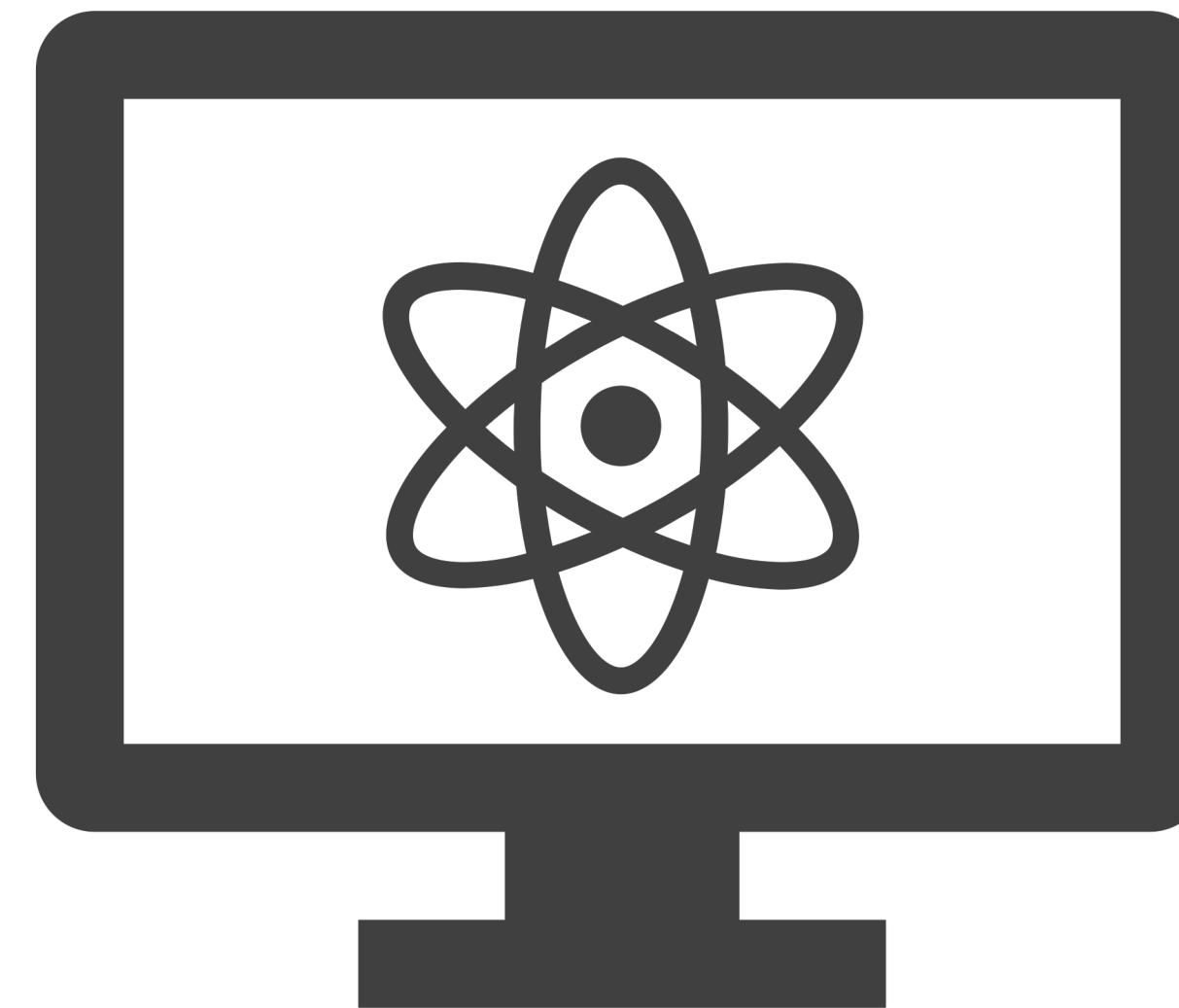


# Bioinformatics on the Command Line



**Instructor: Daryl Domman, PhD**

# Timetable for the week

Sunday	Monday	Tuesday	Wednesday	Thursday
Intro & VM	Data QC	Assembly and annotation	SARS-CoV-2 nextflow	Lineage calling and phylogeny
Command Line	Mapping	SARS-CoV-2 Illumina	SARS-CoV-2 nanopore	Nextstrain

# What do we hope to achieve together?

- To provide a broad overview of bioinformatic tools and techniques used for the analysis of pathogen genomes
- Some theory, but mostly hands on training
- Build confidence in applying these concepts and approaches to your own datasets
- Form networks and help each other!

# Questions genomics can help answer

What is causing the outbreak?

Where did the outbreak begin?

When did it begin?

How many introductions have there been?

Are outbreaks / cases linked?

What is the resistance profile?

How is the pathogen evolving – new variants?

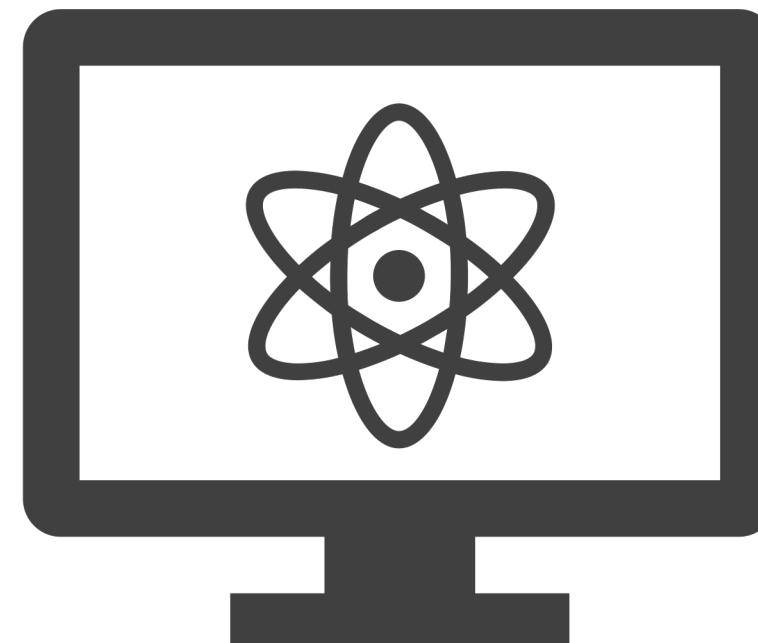
# Why the command line?

## Pro

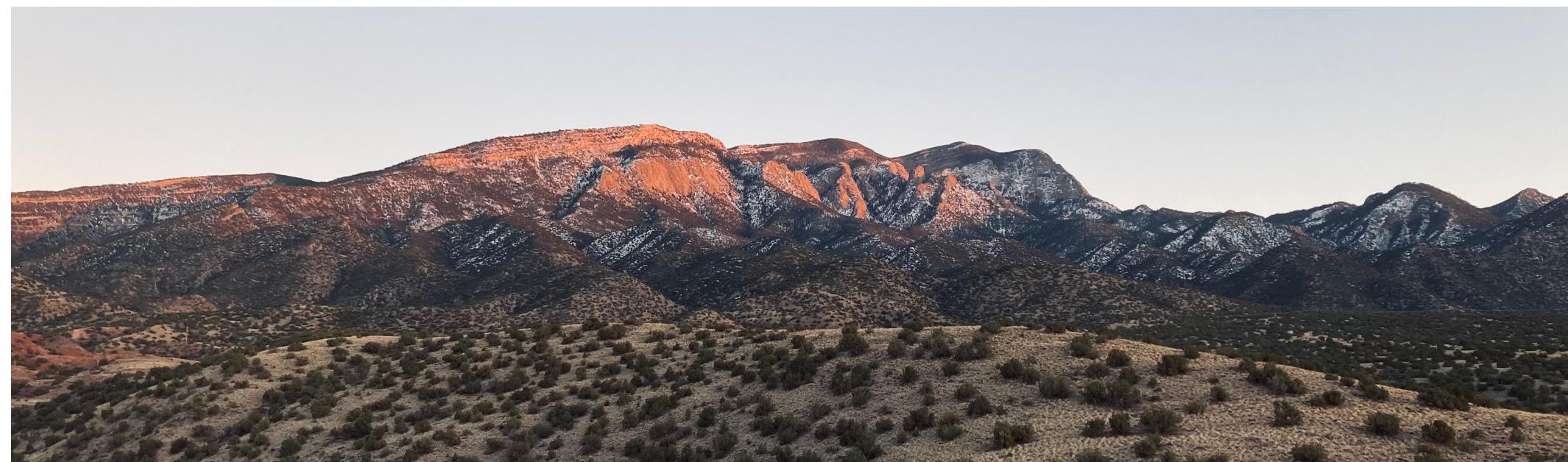
- Ultimate control over workflow
- Doesn't depend on internet
- Data privacy
- Can be cheaper
- Use latest tools

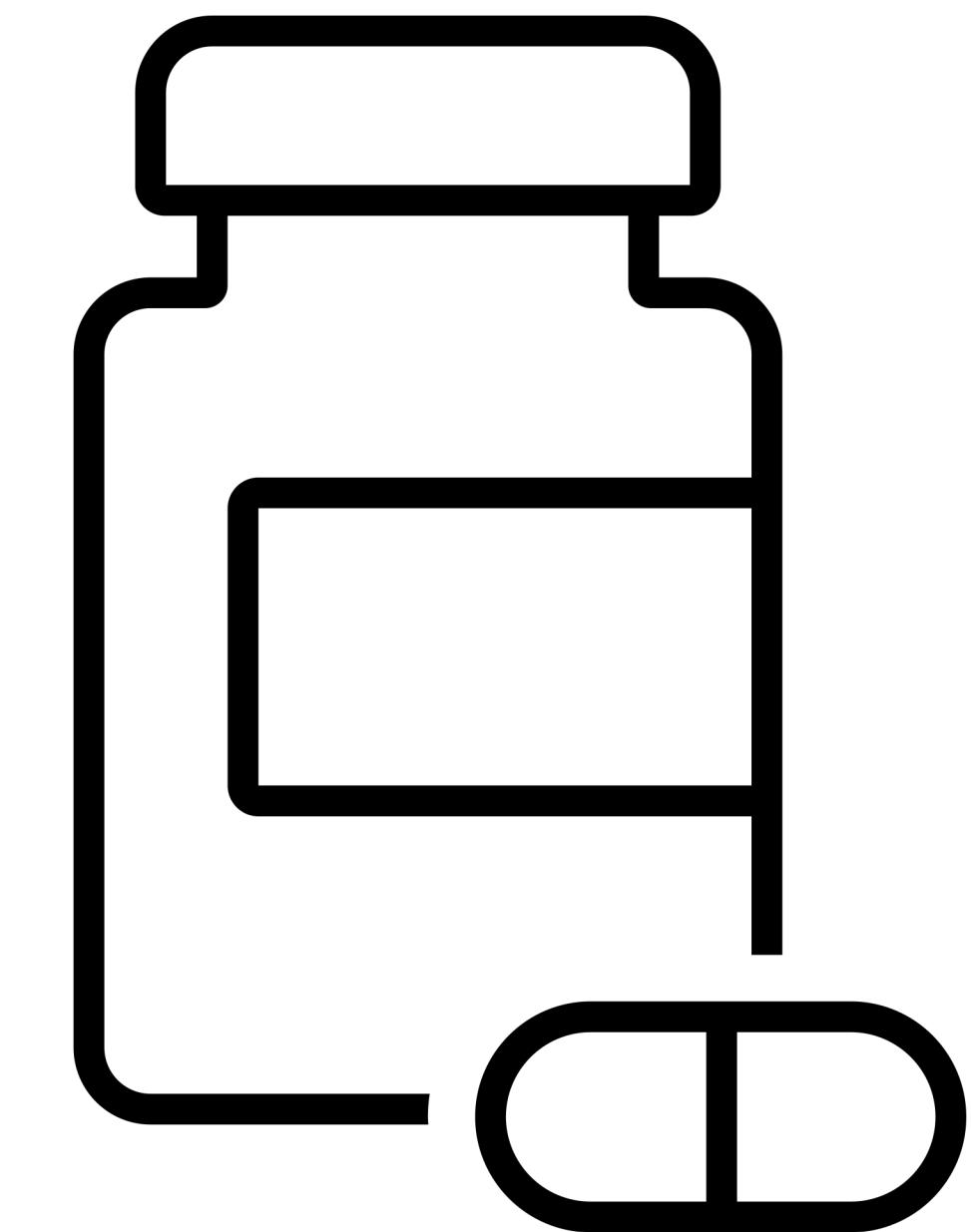
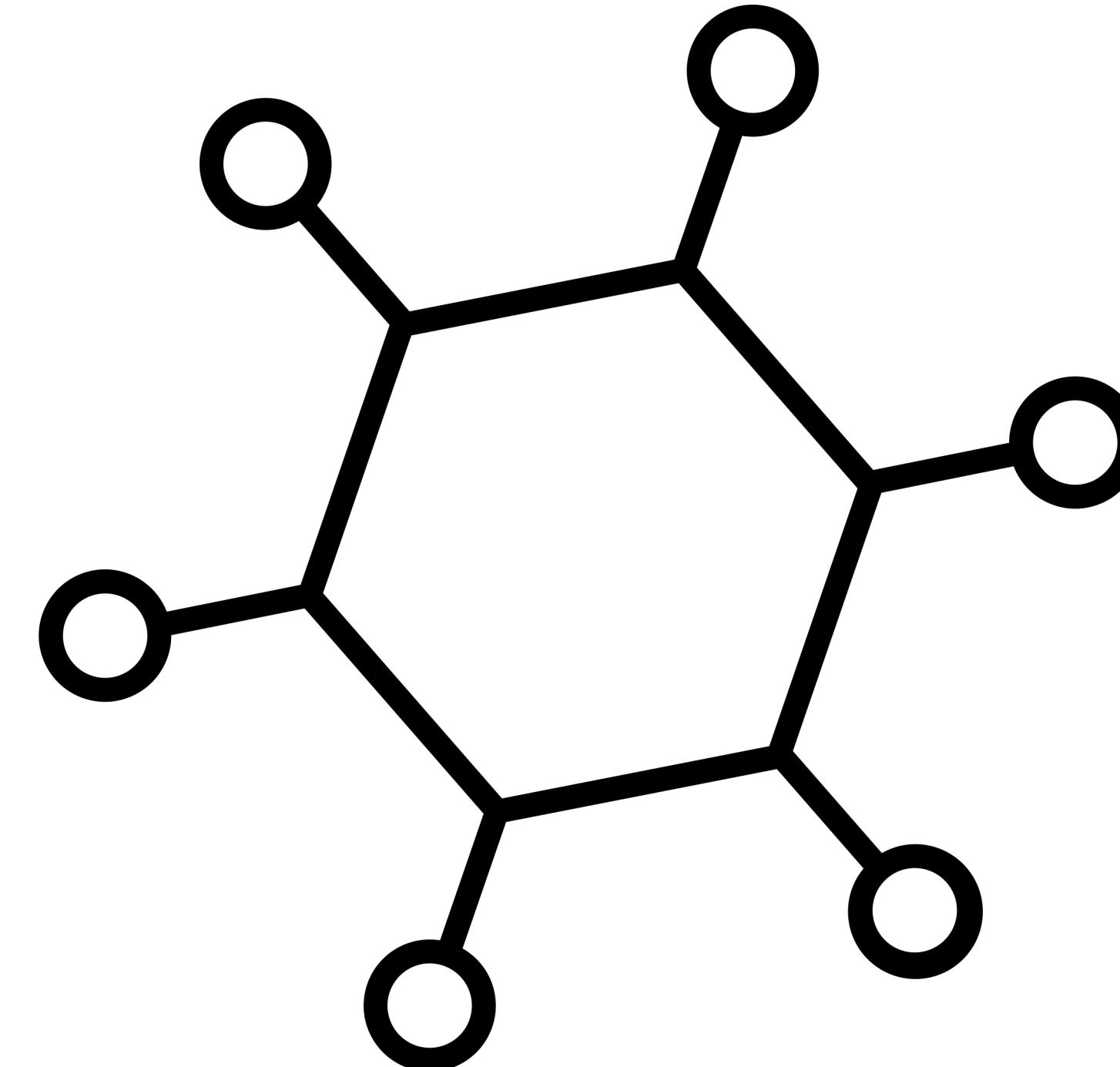
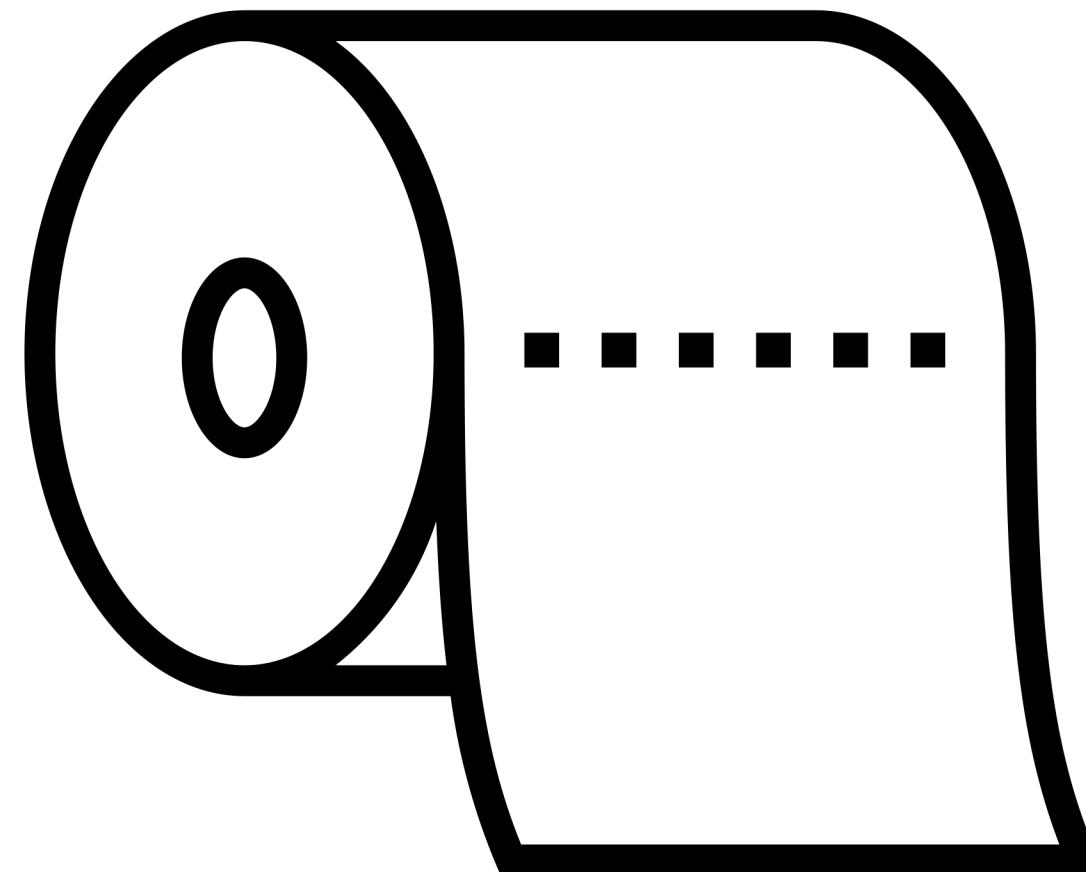
## Con

- May be a steep learning curve
- You must make the decisions
- Errors and troubleshooting
- Output not always in easy to read and interpret format



