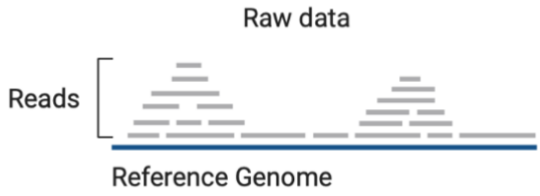
UNIVERSIDADE DA CORUÑA

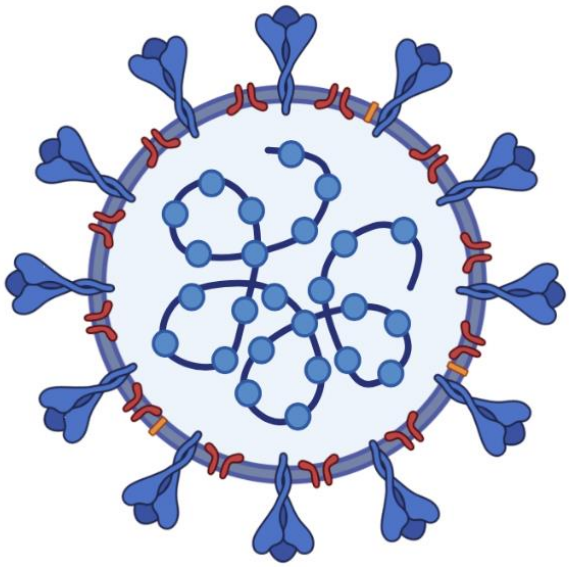
FACULTAD DE INFORMÁTICA

MUBICS

Cuaderno de prácticas

Estructuras de datos y algoritmia para secuencias biológicas





Pedro Sánchez García

Curso 2021-2022

Profesora:
Dra. Susana Ladra González

PRÁCTICA 5. IDENTIFICACIÓN DE CLADOS Y LINAJES

OBJETIVO

Utilizar herramientas existentes para identificación de clados/linajes a partir de lecturas de secuenciación.

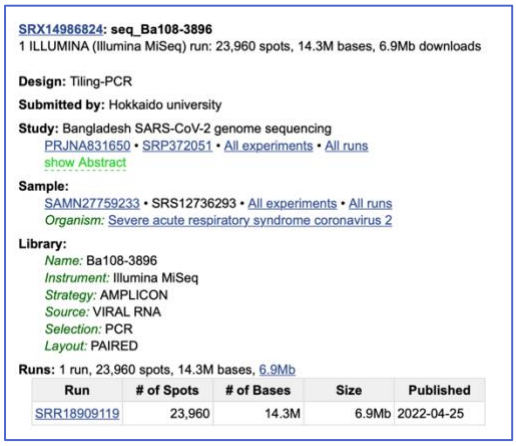
TAREAS (0,5 puntos)

Obtener la identificación de clados y linajes de dos secuencias de lecturas de SARS-CoV-2 (deben usar técnica de amplicones). Se puede utilizar un conjunto de lecturas de NCBI. Posteriormente, se deben obtener las variantes (ivar), clado (nextclade) y linaje (pangolín).

1. Obtención de los conjuntos de lecturas en el NCBI

Se accede al NCBI para obtener los 2 conjuntos datos de secuenciación procedentes de la técnica de amplicones. Esta conforma una amplificación de un fragmento determinado del genoma de interés y su posterior secuenciación, mediante la cual se obtiene una elevada cobertura en el análisis. De manera aleatoria, se procede a la descarga de las *reads forward* y *reverse* para cada uno de los conjuntos elegidos.

