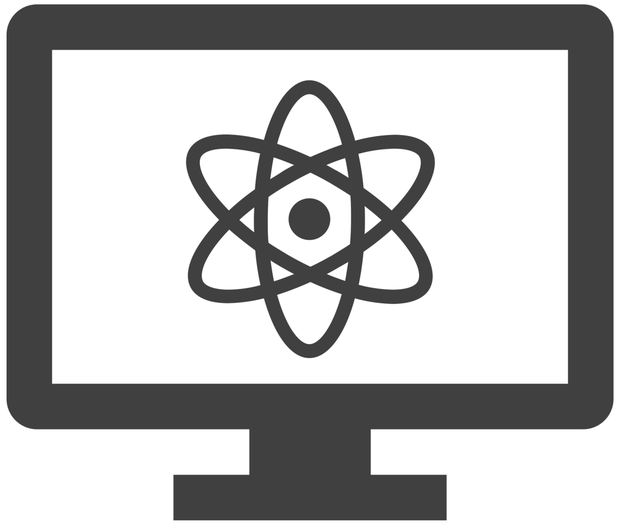
Bioinformatics on the Command Line



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?

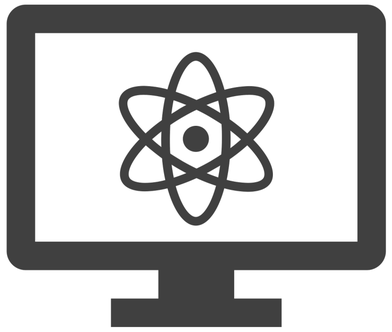
Why the command line?

Pro

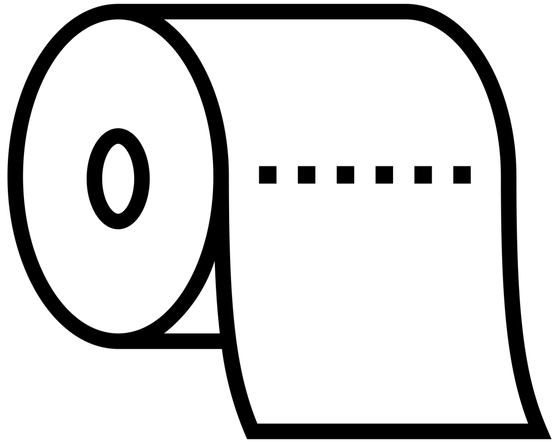
- Ultimate control over workflow
- Doesn’t depend on internet
- Data privacy
- Can be cheaper
- Use latest tools

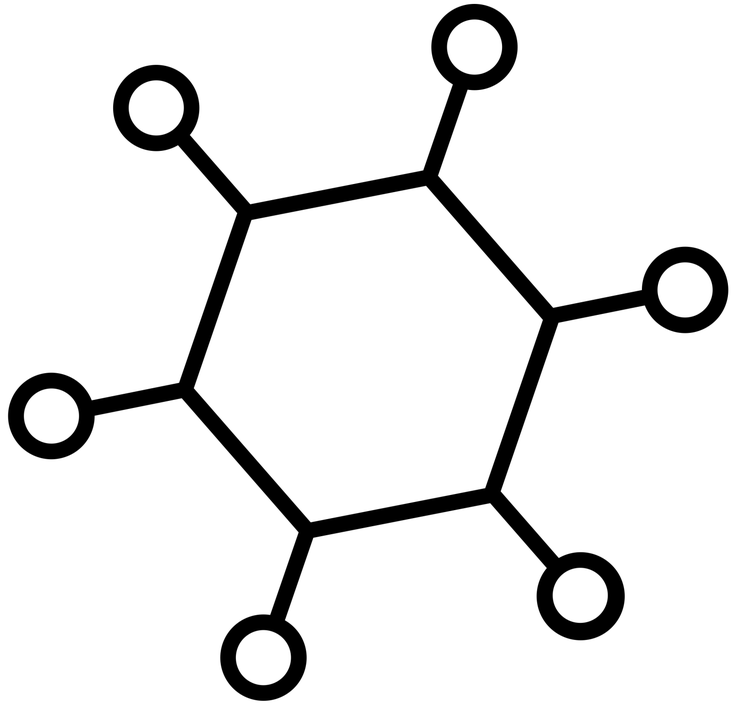
Con

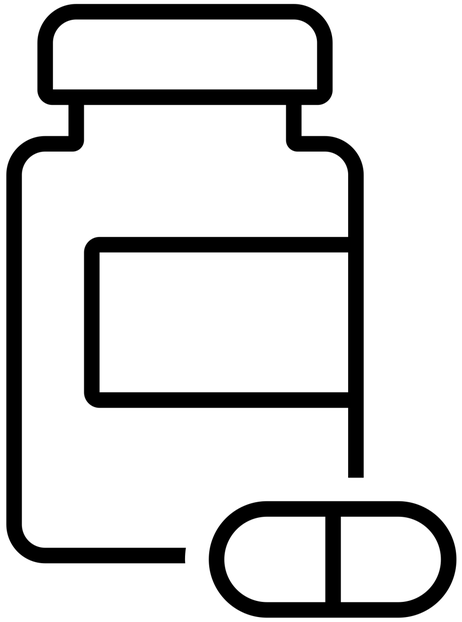
- May be a steep learning curve
- You must make the decisions
- Errors and troubleshooting
- Output not always in easy to read and interpret format



A bit about me

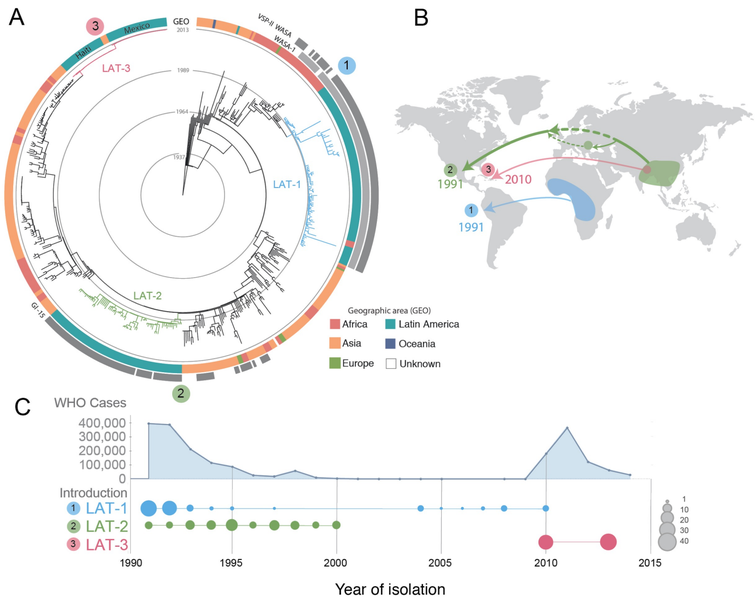






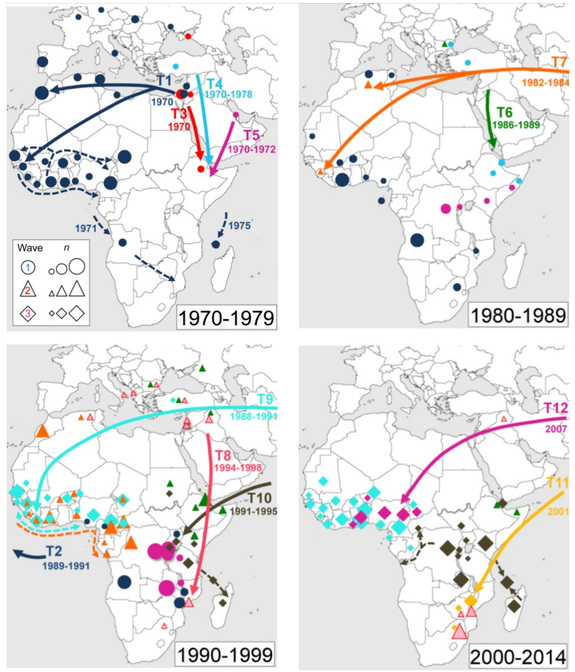
Daryl Domman, PhD
Assistant Professor
Center for Global Health
Department of Internal Medicine
University of New Mexico School of Medicine

