TRABAJO DE PRÁCTICAS

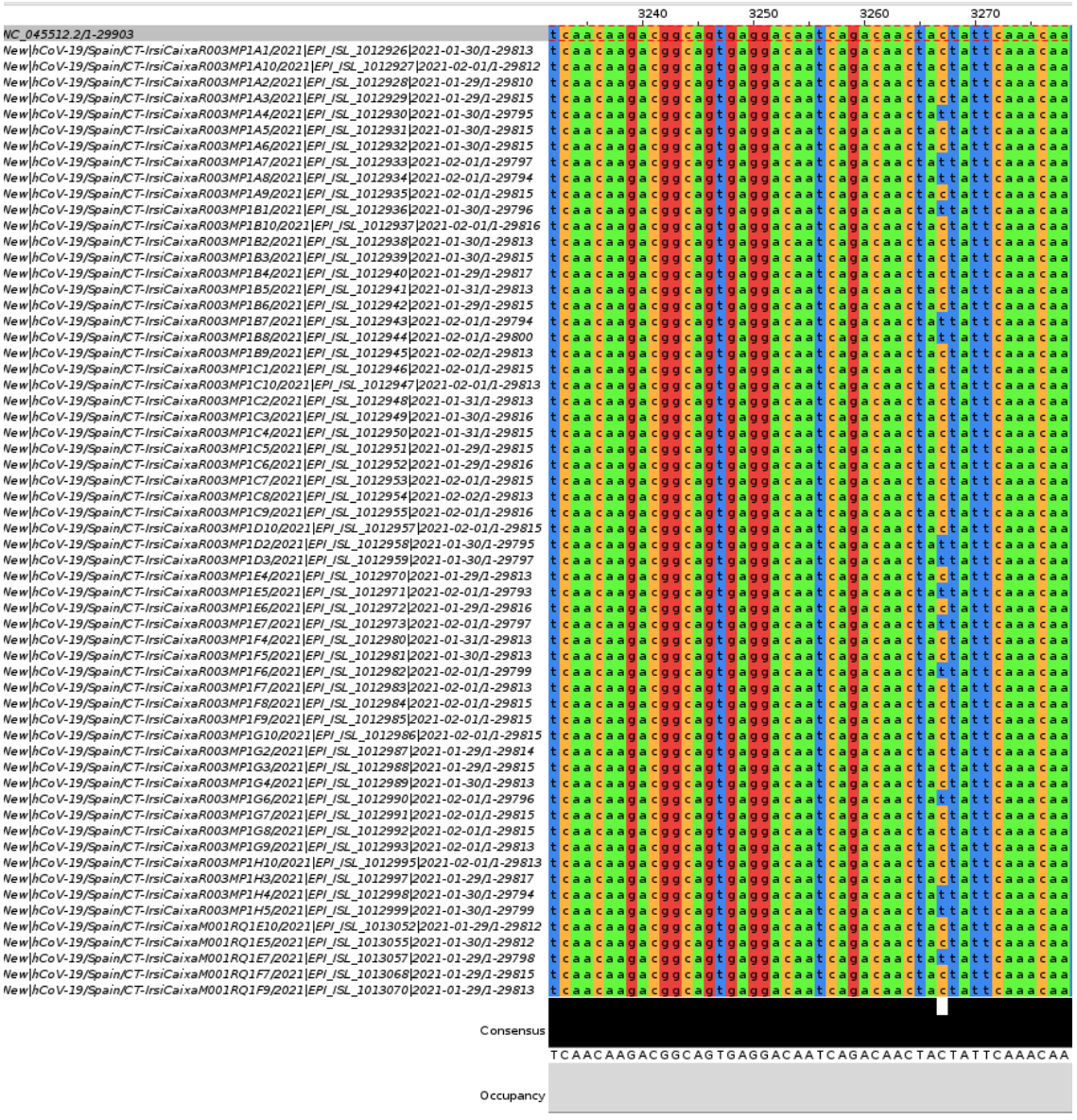
El objetivo de este trabajo de prácticas es estudiar con las herramientas tratadas en las anteriores prácticas la variedad y relación entre las secuencias proporcionadas (secuencias de nucleótidos de SARS-CoV-2, en concreto el grupo catHC4).

