954H, spike:N969K, e:T9I, m:Q19E, m:A63T, nuc:A27259C, nuc:C27807T, nuc:A28271T, n:Pi3L, del:28362:9	default: [object Object] Probable: [object Object]
Alpha (B.I.1.7-like)	B.I.1.7 lineage defining mutations	<a href="https://virological.org/t/563">https://virological.org/t/563</a>	variant	Pango_lineages: B.I.1.7 mrca_lineage: B.I.1.7 PHE_label: VOC-20DEC-01 WHO_label: Alpha representative_genome:	B.I.1.7, VOC 202012/01	nuc:C913T, lab:T100I, lab:AI708D, nuc:C5986T, lab:I2230T, lab:SGF3675-, nuc:C14676T, nuc:C15279T, nuc:T16176C, s:HV69-, s:Y144-, s:N501Y, s:A570D, s:P681H, s:T716I, s:S982A, s:DII18H, nuc:C2680I, 8:Q27*, 8:R52I, 8:Y73C, N:D3I, NS:235F	min_alt: 15 max_ref: 3
Beta (B.I.351-like)	Defining of lineage B.I.351	<a href="https://www.medrxiv.org/content/10.1101/2020.12.21.20248640v1">https://www.medrxiv.org/content/10.1101/2020.12.21.20248640v1</a>	variant	Pango_lineages: B.I.351 mrca_lineage: B.I.351 PHE_label: VOC-20DEC-02 WHO_label: Beta representative_genome:	B.I.351, VOC-20DEC-02, VOC202012/02, Beta, 50I.V2, 20H, GH	NSP2:T85I, ORFlab:K1655N, ORFlab:K3353R, S:D80A, S:D215G, S:E484K, S:N501Y, S:A701V, ORF3a:Q57H, ORF3a:S171L, E:P71L, N:T205I, del:22280:9, del:II287:9	min_alt: 6 max_ref: 3
Delta (B.I.617.2-like)	Defining constellation for lineage B.I.617.2	<a href="https://github.com/cov-lineages/pango-designation/issues/49">https://github.com/cov-lineages/pango-designation/issues/49</a>	variant	Pango_lineages: B.I.617.2,AY.1,AY.2 mrca_lineage: B.I.617.2 lineage_name: B.I.617.2 incompatible_lineage_calls: AY.1,AY.2,AY.4,AY.4.2	B.I.617.2, VOC-21APR-02, Delta, VUI-21APR-02	S:T19R, S:G142D, S:L452R, S:T478K, S:P681R, S:D950N, ORF3a:S26L, M:J82T, ORF7a:V82A, ORF7a:T120I, N:D63G, N:R203M, N:D377Y	min_alt: 5 max_ref: 3

### 6.3.4. Clasificación de linajes y constelaciones de mutaciones: PANGO

Herramientas software que incluye PANGO y que ayudan a los investigadores en el análisis y clasificación de sus secuencias: <https://cov-lineages.org/resources.html>

Finalmente, PANGO incluye una serie de herramientas software que ayudan a los investigadores en el análisis y clasificación de sus secuencias. Entre ellas destacan:

- **Pangolin** (Phylogenetic Assignment of Named Global Outbreak Lineages) (O'Toole et al., 2021). Esta herramienta fue diseñada para implementar esta nomenclatura dinámica, bien en una herramienta por la línea de comandos (CLI) como con una aplicación web. **Asiste al usuario en la asignación del linaje más probable.** Disponible en: <https://cov-lineages.org/resources/pangolin.html>
- **Scorpio** (Serious Constellations of Reoccurring Phylogenetically-Independent Origin). Es una **herramienta, contenida en el programa Pangolin por línea de comandos que permite analizar los SNP de las variantes de preocupación y asignar la constelación a la que pertenecen.** Disponible en: <https://github.com/cov-lineages/scorpio>
- **Pando.** **Permite la interacción con el árbol filogenético de SARS-CoV-2 global.** Disponible en: <http://pando.tools/>
- **Civet** (Cluster Investigation and Virus Epidemiology Tool). Este software se ha diseñado pensando en un **análisis genómico a tiempo real. Utilizando una filogenia de fondo,** como la disponible a través del consorcio COG-UK en CLIMB, **este software genera un informe para un conjunto de secuencias de interés, permitiendo el análisis de un brote en detalle.** Civet coloca las nuevas secuencias en el contexto de la diversidad de fondo, que es conocida. Disponible en: <https://cov-lineages.org/resources/civet.html>
- **Polecat** (Phylogenetic Overview & Local Epidemiological Cluster Analysis Tool). De manera similar a Civet, **utilizando una filogenia de fondo, identificará y marcará los grupos o clústeres.** Disponible en: <https://github.com/artic-network/polecat>

