

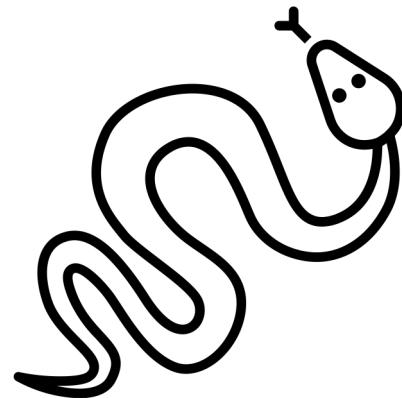
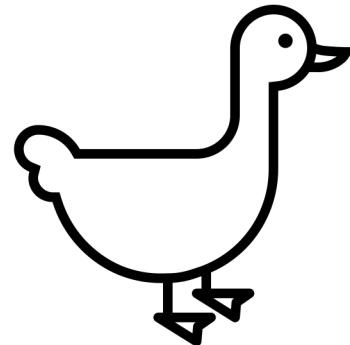
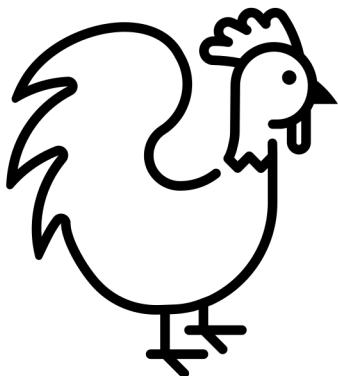
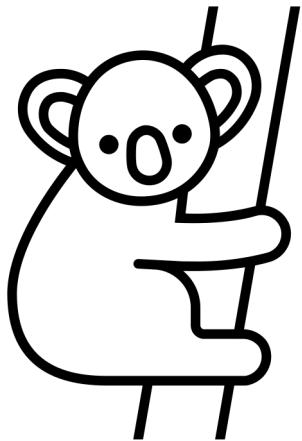
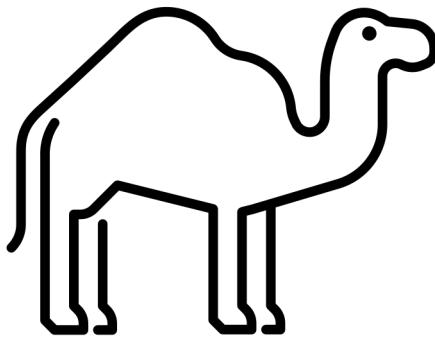
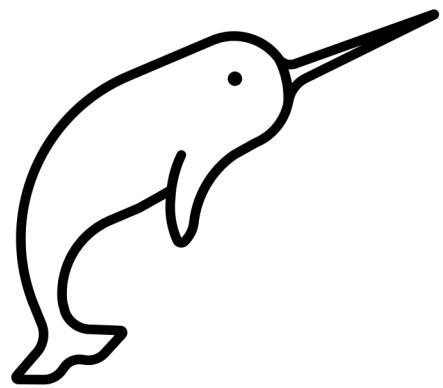
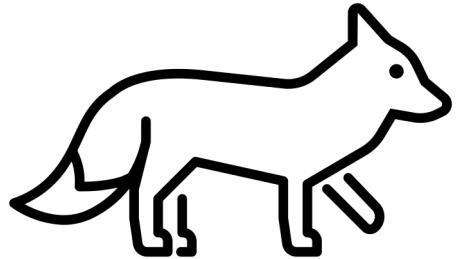