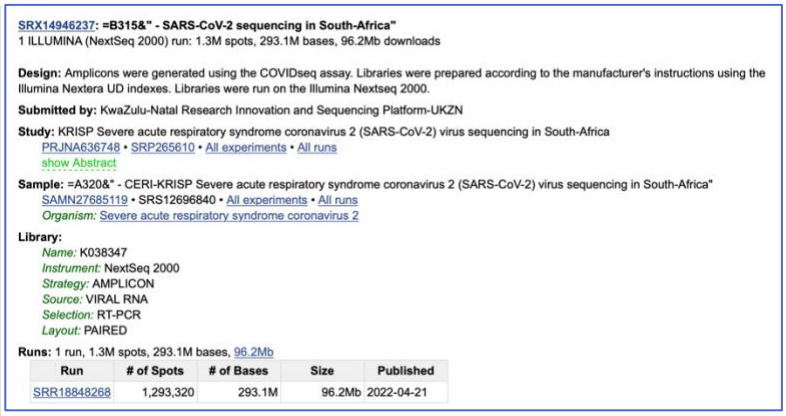
En este caso, se ha optado por realizar la práctica con los siguientes conjuntos:



SRX14986824: seq_Ba108-3896
1 ILLUMINA (Illumina MiSeq) run: 23,960 spots, 14.3M bases, 6.9Mb downloads

Design: Tiling-PCR
Submitted by: Hokkaido university
Study: Bangladesh SARS-CoV-2 genome sequencing
PRJNA831650 • SRP372051 • All experiments • All runs
show Abstract
Sample:
SAMN27759233 • SRS12736293 • All experiments • All runs
Organism: Severe acute respiratory syndrome coronavirus 2
Library:
Name: Ba108-3896
Instrument: Illumina MiSeq
Strategy: AMPLICON
Source: VIRAL RNA
Selection: PCR
Layout: PAIRED
Runs: 1 run, 23,960 spots, 14.3M bases, 6.9Mb

Run	# of Spots	# of Bases	Size	Published
SRR18909119	23,960	14.3M	6.9Mb	2022-04-25



SRX14946237: =B315&" - SARS-CoV-2 sequencing in South-Africa"
1 ILLUMINA (NextSeq 2000) run: 1.3M spots, 293.1M bases, 96.2Mb downloads

Design: Amplicons were generated using the COVIDseq assay. Libraries were prepared according to the manufacturer's instructions using the Illumina Nextera UD indexes. Libraries were run on the Illumina Nextseq 2000.
Submitted by: KwaZulu-Natal Research Innovation and Sequencing Platform-UKZN
Study: KRISP Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus sequencing in South-Africa
PRJNA636748 • SRP265610 • All experiments • All runs
show Abstract
Sample: =A320&" - CERI-KRISP Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus sequencing in South-Africa"
SAMN27685119 • SRS12696840 • All experiments • All runs
Organism: Severe acute respiratory syndrome coronavirus 2
Library:
Name: K038347
Instrument: NextSeq 2000
Strategy: AMPLICON
Source: VIRAL RNA
Selection: RT-PCR
Layout: PAIRED
Runs: 1 run, 1.3M spots, 293.1M bases, 96.2Mb

Run	# of Spots	# of Bases	Size	Published
SRR18848268	1,293,320	293.1M	96.2Mb	2022-04-21

En base a la información proporcionada en el NCBI, se trata de unos conjuntos de secuenciación que se han generado en Bangladesh y en Sudáfrica bajo la tecnología de Illumina, siguiendo un proceso de RT-PCR a partir de muestras nasofaríngeas tomadas de pacientes. Esta metodología ha sido la empleada desde el comienzo de la pandemia de la COVID-19 para el seguimiento de la expansión y aparición de nuevas variantes del SARS-CoV-2.

2. Alineamiento de las lecturas con el genoma de referencia de Wuhan

En Galaxy, mediante la herramienta *BWA-MEM*, se efectúa el alineamiento de las *reads forward* y *reverse* de cada conjunto obtenido con el genoma de referencia, que es el correspondiente con el del SARS-CoV-2 original de Wuhan. Como resultado, se genera un archivo en formato BAM que contiene las *reads* ensambladas/mapeadas con el genoma de referencia.

3. Purgado de las *reads* con ivar trim

A continuación, tomando como partida el archivo en formato BAM ordenado del alineamiento de las *reads* contra el genoma de referencia, se procede a eliminar las secuencias que corresponden con los adaptadores empleados en la aproximación de la técnica de amplicones. Para ello, se emplea la herramienta ivar trim, a la cual se le debe proporcionar, aparte del archivo mencionado, otro archivo en formato BED. Este último contiene información en un formato tabular de todas las *reads* del proceso de secuenciación realizado y es necesario para esta fase.