## Resources

### Internally Developed Tools



SOFTWARE

#### Pangolin

Pangolin was developed to implement the dynamic nomenclature of SARS-CoV-2 lineages, known as the Pango nomenclature. It allows a user to assign a SARS-CoV-2 genome sequence the most likely lineage (Pango lineage) to SARS-CoV-2 query sequences.

Pangolin assigns lineages to query sequences as described in Rambaut et al 2020.

[View](#)

SOFTWARE

#### Scorpio

Serious constellations of reoccurring phylogenetically-independent origin. A tool for SNP-based calling of variants of concern.

[View](#)

SOFTWARE

#### Pando

View and interact with the latest global SARS-CoV-2 phylogenetic tree

[View](#)

SOFTWARE

#### Civet

Civet is a tool developed with 'real-time' genomics in mind.

Using a background phylogeny, such as the large phylogeny available through the COG-UK infrastructure on CLIMB, civet will generate a report for a set of sequences of interest i.e. an outbreak investigation.

[View](#)

SOFTWARE

#### Polecat

Using a background phylogeny, such as the large phylogeny available through the COG-UK infrastructure on CLIMB, polecat will identify and flag clusters based on various configurable statistics.

[View](#)

WEBSITE

#### pango.network

A website documenting the Pango nomenclature as well as the policies involved with designating new lineages.

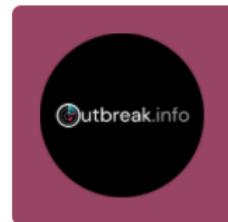
[View](#)

### 6.3.5. El análisis de mutaciones de interés. Bases de datos de monitorización

Existen varios recursos web que actúan como cuadros de mando para la monitorización de los linajes de interés VOC, VOI y VUM, así como de los datos de rastreo epidemiológico disponibles. Algunas de ellas se mencionan a continuación, como referencia para los estudios de los genomas:

- Outbreak.info. Especialmente útil para visualizar la comparativa entre las mutaciones de distintos linajes de interés, así como datos sobre su expansión en todo el mundo. Disponible en: <https://outbreak.info/>
- CoVariants (Figura 41). Además de la monitorización ofrecida por país y por variante, muestra una tabla interesante sobre la conversión de las nomenclaturas de los linajes entre los clados Nextstrain, linaje PANGO y la etiqueta de la OMS. Asimismo, se encuentra una tabla muy útil sobre las mutaciones compartidas entre distintos linajes (Figura 43). Disponible en: <https://covariants.org/>
- covSPECTRUM. Es un cuadro de mandos interactivo donde se muestran las variantes del virus circulante en cada país. Disponible en: <https://cov-spectrum.org/explore/Spain/AllSamples/AllTimes>
- JHU Covid-19 Dashboard. Cuadro de mandos desarrollado por la universidad Johns Hopkins, donde se encuentran datos epidemiológicos actualizados de todo el mundo, incluyendo el acumulado de casos totales, fallecimientos totales, vacunas administradas e incidencia. Disponible en: <https://coronavirus.jhu.edu/map.html>
- COG Mutation Explorer. Desarrollado por el consorcio británico de secuenciación de SARS-CoV-2, líder indiscutible en la monitorización genómica del virus, proponen una herramienta visual para explorar las variantes y mutaciones relevantes. Disponible en: <https://sars2.cvr.gla.ac.uk/cog-uk/>

## Dashboards



SU, WU, AND ANDERSEN LABS AT SCRIPPS RESEARCH

**Outbreak.info**

A data dashboard for viewing more information about the various SARS-CoV-2 variants

[View](#)

EMMA HODCROFT, UNIVERSITY OF BERN, SWITZERLAND

**CoVariants**

A dashboard giving an overview of SARS-CoV-2 variants and interesting mutations

[View](#)

COMPUTATIONAL EVOLUTION GROUP AT ETH ZURICH, SWITZERLAND

**covSPECTRUM**

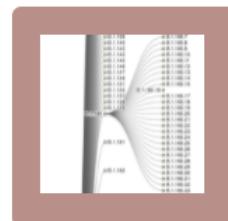
An interactive dashboard SARS-CoV-2 variants with sophisticated queries

[View](#)

CENTER FOR SYSTEMS SCIENCE AND ENGINEERING AT JOHNS HOPKINS UNIVERSITY

**JHU Covid-19 Dashboard**

An interactive dashboard displaying various information about the general spread of Covid-19.