Pandemic cholera dynamics



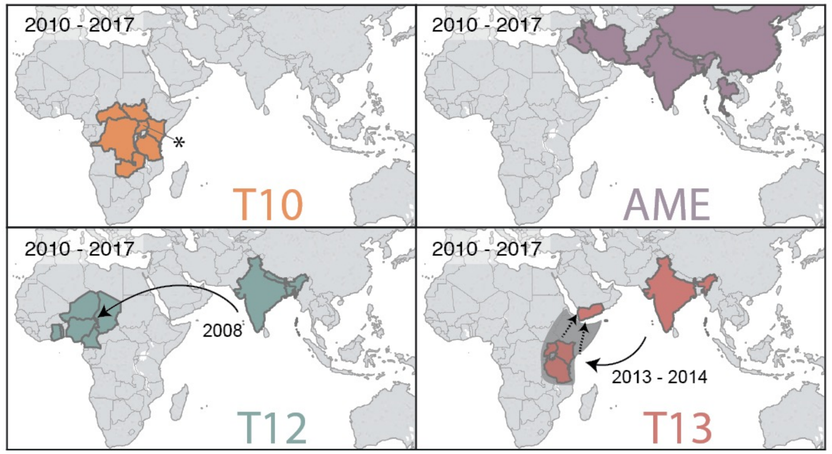
Cholera across Latin America

Domman et al, Science 2017



Cholera across Africa

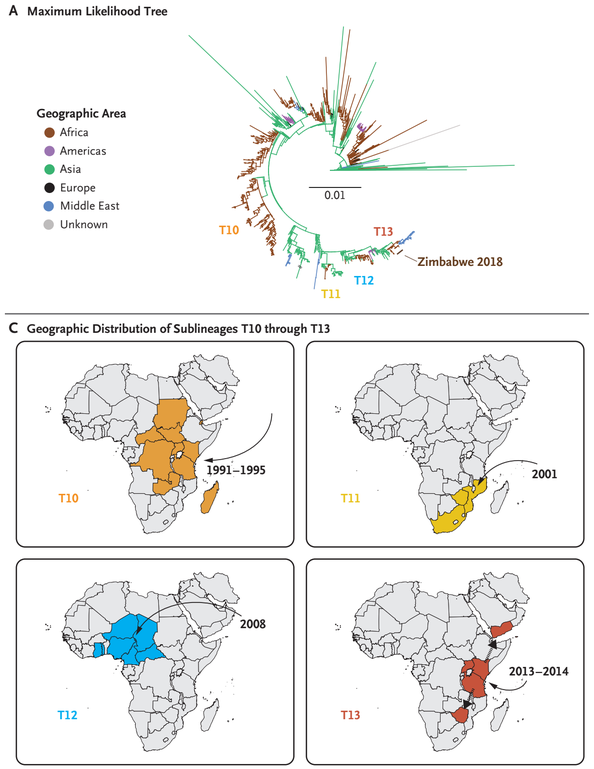
Weill, Domman et al, Science 2017



Yemen epidemic

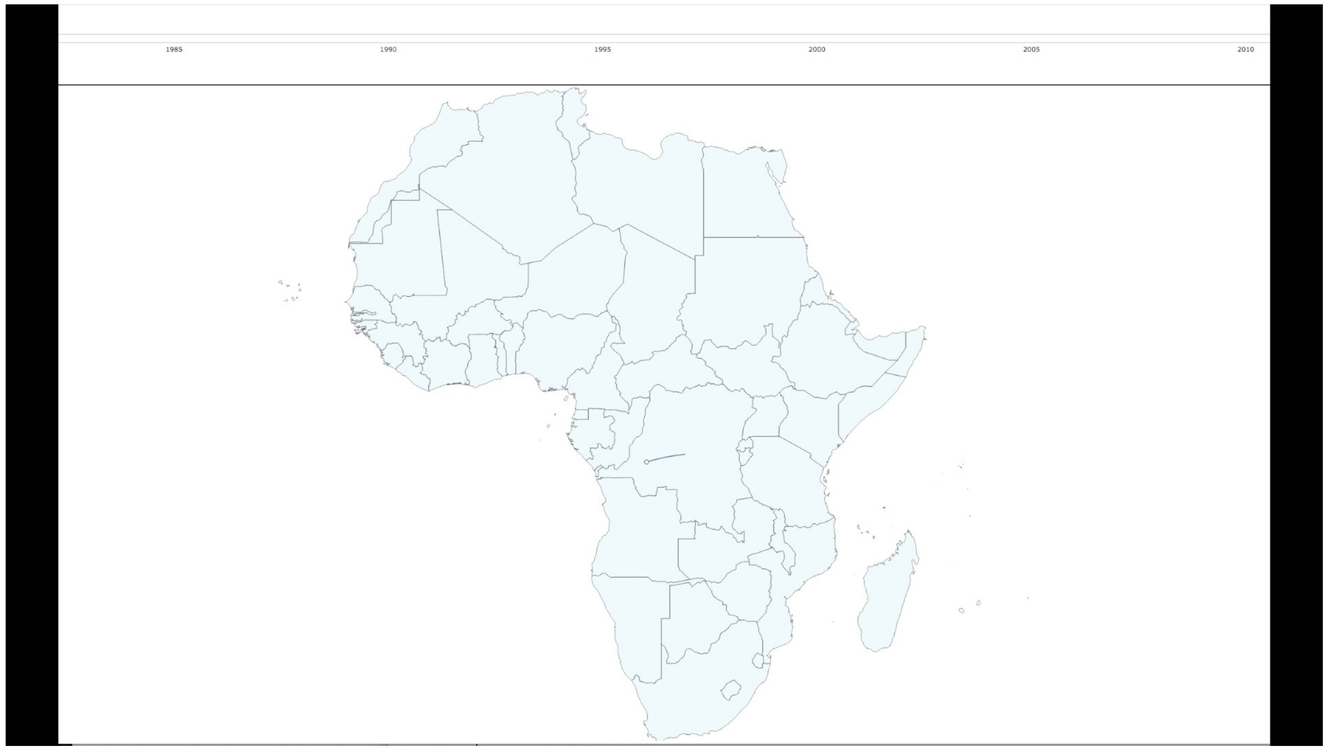
Weill* & Domman* et al, Nature 2019

Cholera transmission in Africa



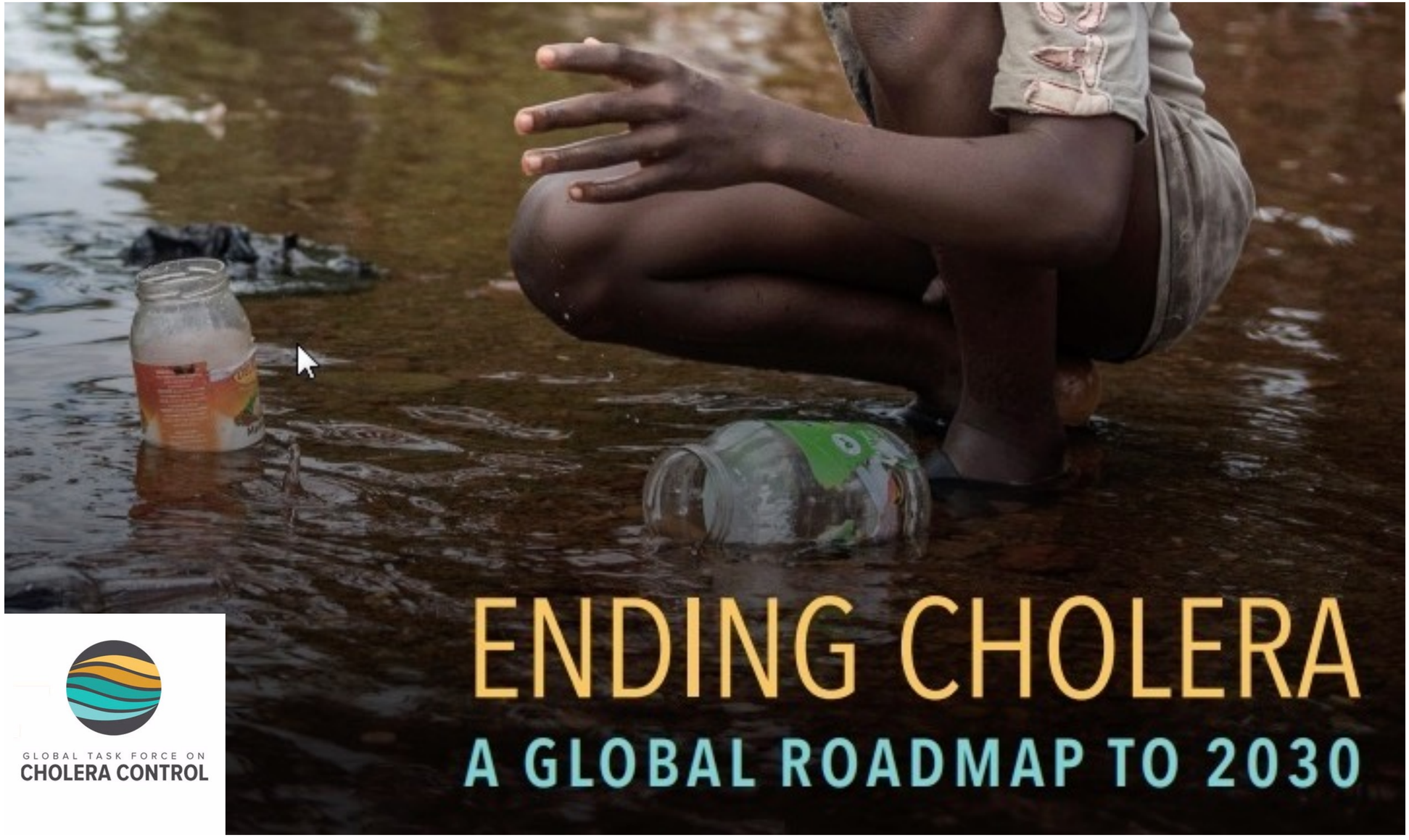
Cholera outbreak Zimbabwe

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



Cholera across DRC

Collaboration with
Johns Hopkins



GLOBAL TASK FORCE ON
CHOLERA CONTROL

Rocky Mountain COVID Consortium + Arkansas

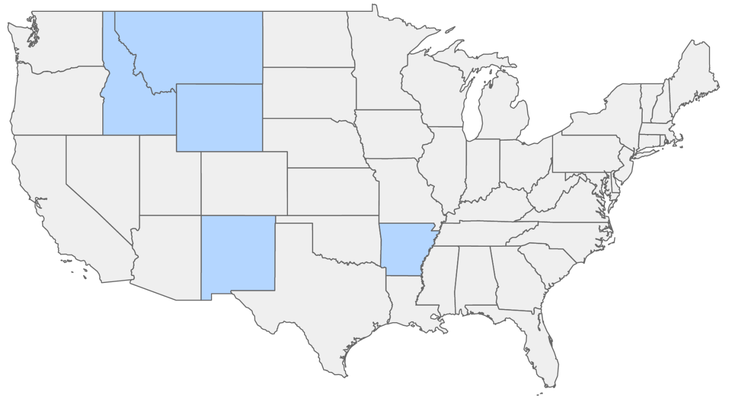
Viral Transport Media
Extracted RNA

SARS-CoV-2 genome

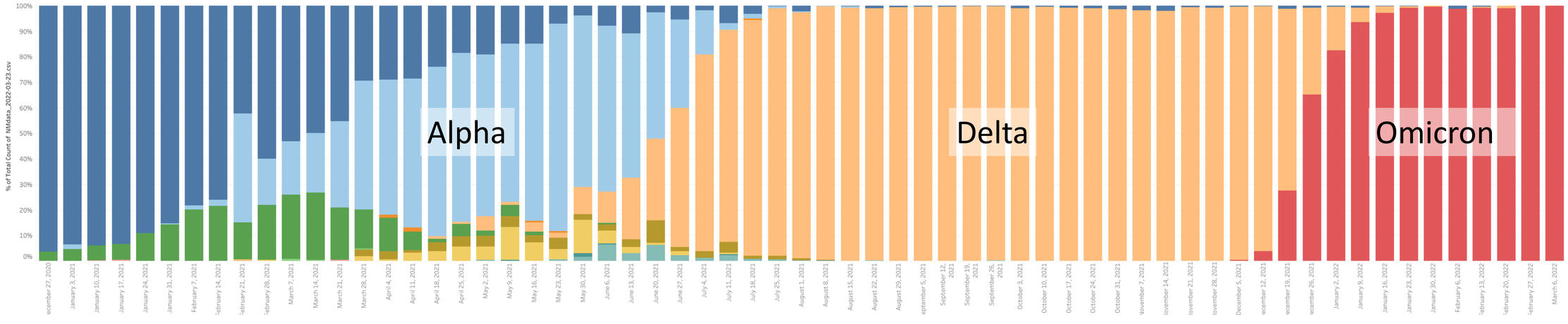
RT-PCR
ARTIC tiling amplicons

Illumina NextSeq2000

Nanopore GridION / MinION



>20,000 genomes sequenced



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

MA

Microsoft Azure <azure-noreply@microsoft.com>

To: domman.genomics@gmail.com

Today at 10:03 AM

Microsoft Azure

Daryl Domman invited you to the lab: JordanNGS

Register now to access the virtual machines in the lab.