4. Obtención de las variantes con ivar variant

En esta fase, se toma como partida el archivo en formato BAM ordenado del alineamiento de las *reads* purgadas contra el genoma de referencia. Se ejecuta la herramienta con los parámetros por defecto y cargando el genoma de referencia empleado anteriormente. Como resultado, se logra una salida en formato tabular que recoge un amplio conjunto de mutaciones detectadas, sobre las que se ha centrado la atención desde el comienzo de la pandemia.

4.1 Conjunto de secuenciación: *SRX14986824*

Tal y como se muestra a continuación, para el primer conjunto obtenido del NCBI, se aprecian numerosas sustituciones nucleotídicas, de las cuales será preciso obtener información de aquellas que resultan más relevantes en la siguiente fase:

REGION	POS	REF	ALT	REF_DP	REF_RV	REF_QUAL	ALT_DP	ALT_RV	ALT_QUAL	ALT_FREQ	TOTAL_DP	PVAL	PASS	GFF_FEATURE
MN908947.3	210	G	T	0	0	0	5	0	38	1	5	0.00793651		TRUE
MN908947.3	241	C	T	0	0	0	4	0	37	1	4	0.0142857		FALSE
MN908947.3	4655	C	T	4	4	38	1	1	38	0.2	5	0.555556	FALSE	NA
MN908947.3	11030	A	T	2	2	38	1	0	38	0.333333	3	0.6	FALSE	NA
MN908947.3	11618	T	C	3	3	38	1	1	38	0.25	4	0.571429	FALSE	NA
MN908947.3	12257	C	A	4	4	37	1	1	21	0.2	5	0.555556	FALSE	NA
MN908947.3	12439	C	T	4	4	38	1	1	38	0.2	5	0.555556	FALSE	NA
MN908947.3	12519	A	G	3	0	38	1	0	38	0.25	4	0.571429	FALSE	NA
MN908947.3	12521	G	A	3	0	38	1	0	38	0.25	4	0.571429	FALSE	NA
MN908947.3	12620	T	C	3	0	38	1	0	23	0.25	4	0.571429	FALSE	NA
MN908947.3	12869	A	G	1	1	35	1	1	38	0.5	2	0.5	FALSE	NA
MN908947.3	13657	A	G	7	1	38	1	1	38	0.125	8	0.533333	FALSE	NA
MN908947.3	14298	T	C	4	0	37	1	0	38	0.2	5	0.555556	FALSE	NA
MN908947.3	14334	T	G	5	0	38	1	0	20	0.2	5	0.6	FALSE	NA
MN908947.3	14408	C	T	0	0	0	5	0	38	1	5	0.00793651		TRUE
MN908947.3	16392	T	C	3	3	38	1	1	38	0.25	4	0.571429	FALSE	NA
MN908947.3	16408	A	G	3	3	38	1	1	38	0.25	4	0.571429	FALSE	NA
MN908947.3	24600	T	A	5	0	37	1	0	38	0.166667	6	0.545455	FALSE	NA
MN908947.3	25697	A	T	4	0	38	1	0	38	0.2	5	0.555556	FALSE	NA
MN908947.3	25804	C	T	4	0	38	1	0	37	0.2	5	0.555556	FALSE	NA
MN908947.3	25812	T	C	4	0	38	1	0	38	0.2	5	0.555556	FALSE	NA
MN908947.3	25832	T	C	3	0	38	1	0	38	0.25	4	0.5	FALSE	NA
MN908947.3	26735	C	T	0	0	0	6	6	36	1	6	0.0021645		TRUE
MN908947.3	26746	G	A	5	5	37	1	1	38	0.166667	6	0.545455	FALSE	NA
MN908947.3	26806	A	T	5	5	38	1	1	38	0.166667	6	0.545455	FALSE	NA
MN908947.3	29442	A	G	1	1	37	1	1	34	0.5	2	0.5	FALSE	NA