[View](#)

KIERAN LAMB @ CVR GLASGOW

**Lineage Tree**

An interactive visualisation representing the Pango lineage system through a collapsable tree structure.

[View](#)

COG-UK

**COG Mutation Explorer**

A data visualisation tool for exploring important SARS-CoV-2 variants and mutations

[View](#)

## Software Tools



ZUHER JAHSHAN, LEONID YAVITS

### CoViT

A tool that enables real-time phylogenetics for the SARS-CoV-2 pandemic using Vision Transformers

[View](#)

CENTRE FOR GENOMIC PATHOGEN SURVEILLANCE

### PanGULLin

A web application version of the Pangolin Tool

[View](#)

TREVOR BEDFORD ET AL

### Nextstrain

Powerful, interactive tools for exploring virus data. Nextstrain provides a large database of viral genomes and bioinformatics pipelines for phylogenetic analysis, including interactive tree visualisations.

[View](#)

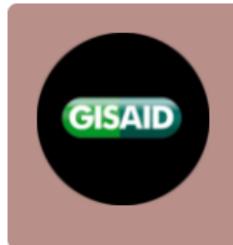
ANGIE HINRICHES, UC SANTA CRUZ GENOMICS INSTITUTE

### UShER Phylogenetic Placement

Upload your SARS-CoV-2 sequence (FASTA or VCF file) to find the most similar complete, high-coverage samples from GISAID or from public sequence databases (NCBI Virus / GenBank, COG-UK and the China National Center for Bioinformation), and your sequence's placement in the UCSC/UShER phylogenetic tree.

[View](#)

## Data Sources



GISAID

### GISAID

Important, global data repository for SARS-CoV-2 genomes as well as for other pathogens

[View](#)

ENA

### European Nucleotide Archive

The ENA provides a high quality, downloadable data set of SARS-CoV-2 genomes. ENA data is available as a data source on Galaxy.

[View](#)

COG-UK

### COG-UK

COG-UK provides a large number of data sets and bioinformatics tools for SARS-CoV-2.

[View](#)

## 6.4.

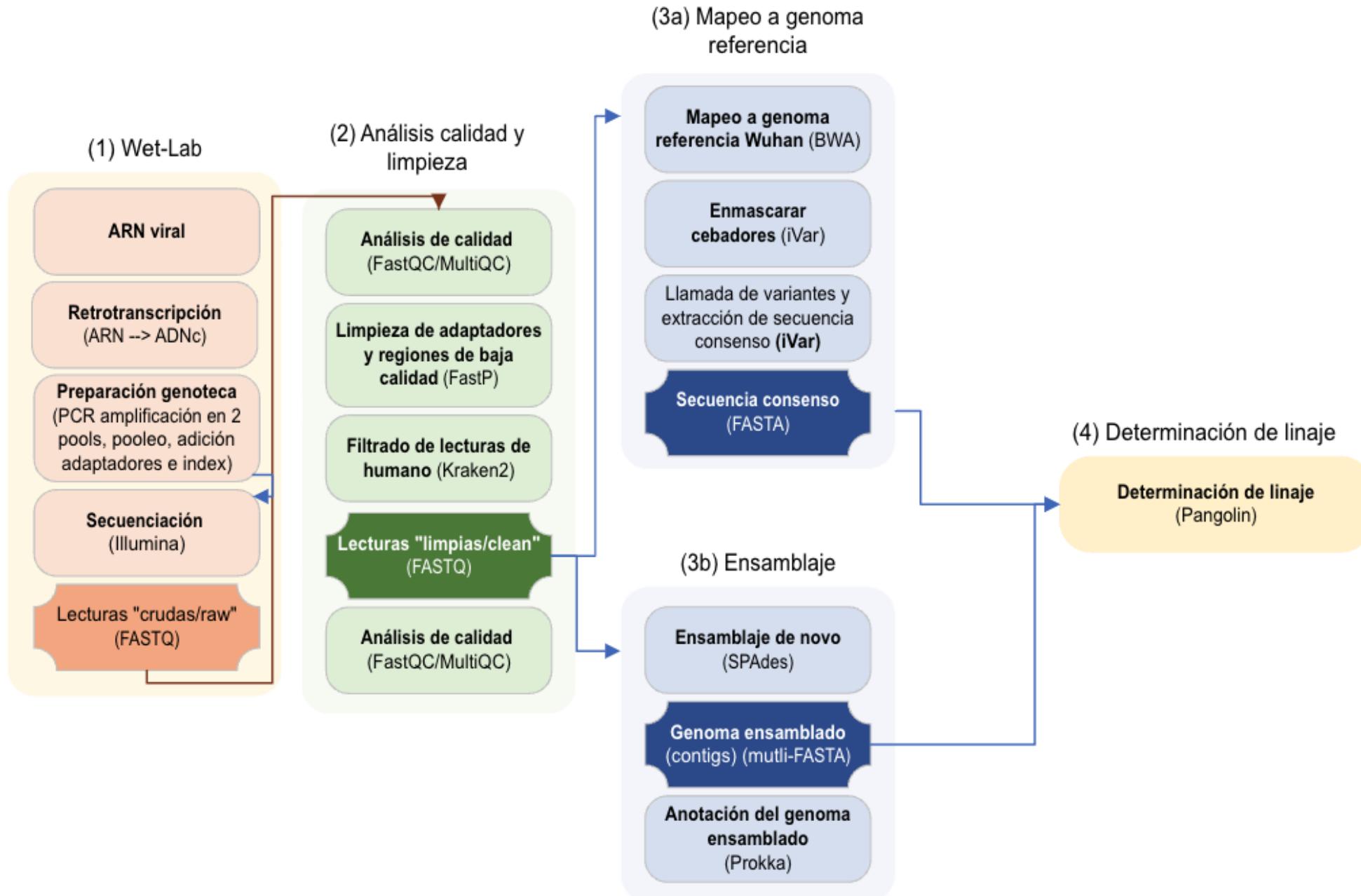
Análisis bioinformático de linajes de  
SARS-CoV-2