Tarea 1: Multialineamiento

Utilizando mafft se va a crear un multialineamiento entre las secuencias del fichero 'gisaid_hcov-19_2021_03_23_catHC4.fasta', la secuencia de la variante británica ('gisaid_hcov-19_2021_03_23_B117.fasta') y la secuencia de referencia ('RefSeqWuhan.fasta'). Se ha realizado el alineamiento en la versión online de la herramienta ya que es mucho más rápida (22 segundos), el comando equivalente en la terminal es:

```
% mafft --reorder --maxambiguous 0.05 --addfragments fragments --auto input
```

El método utilizado ha sido FFT-NS-fragment que está por defecto y una estrategia elegida automáticamente que ha sido multipair ya que son 411 secuencias de 29903 bases, un tamaño razonable para mafft. El alineamiento final tiene una longitud de 29906 y está guardado en el fichero 'alineamiento.fasta' y es posible visualizarlo con Jalview.



Tarea 2: Detección de mutaciones y errores de secuenciación

Para la detección de mutaciones se ha creado un programa en python que lee el alineamiento del fichero 'alineamiento.fasta' utilizando el módulo Biopython y posteriormente calcula la medida de conservación para cada posición en el alineamiento mediante la frecuencia de cada nucleótido.

$$f_v(i) = \frac{n_v(i)}{n(i)} \Rightarrow C(i) = -\sum_{v=1}^{4} f_v(i)\, ln(f_v(i))$$

Tras calcular la medida de conservación se busca secuencias que contengan mutaciones en posiciones cuya medida de conservación sea menor a 0.05 (posiciones con pocas variaciones). En concreto se van a buscar estas mutaciones en las siguientes posiciones: 266 a 13484 (Gen ORF1A), 21563 a 25384(Proteína Spike), 25393 a 26220(Gen ORF3A) y 28274 a 29533(Gen N).

Para la ejecución del programa es necesario tener instalada una versión de python3 así como el módulo Biopython.

```
$ sudo apt-get update
$ sudo apt-get install python3.6
$ pip install biopython
```

Con el siguiente comando se pueden obtener tanto un fichero con las medidas de conservación 'conservaciones.txt' como las variaciones con más mutaciones respecto a la secuencia de referencia en 'mutacionesImportantes.txt':

```
$ python3 mutaciones.py
```

Los resultados de este programa muestran algunas secuencias con muchas mutaciones en estas posiciones (en las casillas se indica el número de mutaciones de cada grupo de cada secuencia):

Identificador de secuencia[1]	**Posiciones con mutaciones**			
	ORF1A	**Spike**	**ORF3A**	**N**
EPI_ISL_1020573	8	6	0	2
EPI_ISL_1012978	5	1	1	8
EPI_ISL_1013002	3	0	0	12
EPI_ISL_1208626	15	2	0	1
EPI_ISL_1208631	15	1	0	1
EPI_ISL_1208653	8	8	2	1

Estas mutaciones pueden afectar de forma muy negativa al individuo al estar en posiciones con un índice de conservación alto.

[1] El identificador de las secuencias se ha acortado por legibilidad, tiene el siguiente formato: Secuencia New|hCoV-19/Spain/CT-HUVH-09286/2021|EPI_ISL_xxxxxxxx|2021-MM-DD

Tarea 3: Construcción de una filogenia

Para la construcción de un árbol filogenético se han probado dos herramientas: RAxML y FastTree. Se ha clonado el repositorio de GitHub https://github.com/stamatak/standard-RAxML y se han seguido las instrucciones de instalación de la versión Pthreads SSE3 (fichero README). También se ha descargado el programa precompilado para Linux de la página inicial de FastTree en el apartado de Instalación(http://www.microbesonline.org/fasttree/#Install) .

Utilizando varios modelos distintos se ha buscado el árbol filogenético de máxima verosimilitud comparando la puntuación Log-likelihood en un modelo Gamma(20). Los resultados obtenidos son del mejor árbol para cada modelo y aparecen en la siguiente tabla. En tiempo de ejecución los resultados han sido muy similares.

	Modelo	Puntuación Log-likelihood
RAMxML	GTR+CAT	-54510.497986
	GTR+GAMMA	-54509.235990
	GTR+CATI	-54443.181115
	GTR+GAMMAI	-54443.669419
FastTree (Gamma20-based likelihood)	GTR+CAT	-56494.199000
	WAG+CAT	-57594.642000
	LG+CAT	-57594.642000
	JTT+CAT	-57594.642000

Los modelos que se han probado para la matriz de sustituciones y optimizaciones son:

- GTR(Generalized time-reversible)
- WAG(Whelan and Goldman)
- JTT(Jones-Taylor-Thornton)
- LG(Le and Gascuel)