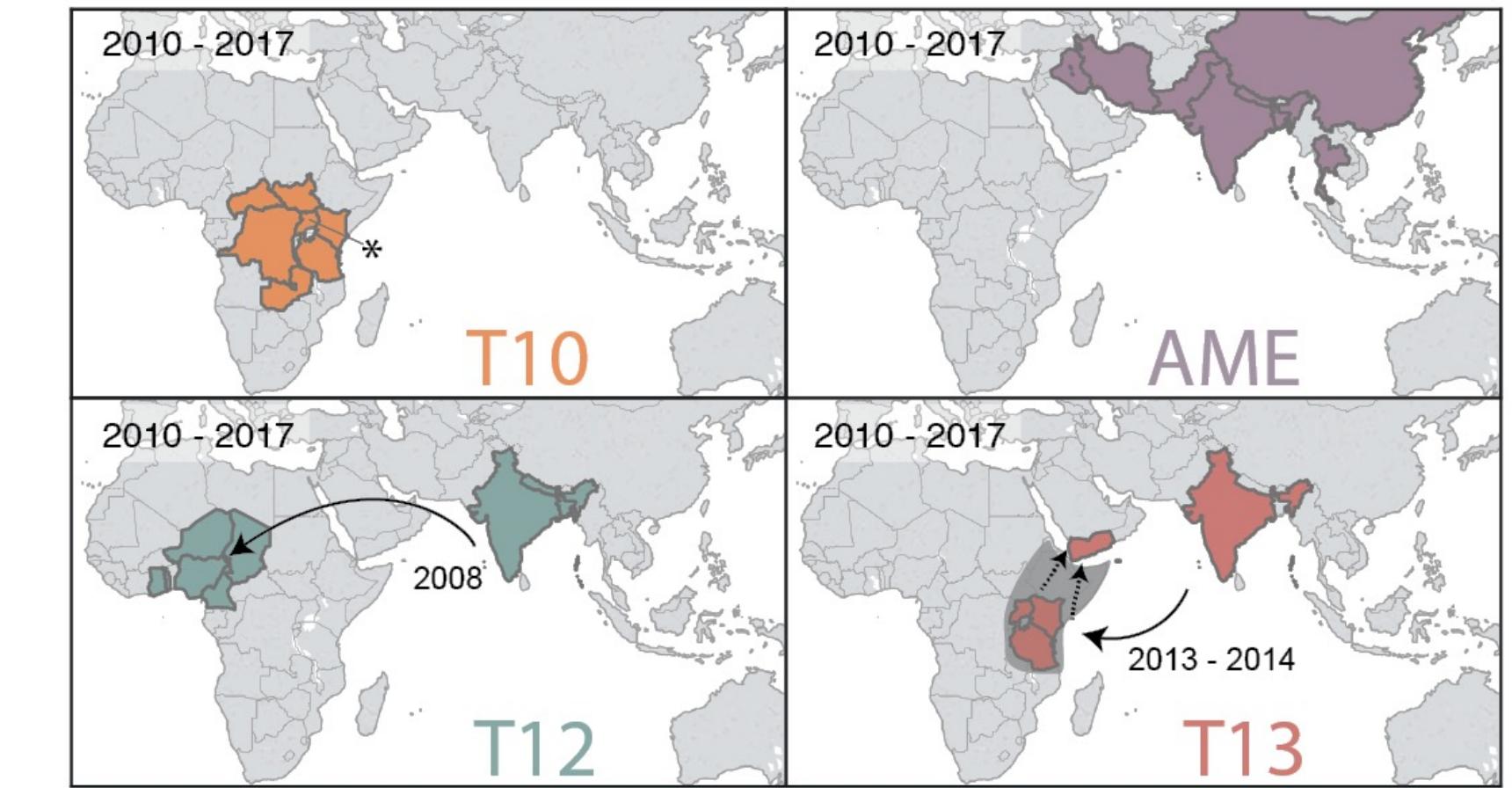
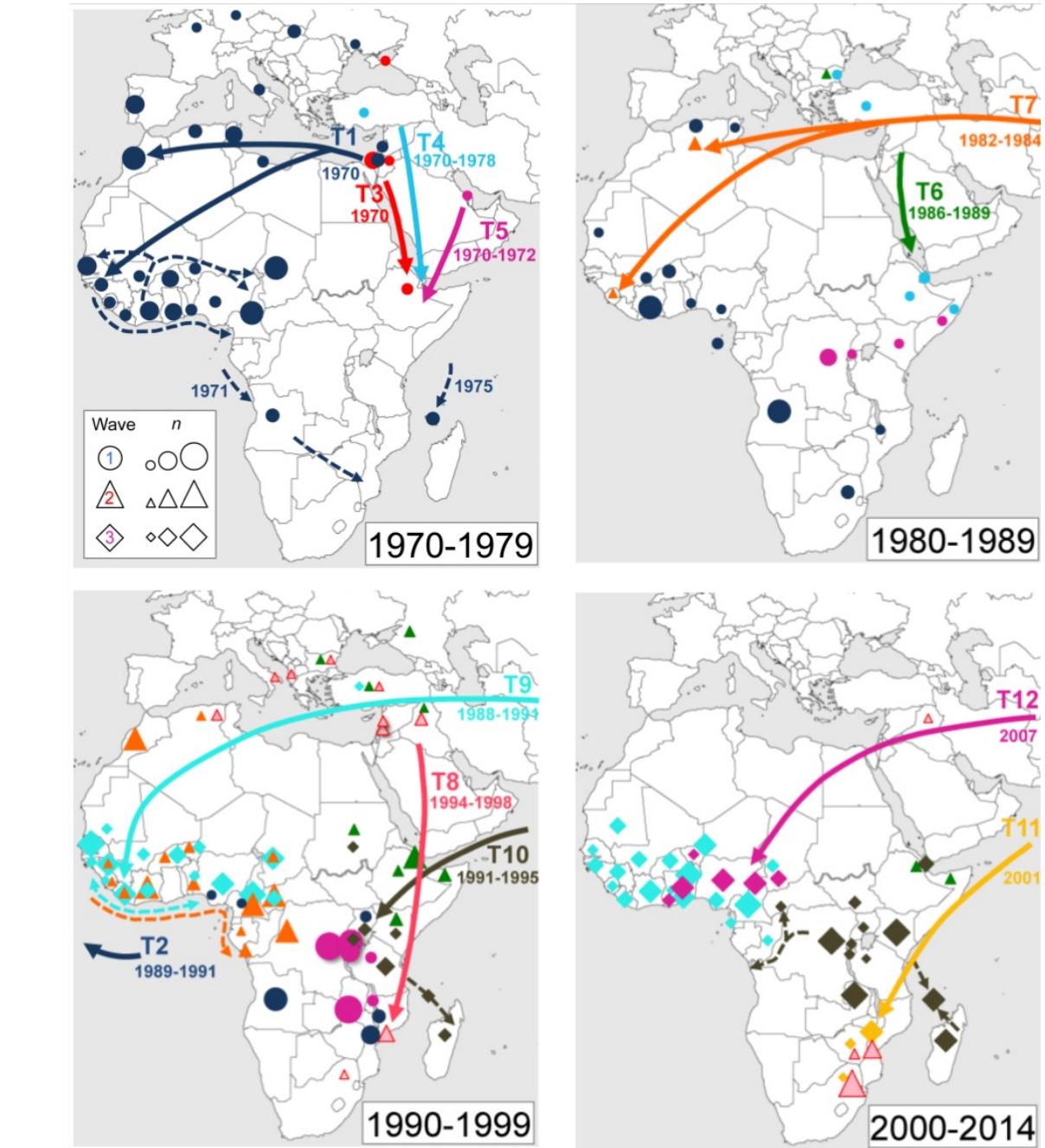
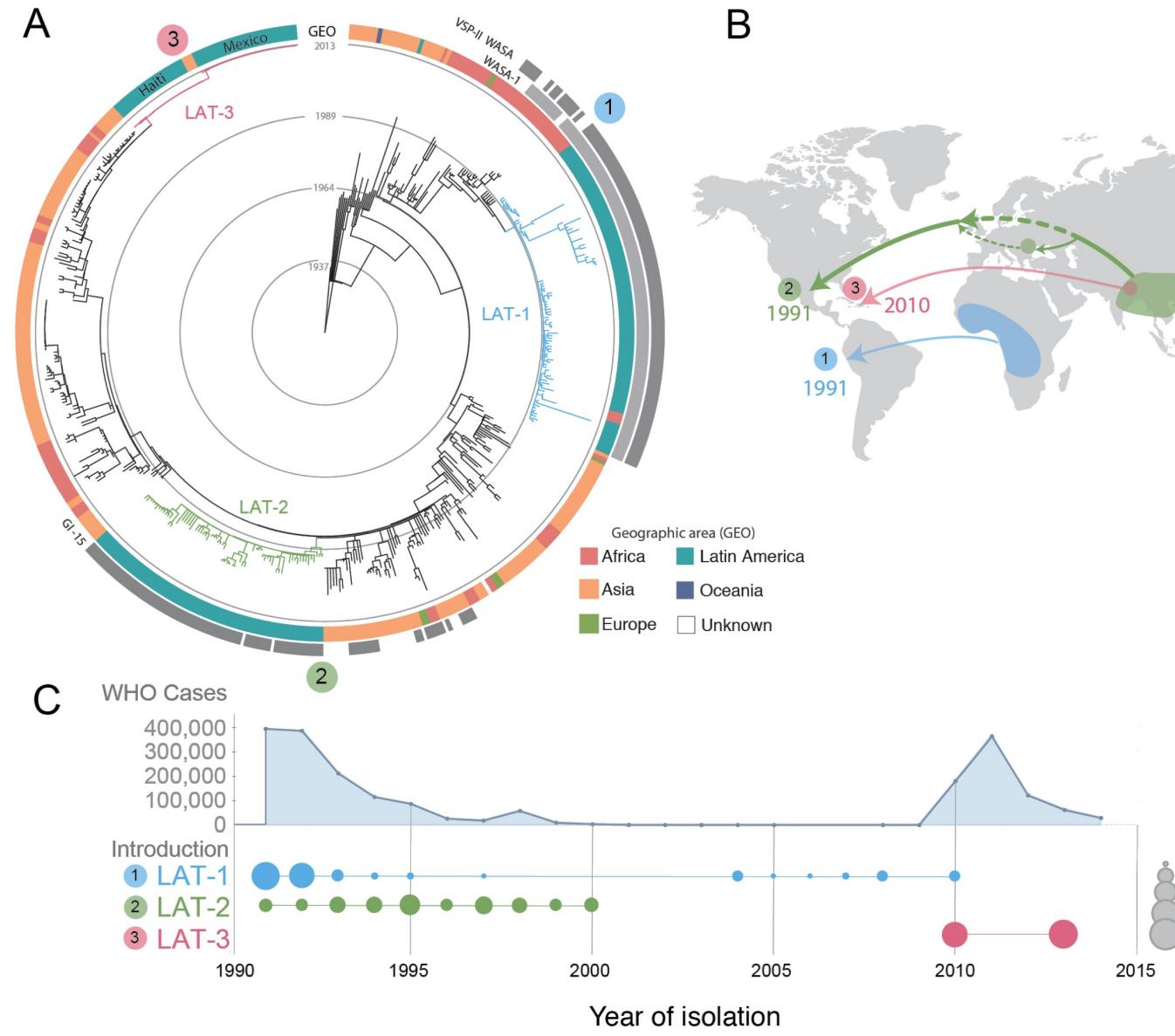
# A bit about me





Daryl Domman, PhD  
Assistant Professor  
Center for Global Health  
Department of Internal Medicine  
University of New Mexico School of Medicine

# Pandemic cholera dynamics



## Cholera across Latin America

Domman et al, Science 2017

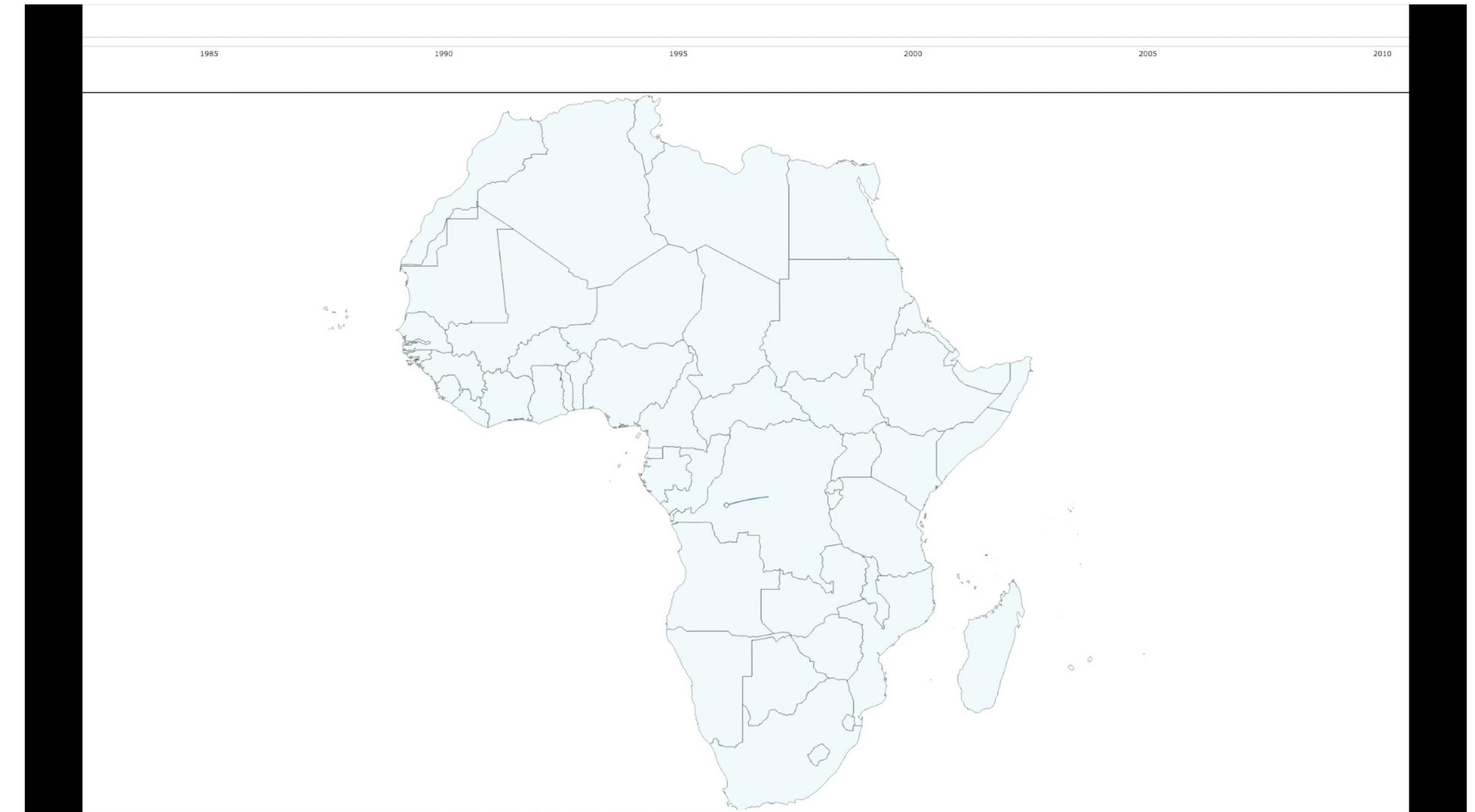
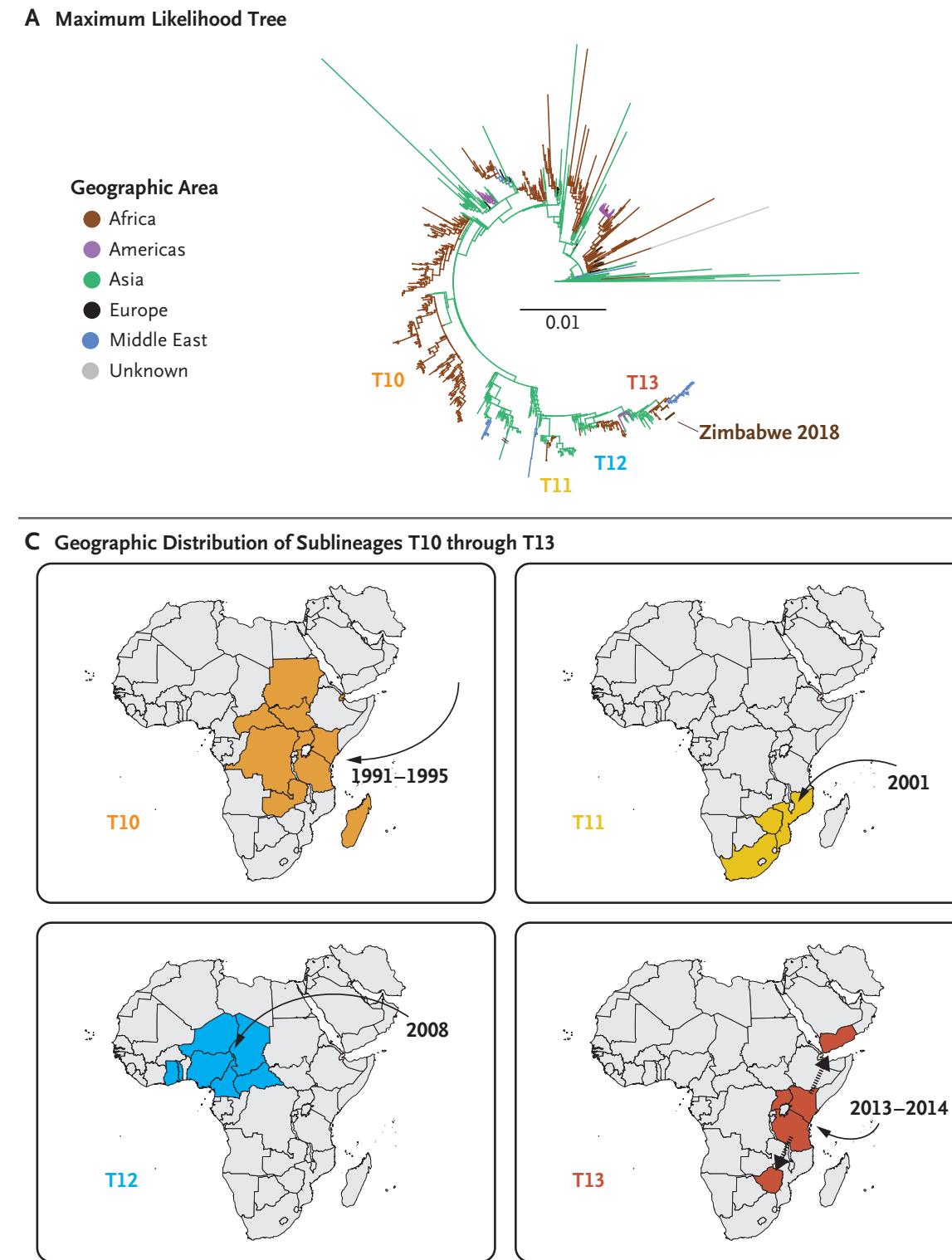
## Cholera across Africa