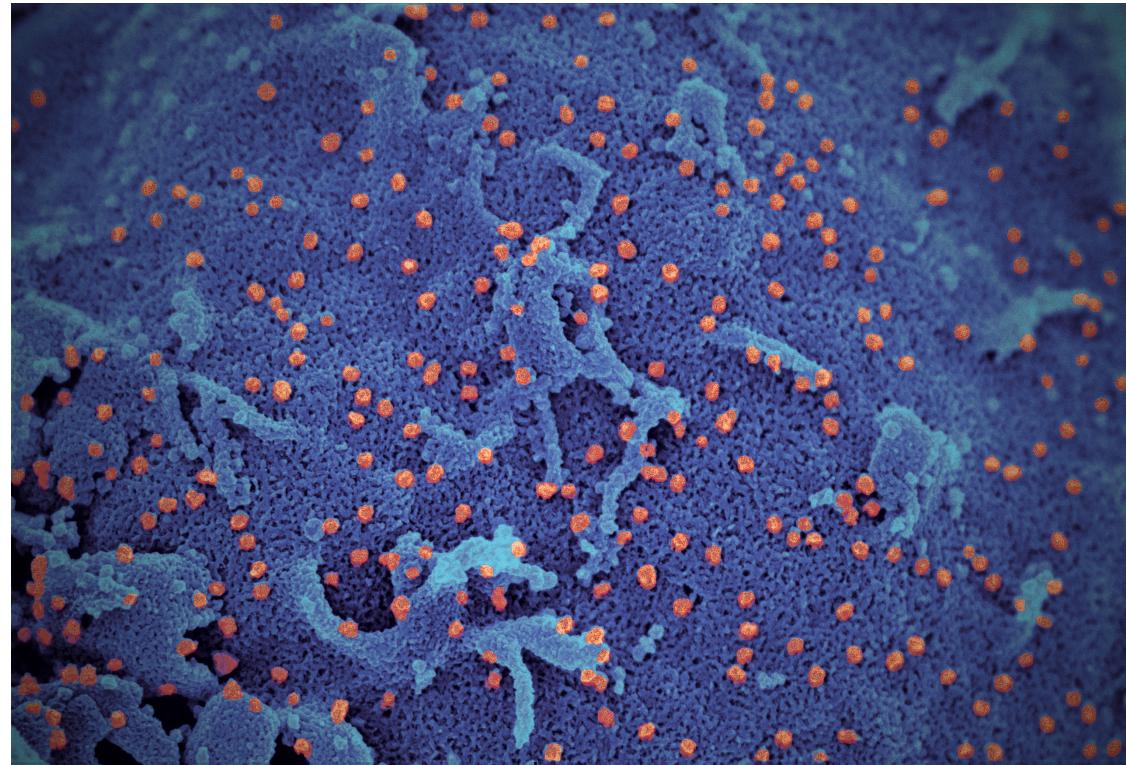
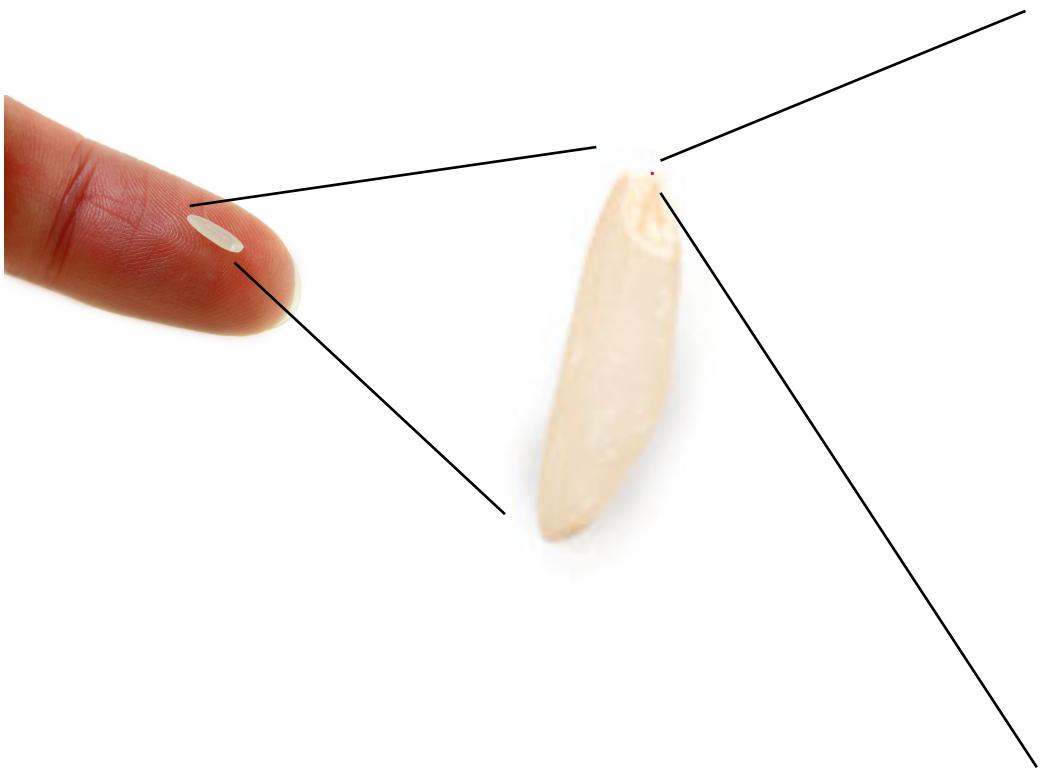
# How sequencing is used to identify SARS-CoV2 strains