4.2 Conjunto de secuenciación: *SRX14946237*

En lo que respecta al segundo conjunto obtenido del NCBI, se puede observar una tendencia similar al caso anterior, con varias sustituciones nucleotídicas en diversas posiciones:

REGION	POS	REF	ALT	REF_DP	REF_RV	REF_QUAL	ALT_DP	ALT_RV	ALT_QUAL	ALT_FREQ	TOTAL_DP	PVAL	PASS	ALT_AA
MN908947.3	2549	G	T	36	15	34	4	2	34	0.1	40	0.0549305	FALSE	NA	NA
MN908947.3	5230	G	T	0	0	0	7	3	40	1	7	0.0001554	TRUE	NA	NA
MN908947.3	5916	T	C	54	23	44	2	0	68	0.0357143	56	0.158273	FALSE	NA	NA
MN908947.3	5922	G	A	58	25	48	2	1	30	0.0333333	60	0.171911	FALSE	NA	NA
MN908947.3	10029	C	T	0	0	0	2	1	51	1	2	0.166667	FALSE	NA	NA
MN908947.3	10059	G	T	2	1	51	1	0	68	0.333333	3	0.428571	FALSE	NA	NA
MN908947.3	11139	C	T	3	1	34	2	1	30	0.4	5	0.277778	FALSE	NA	NA
MN908947.3	11266	G	T	0	0	0	4	2	42	1	4	0.0142857	FALSE	NA	NA
MN908947.3	11332	A	G	0	0	0	9	4	41	1	9	1.08251e-05	TRUE	NA	NA
MN908947.3	11404	T	G	0	0	0	2	1	34	1	2	0.333333	FALSE	NA	NA
MN908947.3	14364	G	T	1	0	34	1	0	34	0.5	2	0.666667	FALSE	NA	NA
MN908947.3	19575	A	G	2	1	30	2	1	34	0.5	4	0.285714	FALSE	NA	NA
MN908947.3	26767	T	C	9	6	41	6	4	45	0.4	15	0.00451891	TRUE	NA	NA
MN908947.3	27807	C	T	0	0	0	3	0	68	1	3	0.0178571	FALSE	NA	NA
MN908947.3	28372	T	C	8	3	55	1	0	24	0.111111	9	0.391304	FALSE	NA	NA
MN908947.3	28551	G	C	2	1	34	8	4	33	0.8	10	0.000357228	TRUE	NA	NA

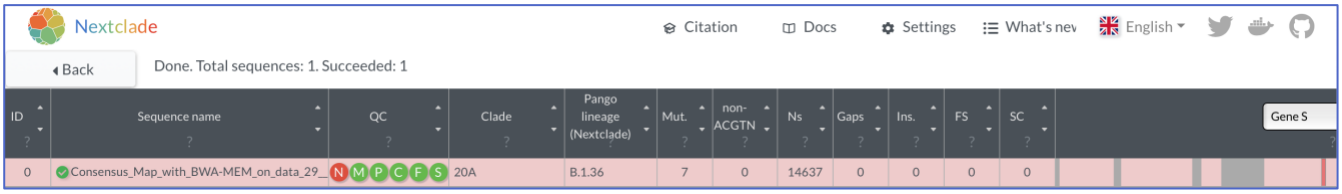
5. Determinación del clado y linaje

Para esta fase, debemos tener en cuenta que la aproximación necesita alcanzar la secuencia consenso. De esta forma, se emplea la herramienta ivar consensus, a la que se proporciona el archivo en formato BAM ordenado del alineamiento de las *reads* purgadas contra el genoma de referencia. De nuevo, manteniendo los parámetros por defecto, se ejecuta la herramienta y se obtiene una salida que consiste en un fichero con formato FASTA. Este contiene la secuencia consenso que se ha extrapolado del alineamiento de las *reads* purgadas con el genoma de referencia.

A continuación, se exporta el archivo en Galaxy, para subirlo a los recursos web de nextclade y pangolín, con el fin de obtener información acerca del clado y linaje del aislado con el que se está trabajando. Con respecto a nextclade, simplemente se indica el SARS-CoV-2 original de Wuhan como referencia, se sube el archivo FASTA con la secuencia consenso y se procede a la ejecución sin necesidad de modificar parámetros.