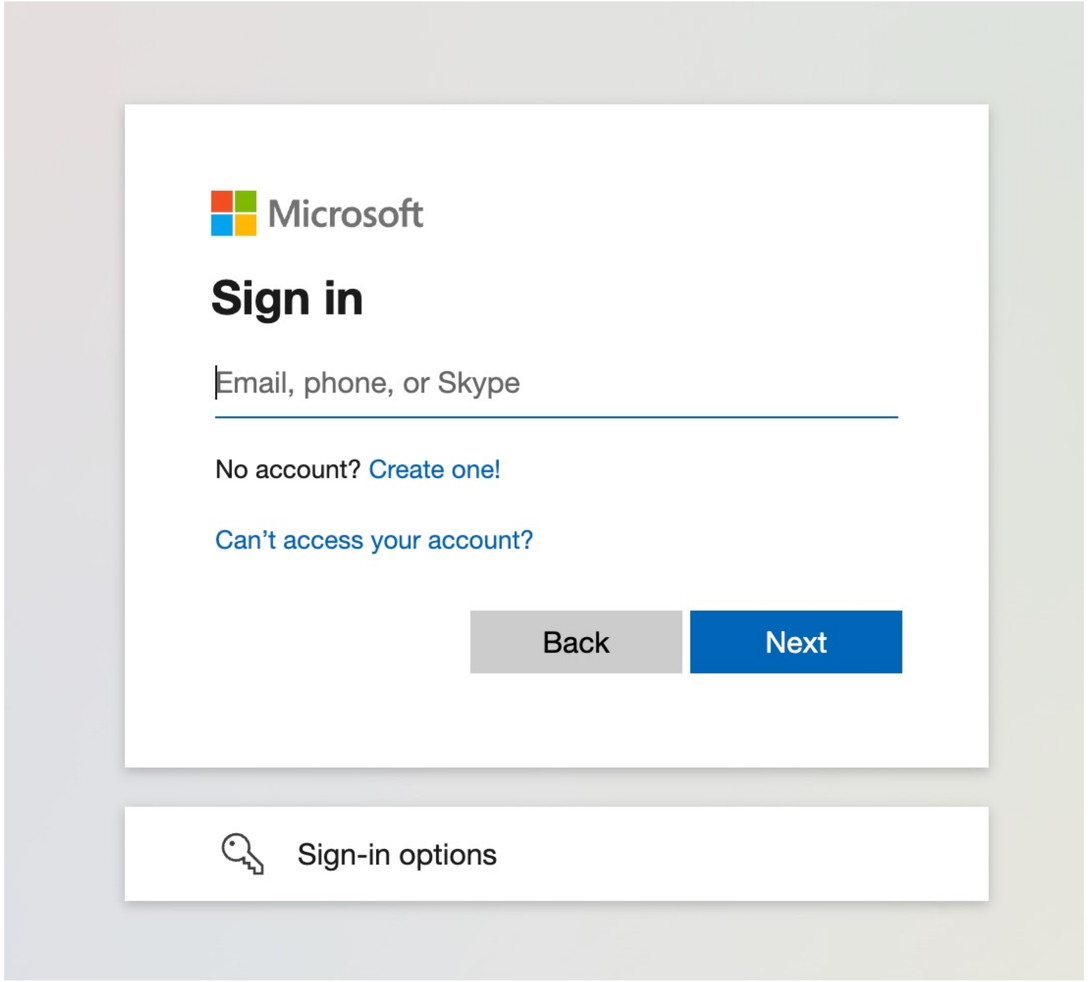
Register for the lab >

Privacy Statement

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

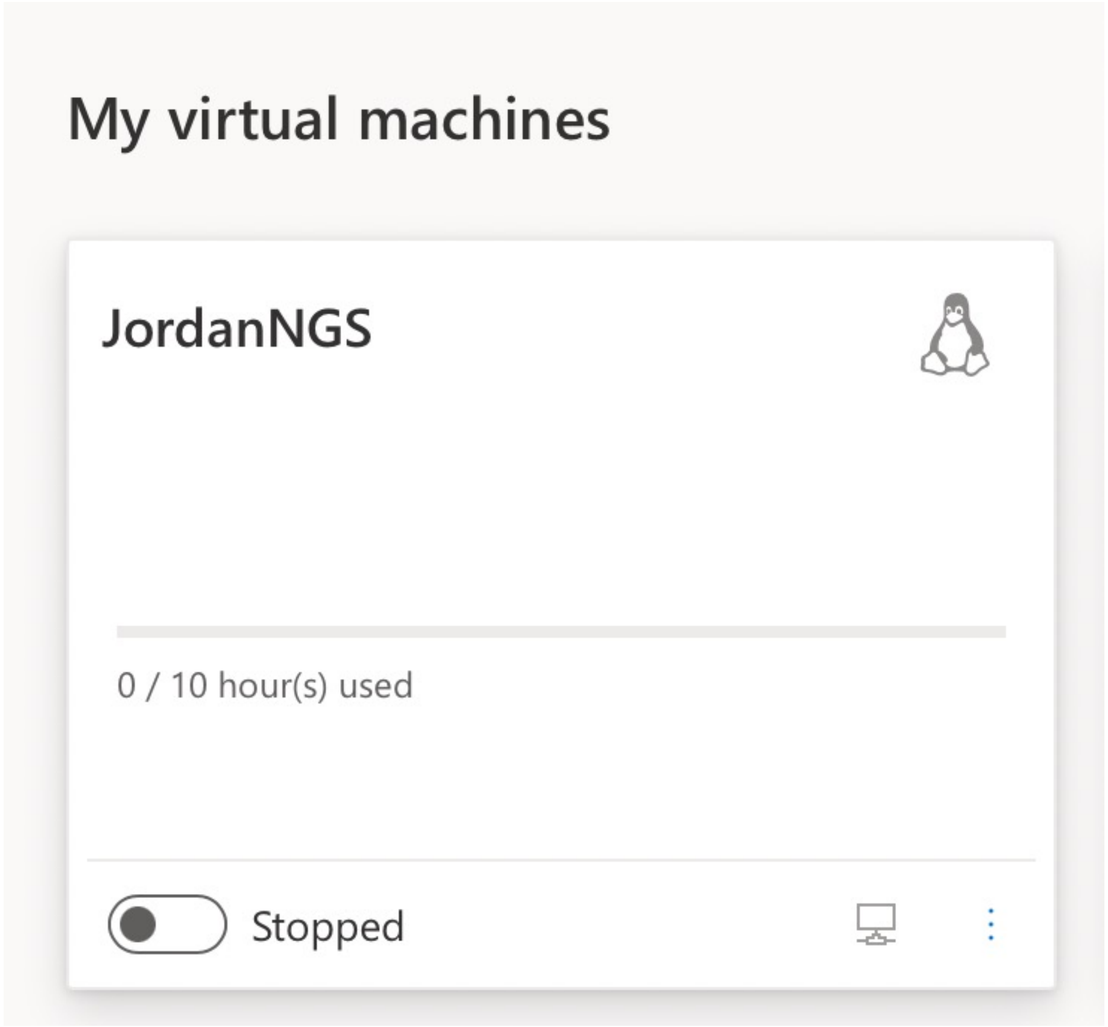
Microsoft

Step 2 : Microsoft Account

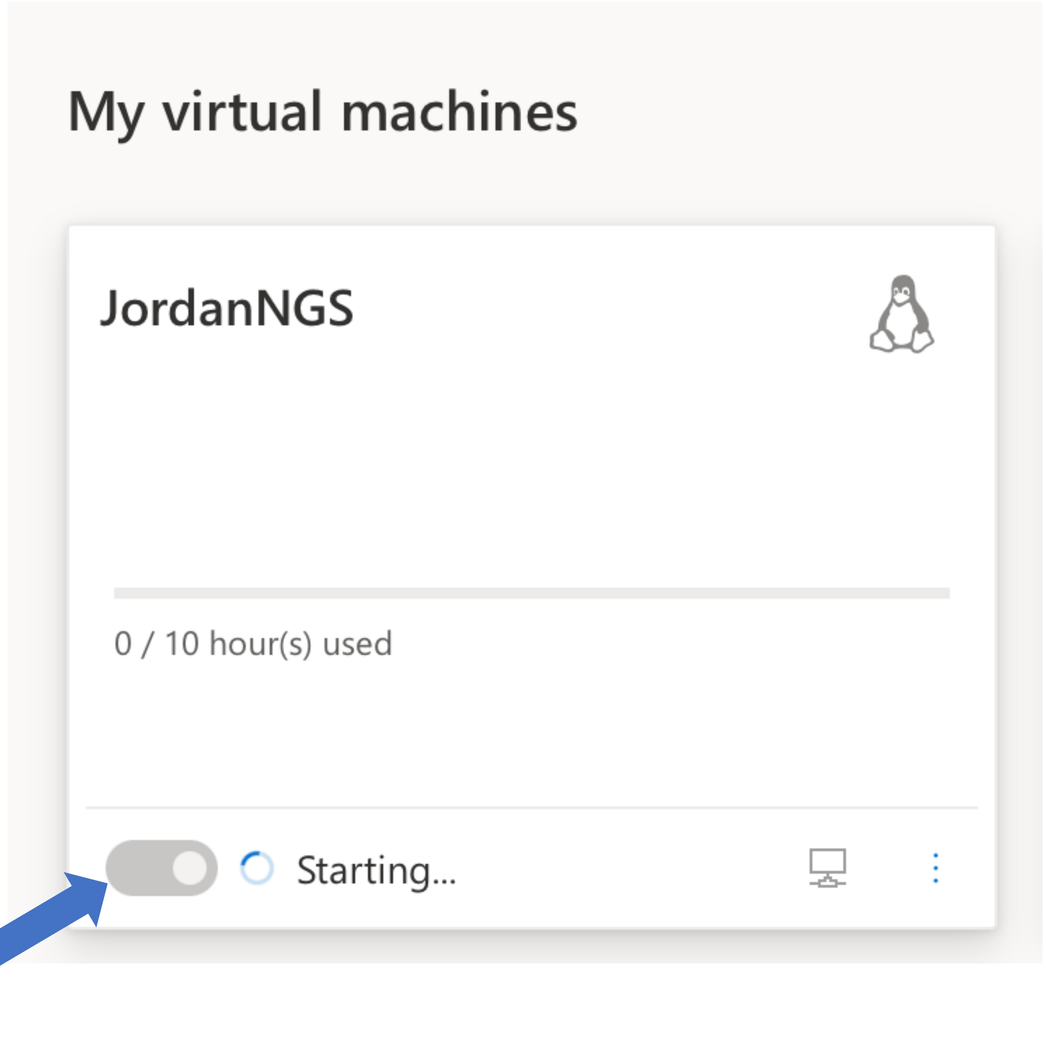


For Azure Lab — needs to be same as invite email

Step 3 : Azure Lab main screen

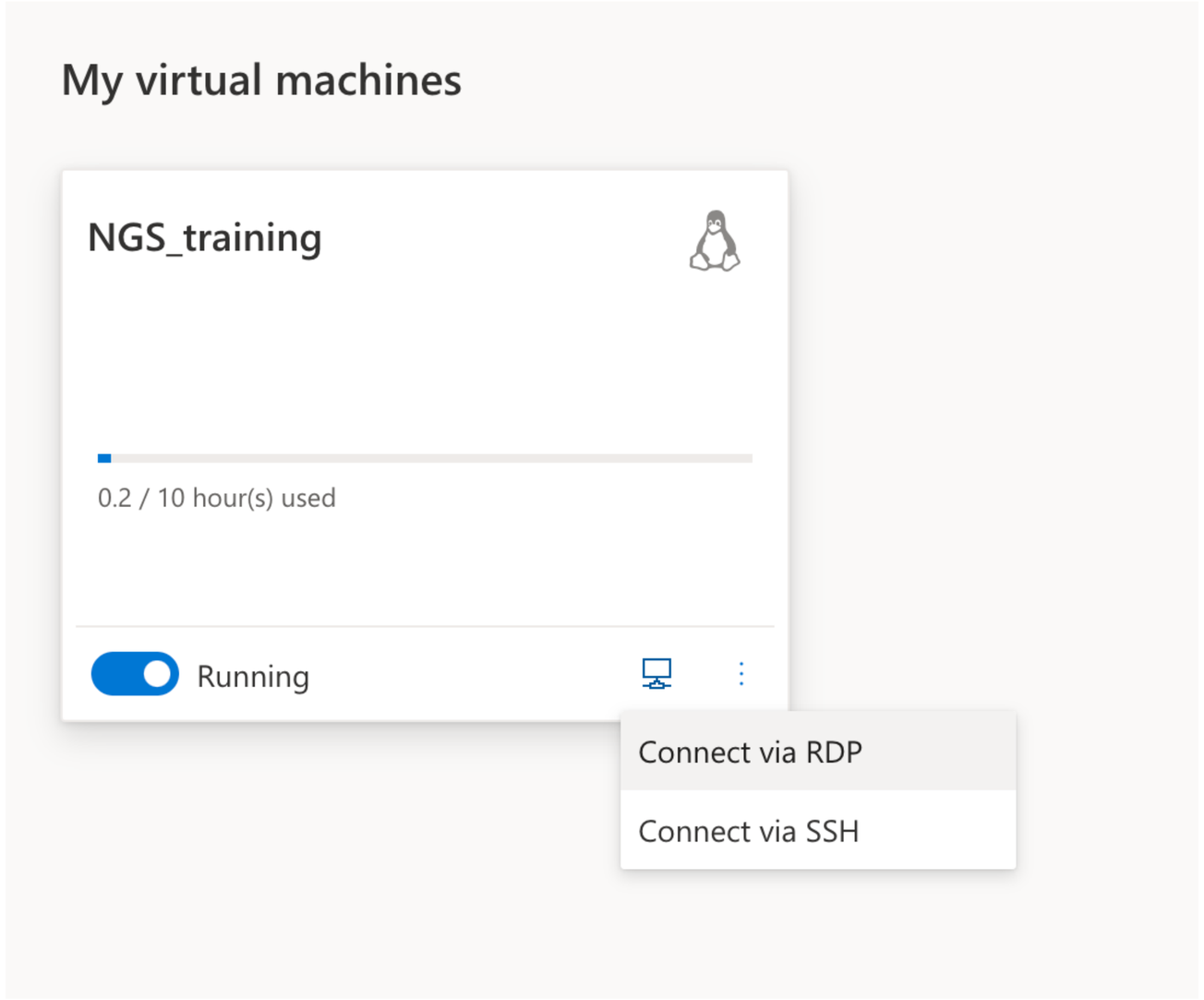