Lo primero que debemos tener en cuenta es que existen dos maneras principales de analizar una muestra nasofaríngea de un paciente infectado por SARS-CoV-2.

- En primer lugar, se trataría de realizar una **secuenciación completa del ARN total** contenido en la muestra, **para posteriormente ensamblar de novo e identificar los contigs ensamblados**. En este método, se llama **secuenciación shotgun o en este caso, metatranscriptoma**, por analizar el ARN total de una muestra. La mayor problemática es que la mayor parte de la muestra analizada corresponderá a ARN del hospedador (humano), e incluso a otros virus que puedan estar contenidos en la muestra y formar parte del viroma del individuo. Esta fue la técnica que se empleó para el análisis de las primeras muestras de SARS-CoV-2 detectadas, a partir de las cuales se obtuvo este nuevo patógeno.
- Actualmente, y de manera rutinaria, sometemos el ARN total extraído del hisopo nasofaríngeo a un **preprocesamiento para enriquecer nuestra secuenciación sobre el virus de interés**. A grandes rasgos, se realiza en primer lugar una retrotranscripción para obtener el ADN complementario, seguido de **dos PCR independientes**, con cebadores diseñados sobre el genoma de referencia modelo. De esta forma, **se amplifica el genoma del virus en dos pools independientes que son purificados y unidos en un único pool**. En este paso se procura enriquecer la mezcla en el genoma diana de estudio, para evitar contaminaciones por ARN humano u otros virus. Finalmente, **se adicionan adaptadores y los índices de secuenciación**. Este es conocido como **protocolo ARTIC** y está siendo ampliamente utilizado por equipos de análisis genómico de todo el mundo, por su bajo coste, versatilidad y adaptabilidad (se van adaptando los cebadores para no perder sensibilidad con las nuevas mutaciones que van apareciendo).