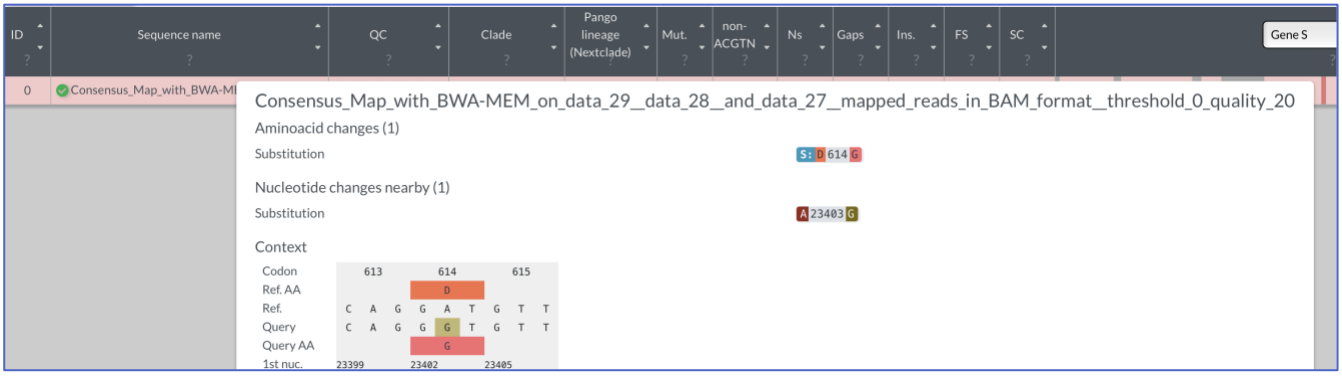
5.1 Conjunto de secuenciación: *SRX14986824*

Para el primer conjunto de secuenciación, se alcanzó el siguiente resultado:

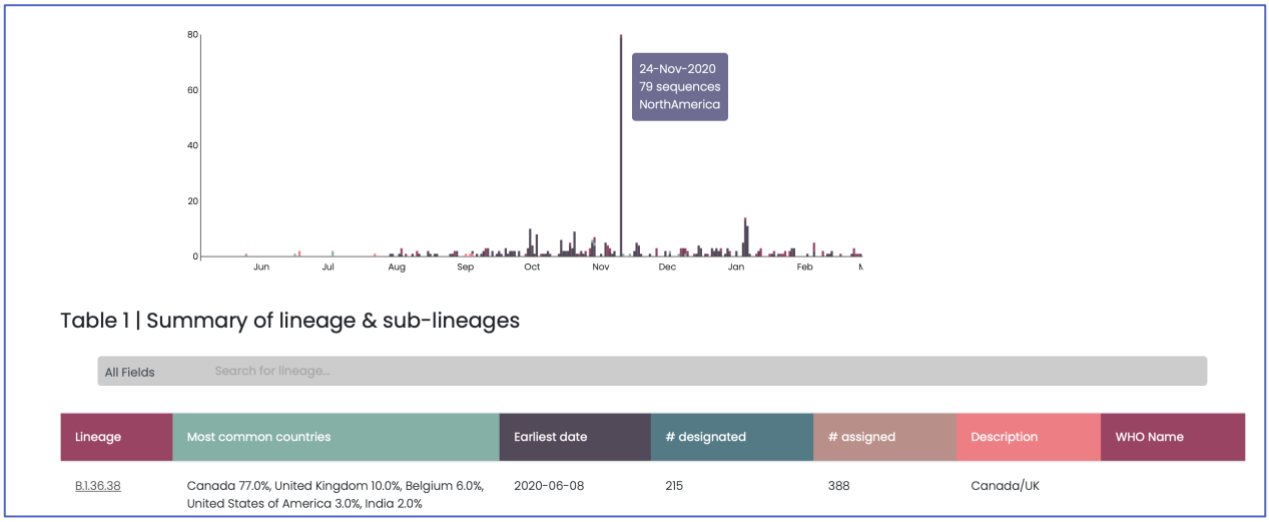


En él, la interfaz nos muestra diferentes campos, donde centramos la atención para extraer información relevante. Por ejemplo, el primer campo (QC) hace referencia al control de calidad, donde en este caso se aprecia una advertencia en rojo, la cual corresponde con un amplio número de regiones indeterminadas en el proceso de secuenciación, que aparecen denotadas como Ns en la