Step 4 : Start Virtual Machine

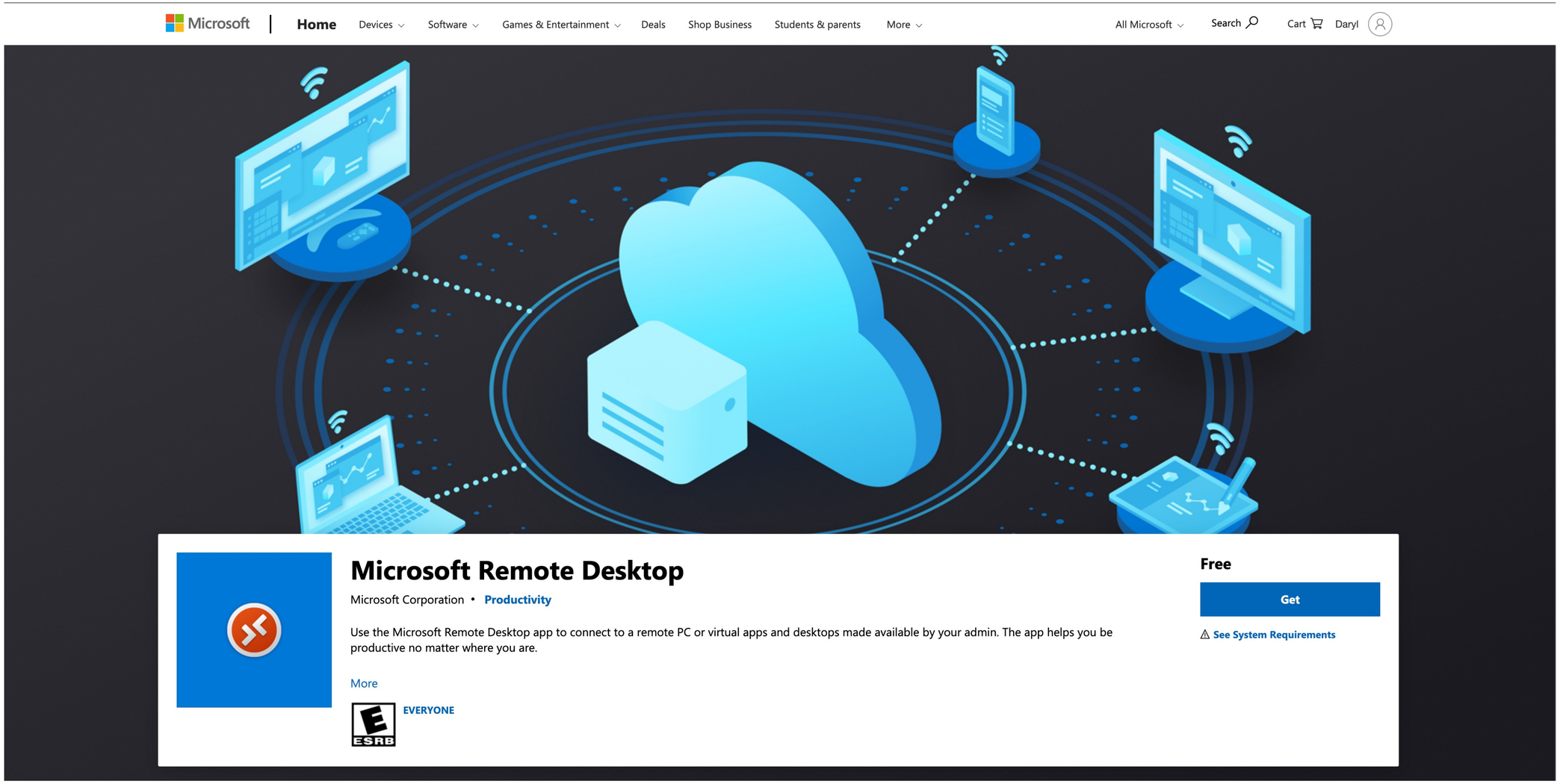


Click the slider to “Start” the VM

Step 5 : Virtual Machine Starting

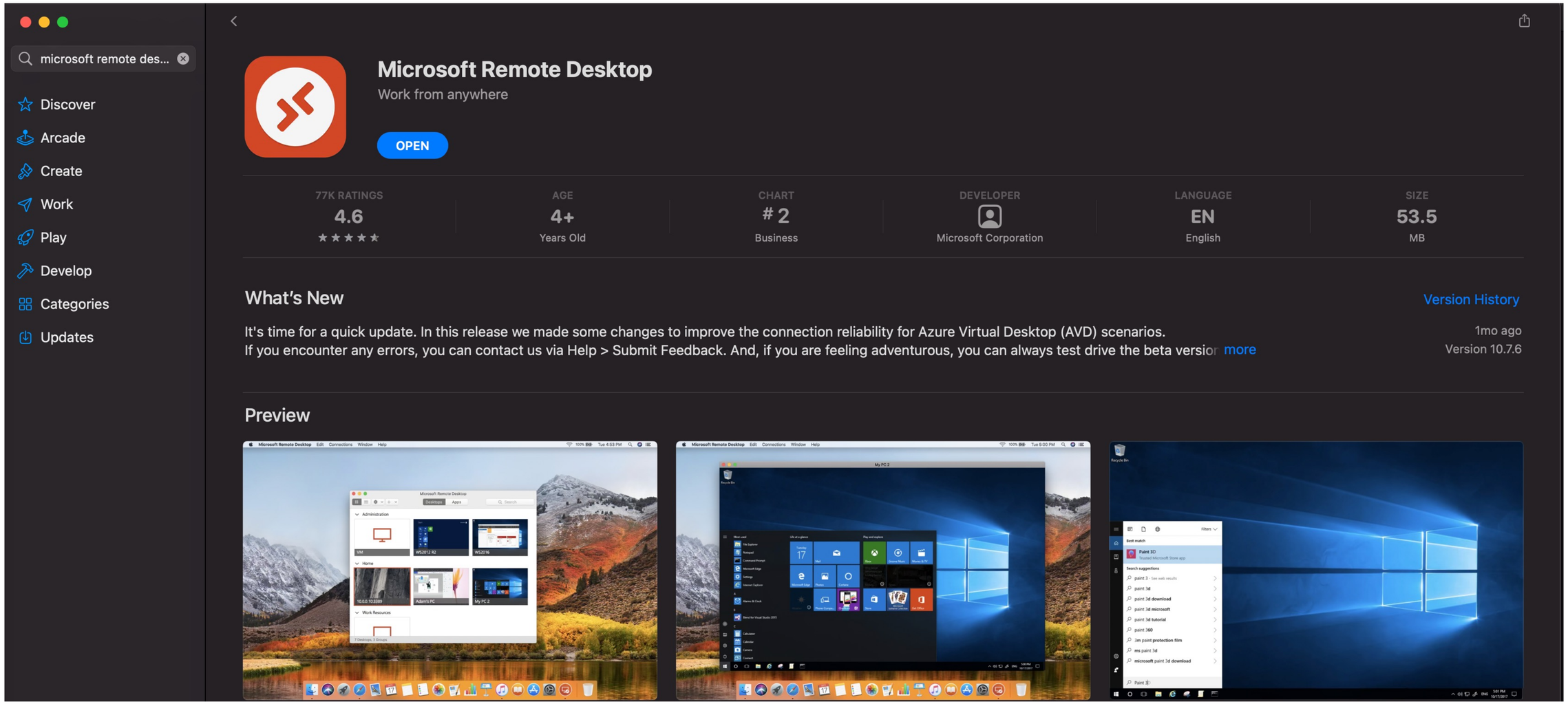


Step 6 : Install Microsoft Remote Desktop



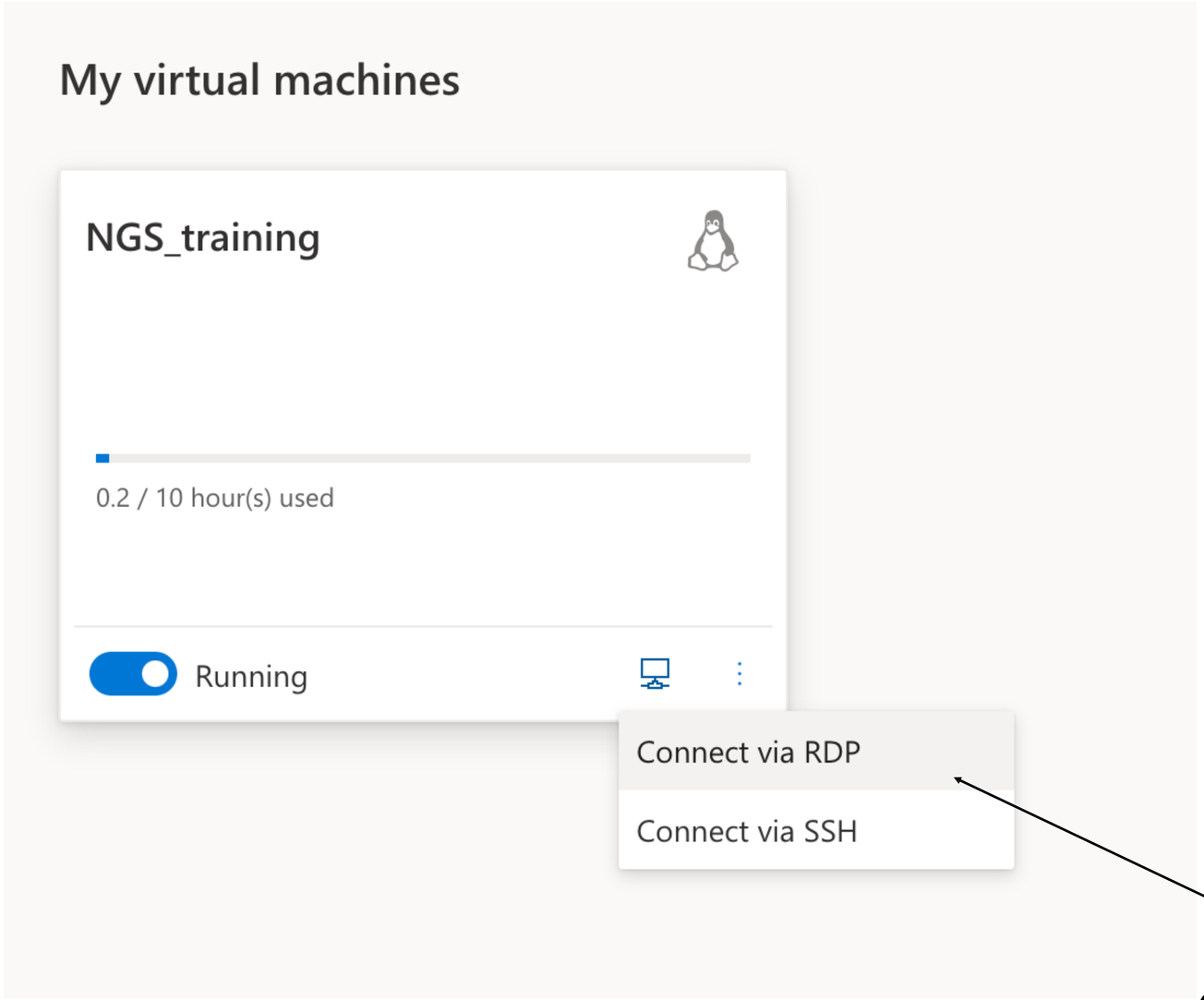
Does not matter what email you use to download !!!

Step 6 : Install Microsoft Remote Desktop (Mac)