Weill, Domman et al, Science 2017

## Yemen epidemic

Weill\* & Domman\* et al, Nature 2019

# Cholera transmission in Africa

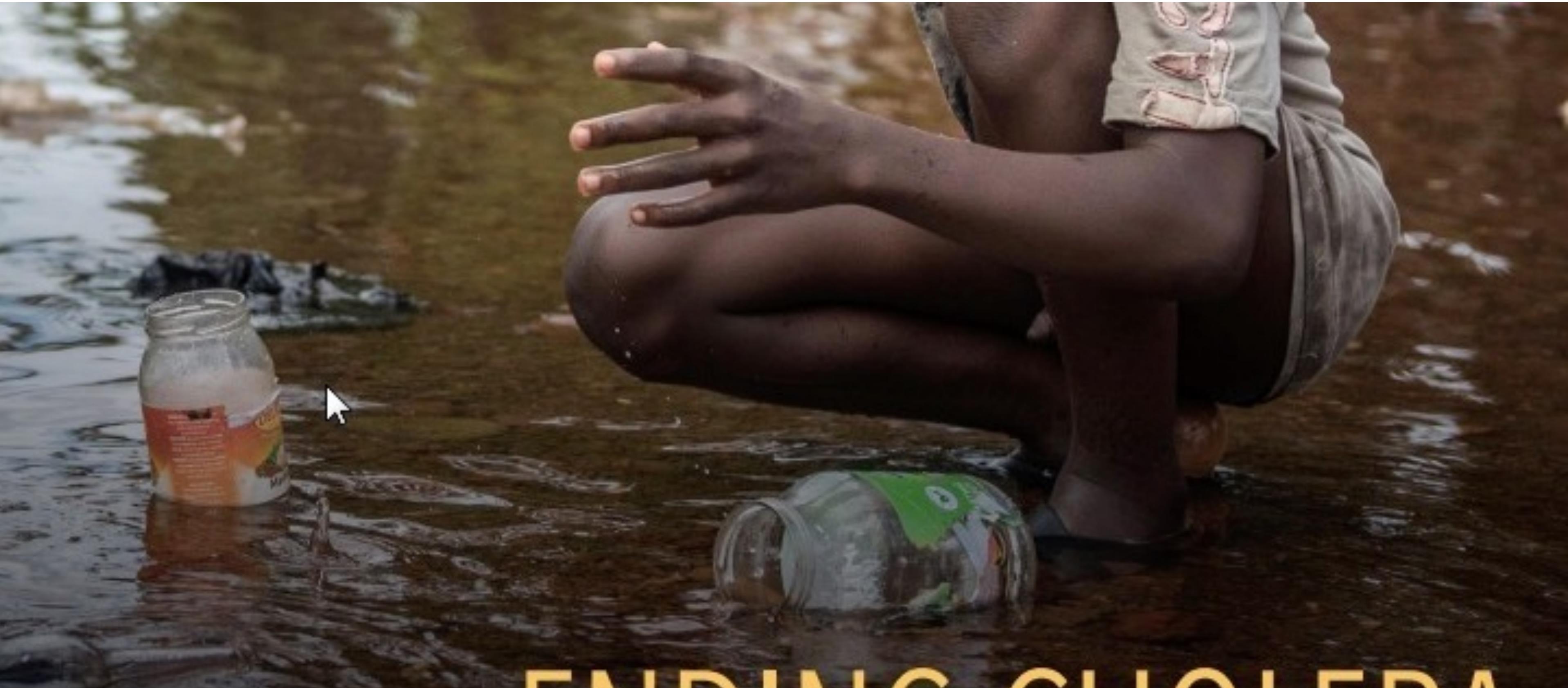


## Cholera outbreak Zimbabwe

Mashe\* & Domman\* et al,  
New England Journal of Medicine, 2020

## Cholera across DRC

Collaboration with  
Johns Hopkins

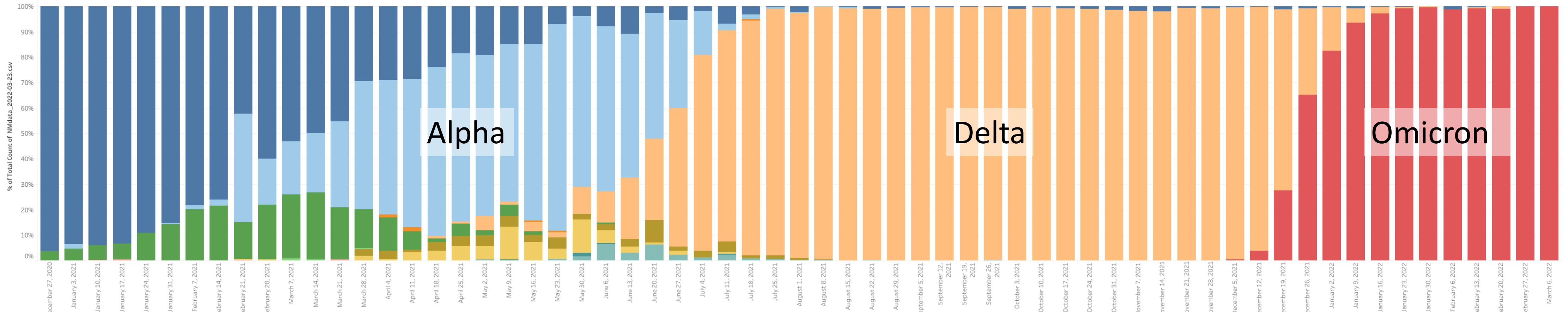
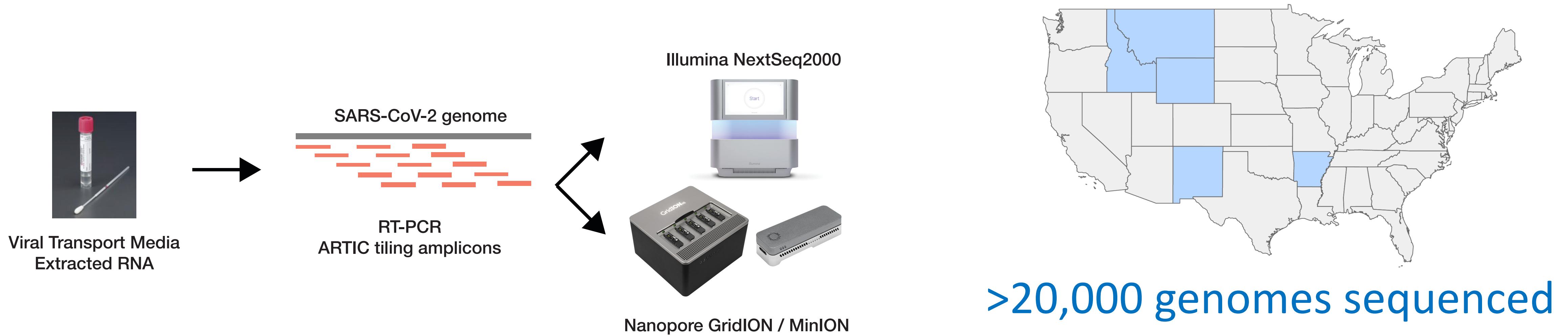


GLOBAL TASK FORCE ON  
CHOLERA CONTROL

# ENDING CHOLERA

## A GLOBAL ROADMAP TO 2030

# Rocky Mountain COVID Consortium + Arkansas



# Introductions for group

Questions?

# Today's Agenda

-  **VM installation and setup**
  -  **Bioinformatics data types**
  -  **Module 1 : SARS-CoV-2 web tools in VM**
  -  **Module 2 : Using the command line**
-



VIRTUAL MACHINE



# Step 1 : Email invite



Microsoft Azure <azure-noreply@microsoft.com>

To: domman.genomics@gmail.com

Today at 10:03 AM



**Daryl Domman invited you to the lab:  
JordanNGS**

Register now to access the virtual machines in the lab.

[Register for the lab >](#)

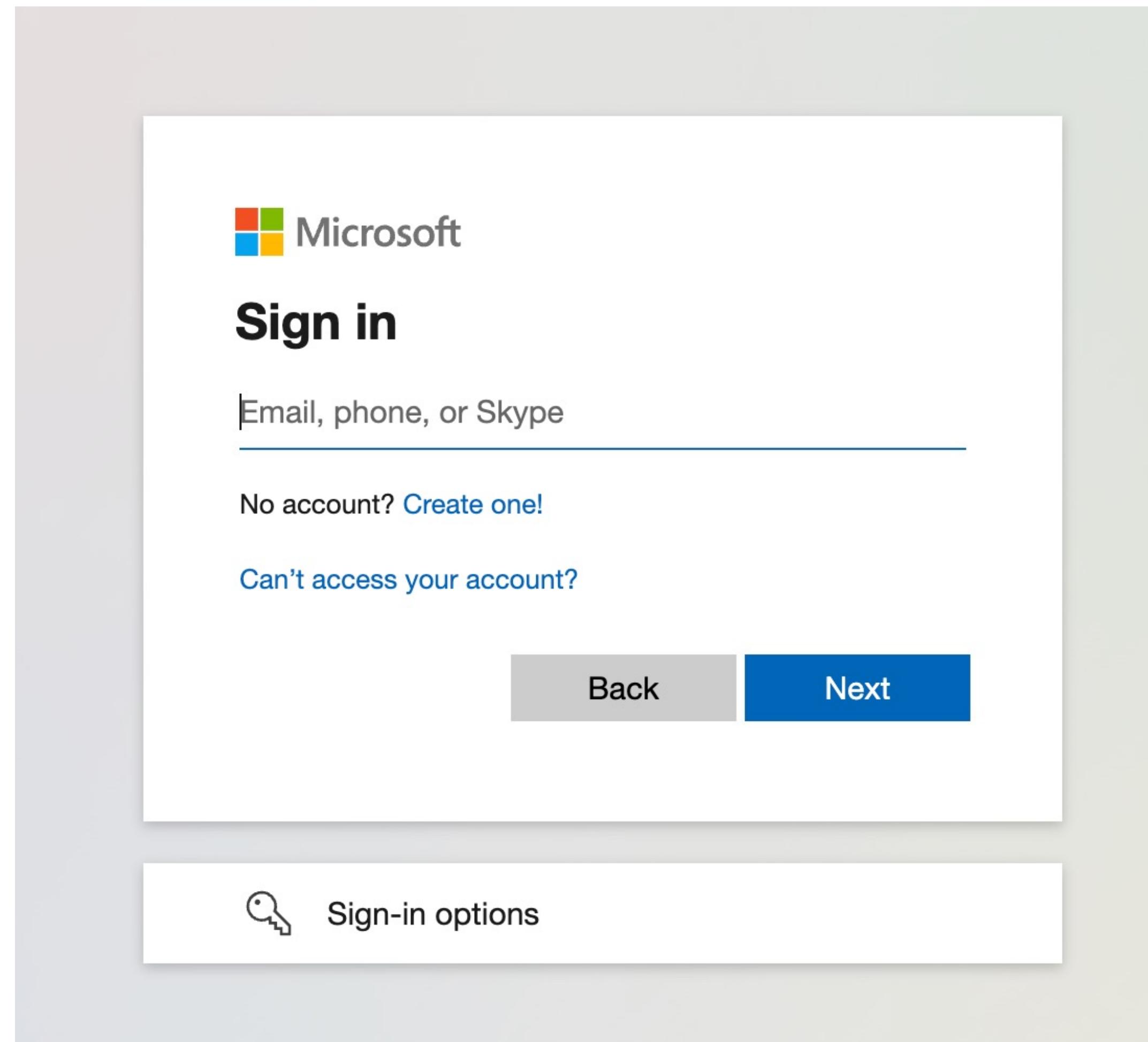


[Privacy Statement](#)

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

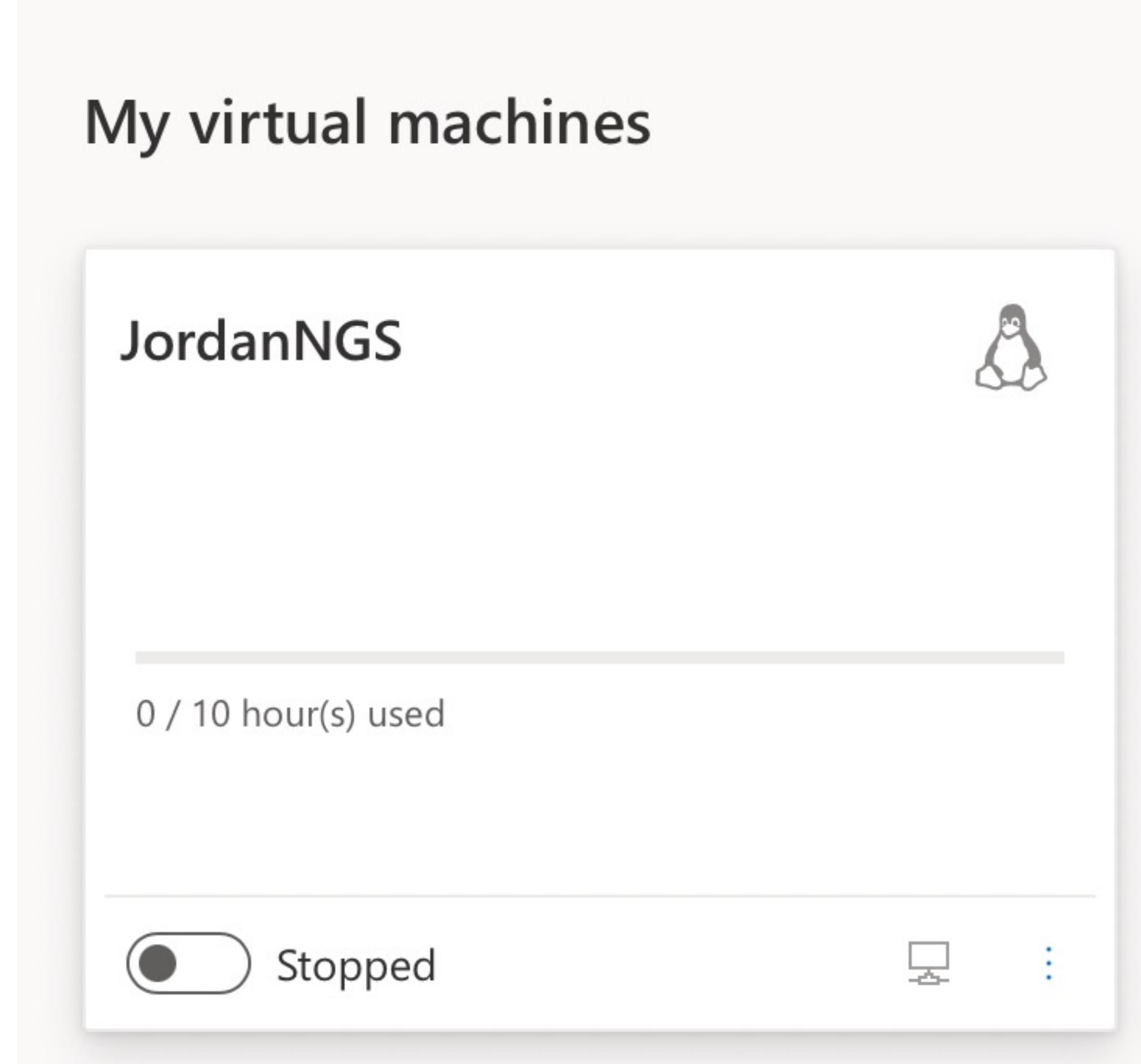


# Step 2 : Microsoft Account

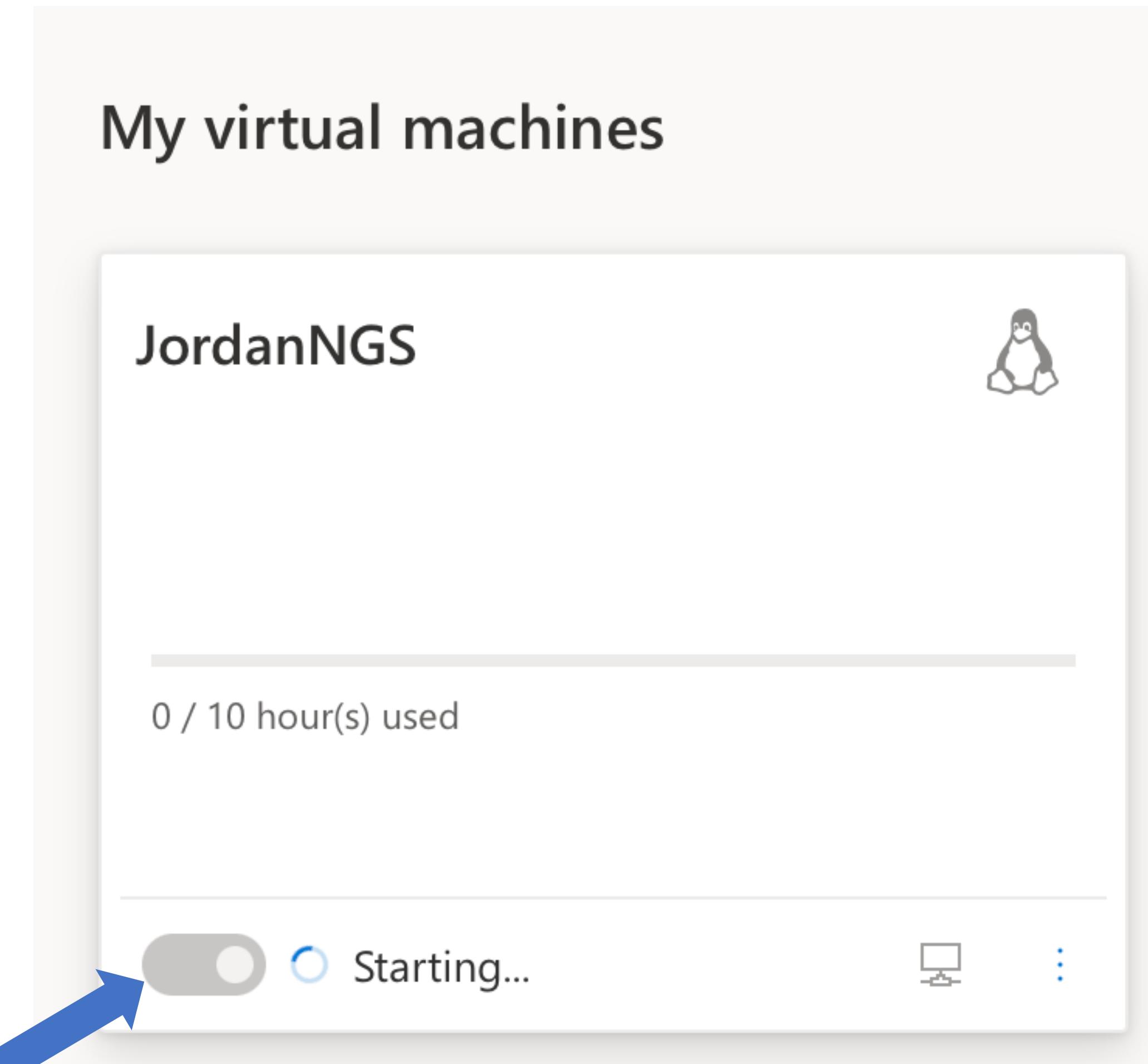


For Azure Lab – needs to be same as invite email

# Step 3 : Azure Lab main screen

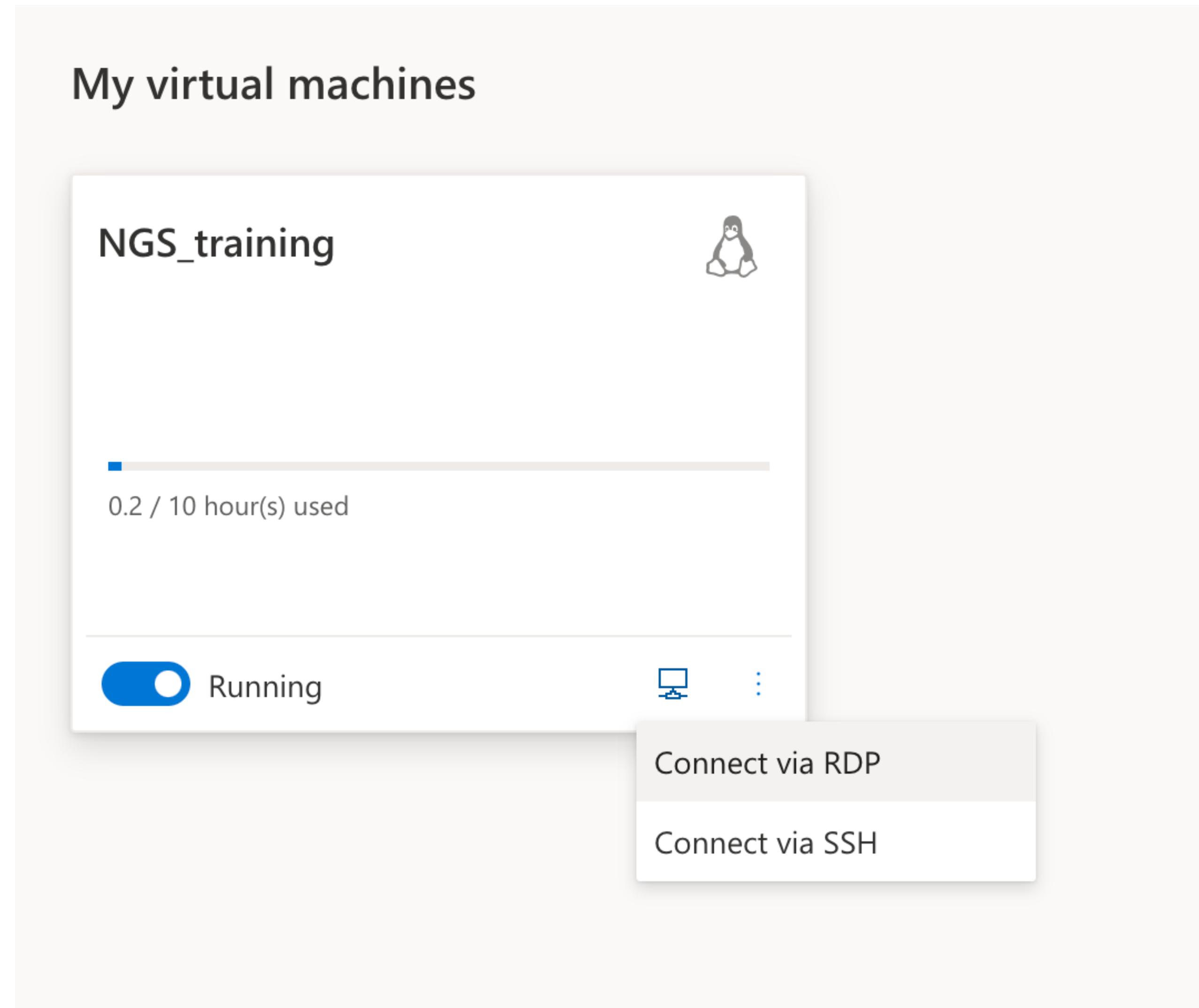


# Step 4 : Start Virtual Machine



Click the slider to “Start” the VM

# Step 5 : Virtual Machine Starting



# Step 6 : Install Microsoft Remote Desktop

The image shows a screenshot of the Microsoft website. At the top, there is a navigation bar with links for Home, Devices, Software, Games & Entertainment, Deals, Shop Business, Students & parents, More, All Microsoft, Search, Cart, and a user profile icon. Below the navigation bar is a large, stylized illustration of a central blue cloud connected to various devices: a desktop monitor, a laptop, a smartphone, a tablet, and a server tower, all represented in a glowing blue color. In the foreground, there is a white rectangular box containing information about the Microsoft Remote Desktop app. The box includes a blue square icon with a white 'K' logo, the app's name 'Microsoft Remote Desktop', the developer 'Microsoft Corporation', the category 'Productivity', a brief description of the app's purpose, a 'More' link, an ESRB rating of 'EVERYONE', and a 'Free' download button labeled 'Get'. There is also a link to 'See System Requirements'.