secuencia consenso. Con respecto al clado, en este caso sería el 20A, mientras que el linaje es el B.1.36. Finalmente, cabe destacar que el número de mutaciones relevantes son 7, lo que varía con el resultado obtenido anteriormente con ivar variant. A modo de complemento, estas mutaciones presentan una representación gráfica, con la que se alcanza información más detallada sobre cada una de ellas, tal y como se refleja en la siguiente imagen:



Una vez que se ha explorado nextclade, se lleva a cabo un análisis en la interfaz web de pangolín, donde se sube el archivo FASTA con la secuencia consenso y se ejecuta el análisis sin necesidad de cambiar parámetros. Tal y como se esperaba el linaje que se muestra es el B.1.36.38, del que se proporciona más información con una amplia sección de resultados gráficos:



En general, podemos destacar que ha sido predominante en Canadá (77%) frente a la India (2%), tal y como reflejan los porcentajes de asociación al número de secuencias que progresivamente se han ido depositando en bases de datos. Además, resulta interesante el pico que corresponde al 24 de noviembre de 2020, pues se depositaron 79 secuencias asociadas a casos en Estados Unidos, lo que da una visión de su expansión por el número de casos.

5.2 Conjunto de secuenciación: *SRX14946237*

Para el segundo conjunto, se alcanzó el siguiente resultado con nextclade:

Nextclade Citation Docs Settings What's

Back Done. Total sequences: 1. Succeeded: 1

ID	Sequence name	QC	Clade	Pango lineage (Nextclade)	Mut.	non-ACGTN	Ns	Gaps	Ins.	FS	SC
0	Consensus_Map_with_BWA-MEM_on_data_36_	N M P C F S	20H (Beta, V2)	B.1.351	9	2	21630	0	0	0	0

Es preciso destacar un elevado número de nucleótidos indeterminados, lo que podría ser debido a un protocolo inadecuado de secuenciación que conduce finalmente a una secuencia consenso de este tipo. Por otro lado, se observa que el clado es el 20H (Beta, V2) y el linaje es el B.1.351. Se determinan 9 mutaciones relevantes.

Posteriormente, con la ampliación de datos sobre el linaje, se acudió a pangolín, donde se aprecia que se trata de un linaje con expansión entre Estados Unidos (74%) y México (23%), así como un patrón más disperso en torno a la representación gráfica de las secuencias depositadas al respecto en las bases de datos:

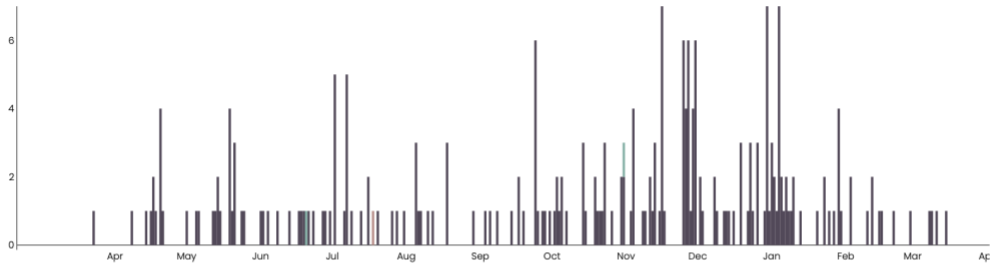


Table 1 | Summary of lineage & sub-lineages

All Fields Search for lineage...

Lineage	Most common countries	Earliest date	# designated	# assigned	Description	WHO Name
B.1.558	United States of America 74.0%, Mexico 23.0%, Guatemala 1.0%, Canada 1.0%, Japan 0.0%	2020-04-06	88	244	USA/ Mexico lineage	