*El protocolo ARTIC de secuenciación está disponible en el siguiente enlace, con actualizaciones semanales.*

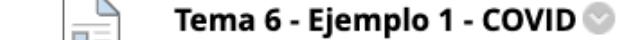
<https://www.protocols.io/view/ncov-2019-sequencing-protocol-bp2l6n26rgqe/v1>



<b>Instalación Mamba:</b>	conda install mamba -n base -c conda-forge				
Environment*	Programa	Comando descarga	Versión	Utilidad	Anotaciones (extras a instalar, referencias...)
<b>04MBIF_COVID</b>	<b>mamba create -n 04MBIF_COVID python=3.8</b> <b>mamba activate 04MBIF_COVID</b>				
	FastQC	mamba install -c bioconda fastqc	v0.12.1	Análisis primario de calidad	
	FastP	mamba install -c bioconda fastp	v0.23.4	Limpieza de adaptadores y trimming	
	kraken2	mamba install -c bioconda kraken2	v2.1.3	mapeo a base de datos de genomas	
	seqtk	mamba install -c bioconda seqtk	v1.4	herramientas de manejo de secuencias fasta	<a href="https://github.com/lh3/seqtk">https://github.com/lh3/seqtk</a>
	pangolin	mamba install -c bioconda pangolin	v4.3	asignación linajes SARS-CoV-2	
	BWA	mamba install -c bioconda bwa	v0.7.17	mapeo de lecturas sobre genoma de referencia	
	iVar + samtools	mamba install -c bioconda ivar	v1.4.2 + v1.19		
	Spades	mamba install -c bioconda spades	v3.15.5	Ensamblaje de novo	
	Prokka	mamba install -c conda-forge -c bioconda -c defaults prokka	v1.14.6	Anotación de genomas	Antes de utilizarlo <i>puede</i> que sea necesario configurar las variables export PERL5LIB=\$CONDA_PREFIX/lib/perl5/site_perl/5.22.0/ conda env config vars set PERL5LIB=\$CONDA_PREFIX/lib/perl5/site_perl/5.22.0/-n 04MBIF_COVID conda deactivate conda activate 04MBIF_COVID
<b>conda env export --file=04MBIF_COVID.yml</b>					



## Tema 6 - Ejemplo 1 - COVID



Como en cualquier protocolo bioinformático, los pasos iniciales constan del análisis de calidad de las lecturas, así como su limpieza de adaptadores y regiones de baja calidad. En este caso, debido a la naturaleza de la muestra, en la que se ha podido arrastrar ARN del hospedador (humano), se realiza un paso de eliminación de éste, mediante la herramienta bioinformática Kraken2 (Kraken2, n.d.; Wood et al., 2019)

Las lecturas para iniciar este ejemplo están disponibles [Tema6\\_Ejemplo1\\_Qualify](#)

Los pasos iniciales para realizar son los siguientes:

Entorno de trabajo: 04MBIF\_COVID

**1) Análisis de calidad con FastQC sobre las lecturas crudas obtenidas del secuenciador:** ¿De qué longitud son las secuencias? ¿tienen adaptadores?

```
fastqc *.fastq.gz
```

**2) Limpieza de adaptadores y regiones de baja calidad con FastP.** Revisa el archivo de salida de FastP. ¿se han detectado adaptadores?

```
fastp -i 2022_sample1.R1.fastq.gz -I 2022_sample1.R2.fastq.gz -o out1.fastq.gz -O out2.fastq.gz --detect_adapter_for_pe --cut_tail 25 --cut_front 25 --cut_mean_quality 25 --html out.fastp.html
```

**3) Análisis de calidad con FastQC sobre las lecturas limpias obtenidas del secuenciador:** ¿De qué longitud son las secuencias? ¿tienen adaptadores? ¿ha mejorado la calidad?

```
fastqc out*.fastq.gz
```

**4) Filtrado con Kraken2 del genoma humano.** La base de datos a utilizar es la que está precompilada en la web de Kraken2 (Kraken2, n.d.), disponible a través del link de descarga. Su tamaño es de 5.5 Gb. Para facilitar, usad la base de datos kraken\_h+v suministrada en el link de descarga. Esta base solo tiene el genoma humano GRCh38 y una base de datos de virus (incluyendo SARS-CoV-2, por ello es más ágil en el cómputo). Esta base de datos se ha construido de la siguiente forma:

<https://genexa.ch/sars2-bioinformatics-resources/>

Kraken version 2.0.8-beta:

Wood, D.E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 20, 257 (2019). <https://doi.org/10.1186/s13059-019-1891-0>