**Microsoft Remote Desktop**

Microsoft Corporation • Productivity

Use the Microsoft Remote Desktop app to connect to a remote PC or virtual apps and desktops made available by your admin. The app helps you be productive no matter where you are.

More

EVERYONE

Free

Get

△ See System Requirements

**Does not matter what email you use to download !!!**

# Step 6 : Install Microsoft Remote Desktop (Mac)

The screenshot shows the Microsoft Remote Desktop page in the Mac App Store. The app icon is orange with a white 'K' logo. The title is 'Microsoft Remote Desktop' with the tagline 'Work from anywhere'. A large blue 'OPEN' button is visible. Below it, the app has 77K ratings (4.6), is suitable for ages 4+, and is ranked #2 in the Business category. It was developed by Microsoft Corporation and is available in English (53.5 MB). The 'What's New' section notes a quick update to improve connection reliability for Azure Virtual Desktop (AVD) scenarios. The 'Preview' section shows three screenshots of the app interface running on a Mac, displaying a desktop environment with multiple windows and a search bar.

microsoft remote des...

**Discover**

**Arcade**

**Create**

**Work**

**Play**

**Develop**

**Categories**

**Updates**

**Microsoft Remote Desktop**  
Work from anywhere

**OPEN**

77K RATINGS  
**4.6**  
★ ★ ★ ★

AGE  
**4+**  
Years Old

CHART  
**# 2**  
Business

DEVELOPER  
Microsoft Corporation

LANGUAGE  
**EN**  
English

SIZE  
**53.5**  
MB

**What's New**

It's time for a quick update. In this release we made some changes to improve the connection reliability for Azure Virtual Desktop (AVD) scenarios.  
If you encounter any errors, you can contact us via Help > Submit Feedback. And, if you are feeling adventurous, you can always test drive the beta version [more](#)

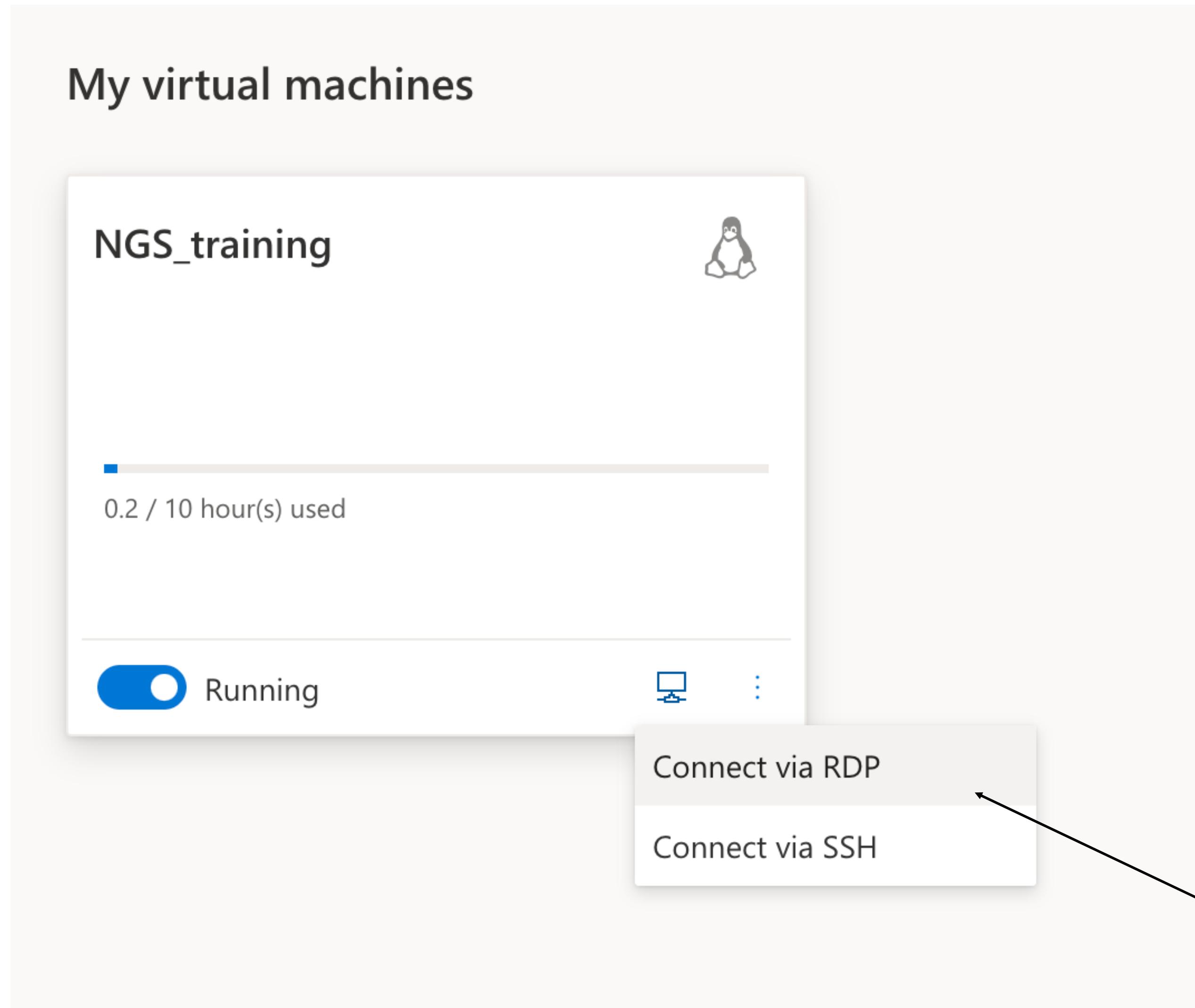
1mo ago  
Version 10.7.6

**Preview**

The preview section displays three screenshots of the Microsoft Remote Desktop application. The first screenshot shows the main interface with a sidebar for 'Administration' and 'Work Resources'. The second screenshot shows a Windows desktop environment with pinned apps like Microsoft Edge, Photos, and File Explorer. The third screenshot shows a search results window for 'Paint 3D' in the Windows Start menu.

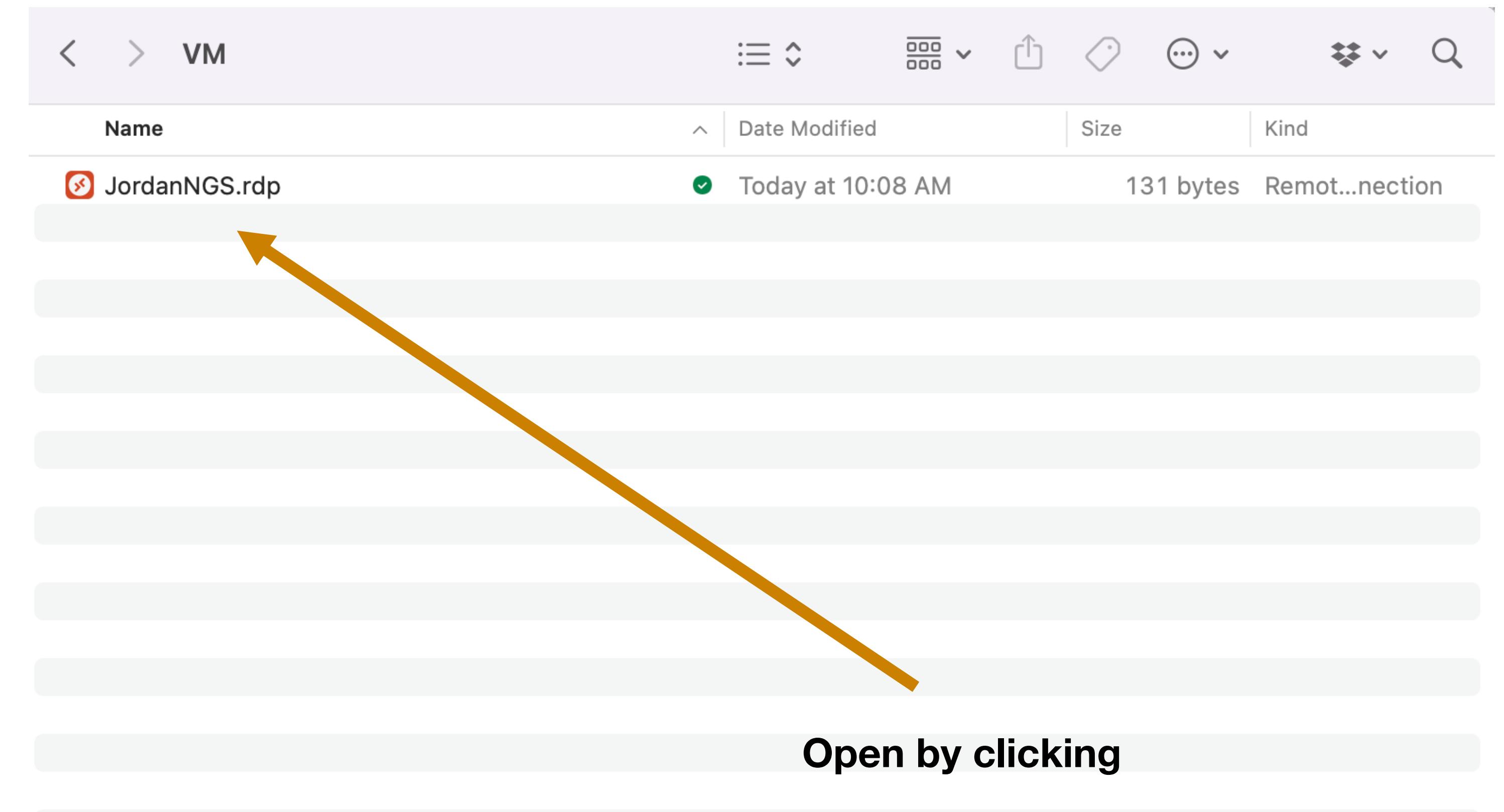
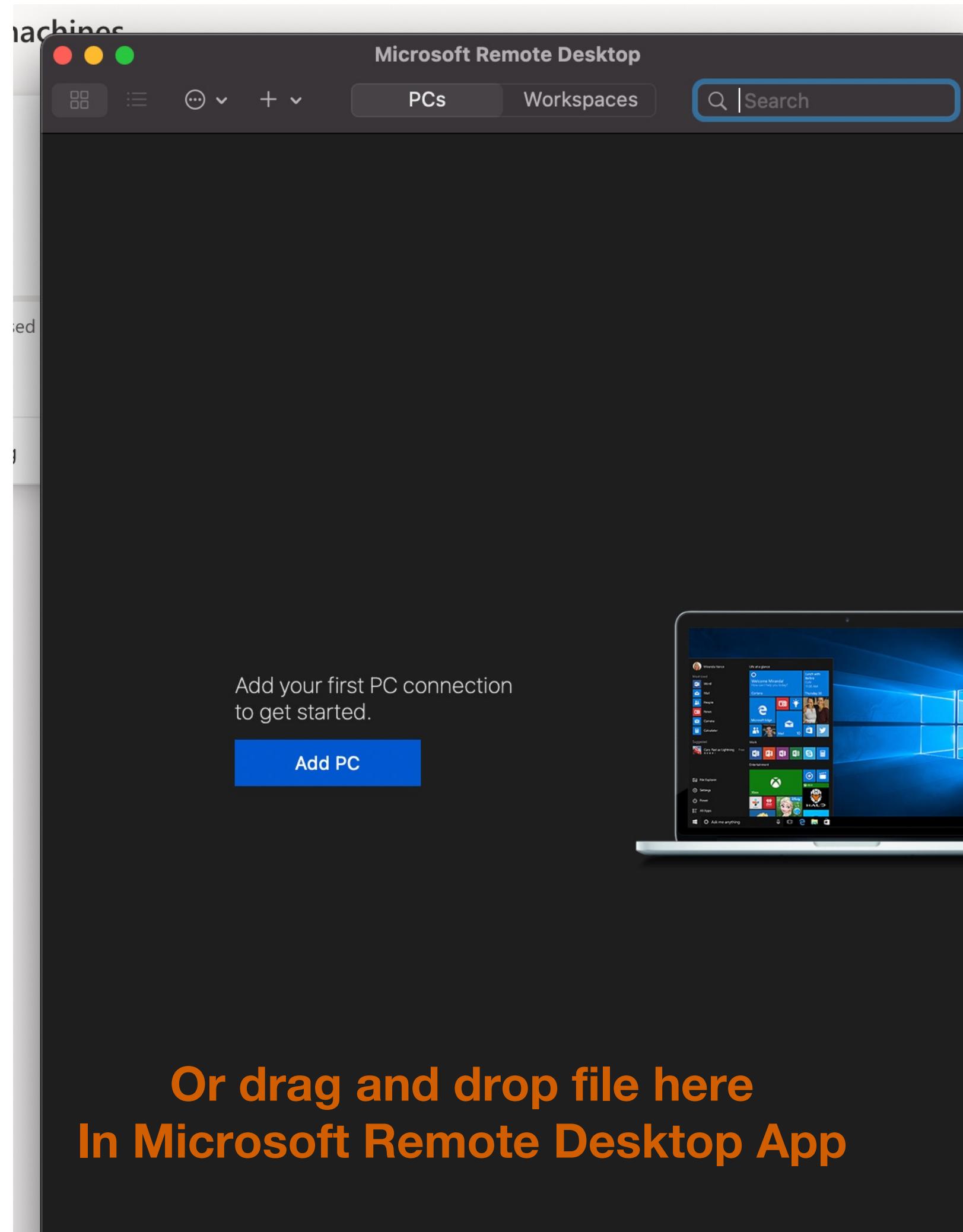
[Download in App Store](#)