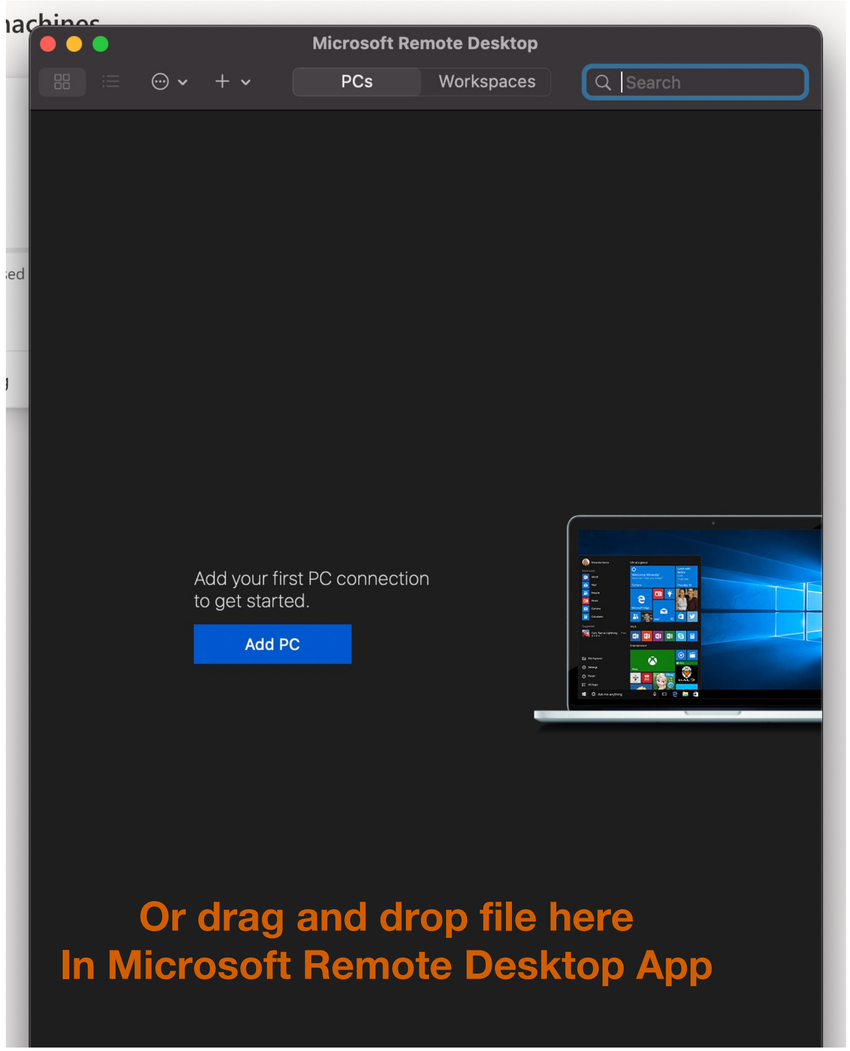
Download in App Store

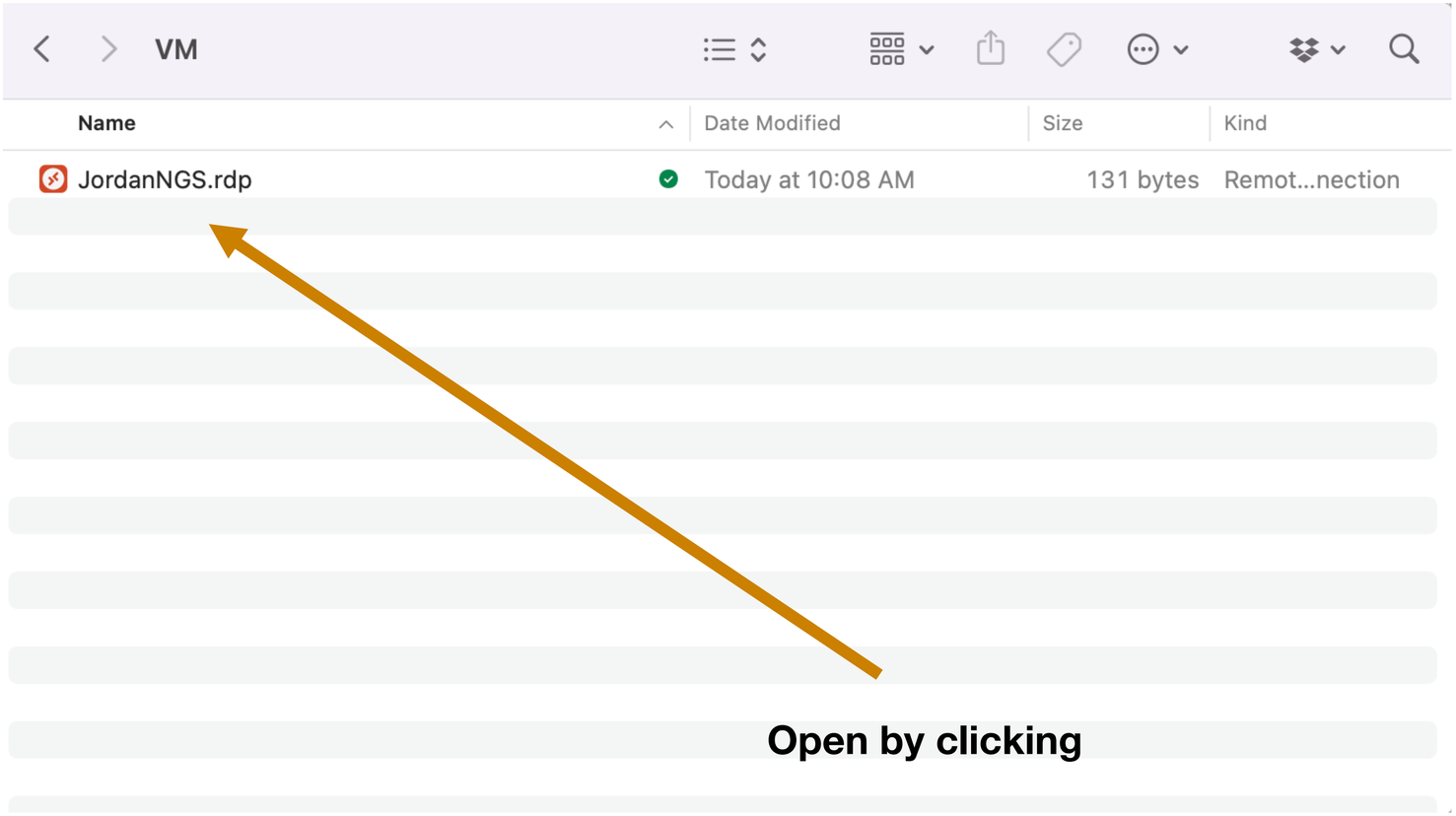
Step 5 : Virtual Machine Starting



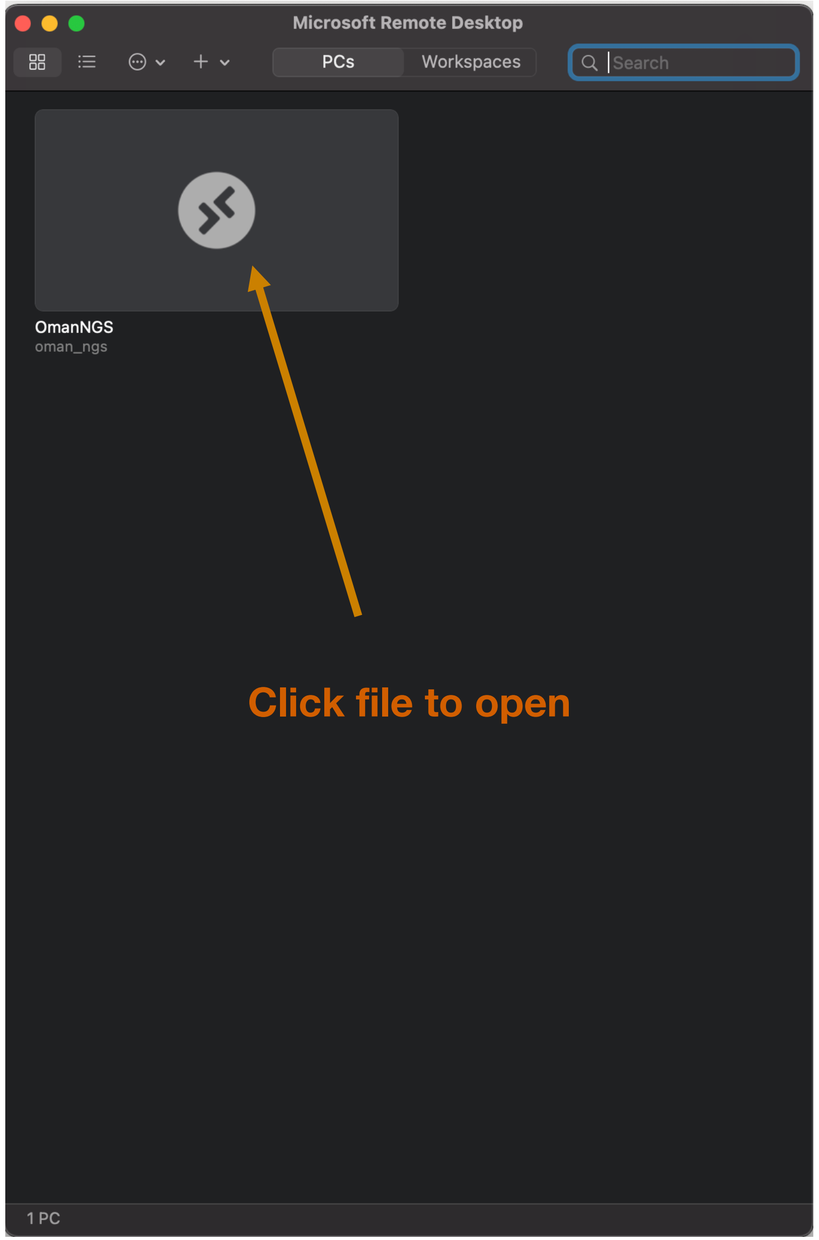
Click “Connect via RDP”
Save file