HKU STEM Data Science Lab  
11<sup>th</sup> July 2022

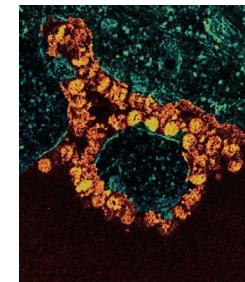
Dr Jason Wong & Dr Joshua Ho  
The University of Hong Kong

# How do we identify animals?





Virus are very very very small  
(0.00001 cm or 100 nm)

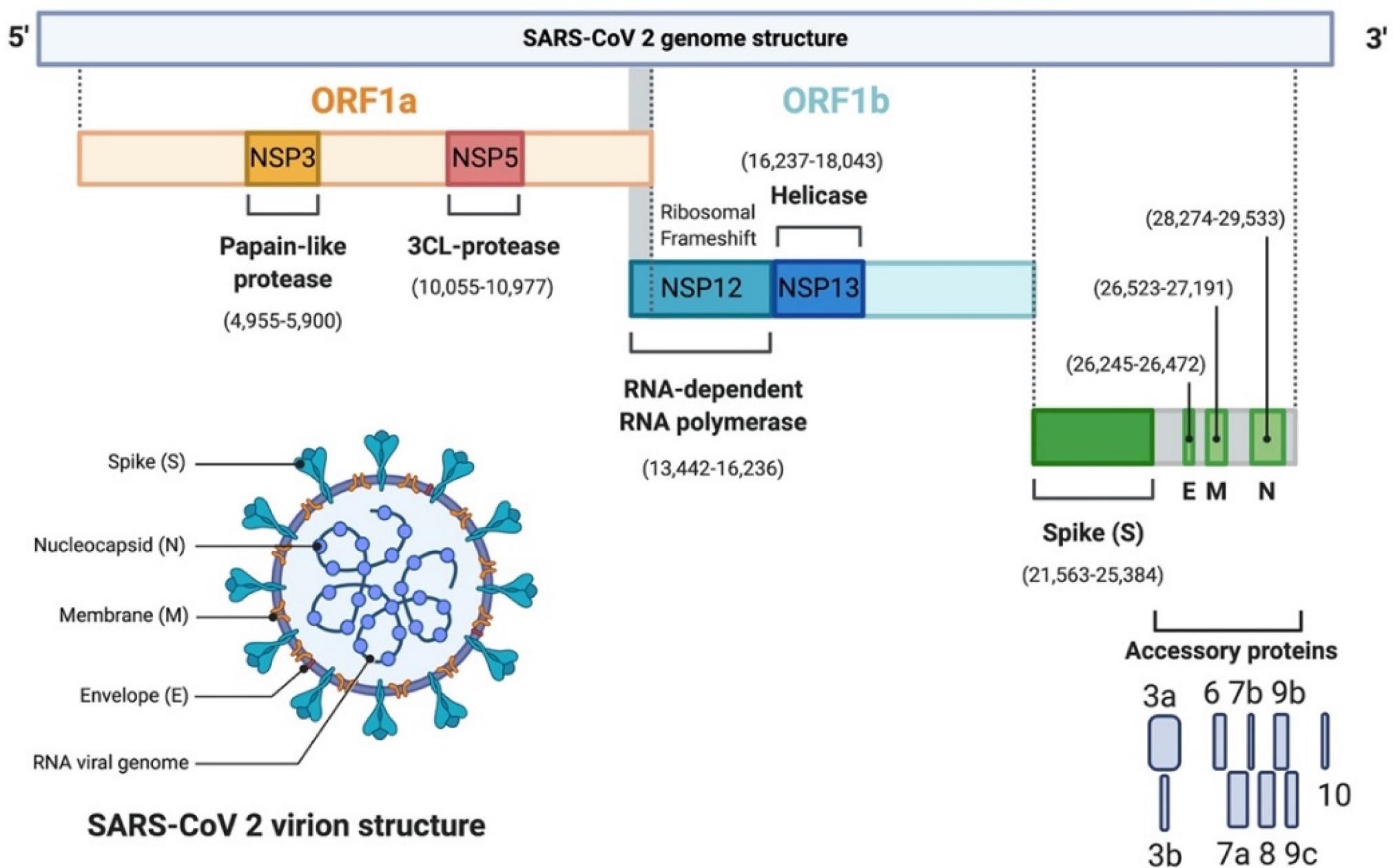


266

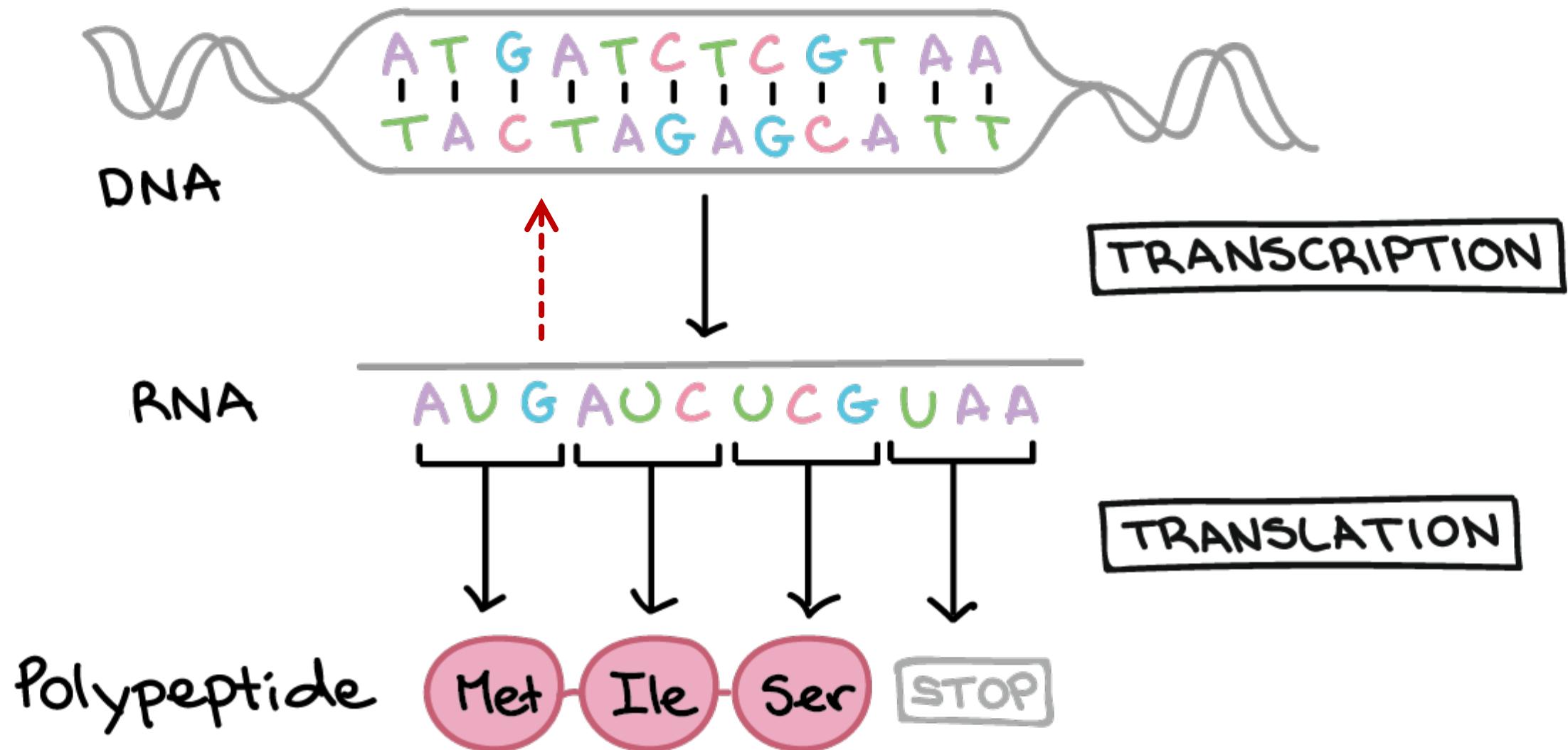
13,468

21,563

29,674



# THE CENTRAL DOGMA



# Reference SARS-CoV2 sequence (Wuhan strain)

- Open NCBI virus webpage in your browser  
(<https://www.ncbi.nlm.nih.gov/labs/virus/>)
- Click on Refseq genome  
([https://www.ncbi.nlm.nih.gov/nuccore/NC\\_045512.2](https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2))

The screenshot shows the NCBI Virus website. At the top, there is a navigation bar with links for "About Us", "Find Data", "Help", "How to Participate", "Submit Sequences", and "Contact Us". Below the navigation bar, a banner highlights "Quick Access to SARS-CoV-2 