# Step 5 : Virtual Machine Starting

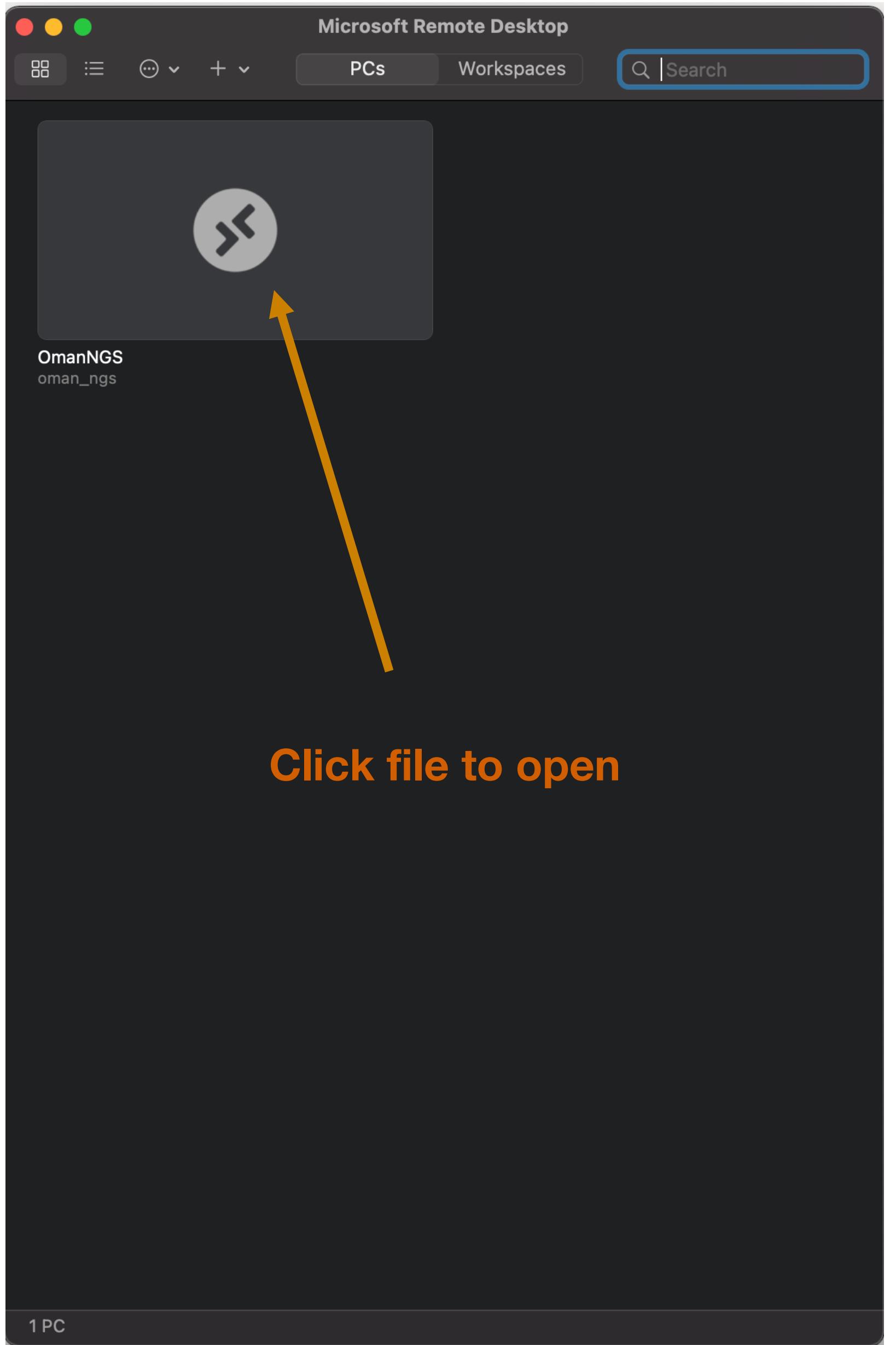


**Click “Connect via RDP”  
Save file**

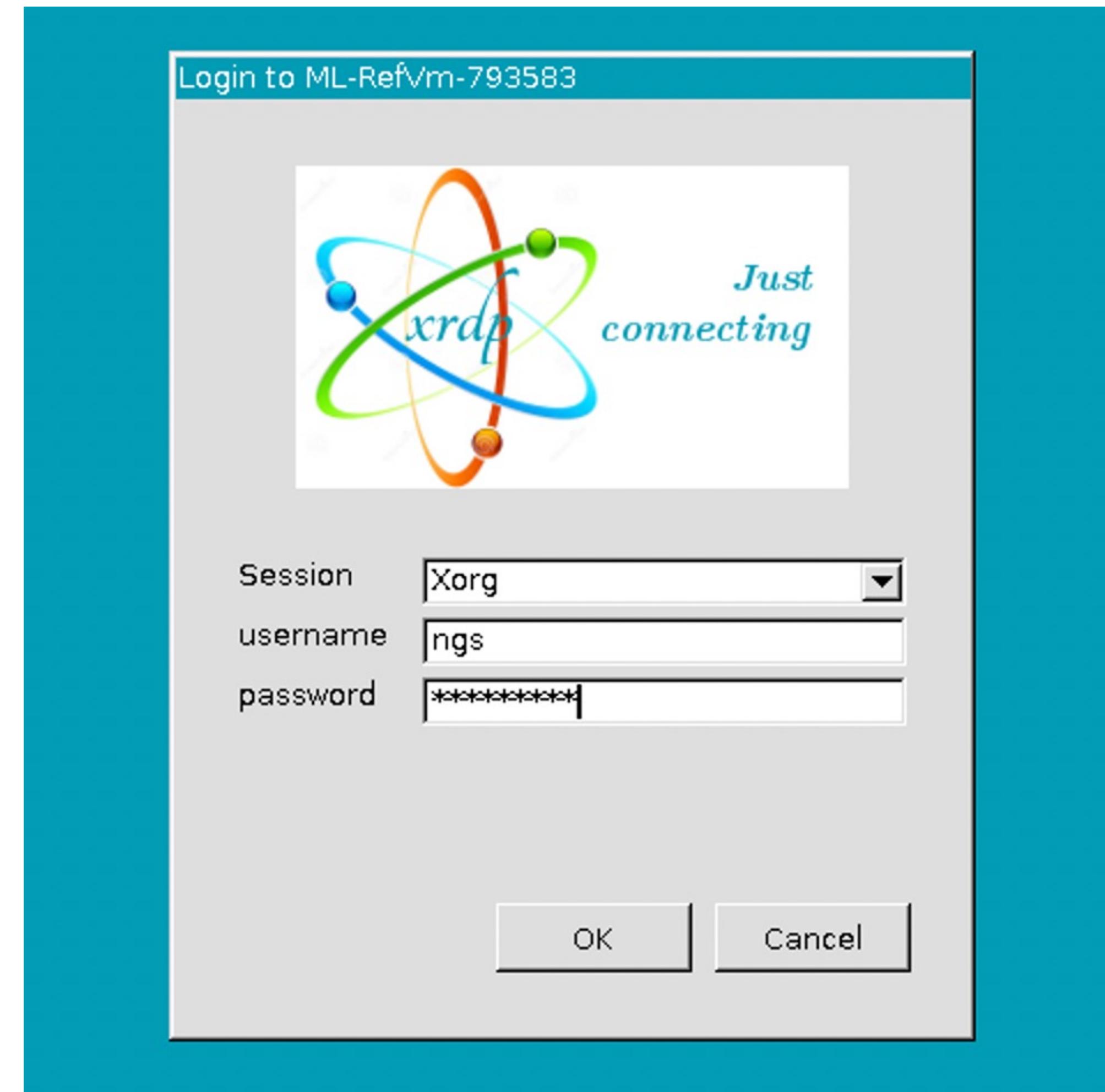
# Step 6 : Open RDP file



# Step 6 : Open RDP file

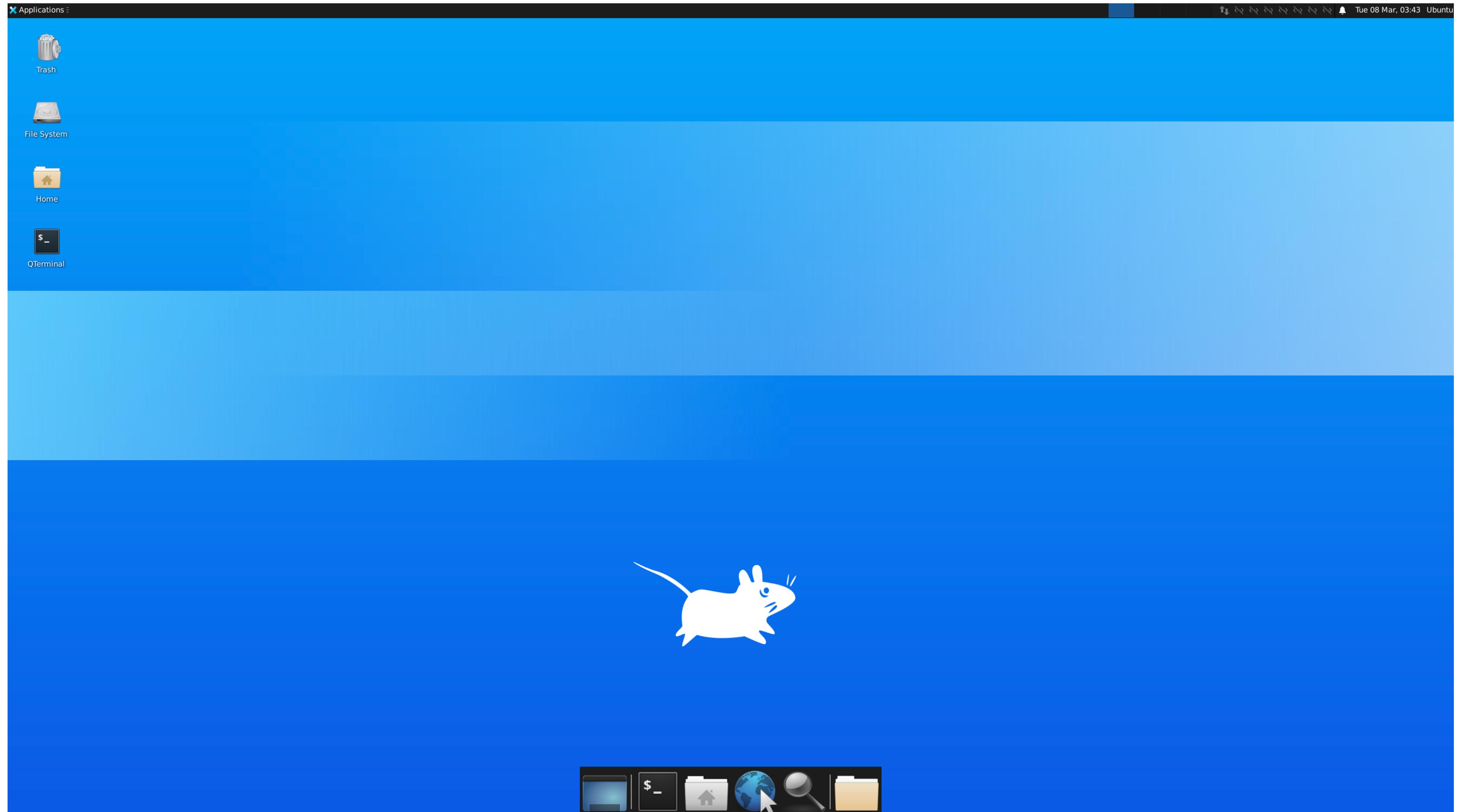


# Step 7 : VM Start Screen

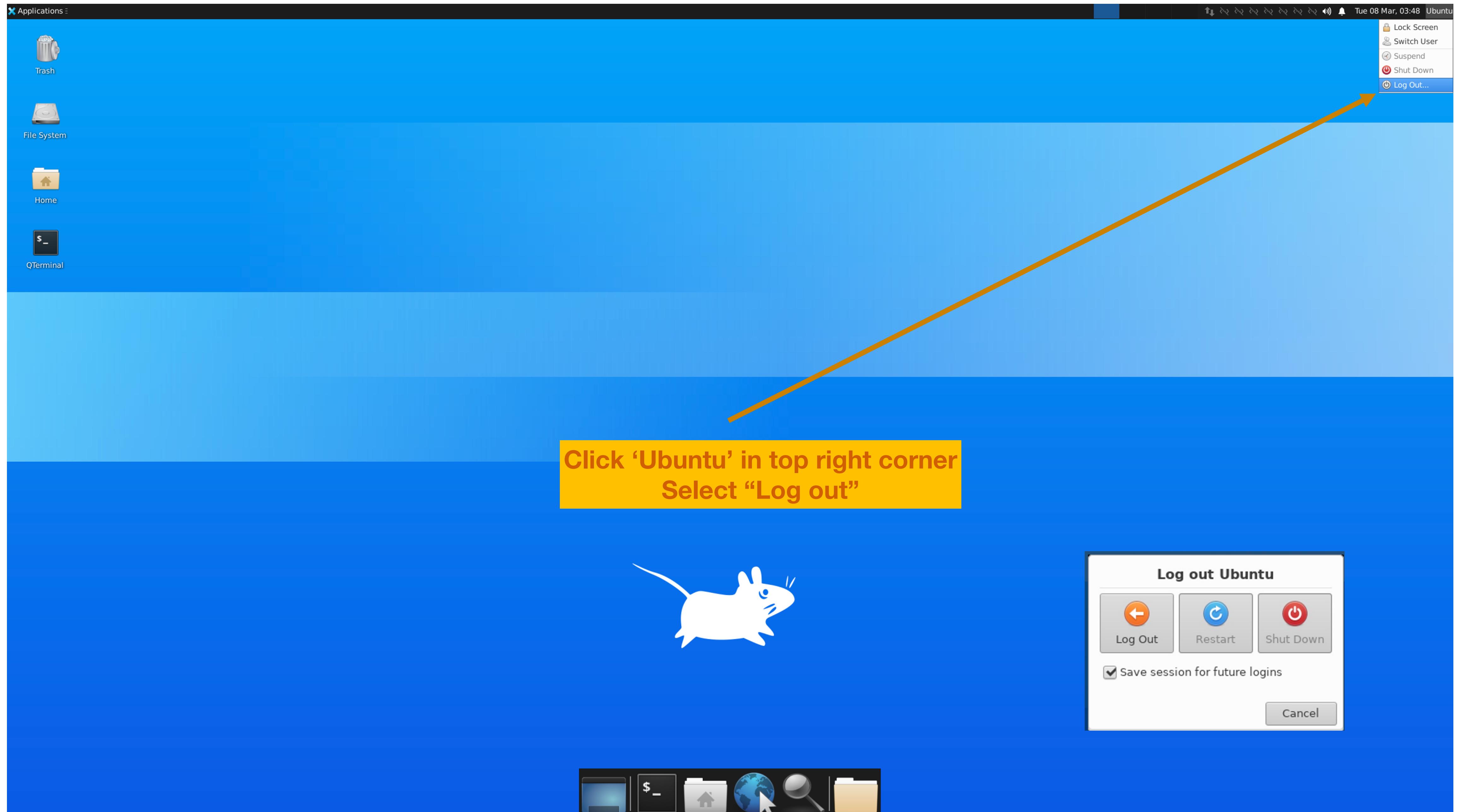


**password: NGSticks!**

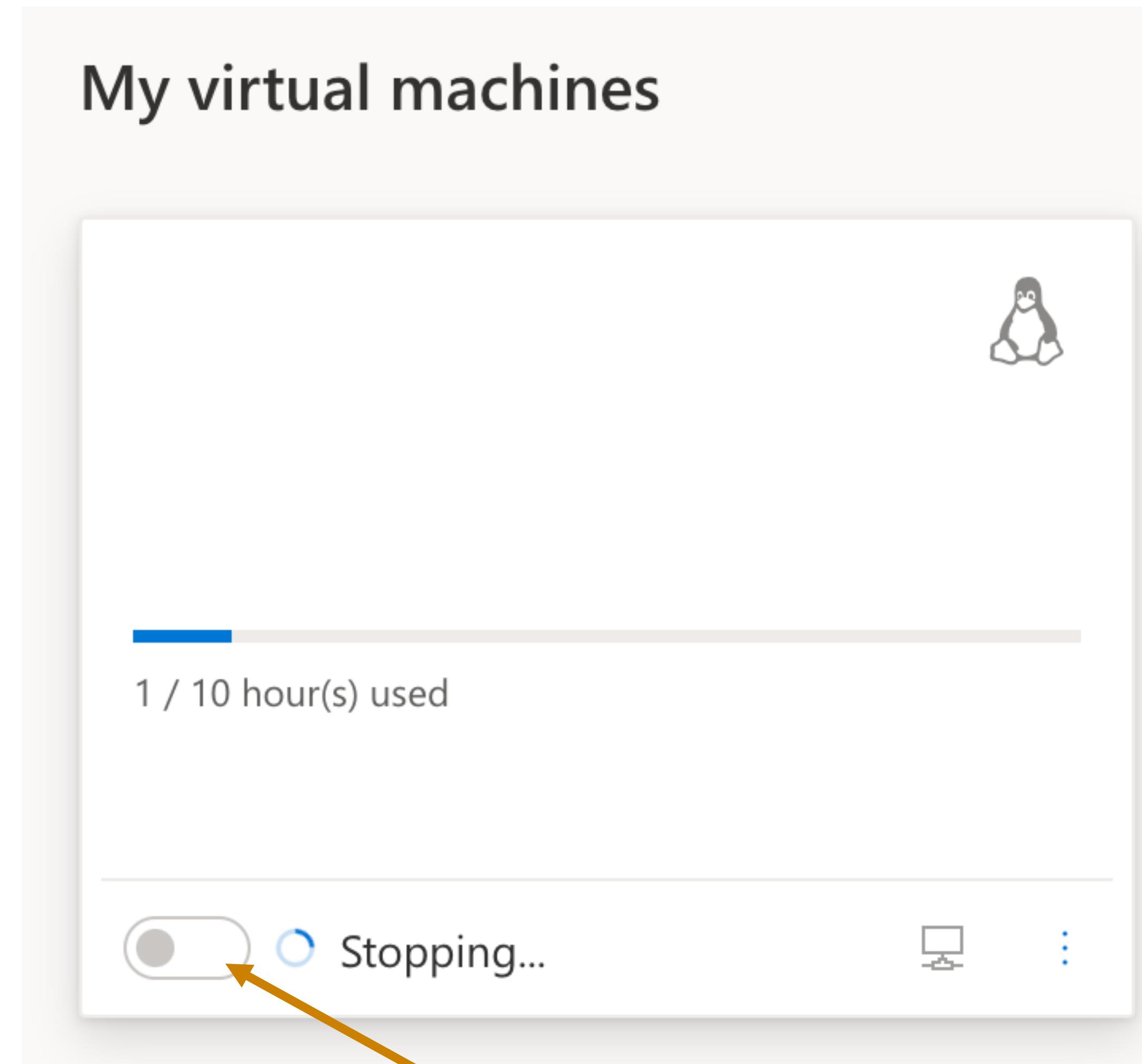
# Step 8 : Have Fun !!!



# Shutting Down the VM



# Shutting Down the VM



**Stop the VM here !!!**

# What is Unix / Linux?

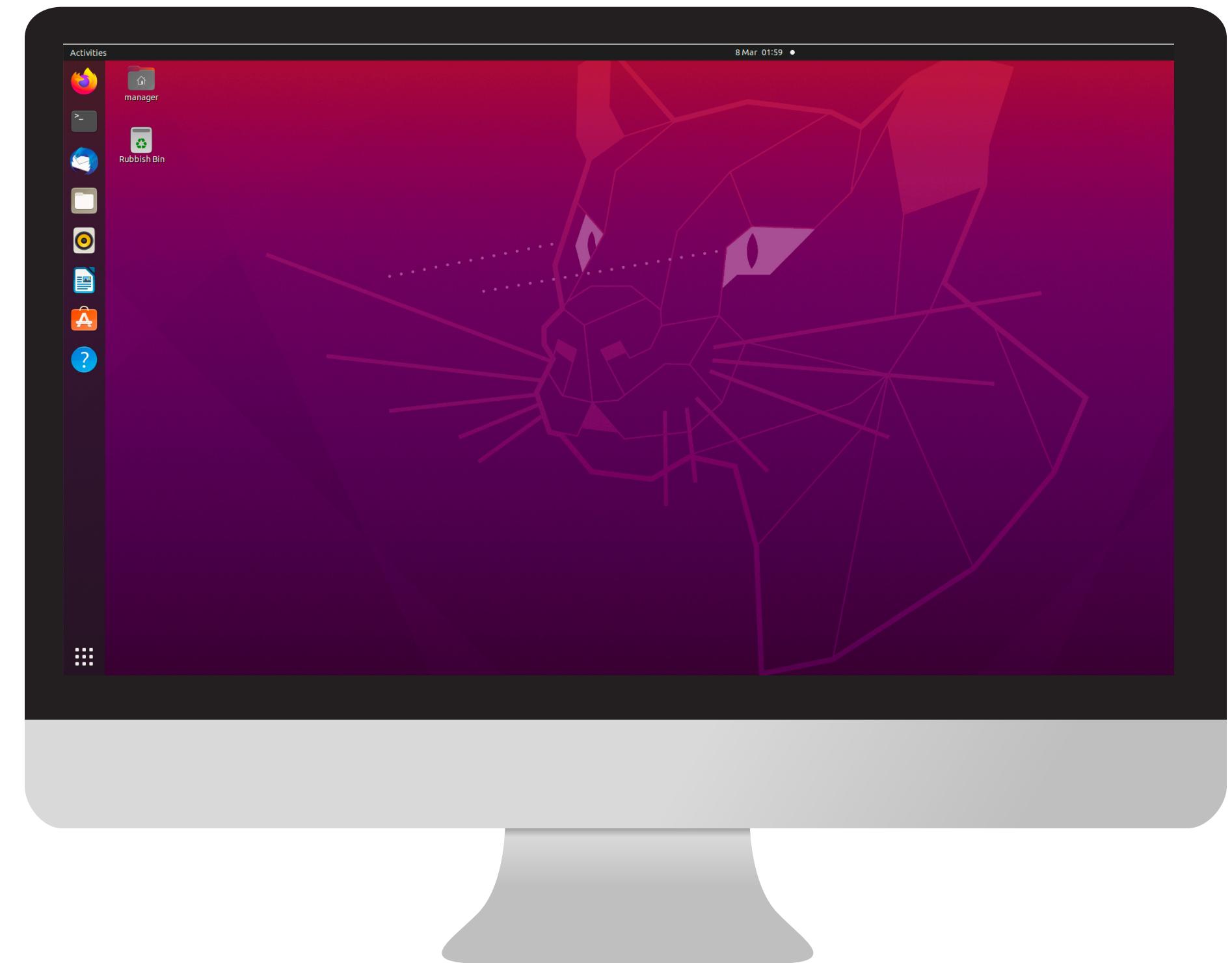
Standard operating system  
(alternative to MS Windows, Mac OS)