Step 6 : Open RDP file

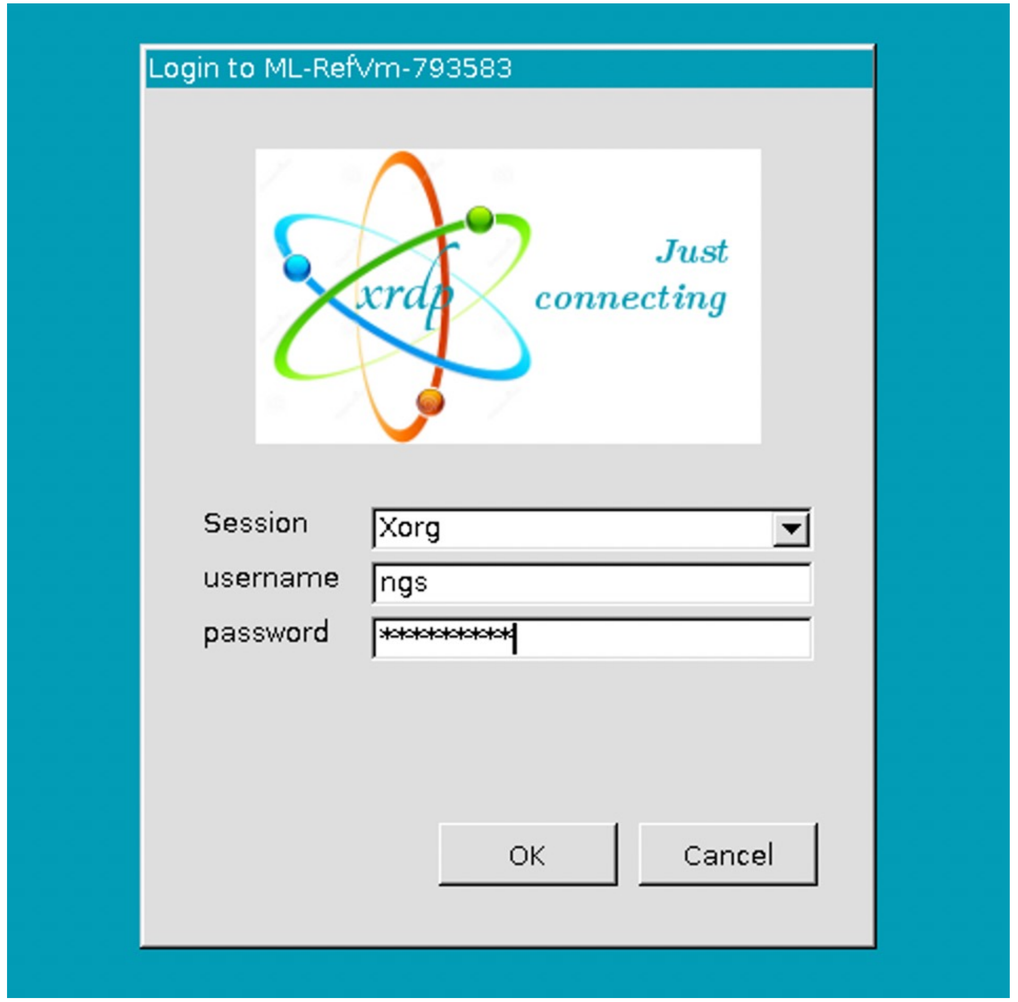




Step 6 : Open RDP file

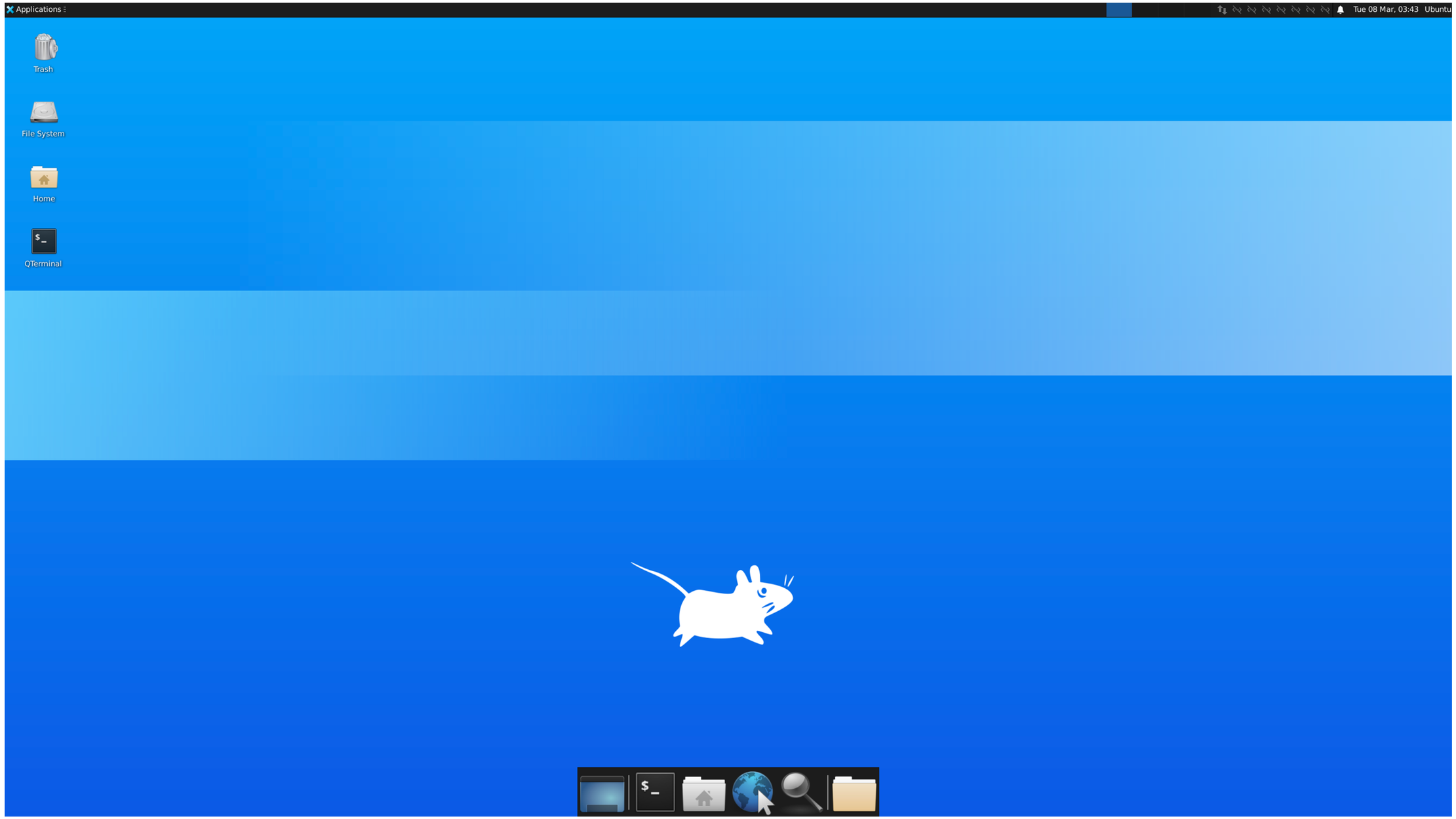


Step 7 : VM Start Screen

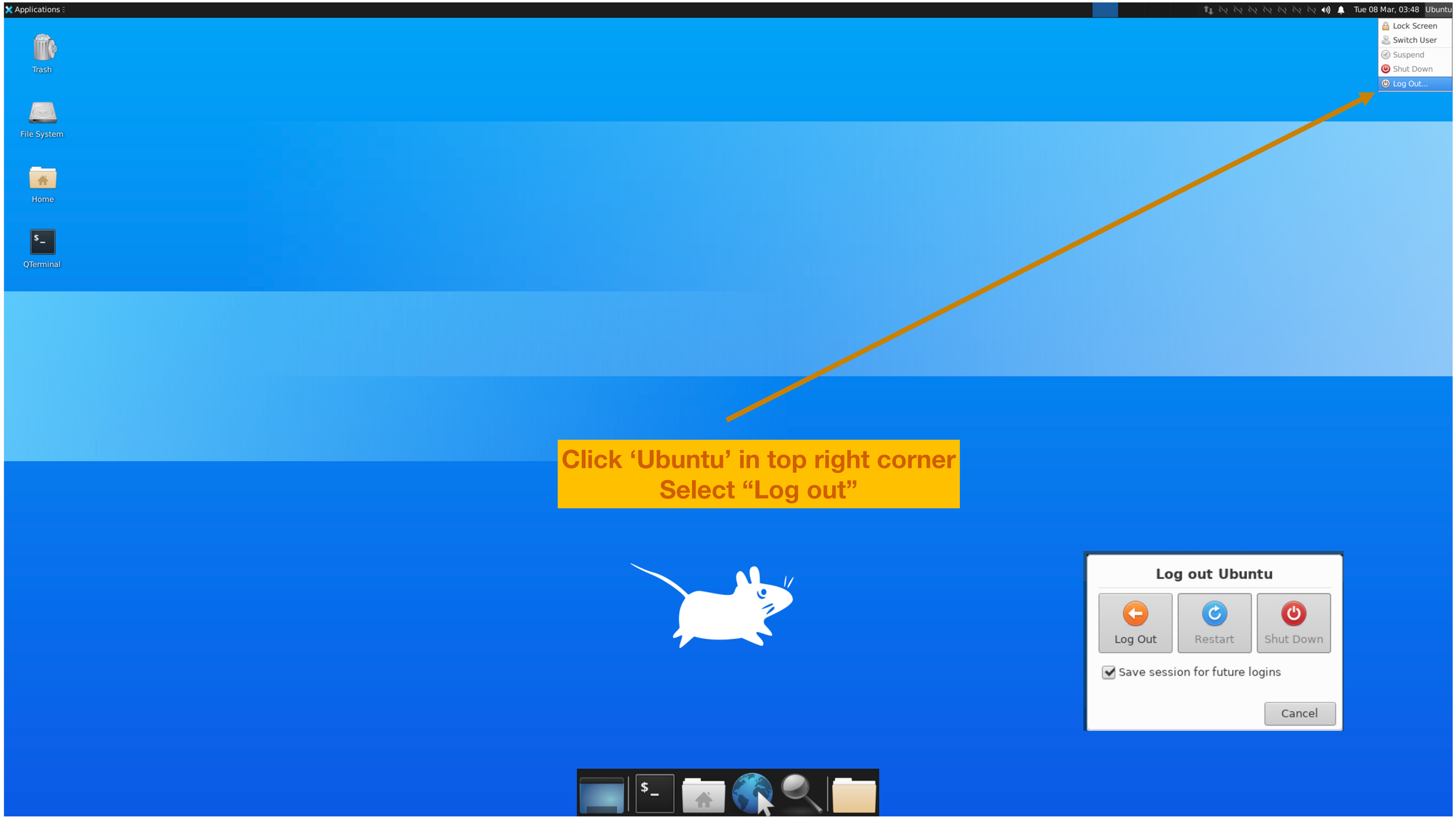


password: NGSrocks!