Data!" with four bullet points: "Novel Severe acute respiratory syndrome coronavirus 2 RefSeq genomes, nucleotide, and protein sequences.", "View our new SARS-CoV-2 interactive dashboard.", "A new page to submit SARS-CoV-2 sequences is now available.", and "Visit our new SARS-CoV-2 Variants Overview New!". The main content area is titled "NCBI Virus" and describes it as a community portal for viral sequence data from RefSeq, GenBank and other NCBI repositories. It provides two search options: "Search by sequence" (using NCBI BLAST™) and "Search by virus" (using virus name or taxid).

**NCBI Virus** is a community portal for viral sequence data from RefSeq, GenBank and other NCBI repositories. To find, retrieve and analyze data, please select an option below.

**Search by sequence**  
Use the NCBI BLAST™ tool to find similar viral nucleotide and protein sequences.

**Search by virus**  
Use virus name or taxid to find viral nucleotide and protein sequences.

# 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: [PRJNA485481](#)  
KEYWORDS RefSeq.  
SOURCE Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)  
ORGANISM [Severe acute respiratory syndrome coronavirus 2](#)  
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 [32015508](#)  
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 [15680415](#)  
REFERENCE 3 (bases 29728 to 29768)  
AUTHORS Robertson,M.P., Igel,H., Baertsch,R., Haussler,D., Ares,M. Jr. and Scott,W.G.  
TITLE The structure of a rigorously conserved RNA element within the SARS virus genome  
JOURNAL PLoS Biol. 3 (1), e5 (2005)  
PUBMED [15630477](#)  
REFERENCE 4 (bases 29609 to 29657)  
AUTHORS Williams,G.D., Chang,R.Y. and Brian,D.A.

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

NCBI Virus

Retrieve, view, and download SARS-CoV-2 coronavirus genomic and protein sequences.