Provides a way for you to interact with  
the computer

Many 'flavors' of Unix, using Linux

- Ubuntu
- Centos
- Debian

Typically free



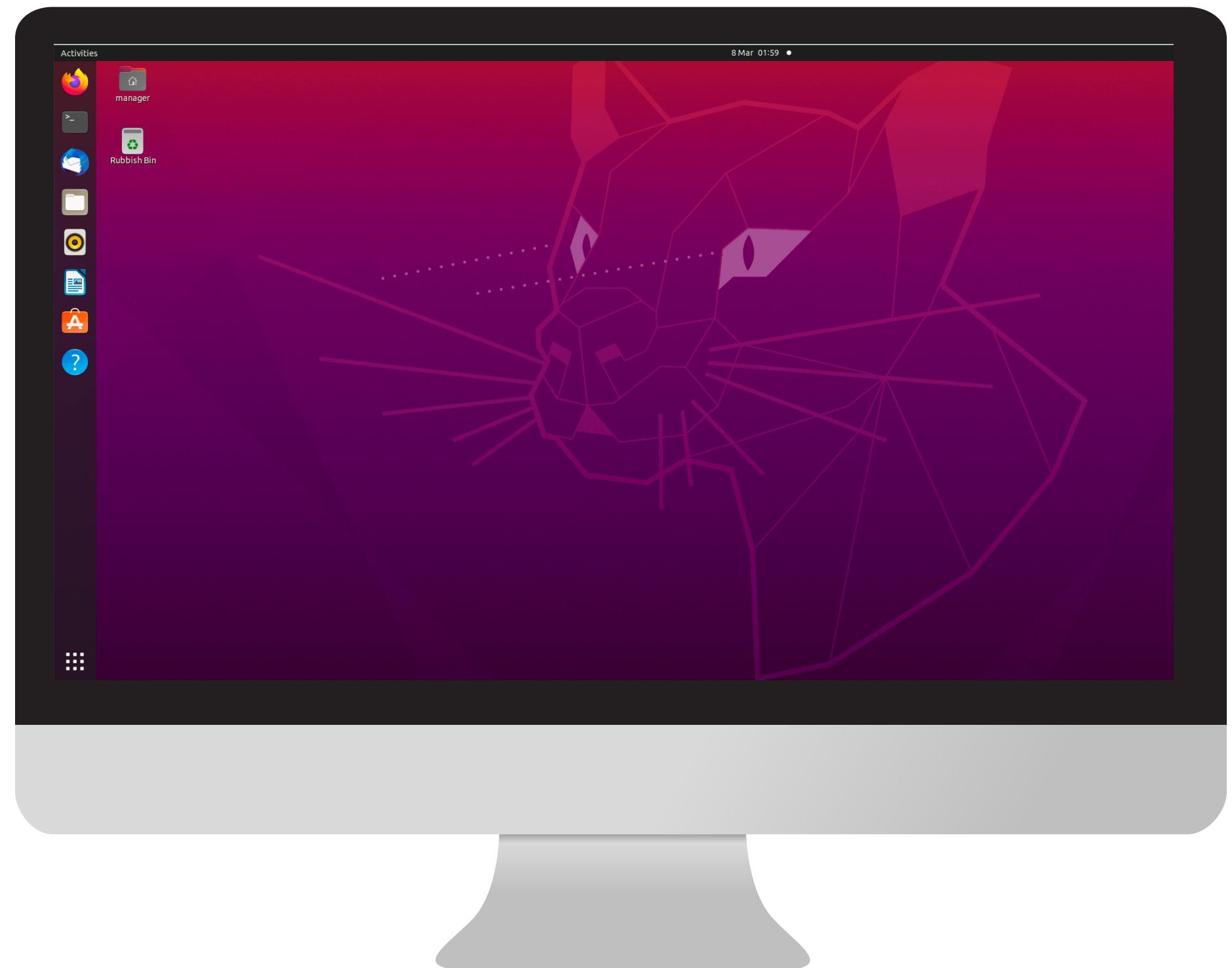
# Why use Linux / Unix ?

Many bioinformatics programs produce outputs that are large text files

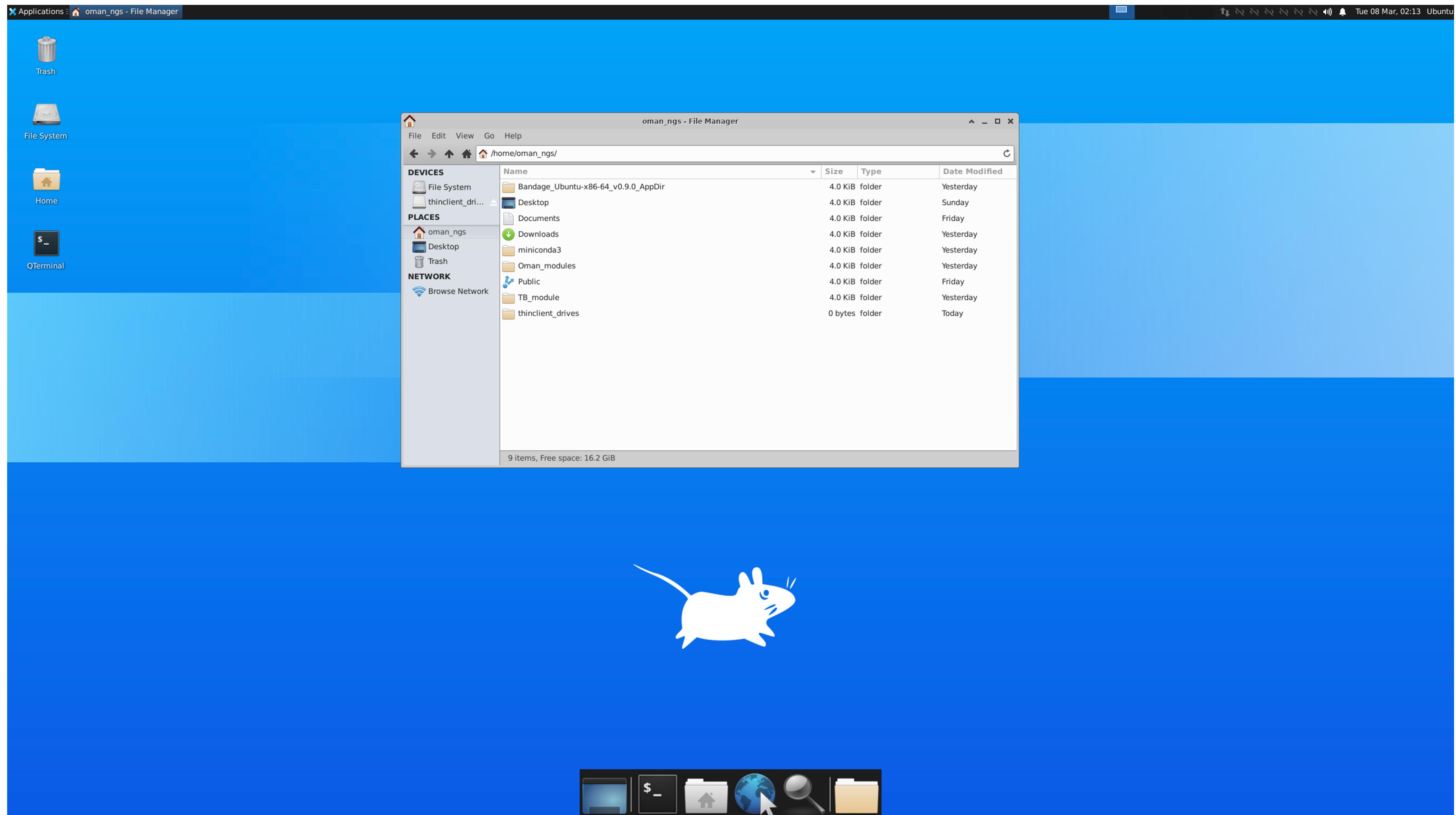
Unix is very suitable for dealing with such files

Powerful and flexible commands for processing large text files

Widely used in scientific community



# Using Linux



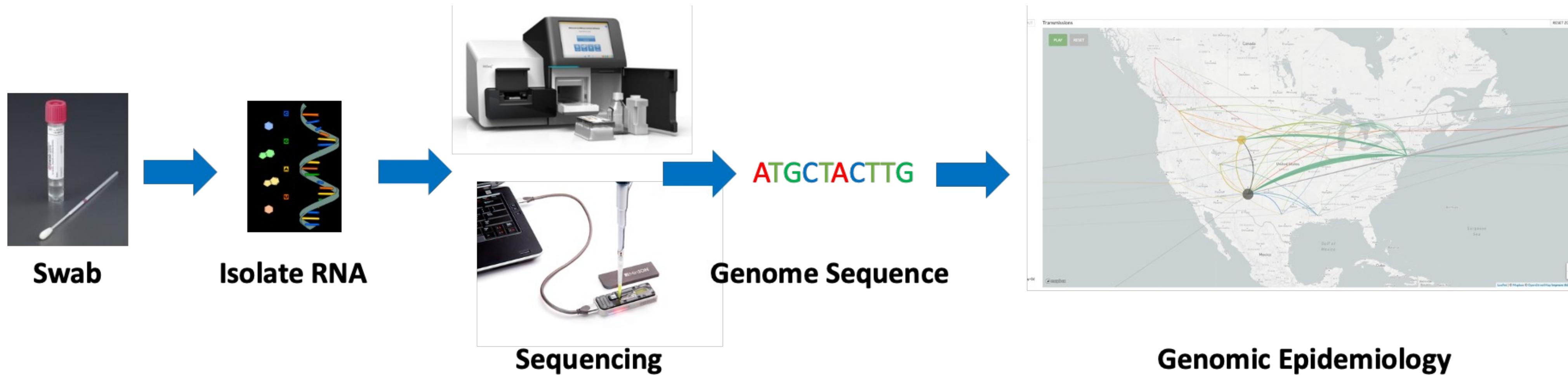
# INTRODUCTION TO NGS DATA TYPES



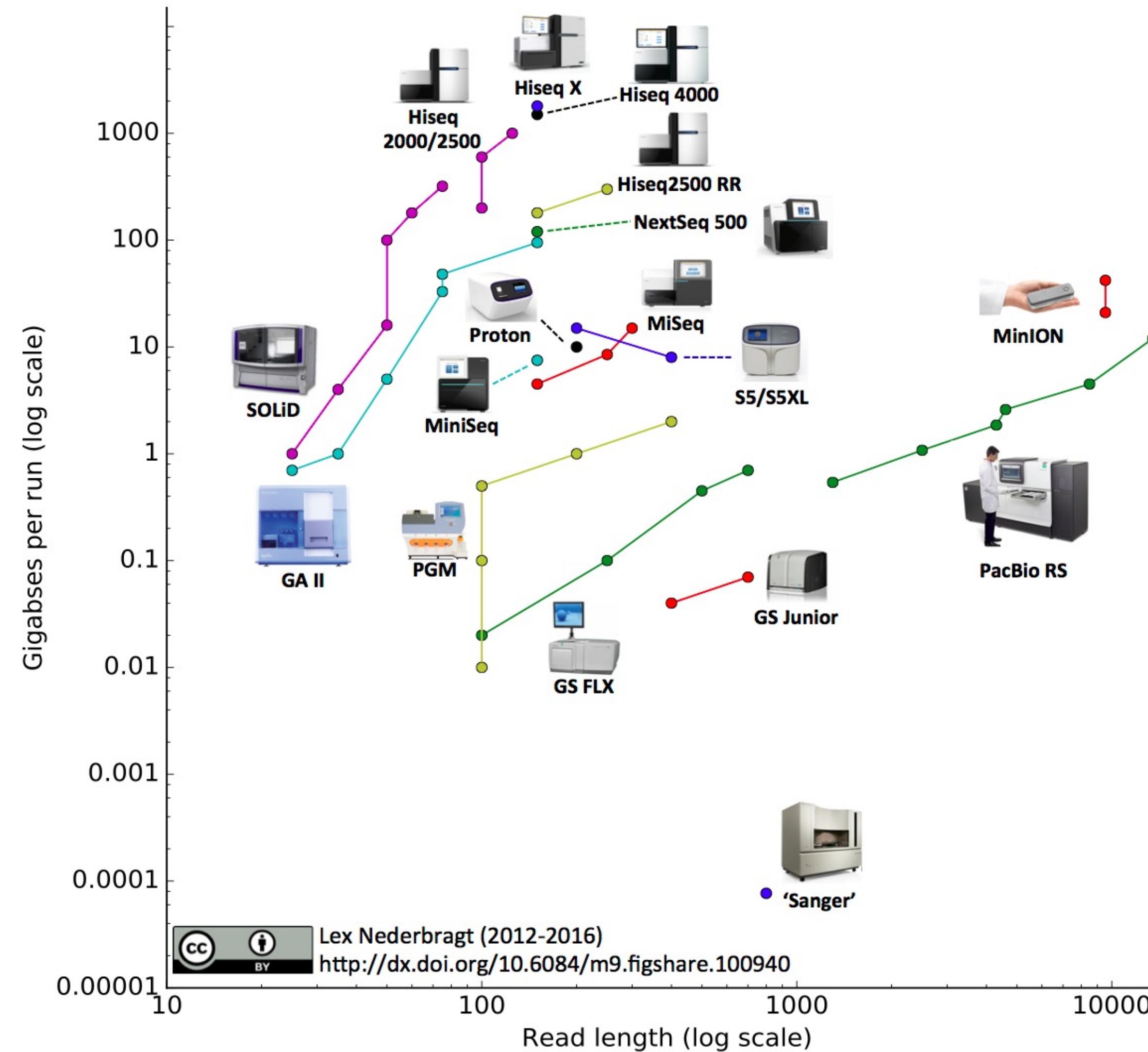
# We have a sequence, now what?



# How do we go from sample to actionable data?



# Sequencing is likely no longer the bottleneck – it is analysis



# Bioinformatics platforms

## Commercial “Point and click”

- ✓ - CLC Genomics
- Geneious Prime

## Free “Point and Click”

- ✓ - UGENE
- MEGA

## Web-based pipelines

- ✓ - Galaxy Server
- Terra (Google cloud)
- Illumina Basespace

## Command line

Thousands of individual programs





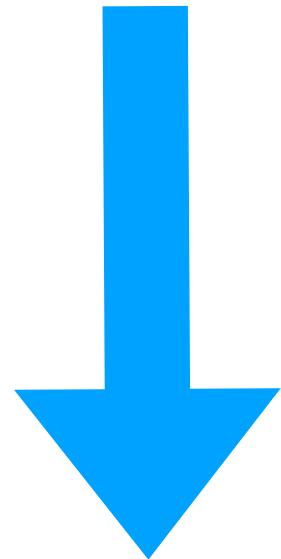
## RAW SEQUENCE DATA FILES



# Two common paths to generating genome fasta files

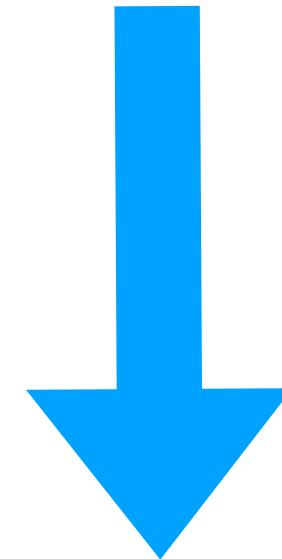


# Illumina



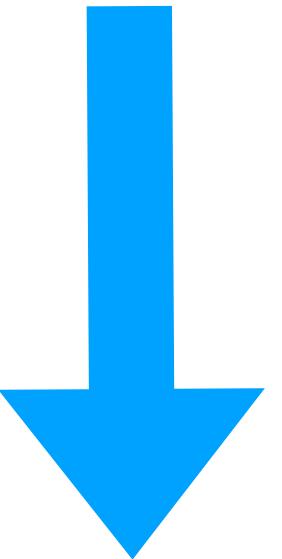
fastq

# Nanopore



fast5

# Ion Torrent



bam

# Fasta format

1. Each entry begins with '>'

2. Name of sequence directly after '>'

3. Everything after name is called **Description**

4. Sequence (nucleotides or AA)

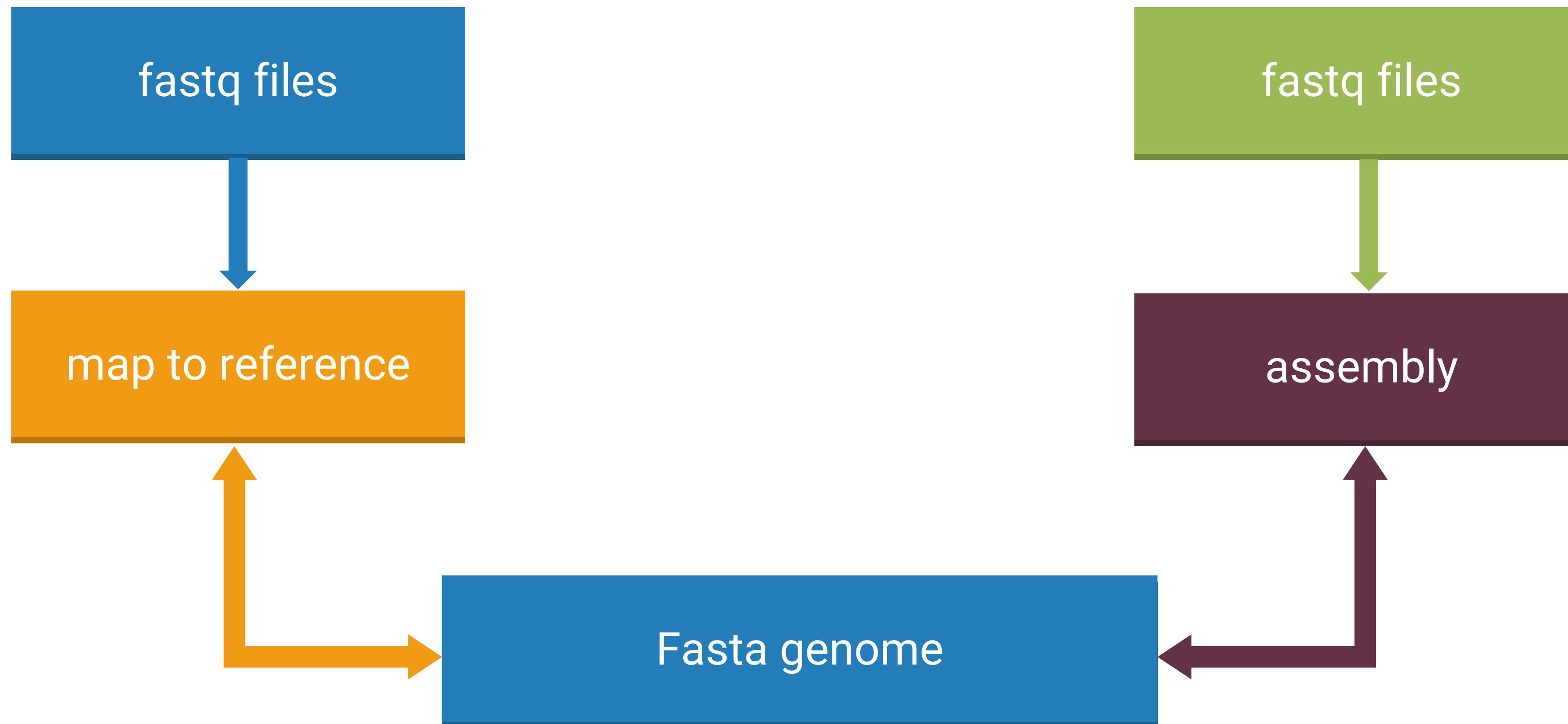
Can be one line or have line break every 60-80 characters (like here)

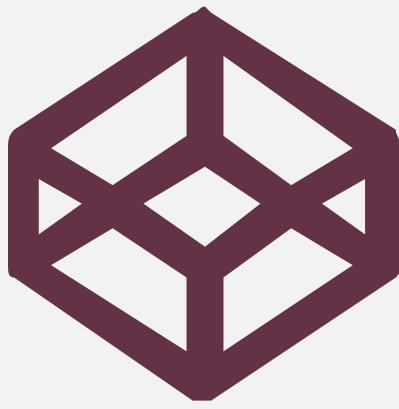
```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ATTAAAGGTTTATACCTTCCCAGGTAAACAAACCAACTTCGATCTCTTAGATCTGTTCTCAA
CGAACTTTAAAATCTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGAGTATAATTAAAC
TAATTACTGTGTTGACAGGACACGAGTAACCTCGTCTATCTTCTGCAGGCTGCTACGGTTCTG
TTGCAGCCGATCATCAGCACATCTAGGTTCTGCCGGGTGACCGAAAGGTAAGATGGAGAGCCTTG
CCTGGTTTCAACGAGAAAACACACGTCACACTCAGTTGCCGTGTTTACAGGTTCGCACGTGCTCGTAC
GTGGCTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTAAAGATGGCACTTGTGG
CTTAGTGAAGTTGAAAAGGCCTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTGGAT
GCTCGAAGTGCACCTCATGGTCATGTTATGGTGTAGCTGGTAGCAGAACTCGAACGGCATTCACTGGTC
GTAGTGGTGAGACACTTGGTGTCCCTGTCCCTCATGTGGCGAAATACCAAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGAATAAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTGACTTA
GGCGACGGAGCTTGGCACTGATCCTTATGAAGATTTCAAGAAAACGGAAACACTAAACATAGCAGTGGT
TTACCCGTGAACACTATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACCTCTGTGG
CCCTGATGGCTACCCCTTGTGGCTGATGTCATTAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTAACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAAGACACCTTTGAAATTAAATTGGCAAAGAA
ATTGACACCTTCAATGGGAATGTCAAATTGTATTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGTGGCTTATGGTAGAATTGATCTGCTATCCAGTTGGTCAC
CAAATGAATGCAACCAATGTGCCTTCAACTCTCATGAAGTGTGATCATTGTGGTAAACCTTCACTGGCA
GACGGCGATTTGTTAAAGCCACTTGCAGATTGTGGCACTGAGAATTGACTAAAGAAGGTGCCACT
ACTTGTGGTTACTTACCCCCAAATGCTGTTAAAATTATTGTCAGCATGTCACAATTCAAGAAGTAG
GACCTGAGCATAGTCTGCCGAATACCATAATGAATCTGGCTGAAAACCATTCTCGTAAGGGTGGTCG
CACTATTGCCCTTGGAGGCTGTGTTCTTATGTTGGTGCATAACAAGTGTGCCTATTGGTTCCA
```

## File name extensions:

- .fnt (nucleotide)
- .fna (nucleotide)
- .faa (amino acid)
- .fasta
- .fa
- .fas

# Two common paths to generating genome fasta files





## GENOME FILES



# Genome annotation files



**European Nucleotide Archive (ENA) :  
EMBL**

The screenshot shows the INSDC homepage. At the top, there's a banner with the INSDC logo and the text "International Nucleotide Sequence Database Collaboration". Below the banner, there's a navigation bar with links for "ABOUT INSDC", "POLICY", "ADVISORS", and "DOCUMENTS". On the left side of the main content area, there are logos for ENA, NCBI, and DDBJ. The main content area contains a section titled "International Nucleotide Sequence Database Collaboration" which includes a bulleted list about the collaboration. To the right of this is a table mapping data types to their respective databases across three organizations: ENA, EMBL-EBI, and NCBI.