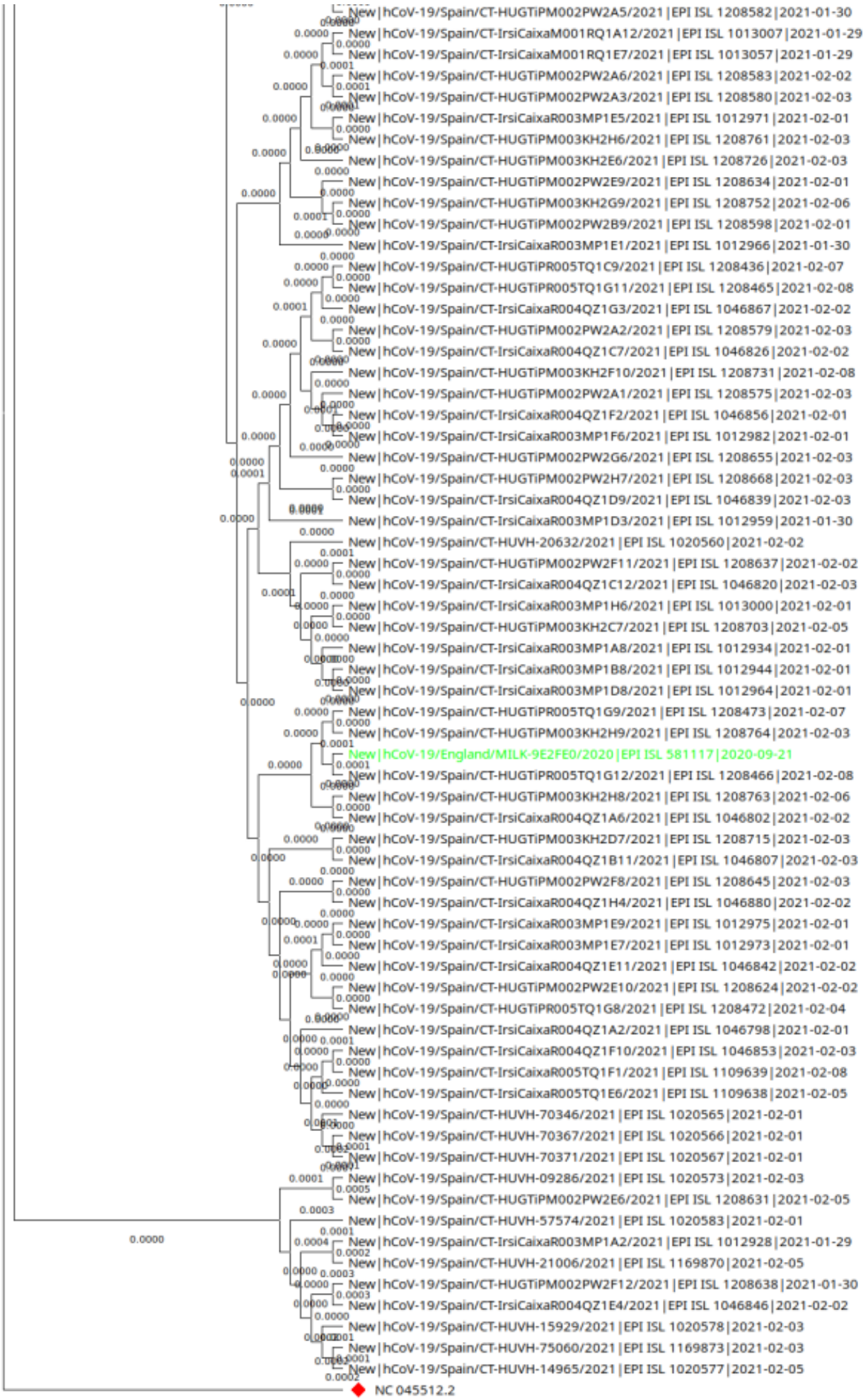
Los modelos evolutivos que se han probado son:

- CAT: Más rápido al realizar aproximadamente ¼ de las operaciones que el modelo GAMMA precisa. Solución menos precisa pero bastante aproximada.
  *CATI: después de la búsqueda con CAT, se evaluan los árboles finales bajo una estimación de GAMMA en vez de la GAMMA predeterminada.
- GAMMA: Más lento, pero más preciso. Puede no funcionar para árboles muy grandes con más de 10.000 taxones.
  *GAMMAI: igual que GAMMA pero con una estimación de la proporción de los sitios invariables.