<https://ccb.jhu.edu/software/kraken2/index.shtml>

Commands generating human + virus db:  
\*\*\*\*\*

(Run on March 19th 2020)

```
kraken2-build --download-library human --db kraken_h_v
kraken2-build --download-library viral --db kraken_h_v

kraken2-build --download-taxonomy --db kraken_h_v

kraken2-build --build --db kraken_h_v
```

i. Desempaquetamos la base de datos que nos hemos bajado del link (ocupa bastante!!!):

```
tar -xvzf kraken2_h+v_20200319.tar.gz
```

i. Mapeo Kraken2 a la base de datos, donde se clasifican todas las lecturas de la muestra. ¿Cuántas secuencias corresponden a humano? ¿Cuántas al virus?

```
kraken2 --db kraken_db/ --paired out1.fastq.gz out2.fastq.gz --output sample1.kraken --threads 2 --gzip-compressed
```

i. De este archivo de salida de texto, extraemos las lecturas que NO son de humano, con la herramienta awk:

```
awk '$3 != "9606" { print $2 }' sample1.kraken > sample1.kraken.nohuman.ids
```

i. Ahora sacamos las lecturas a un nuevo archivo FASTQ y comprimimos.

```
seqtk subseq out1.fastq.gz sample1.kraken.nohuman.ids > sample1.R1.nonhuman.fq
```

```
seqtk subseq out2.fastq.gz sample1.kraken.nohuman.ids > sample1.R2.nonhuman.fq
```

```
gzip *.fq
```

Estas lecturas sin restos de humano son nuestros archivos limpios para continuar el proceso.

## Standard Kraken Output Format

Each sequence (or sequence pair, in the case of paired reads) classified by Kraken 2 results in a single line of output. Kraken 2's output lines contain five tab-delimited fields; from left to right, they are:

1. "C"/"U": a one letter code indicating that the sequence was either classified or unclassified.
2. The sequence ID, obtained from the FASTA/FASTQ header.
3. The taxonomy ID Kraken 2 used to label the sequence; this is 0 if the sequence is unclassified.
4. The length of the sequence in bp. In the case of paired read data, this will be a string containing the lengths of the two sequences in bp, separated by a pipe character, e.g. "98|94".
5. A space-delimited list indicating the LCA mapping of each *k*-mer in the sequence(s). For example, "562:13 561:4 A:31 0:1 562:3" would indicate that:
  - the first 13 *k*-mers mapped to taxonomy ID #562
  - the next 4 *k*-mers mapped to taxonomy ID #561
  - the next 31 *k*-mers contained an ambiguous nucleotide
  - the next *k*-mer was not in the database
  - the last 3 *k*-mers mapped to taxonomy ID #562

Note that paired read data will contain a " | : | " token in this list to indicate the end of one read and the beginning of another.

#### 6.4.2. Mapeo al genoma de referencia y determinación del linaje

La metodología a continuación expuesta trata de un mapeo sobre el genoma de referencia de SARS-CoV-2, para posteriormente realizar una llamada de variantes y extraer una secuencia consenso mediante el programa iVar (iVar: Manual, n.d.). Finalmente, se utilizará la herramienta Pangolin (GitHub - cov-lineages/pangolin: Software package for assigning SARS-CoV-2 genome sequences to global lineages., n.d.; O'Toole et al., 2021) para determinar el linaje del virus secuenciado.



## Tema 6 - Ejemplo 2 - Mapeo a genoma referencia

La metodología a continuación expuesta trata de un mapeo sobre el genoma de referencia de SARS-CoV-2, para posteriormente realizar una llamada de variantes y extraer una secuencia consenso mediante el programa iVar (*iVar: Manual*, n.d.). Finalmente, se utilizará la herramienta Pangolin (*Github - cov-lineages/pangolin: Software package for assigning SARS-CoV-2 genome sequences to global lineages.*, n.d.; O'Toole et al., 2021) para determinar el linaje del virus secuenciado.

Tienes disponibles las lecturas limpias para el análisis [Tema6\\_Ejemplo2\\_MapeoVariantes](#)

**1.- Descarga del genoma de referencia, desde NCBI** ([https://www.ncbi.nlm.nih.gov/assembly/GCF\\_009858895.2/](https://www.ncbi.nlm.nih.gov/assembly/GCF_009858895.2/)) en versión FASTA. Disponible en el link de descarga.