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	<a href="#">Sequence Read Archive</a>	European Nucleotide Archive (ENA)	<a href="#">Sequence Read Archive</a>
Capillary reads	<a href="#">Trace Archive</a>		<a href="#">Trace Archive</a>
Annotated sequences	<a href="#">DDBJ</a>		<a href="#">GenBank</a>
Samples	<a href="#">BioSample</a>		<a href="#">BioSample</a>
Studies	<a href="#">BioProject</a>		<a href="#">BioProject</a>

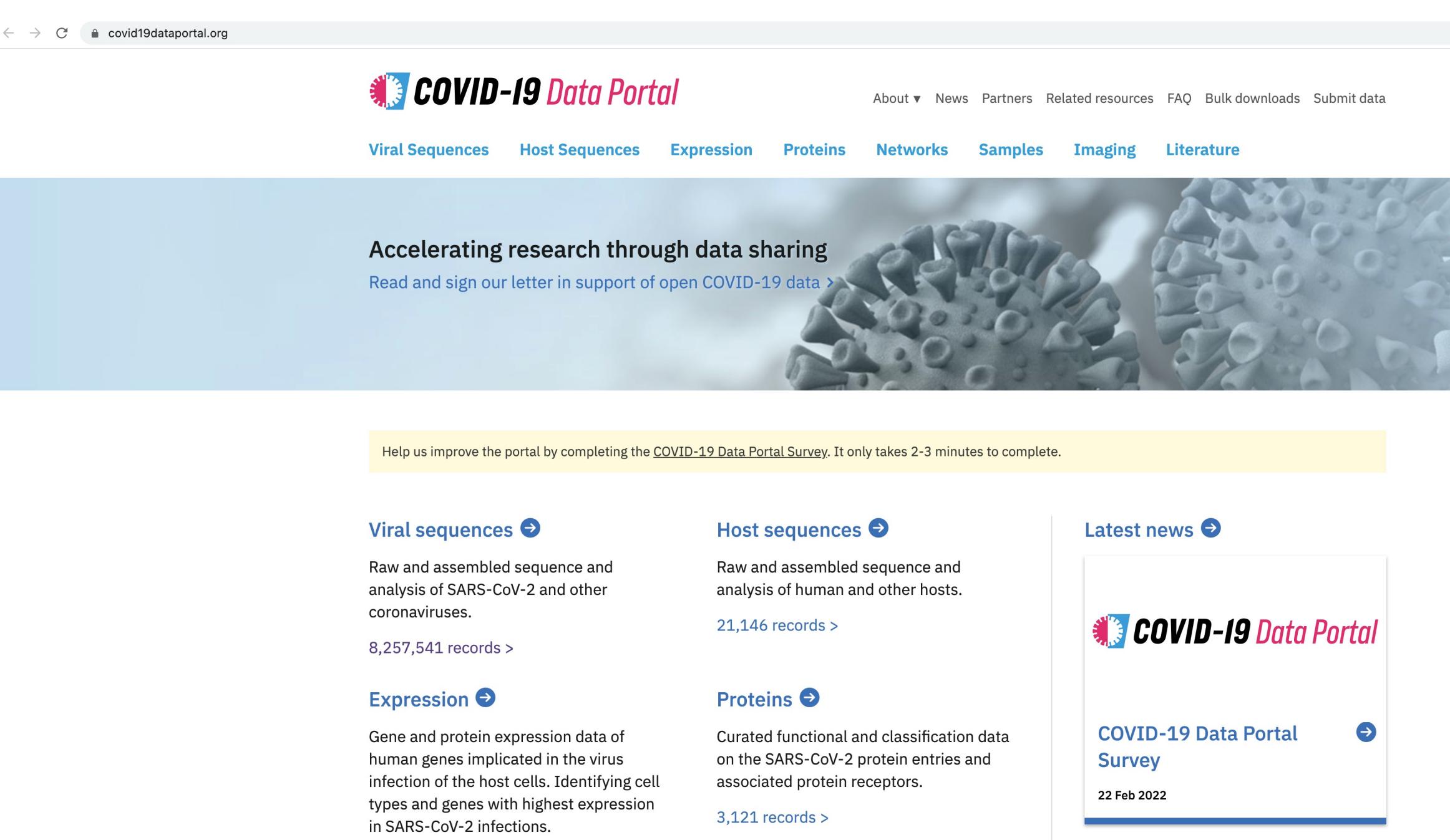


**NIH / NCBI : GenBank**



**General Feature File (GFF)**

# COVID-19 Data Portal – ENA



Accelerating research through data sharing  
Read and sign our letter in support of open COVID-19 data >

Viral Sequences Host Sequences Expression Proteins Networks Samples Imaging Literature

**Viral sequences**  
Raw and assembled sequence and analysis of SARS-CoV-2 and other coronaviruses

**Host sequences**  
Raw and assembled sequence and analysis of human and other hosts.  
21,146 records >

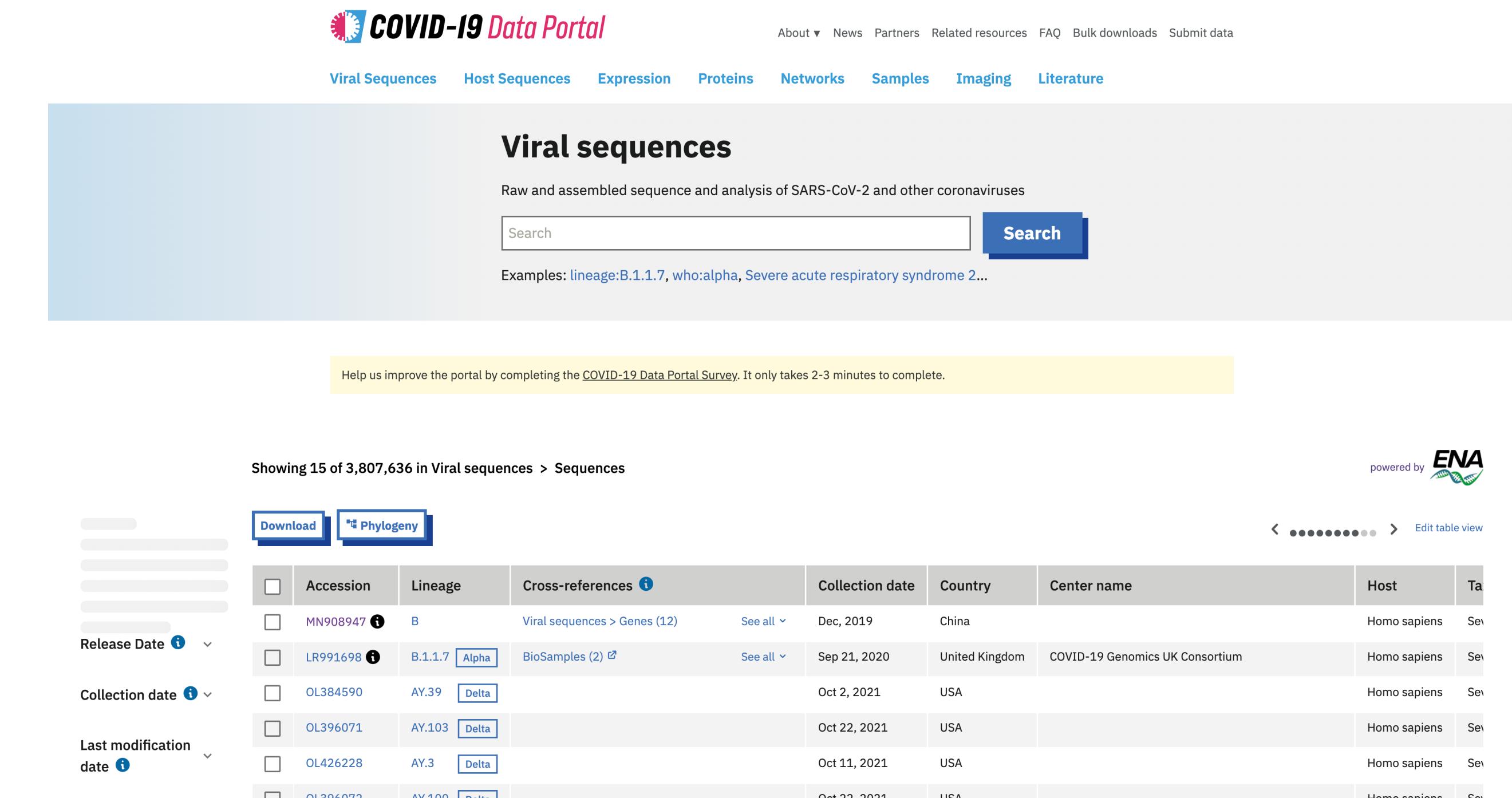
**Latest news**

**COVID-19 Data Portal Survey**  
22 Feb 2022

**Download** **Phylogeny**

Release Date ▾ Collection date ▾ Last modification date ▾

Accession	Lineage	Cross-references	Collection date	Country	Center name	Host	Type
MN908947	B	Viral sequences > Genes (12)	Dec, 2019	China		Homo sapiens	Serotype
LR991698	B.1.1.7 Alpha	BioSamples (2)	Sep 21, 2020	United Kingdom	COVID-19 Genomics UK Consortium	Homo sapiens	Serotype
OL384590	AY.39 Delta		Oct 2, 2021	USA		Homo sapiens	Serotype
OL396071	AY.103 Delta		Oct 22, 2021	USA		Homo sapiens	Serotype
OL426228	AY.3 Delta		Oct 11, 2021	USA		Homo sapiens	Serotype
OL396072	AY.100 Delta		Oct 22, 2021	USA		Homo sapiens	Serotype



**COVID-19 Data Portal**

About ▾ News Partners Related resources FAQ Bulk downloads Submit data

Viral Sequences Host Sequences Expression Proteins Networks Samples Imaging Literature

**Viral sequences**

Raw and assembled sequence and analysis of SARS-CoV-2 and other coronaviruses

Search **Search**

Examples: lineage:B.1.1.7, who:alpha, Severe acute respiratory syndrome 2...

Help us improve the portal by completing the COVID-19 Data Portal Survey. It only takes 2-3 minutes to complete.

Showing 15 of 3,807,636 in Viral sequences > Sequences

Download Phylogeny

Release Date ▾ Collection date ▾ Last modification date ▾

Accession	Lineage	Cross-references	Collection date	Country	Center name	Host	Type
MN908947	B	Viral sequences > Genes (12)	Dec, 2019	China		Homo sapiens	Serotype
LR991698	B.1.1.7 Alpha	BioSamples (2)	Sep 21, 2020	United Kingdom	COVID-19 Genomics UK Consortium	Homo sapiens	Serotype
OL384590	AY.39 Delta		Oct 2, 2021	USA		Homo sapiens	Serotype
OL396071	AY.103 Delta		Oct 22, 2021	USA		Homo sapiens	Serotype
OL426228	AY.3 Delta		Oct 11, 2021	USA		Homo sapiens	Serotype
OL396072	AY.100 Delta		Oct 22, 2021	USA		Homo sapiens	Serotype

powered by 

# EMBL format from EBI

Two-character line code  
indicates the type of information  
contained in the line

Feature  
Key

ID MN908947; SV 3; linear; genomic RNA; STD; VRL; 29903 BP.  
XX  
AC MN908947;  
XX  
DT 13-JAN-2020 (Rel. 143, Created)  
DT 19-MAR-2020 (Rel. 144, Last updated, Version 6)  
XX  
DE Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,  
DE complete genome.  
XX  
KW .  
XX  
OS Severe acute respiratory syndrome coronavirus 2  
OC Viruses; Riboviria; Nidovirales; Cornidovirinae; Coronaviridae;  
OC Orthocoronavirinae; Betacoronavirus; Sarbecovirus.  
XX  
RN [1]  
RP 1-29903  
RX PUBMED; 32015508.  
RA Wu F., Zhao S., Yu B., Chen Y.M., Wang W., Song Z.G., Hu Y., Tao Z.W.,  
RA Tian J.H., Pei Y.Y., Yuan M.L., Zhang Y.L., Dai F.H., Liu Y., Wang Q.M.,  
RA Zheng J.J., Xu L., Holmes E.C., Zhang Y.Z.;  
RT "A new coronavirus associated with human respiratory disease in China";  
RL Nature 579(7798):265-269(2020).  
XX  
RN [2]  
RP 1-29903  
RA Wu F., Zhao S., Yu B., Chen Y.-M., Wang W., Hu Y., Song Z.-G., Tao Z.-W.,  
RA Tian J.-H., Pei Y.-Y., Yuan M.L., Zhang Y.-L., Dai F.-H., Liu Y.,  
RA Wang Q.-M., Zheng J.-J., Xu L., Holmes E.C., Zhang Y.-Z.;  
RT ;  
RL Submitted (05-JAN-2020) to the INSDC.  
RL Shanghai Public Health Clinical Center & School of Public Health, Fudan  
University, Shanghai, China  
XX  
DR MD5: 105c82802b67521950854a851fc6eefd.  
XX  
CC On Jan 17, 2020 this sequence version replaced MN908947.2.  
CC ##Assembly-Data-START##  
CC Assembly Method :: Megahit v. V1.1.3  
CC Sequencing Technology :: Illumina  
CC ##Assembly-Data-END##  
XX  
FH Key Location/Qualifiers  
FH →  
FT source 1..29903  
FT /organism="Severe acute respiratory syndrome coronavirus 2"  
FT /host="Homo sapiens"  
FT /isolate="Wuhan-Hu-1"  
FT /mol\_type="genomic RNA"  
FT /country="China"  
FT /collection\_date="Dec-2019"  
FT /db\_xref="taxon:2697049"  
FT 5' UTR 1..265  
FT gene 266..21555  
FT /gene="orf1ab"  
FT CDS join(266..13468,13468..21555)  
FT /codon\_start=1  
FT /ribosomal\_slippage  
FT /gene="orf1ab"  
FT /product="orf1ab polyprotein"  
FT /note="pplab; translated by -1 ribosomal frameshift"  
FT /protein\_id="QHD43415.1"  
FT /translations="MESLVPGEFNEKTHVQLSLPVLQVRDVLRGVFGDSVEEVLSearqh  
LKDGTGCLVEVEKGVLpqLEQPYVFIKRSDARTAPHGHVMVELVAELEGIYGRSGETL  
GVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFOENWNT  
FT KHSSGVTRELMRELNGGAYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFI  
XX  
SQ Sequence 29903 BP; 8954 A; 5492 C; 5863 G; 9594 T; 0 other;  
attaaagggtt tatacccttc caggtaacaa accaaccaac ttccatctc ttgttagatct 60  
gttctctaaa cgaactttaa aatctgttg gctgtactc ggctgcgtc ttatgtcact 120  
cacgcagttt aattaataac taattactgt cggtgacagg acacgatcaa ctctgtatc 180  
ttctgcaggc tgcttacgtt ttctgcgtt ttgcagccga tcattcggcac atcttagttt 240  
cggtccgggtg tgaccgaaag gtaatgttgaa gaggctgtc cctgggttca acggaaaaac 300  
acacgtccaa ctcagtttc ctgttttaca ggttgcgcac gtgtcgatc gtggctttgg 360  
agactccgtt gaggaggtt tatcagaggc acgtcaacat cttaaaggatg gcacttgtgg 420

## Header

## Annotation

## Sequence

# NIH / NCBI Virus

National Library of Medicine  
National Center for Biotechnology Information

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

NCBI Virus Sequences for discovery

About Us | Find Data | Help | How to Participate | Submit Sequences | Contact Us

SARS-CoV-2 Data Hub Download

Tabular View Dashboard Visualizations Mutations in SRA Complete Tree

Selected Results: 0 Align Build Phylogenetic Tree

Refine Results Reset

Virus +

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), taxid:2697049

Accession +

Sequence Length +

Ambiguous Characters +

Sequence Type +

RefSeq Genome Completeness +

Nucleotide Completeness +

Pango lineage +

Random Sampling New +

Isolate +

Proteins +

Provirus +

Geographic Region +

Host +

Submitters +

Isolation Source +

Nucleotide (4,083,187) Protein (24,012,480) RefSeq Genome (1)

Quick Links Betacoronavirus BLAST CDC Outbreak Information SARS-CoV-2 Articles in PubMed SRA Data NCBI SARS-CoV-2 Resources Datasets command line

Accession	Submitters	Release Date	Pangolin	Isolate	Species	Molecule type	Length	Geo Location	USA	Host	Isolation Source	Collection Date
NC_045512	Wu,F., et al.	2020-01-13	B	Wuhan-Hu-1	Severe acute respiratory syndrome-r...	ssRNA(+)	29903	China		Homo sapiens		2019-12
OM840138	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-U1-IIDS-U0847	Severe acute respiratory syndrome-r...	ssRNA(+)	29781	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840139	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-U1-IIDS-U0850	Severe acute respiratory syndrome-r...	ssRNA(+)	29808	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840140	Andrews,K.R., et al.	2022-02-27	B.1	ID-U1-IIDS-U0852	Severe acute respiratory syndrome-r...	ssRNA(+)	29780	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840141	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-U1-IIDS-U0853	Severe acute respiratory syndrome-r...	ssRNA(+)	29717	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840142	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-U1-IIDS-U0856	Severe acute respiratory syndrome-r...	ssRNA(+)	29775	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840143	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-U1-IIDS-U0857	Severe acute respiratory syndrome-r...	ssRNA(+)	29717	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840144	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-U1-IIDS-U0862	Severe acute respiratory syndrome-r...	ssRNA(+)	29779	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840145	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-U1-IIDS-U0863	Severe acute respiratory syndrome-r...	ssRNA(+)	29808	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840146	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-U1-IIDS-U0864	Severe acute respiratory syndrome-r...	ssRNA(+)	29808	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840147	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-U1-IIDS-U0866	Severe acute respiratory syndrome-r...	ssRNA(+)	29780	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840148	Andrews,K.R., et al.	2022-02-27	B.1	ID-U1-IIDS-U0867	Severe acute respiratory syndrome-r...	ssRNA(+)	29780	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840149	Andrews,K.R., et al.	2022-02-27	B.1	ID-U1-IIDS-U0870	Severe acute respiratory syndrome-r...	ssRNA(+)	29779	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840150	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-U1-IIDS-U0872	Severe acute respiratory syndrome-r...	ssRNA(+)	29779	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11
OM840151	Andrews,K.R., et al.	2022-02-27	B.1	ID-U1-IIDS-U0877	Severe acute respiratory syndrome-r...	ssRNA(+)	29808	USA: Moscow, Idaho	ID	Homo sapiens	oronasopharynx	2020-11-11