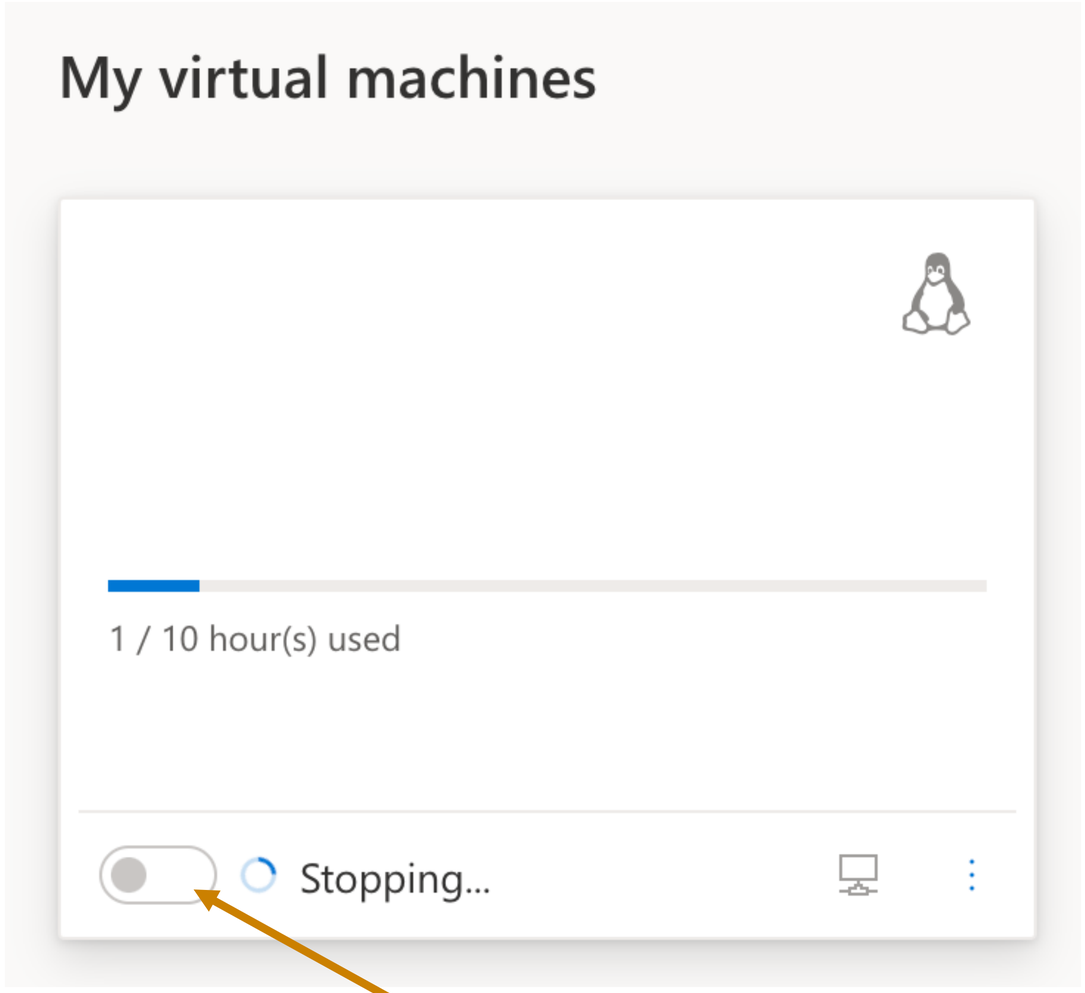
Step 8 : Have Fun !!!



Shutting Down the VM



Shutting Down the VM



Stop the VM here !!!!

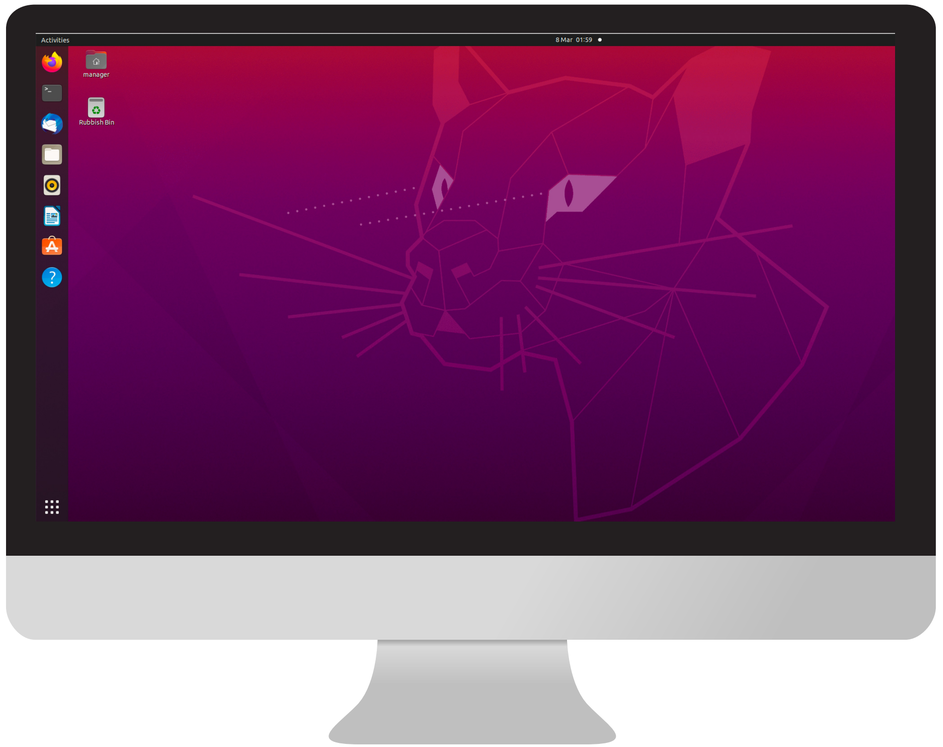
What is Unix / Linux?

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



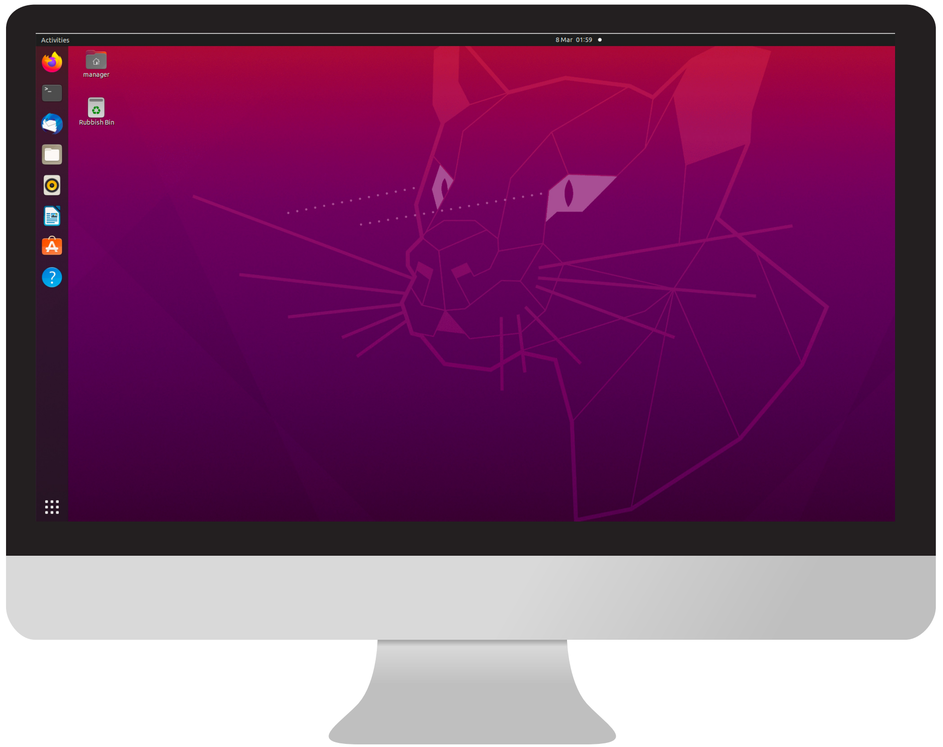
Why use Linux / Unix ?

Many bioinformatics programs produce outputs that are large text files

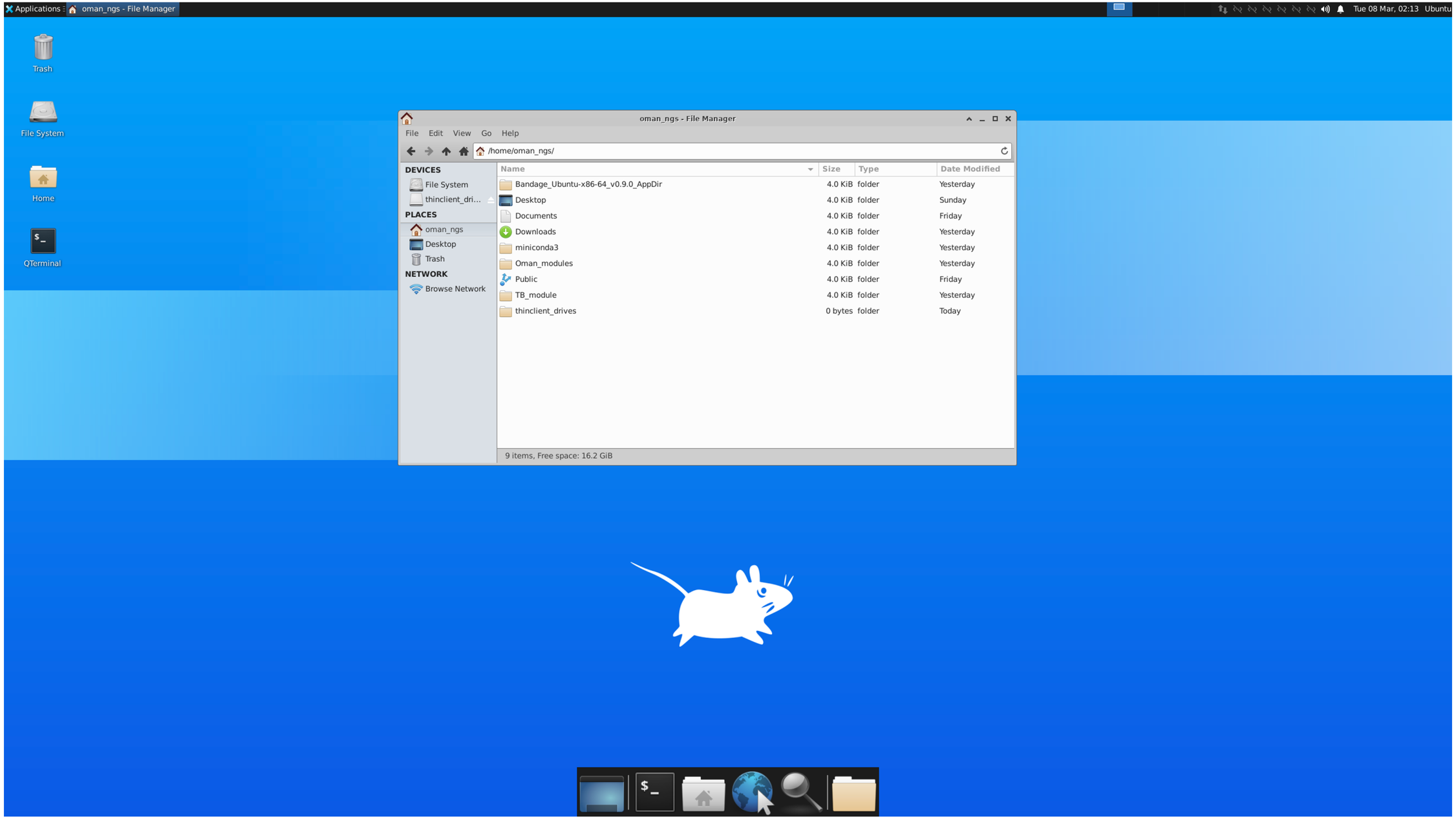
Unix is very suitable for dealing with such files