**2.- Indexado del genoma de referencia**

```
bwa index GCF_009858895.2_ASM985889v3_genomic.fna
```

**3.- Mapeo de las lecturas limpias al genoma de referencia.** En este comando complejo se incluye la reorganización del archivo SAM de salida y su conversión a archivo BAM.

```
bwa mem -Y -M -R '@RG\tID:\tSM:' -t 2 GCF_009858895.2_ASM985889v3_genomic.fna sample1.R1.nonhuman.fq.gz sample1.R2.nonhuman.fq.gz | samtools sort | samtools view -F 4 -b -@ 2 -o sample1.sort.bam
```

**4.- Indexado del archivo BAM**

```
samtools index sample1.sort.bam
```

**5.- Descargamos los cebadores ARTIC v4** (utilizados en esta versión de la preparación del protocolo). Se descarga un archivo de tipo BED desde [https://github.com/artic-network/artic-ncov2019/tree/master/primer\\_schemes/nCoV-2019/V4](https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V4). Lo tenemos disponible en el link.

**6.- Con el programa iVar se realiza la eliminación/enmascaramiento de estos primers para que no interfieran en la determinación de variantes**

```
ivar trim -i sample1.sort.bam -p sample1.trim -m 30 -q 20 -s 4 -b nCoV-2019_v4.primer.bed
```

**7.- Orden e indexado de los archivos de salida.**

```
samtools sort sample1.trim.bam -o sample1.trim.sort.bam
```

```
samtools index sample1.trim.sort.bam
```

**8.- Realizamos la llamada de variants con iVar. Necesitaremos el archivo FASTA de la referencia, pero también su anotación en formato GFF (disponible en NCBI o en el link de descarga). Obtenemos una tabla con las mutaciones.**

```
samtools mpileup -A -d 0 --reference GCF_009858895.2_ASM985889v3_genomic.fna -B -Q 0 sample1.trim.sort.bam | ivar variants -p sample1.ivar
```

**9.- Extraemos la secuencia consenso con iVar.**

```
samtools mpileup -aa -A -d 0 -B -Q 0 sample1.trim.sort.bam | ivar consensus -p sample1.ivar -q 20 -t 0.8 -m 30 -n N
```

**10.- Determinamos el linaje con Pangolin. Antes de correr el programa, actualizamos la versión de la base de datos.**

```
pangolin --update
```

```
pangolin sample1.ivar.fa -o sample1.pangolin --outfile sample1.csv -t 2 --max-ambig 0.3
```

Resultados: [Tema6\\_Ejemplo2\\_MapeoVariantes](#)



## Tema 6 - Ejemplo 3 - Ensamblaje de novo

Finalmente, de manera alternativa, se va a realizar el ensamblaje de novo y anotación del genoma secuenciado. Para ello se utilizará SPAdes en su versión para coronavirus para el ensamblaje y Prokka para la anotación. En ambos casos realizaremos ensamblaje y anotación dirigidas.

Las lecturas para el ejemplo (limpias, del ejemplo 1 de este tema) están [Tema6\\_Ejemplo3\\_Assembly](#)

**1.- Utilizando las lecturas limpias de regiones de baja calidad y restos de ARN de hospedador, realizamos su ensamblaje con SPAdes. Utilizamos como genoma de confianza el de referencia de SARS-CoV-2.**

```
spades.py --corona -1 sample1.R1.nonhuman.fq.gz -2 sample1.R2.nonhuman.fq.gz -t 2 --trusted-contigs  
GCF_009858895.2_ASM985889v3_genomic.fasta -o covid.spades
```

Nota: el archivo de referencia tiene que tener como extensión '.fasta' para que spades lo pueda reconocer como trusted contigs

**2.- Realizaremos la anotación de los scaffolds obtenidos con Prokka. Para poder dirigir la anotación, utilizamos las proteínas del genoma de referencia, que podemos descargar desde NCBI ([https://www.ncbi.nlm.nih.gov/assembly/GCF\\_009858895.2/](https://www.ncbi.nlm.nih.gov/assembly/GCF_009858895.2/)).**

```
prokka --outdir sample1.prokka --addgenes --kingdom Viruses --compliant --proteins GCF_009858895.2_ASM985889v3_protein.faa  
covid.spades/scaffolds.fasta
```

**3.- Podemos determinar el linaje sobre el archivo ensamblado con Spades.**

```
pangolin covid.spades/scaffolds.fasta -o sample1.pangolin_assembly --outfil sample1.csv -t 2 --max-ambig 0.3
```

Resultados: [Tema6\\_Ejemplo3\\_Assembly](#)

# ¡Gracias!

The logo consists of the lowercase letters "viu" in white, centered within a dark orange, rounded rectangular shape.

viu

**Universidad**  
Internacional  
de Valencia

[universidadviu.com](http://universidadviu.com)

De:  
 Planeta Formación y Universidades