Page 1 of 20416

# GenBank format from NCBI

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; Pisoniviricates; 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 Virolology 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.  
Title A phylogenetically conserved hairpin-type 3' untranslated region pseudoknot functions in coronavirus RNA replication  
Journal J. Virol. 73 (10), 8349-8355 (1999)  
Pubmed 10482585  
Reference 5 (bases 1 to 29903)  
Cnstrm NCBI Genome Project  
Title Direct Submission  
Journal Submitted (17-JAN-2020) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA  
Reference 6 (bases 1 to 29903)  
Authors Wu,F., Zhao,S., Yu,B., Chen,Y.-M., Wang,W., Hu,Y., Song,Z.-G., 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 Direct Submission  
Journal Submitted (05-JAN-2020) Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China  
Comment REVIEWED REFSEQ: This record has been curated by NCBI staff. The reference sequence is identical to MN908947. On Jan 17, 2020 this sequence version replaced NC\_045512.1. Annotation was added using homology to SARS-CoV NC\_004718.3. ### Formerly called 'Wuhan seafood market pneumonia virus.' If you have questions or suggestions, please email us at info@ncbi.nlm.nih.gov and include the accession number NC\_045512.### Protein structures can be found at <https://www.ncbi.nlm.nih.gov/structure/?term=sars-cov-2.##> Find all other Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequences at <https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/>  
##Assembly-Data-START##  
Assembly Method :: Megahit v. V1.1.3  
Sequencing Technology :: Illumina  
##Assembly-Data-END##  
Completeness: full length.  
Features source 1..29903  
/organism="Severe acute respiratory syndrome coronavirus 2"  
/mol\_type="genomic RNA"  
/isolate="Wuhan-Hu-1"  
/host="Homo sapiens"  
/db\_xref="taxon:2697049"  
/country="China"  
/collection\_date="Dec-2019"  
5' UTR gene 1..265  
266..21555  
/gene="ORF1ab"  
/locus\_tag="GU280\_gp01"  
/db\_xref="GeneID:43740578"  
join(266..13468,13468..21555)  
/gene="ORF1ab"  
/locus\_tag="GU280\_gp01"  
/ribosomal\_slippage  
/note="pp1ab; translated by -1 ribosomal frameshift"  
/codon\_start=1  
/product="ORF1ab polyprotein"  
/protein\_id="YP\_009724389.1"  
/db\_xref="GeneID:43740578"  
/translation="MESLVPGFNEKTHVOLSLPVLRVDRVLVLRGFGDSVEEVLSearQ  
HLKDGTGGLVEVEKGVLPLQEOPYVFIKRSARTAPHGHVMVAELEGIQYGRSGE  
TLGLVLPHVGEIPVAYRKVLLRKNNGKAGGHSYGAIDLKSFDLGDELTPDYEDFQEN  
WNTKHSSGVTELMLRELNNGAYTRYVDNNFCGPDGYPLECIKDLLARAKGASCTLSEQ  
ORIGIN  
1 attaaagggtt tataccttcc caggttaacaa accaacaacc ttgcgtatctc ttgttagatct  
61 gttctctaa cgaactttaa aatctgtgt gctgtcactc ggctgcattc ttgtgtcact  
121 cacgcgttat aatattacgtt taatattacgtt cttgtacacgg acacggataa ctctgttca  
181 ttctgcggcc tgcttcgttg ttctgttccgtt ttgcgttccgtt tcatttcgttccgtt atcttttgtt  
241 cgtccgggtt tgaccgaaag gtaagatggta gaggcttgc ctttgttca acgagaaaaac

## Header

## Annotation

## Sequence

# General Feature File (GFF)

```

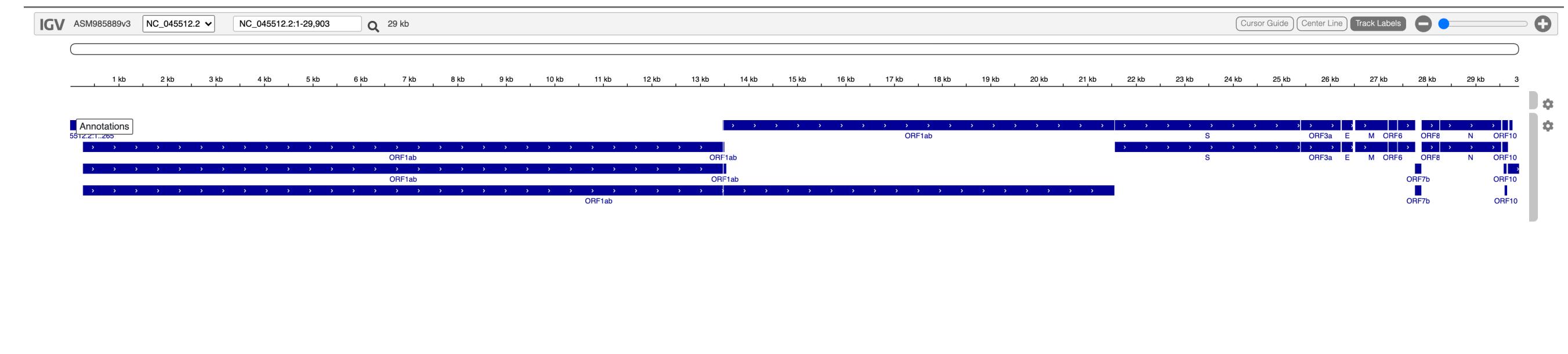
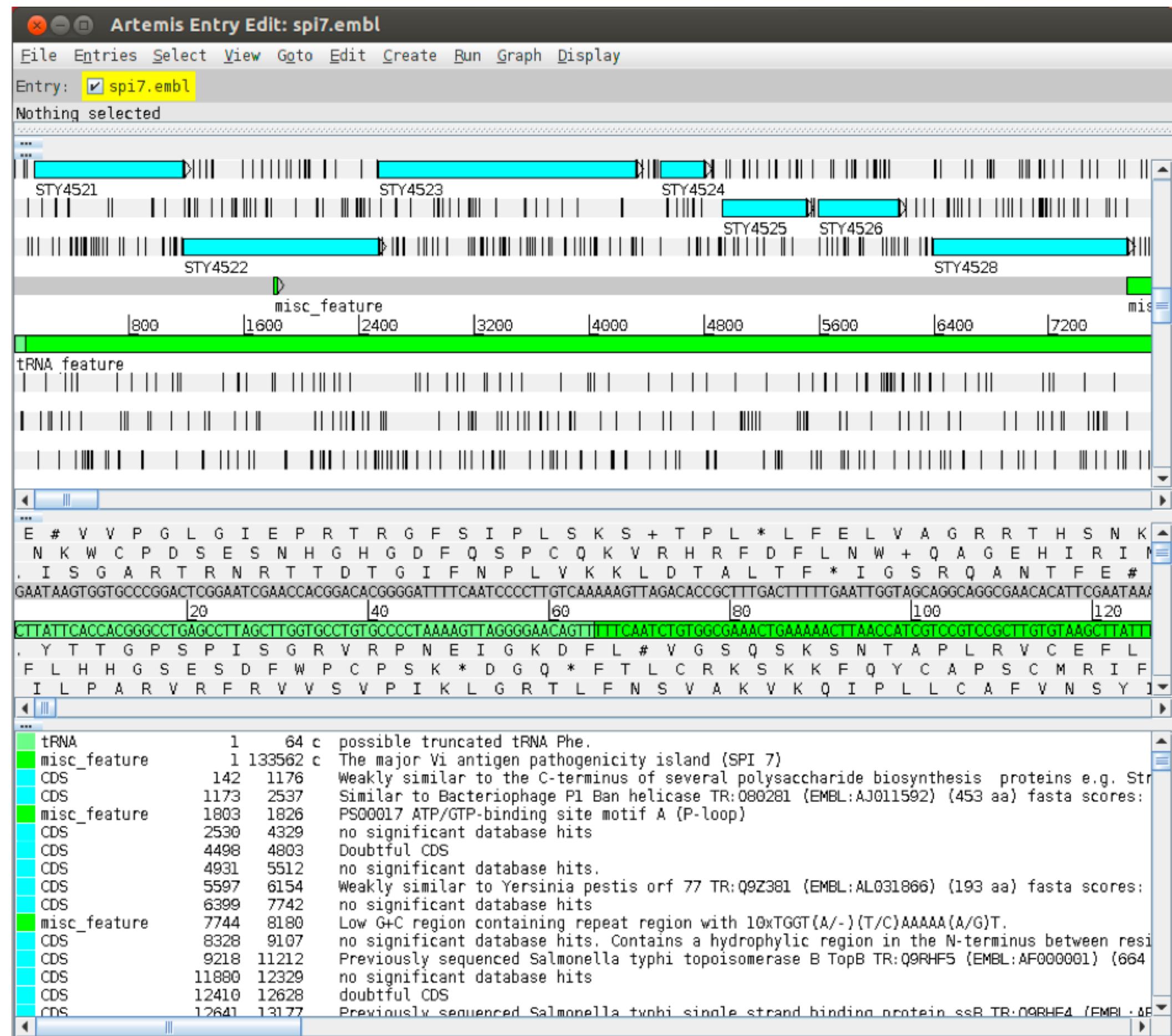
##sequence-region NC_045512.2 1 29903
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=2697049
NC_045512.2 RefSeq region 1 29903 . + . ID=NC_045512.2:1..29903;Dbxref=taxon:2697049;collection-date=Dec-2019;country=China;gb-acronym=SARS-CoV-2;gbkey=Src;genome=genomic;isolate=Wuhan-Hu-1;mol_type=genomic RNA;nat-host=Homo sapiens;old-name=Wuhan seafood market pneumonia virus
NC_045512.2 RefSeq five_prime_UTR 1 265 . + . ID=id-NC_045512.2:1..265;gbkey=5'UTR
NC_045512.2 RefSeq gene 266 21555 . + . ID=Gene-GU280_gp01;Dbxref=GeneID:43740578;Name=ORF1ab;gbkey=Gene;gene=ORF1ab;gene_biotype=protein_coding;locus_tag=GU280_gp01
NC_045512.2 RefSeq CDS 266 13468 . + 0 ID=cds-YP_009724389.1;Parent=Gene-GU280_gp01;Dbxref=Genbank:YP_009724389.1;GeneID:43740578;Name=YP_009724389.1;Note=pp1ab%3B translated by -1 ribosomal frameshift;exception=ribosomal slippage;gbkey=CDS;gene=ORF1ab;locus_tag=GU280_gp01;product=ORF1ab polyprotein;protein_id=YP_009724389.1
NC_045512.2 RefSeq CDS 13468 21555 . + 0 ID=cds-YP_009724389.1;Parent=Gene-GU280_gp01;Dbxref=Genbank:YP_009724389.1;GeneID:43740578;Name=YP_009724389.1;Note=pp1ab%3B translated by -1 ribosomal frameshift;exception=ribosomal slippage;gbkey=CDS;gene=ORF1ab;locus_tag=GU280_gp01;product=ORF1ab polyprotein;protein_id=YP_009724389.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 266 805 . + . ID=id-YP_009724389.1:1..180;Note=nsp1%3B produced by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=leader protein;protein_id=YP_009725297.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 806 2719 . + . ID=id-YP_009724389.1:1..181..181;Note=produced by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=nsp2;protein_id=YP_009725298.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 2720 8554 . + . ID=id-YP_009724389.1:181..2763;Note=former nsp1%3B conserved domains are: N-terminal acidic (Ac)%2C predicted phosphoesterase%2C papain-like proteinase%2C Y-domain%2C transmembrane domain 1 (TM1)%2C adenosine diphosphate-ribose 1''-phosphatase (ADRP)%3B produced by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=nsp4;protein_id=YP_009725300.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 8555 10054 . + . ID=id-YP_009724389.1:2764..3263;Note=nsp4B%3B contains transmembrane domain 2 (TM2)%3B produced by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=nsp4;protein_id=YP_009725300.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 10055 10972 . + . ID=id-YP_009724389.1:3264..3569;Note=nsp5A_3CLpro and nsp5B_3CLpro%3B main proteinase (Mpro)%3B mediates cleavage downstream of nsp4. 3D structure of the SARSr-CoV homolog has been determined (Yang et al. 2003)%3B produced by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=nsp6;protein_id=YP_009725302.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 10973 11842 . + . ID=id-YP_009724389.1:3570..3859;Note=nsp6_TM4%3B putative transmembrane domain%3B produced by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=nsp7;protein_id=YP_009725303.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 11843 12091 . + . ID=id-YP_009724389.1:3860..3942;Note=produced by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=nsp7;protein_id=YP_009725303.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 12092 12685 . + . ID=id-YP_009724389.1:3943..4140;Note=produced by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=nsp8;protein_id=YP_009725304.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 12686 13024 . + . ID=id-YP_009724389.1:4141..4253;Note=ssRNA-binding protein%3B produced by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=nsp9;protein_id=YP_009725305.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 13025 13441 . + . ID=id-YP_009724389.1:4254..4392;Note=nsp10_CysHis%3B formerly known as growth-factor-like protein (GFL)%3B produced by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=nsp10;protein_id=YP_009725306.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 13442 13468 . + . ID=id-YP_009724389.1:4393..5324;Note=nsp12%3B NiRAN and RdRp%3B produced by pp1ab only;Parent=cds-YP_009724389.1;gbkey=Prot;product=RNA-dependent RNA polymerase;protein_id=YP_009725307.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 13468 16236 . + . ID=id-YP_009724389.1:4393..5324;Note=nsp12%3B NiRAN and RdRp%3B produced by pp1ab only;Parent=cds-YP_009724389.1;gbkey=Prot;product=RNA-dependent RNA polymerase;protein_id=YP_009725307.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 16237 18039 . + . ID=id-YP_009724389.1:5325..5925;Note=nsp13_ZBD%2C nspl3_TB%2C and nsp_HEL1core%3B zinc-binding domain (ZD)%2C NTPase/helicase domain (HEL)%2C RNA 5'-triphosphatase%3B produced by pp1ab only;Parent=cds-YP_009724389.1;gbkey=Prot;product=helicase;protein_id=YP_009725308.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 18040 19620 . + . ID=id-YP_009724389.1:5926..6452;Note=nsp14A2_Exo1 and nsp14B_NMT%3B produced by pp1ab only;Parent=cds-YP_009724389.1;gbkey=Prot;product=3'-to-5' exonuclease;protein_id=YP_009725309.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 19621 20658 . + . ID=id-YP_009724389.1:6453..6798;Note=nsp15_A1 and nsp15B_NendoU%3B produced by pp1ab only;Parent=cds-YP_009724389.1;gbkey=Prot;product=endoRNase;protein_id=YP_009725310.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 20659 21552 . + . ID=id-YP_009724389.1:6799..7096;Note=nsp16_0MT%3B 2'-o-MT%3B produced by pp1ab only;Parent=cds-YP_009724389.1;gbkey=Prot;product=2'-o-ribose methyltransferase;protein_id=YP_009725311.1
NC_045512.2 RefSeq CDS 266 13483 . + 0 ID=cds-YP_009725295.1;Parent=Gene-GU280_gp01;Dbxref=Genbank:YP_009725295.1;GeneID:43740578;Name=YP_009725295.1;Note=pp1a;gbkey=CDS;gene=ORF1ab;locus_tag=GU280_gp01;product=ORF1ab polyprotein;protein_id=YP_009725295.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 266 805 . + . ID=id-YP_009725295.1:1..180;Note=nsp1%3B produced by both pp1a and pp1ab;Parent=cds-YP_009725295.1;gbkey=Prot;product=leader protein;protein_id=YP_009742608.1
NC_045512.2 RefSeq mature_protein_region_of_CDS 806 2719 . + . ID=id-YP_009725295.1:181..181;Note=produced by both pp1a and pp1ab;Parent=cds-YP_009725295.1;gbkey=Prot;product=nsp2;protein_id=YP_009742609.1

```

## General GFF3 structure

Position index	Position name	Description
1	seqid	The name of the sequence where the feature is located.
2	source	Keyword identifying the source of the feature, like a program (e.g. <a href="#">Augustus</a> or <a href="#">RepeatMasker</a> ) or an organization (like <a href="#">TAIR</a> ).
3	type	The feature type name, like "gene" or "exon". In a well structured GFF file, all the children features always follow their parents in a single block (so all exons of a transcript are put after their parent "transcript" feature line and before any other parent transcript line). In GFF3, all features and their relationships should be compatible with the <a href="#">standards released by the Sequence Ontology Project</a> .
4	start	Genomic start of the feature, with a <b>1-base offset</b> . This is in contrast with other 0-offset half-open sequence formats, like <a href="#">BED</a> .
5	end	Genomic end of the feature, with a <b>1-base offset</b> . This is the same end coordinate as it is in 0-offset half-open sequence formats, like <a href="#">BED</a> . <small>[citation needed]</small>
6	score	Numeric value that generally indicates the confidence of the source in the annotated feature. A value of "." (a dot) is used to define a null value.
7	strand	Single character that indicates the <a href="#">strand</a> of the feature; it can assume the values of "+" (positive, or 5'->3'), "-", (negative, or 3'->5'), "." (undetermined).
8	phase	phase of CDS features; it can be either one of 0, 1, 2 (for CDS features) or "." (for everything else). See the section below for a detailed explanation.
9	attributes	All the other information pertaining to this feature. The format, structure and content of this field is the one which varies the most between the three competing file formats.

# Annotation files can be visualized and explored



# Integrated Genome Viewer (IGV)

# Artemis

# Questions?

# Today's Agenda



**VM installation and setup**



**Bioinformatics data files**



**Module 1 : SARS-CoV-2 web tools in VM**



**Module 2 : Using the command line**

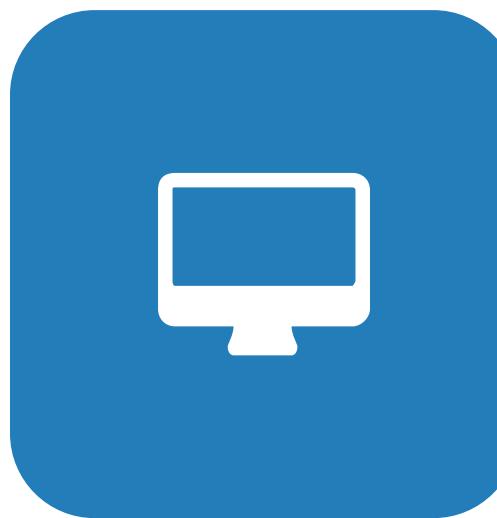


# Bioinformatics Module 1

[https://domman-genomics.github.io/Jordan\\_NGS/manuals/01\\_Intro\\_to\\_NGS/module\\_Intro.html](https://domman-genomics.github.io/Jordan_NGS/manuals/01_Intro_to_NGS/module_Intro.html)

## Explore NextClade

- utilize SARS-CoV-2 genomes in **.fasta** format
- Start building an understanding of looking at genome data



## Call lineages with Pangolin

- use web based Pangolin to designate Pango lineages