Related information

Assembly

BioProject

Protein

PubMed

Taxonomy

Full text in PMC

Gene

Genome

Identical GenBank Sequence

Mature Peptides

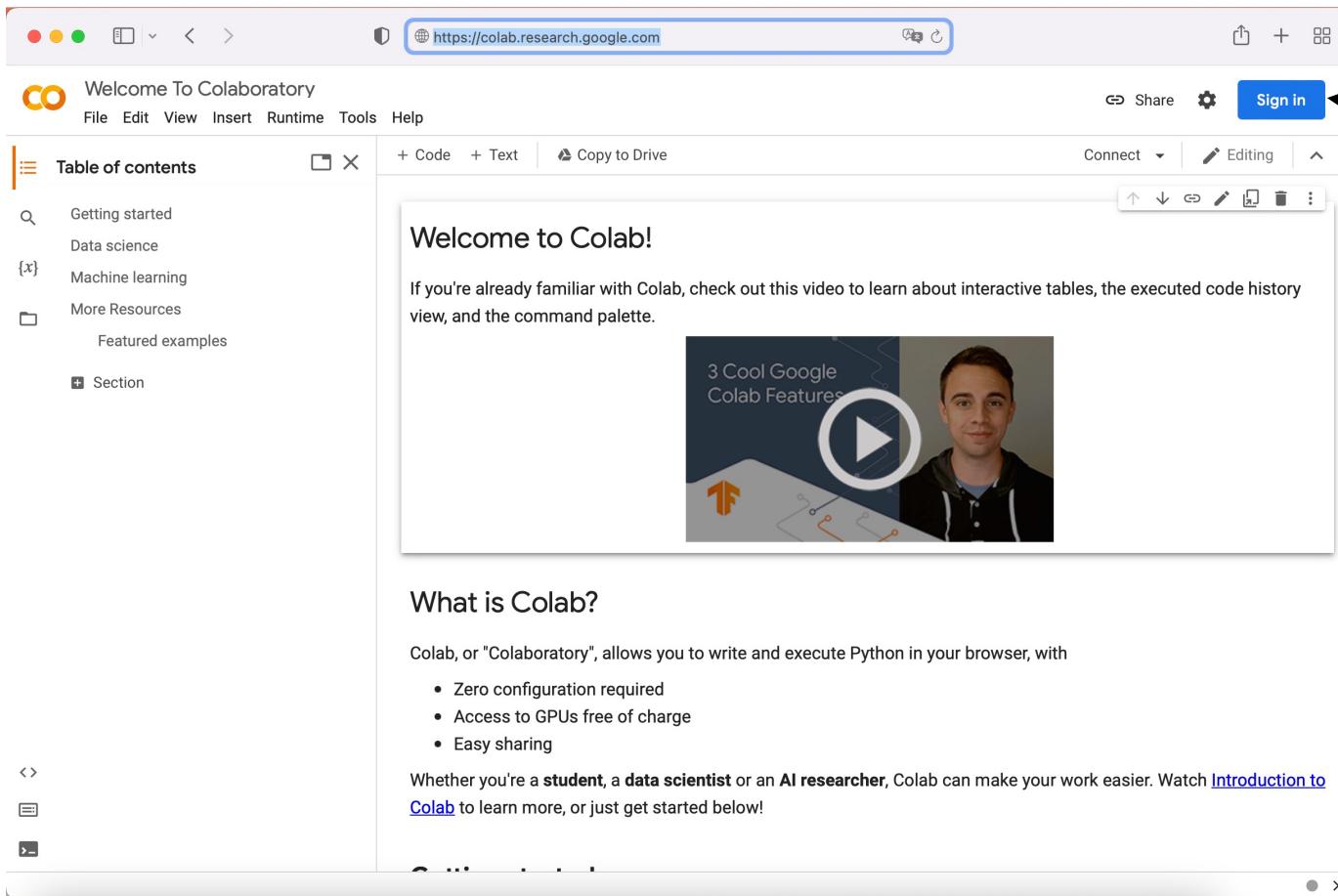
Other INSDC Genome Sequences

PubMed (Weighted)