El mejor árbol filogenético (GTR+CATI con RAxML) se encuentra en el fichero 'RAxML_bestTree.gtrcati.txt'.El comando utilizado para obtener este árbol es:

$ ./raxmlHPC-PTHREADS-SSE3 -s alineamiento.fasta -n gtrcati.txt -m GTRCATI -p 111

El formato de salida de ambas herramientas son árboles Newick por lo que es posible visualizar dichos árboles en MEGA-X o cualquier otro visualizador. En la imagen posterior se muestra una parte del árbol donde se puede ver la secuencia de referencia marcada con un rombo rojo y la secuencia correspondiente a la variante británica en color verde.


New|hCoV-19/Spain/CT-HUGTiPM002PW2A5/2021|EPI ISL 1208582|2021-01-30
New|hCoV-19/Spain/CT-IrsiCaixaM001RQ1A12/2021|EPI ISL 1013007|2021-01-29
New|hCoV-19/Spain/CT-IrsiCaixaM001RQ1E7/2021|EPI ISL 1013057|2021-01-29
New|hCoV-19/Spain/CT-HUGTiPM002PW2A6/2021|EPI ISL 1208583|2021-02-02
New|hCoV-19/Spain/CT-HUGTiPM002PW2A3/2021|EPI ISL 1208580|2021-02-03
New|hCoV-19/Spain/CT-IrsiCaixaR003MP1E5/2021|EPI ISL 1012971|2021-02-01
New|hCoV-19/Spain/CT-HUGTiPM003KH2H6/2021|EPI ISL 1208761|2021-02-03
New|hCoV-19/Spain/CT-HUGTiPM003KH2E6/2021|EPI ISL 1208726|2021-02-03
New|hCoV-19/Spain/CT-HUGTiPM002PW2E9/2021|EPI ISL 1208634|2021-02-01
New|hCoV-19/Spain/CT-HUGTiPM003KH2G9/2021|EPI ISL 1208752|2021-02-06
New|hCoV-19/Spain/CT-HUGTiPM002PW2B9/2021|EPI ISL 1208598|2021-02-01
New|hCoV-19/Spain/CT-IrsiCaixaR003MP1E1/2021|EPI ISL 1012966|2021-01-30
New|hCoV-19/Spain/CT-HUGTiPR005TQ1C9/2021|EPI ISL 1208436|2021-02-07
New|hCoV-19/Spain/CT-HUGTiPR005TQ1G11/2021|EPI ISL 1208465|2021-02-08
New|hCoV-19/Spain/CT-IrsiCaixaR004QZ1G3/2021|EPI ISL 1046867|2021-02-02
New|hCoV-19/Spain/CT-HUGTiPM002PW2A2/2021|EPI ISL 1208579|2021-02-03
New|hCoV-19/Spain/CT-IrsiCaixaR004QZ1C7/2021|EPI ISL 1046826|2021-02-02
New|hCoV-19/Spain/CT-HUGTiPM003KH2F10/2021|EPI ISL 1208731|2021-02-08
New|hCoV-19/Spain/CT-HUGTiPM002PW2A1/2021|EPI ISL 1208575|2021-02-03
New|hCoV-19/Spain/CT-IrsiCaixaR004QZ1F2/2021|EPI ISL 1046856|2021-02-01
New|hCoV-19/Spain/CT-IrsiCaixaR003MP1F6/2021|EPI ISL 1012982|2021-02-01
New|hCoV-19/Spain/CT-HUGTiPM002PW2G6/2021|EPI ISL 1208655|2021-02-03
New|hCoV-19/Spain/CT-HUGTiPM002PW2H7/2021|EPI ISL 1208668|2021-02-03
New|hCoV-19/Spain/CT-IrsiCaixaR004QZ1D9/2021|EPI ISL 1046839|2021-02-03
New|hCoV-19/Spain/CT-IrsiCaixaR003MP1D3/2021|EPI ISL 1012959|2021-01-30
New|hCoV-19/Spain/CT-HUVH-20632/2021|EPI ISL 1020560|2021-02-02
New|hCoV-19/Spain/CT-HUGTiPM002PW2F11/2021|EPI ISL 1208637|2021-02-02
New|hCoV-19/Spain/CT-IrsiCaixaR004QZ1C12/2021|EPI ISL 1046820|2021-02-03