# Genome data science workshop

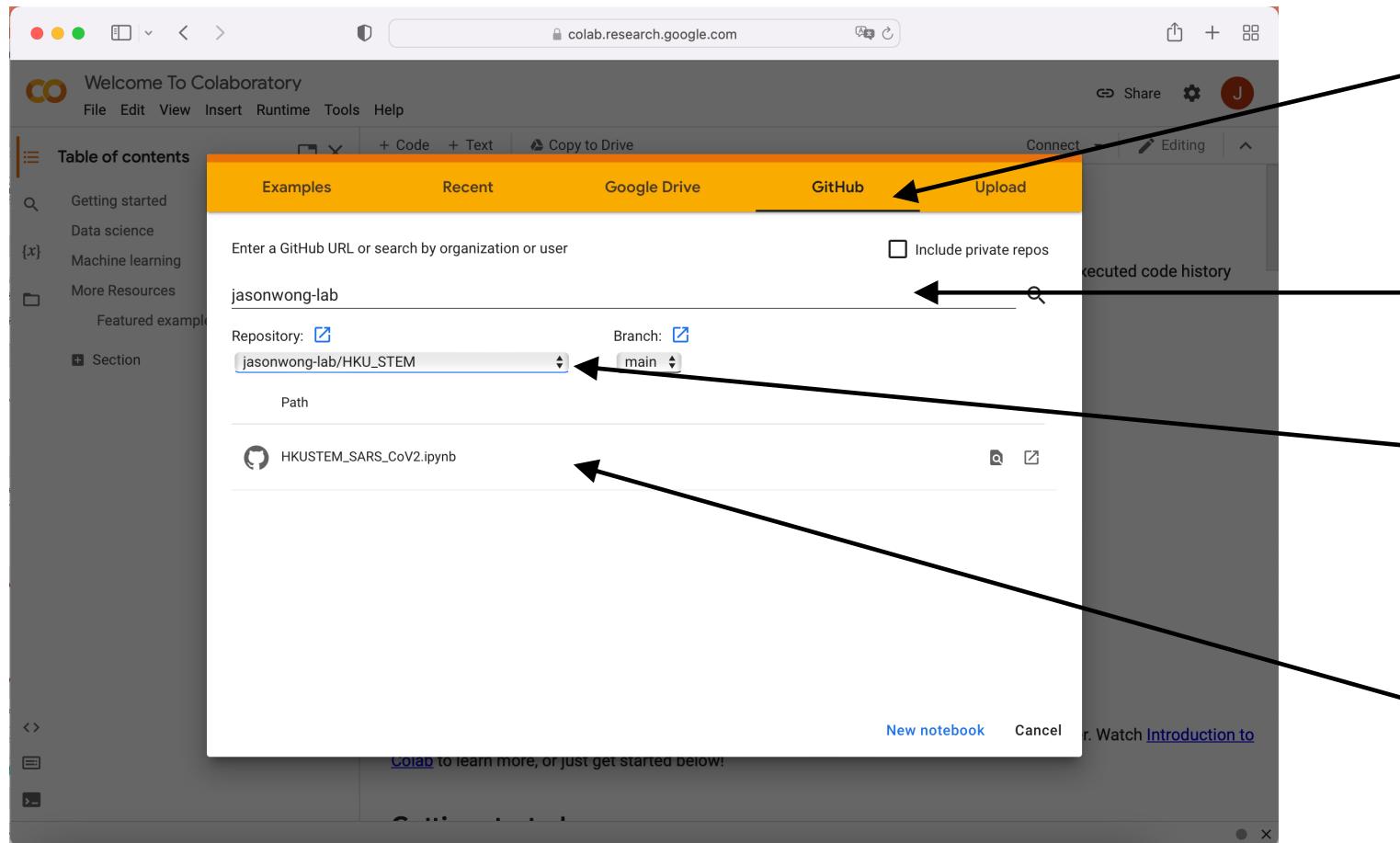
- 1) Set up Google Colab (5 minutes)
- 2) Multiple sequence alignment of Coronaviruses (15 minutes)
- 3) Alignment of Spike protein (10 minutes)
- 4) Visualise Spike protein (10 minutes)

# Google Colab (<https://colab.research.google.com>)



Sign in with your Google account

# Open jupyter notebook from GitHub



After logging in a window should open to open a notebook. If not go to File->Open notebook  
Go to the GitHub tab

Search for “jasonwong-lab”

Select jasonwong-lab/HKU\_STEM

Double click  
HKUSTEM\_SARS\_CoV2.ipynb

HKUSTEM SARS-CoV2.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text Copy to Drive Connect Editing

HKU STEM workshop 2022 - "Data Science Lab": how sequencing is used to identify SARS-CoV2 strains  
by Dr Jason Wong and Dr Joshua Ho

The objective of this Data Science Lab is show demonstrate how sequencing data of viruses are used to identify coronavirus strains and to study their effect on the virus.

▼ Set working directory

By default working directoy will be My Drive/SARS-CoV2

```
[ ] # set working pathway to your own google drive doc (~ 1 min)
from google.colab import drive
drive.mount('/content/gdrive') # if using for the first time, you be requested to grant permission to link

import os
try:
    os.mkdir("/content/gdrive/My Drive/SARS-CoV2") # change this path if necessary
except FileExistsError:
    print("directory already exist. OK to continue")
os.chdir("/content/gdrive/My Drive/SARS-CoV2")

Mounted at /content/gdrive
```

pwd

'/content/gdrive/My Drive/SARS-CoV2'

# Resources

- NCBI SARS-CoV2 resource  
(<https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/sars-cov-2>)
- PDB spike protein (<https://pdb101.rcsb.org/motm/246>)
- Understanding evolution and phylogenetics  
(<https://evolution.berkeley.edu/>)