New|hCoV-19/Spain/CT-IrsiCaixaR003MP1H6/2021|EPI ISL 1013000|2021-02-01
New|hCoV-19/Spain/CT-HUGTiPM003KH2C7/2021|EPI ISL 1208703|2021-02-05
New|hCoV-19/Spain/CT-IrsiCaixaR003MP1A8/2021|EPI ISL 1012934|2021-02-01
New|hCoV-19/Spain/CT-IrsiCaixaR003MP1B8/2021|EPI ISL 1012944|2021-02-01
New|hCoV-19/Spain/CT-IrsiCaixaR003MP1D8/2021|EPI ISL 1012964|2021-02-01
New|hCoV-19/Spain/CT-HUGTiPR005TQ1G9/2021|EPI ISL 1208473|2021-02-07
New|hCoV-19/Spain/CT-HUGTiPM003KH2H9/2021|EPI ISL 1208764|2021-02-03
New|hCoV-19/England/MILK-9E2FE0/2020|EPI ISL 581117|2020-09-21
New|hCoV-19/Spain/CT-HUGTiPR005TQ1G12/2021|EPI ISL 1208466|2021-02-08
New|hCoV-19/Spain/CT-HUGTiPM003KH2H8/2021|EPI ISL 1208763|2021-02-06
New|hCoV-19/Spain/CT-IrsiCaixaR004QZ1A6/2021|EPI ISL 1046802|2021-02-02
New|hCoV-19/Spain/CT-HUGTiPM003KH2D7/2021|EPI ISL 1208715|2021-02-03
New|hCoV-19/Spain/CT-IrsiCaixaR004QZ1B11/2021|EPI ISL 1046807|2021-02-03
New|hCoV-19/Spain/CT-HUGTiPM002PW2F8/2021|EPI ISL 1208645|2021-02-03
New|hCoV-19/Spain/CT-IrsiCaixaR004QZ1H4/2021|EPI ISL 1046880|2021-02-02
New|hCoV-19/Spain/CT-IrsiCaixaR003MP1E9/2021|EPI ISL 1012975|2021-02-01
New|hCoV-19/Spain/CT-IrsiCaixaR003MP1E7/2021|EPI ISL 1012973|2021-02-01
New|hCoV-19/Spain/CT-IrsiCaixaR004QZ1E11/2021|EPI ISL 1046842|2021-02-02
New|hCoV-19/Spain/CT-HUGTiPM002PW2E10/2021|EPI ISL 1208624|2021-02-02
New|hCoV-19/Spain/CT-HUGTiPR005TQ1G8/2021|EPI ISL 1208472|2021-02-04
New|hCoV-19/Spain/CT-IrsiCaixaR004QZ1A2/2021|EPI ISL 1046798|2021-02-01
New|hCoV-19/Spain/CT-IrsiCaixaR004QZ1F10/2021|EPI ISL 1046853|2021-02-03
New|hCoV-19/Spain/CT-IrsiCaixaR005TQ1F1/2021|EPI ISL 1109639|2021-02-08
New|hCoV-19/Spain/CT-IrsiCaixaR005TQ1E6/2021|EPI ISL 1109638|2021-02-05
New|hCoV-19/Spain/CT-HUVH-70346/2021|EPI ISL 1020565|2021-02-01
New|hCoV-19/Spain/CT-HUVH-70367/2021|EPI ISL 1020566|2021-02-01
New|hCoV-19/Spain/CT-HUVH-70371/2021|EPI ISL 1020567|2021-02-01
New|hCoV-19/Spain/CT-HUVH-09286/2021|EPI ISL 1020573|2021-02-03
New|hCoV-19/Spain/CT-HUGTiPM002PW2E6/2021|EPI ISL 1208631|2021-02-05
New|hCoV-19/Spain/CT-HUVH-57574/2021|EPI ISL 1020583|2021-02-01
New|hCoV-19/Spain/CT-IrsiCaixaR003MP1A2/2021|EPI ISL 1012928|2021-01-29
New|hCoV-19/Spain/CT-HUVH-21006/2021|EPI ISL 1169870|2021-02-05
New|hCoV-19/Spain/CT-HUGTiPM002PW2F12/2021|EPI ISL 1208638|2021-01-30
New|hCoV-19/Spain/CT-IrsiCaixaR004QZ1E4/2021|EPI ISL 1046846|2021-02-02
New|hCoV-19/Spain/CT-HUVH-15929/2021|EPI ISL 1020578|2021-02-03
New|hCoV-19/Spain/CT-HUVH-75060/2021|EPI ISL 1169873|2021-02-03
New|hCoV-19/Spain/CT-HUVH-14965/2021|EPI ISL 1020577|2021-02-05
NC 045512.2