Powerful and flexible commands for processing large text files

Widely used in scientific community



Using Linux

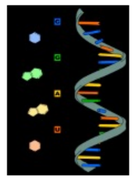

Applications
oman_ngs - File Manager
Tue 08 Mar, 02:13
Ubuntu
Trash
File System
Home
QTerminal
oman_ngs - File Manager
File Edit View Go Help
/home/oman_ngs/
DEVICES
File System
thinclient_dri...
PLACES
oman_ngs
Desktop
Trash
NETWORK
Browse Network
Name
Size
Type
Date Modified
Bandage_Ubuntu-x86-64_v0.9.0_AppDir 4.0 KiB folder Yesterday
Desktop 4.0 KiB folder Sunday
Documents 4.0 KiB folder Friday
Downloads 4.0 KiB folder Yesterday
miniconda3 4.0 KiB folder Yesterday
Oman_modules 4.0 KiB folder Yesterday
Public 4.0 KiB folder Friday
TB_module 4.0 KiB folder Yesterday
thinclient_drives 0 bytes folder Today
9 items, Free space: 16.2 GiB

INTRODUCTION TO NGS DATA TYPES

We have a sequence, now what?

How do we go from sample to actionable data?

Swab

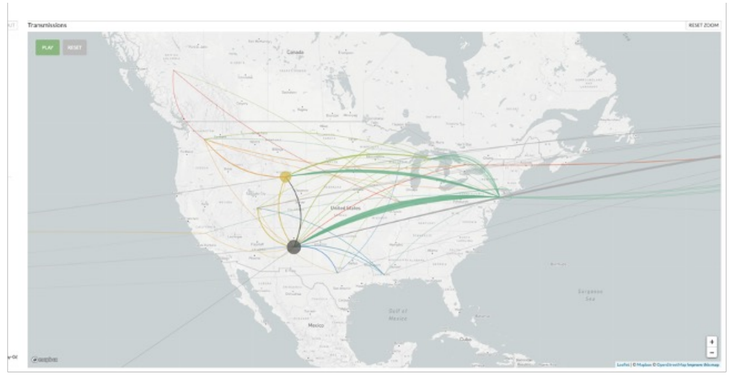


Isolate RNA

Sequencing

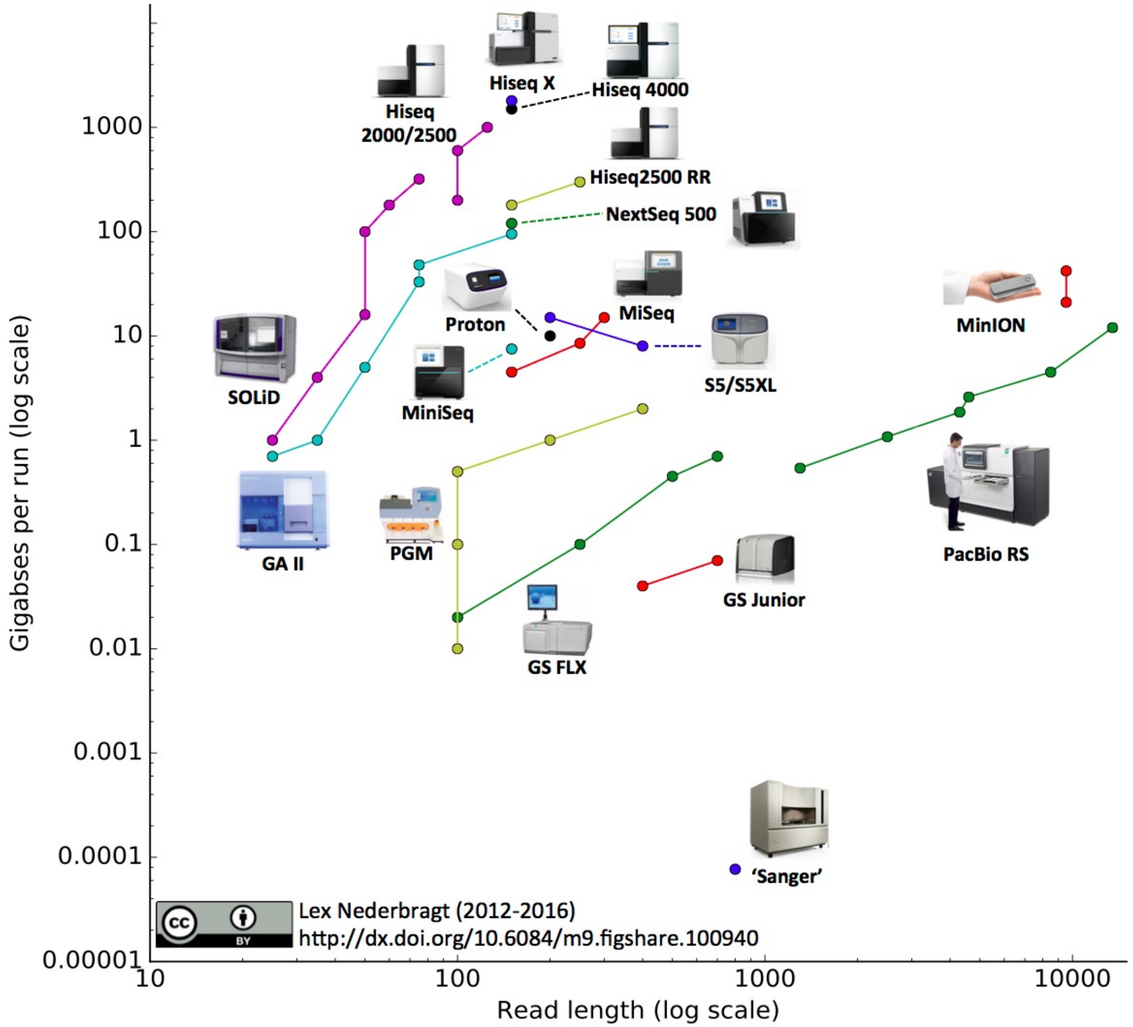
ATGCTACTTG

Genome Sequence



Genomic Epidemiology

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**

```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
ATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCAC
CAAATGAATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA
GACGGGCGATTTTGTTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGAAGGTGCCACT
ACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGCATGTCACAATTCAGAAGTAG
GACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGGCTTGAAAACCATTCTTCGTAAGGGTGGTCG
CACTATTGCCTTTGGAGGCTGTGTGTTCTCTTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCA
```

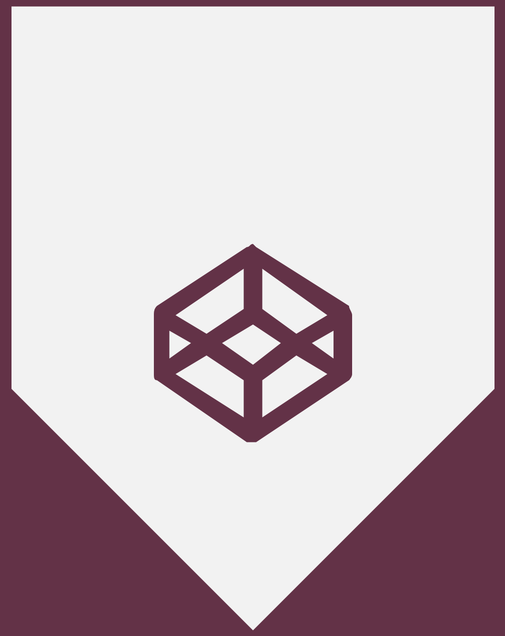
4. Sequence (nucleotides or AA)

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

File name extensions:

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

Two common paths to generating genome fasta files



GENOME FILES

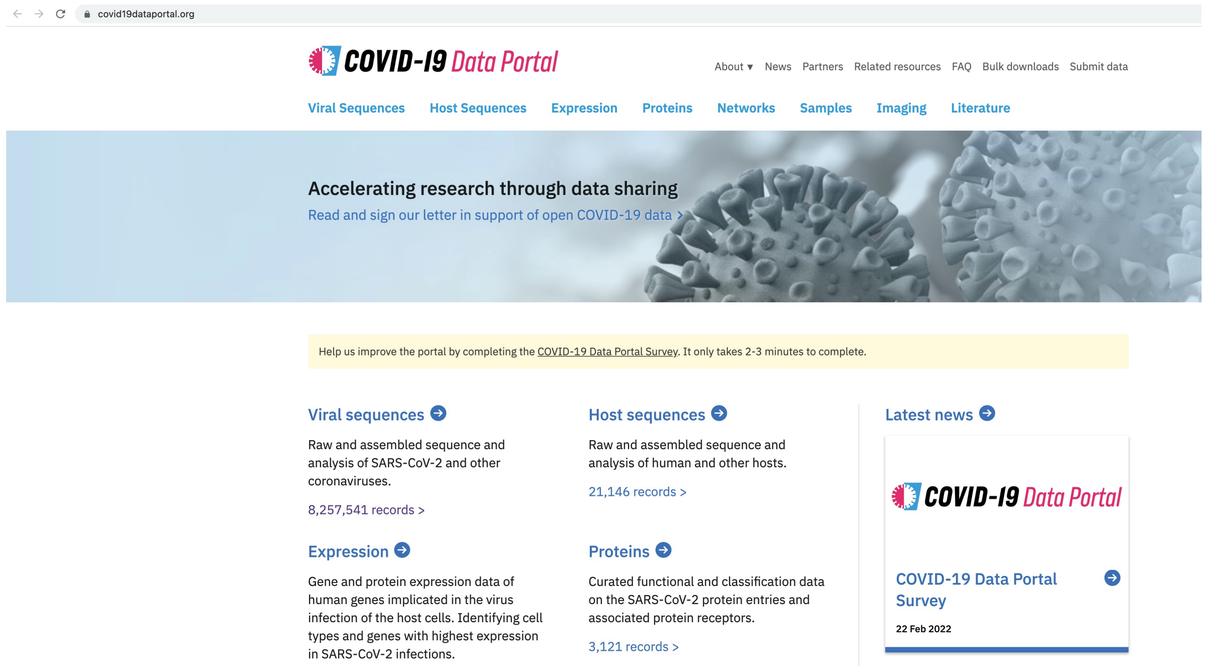
Genome annotation files

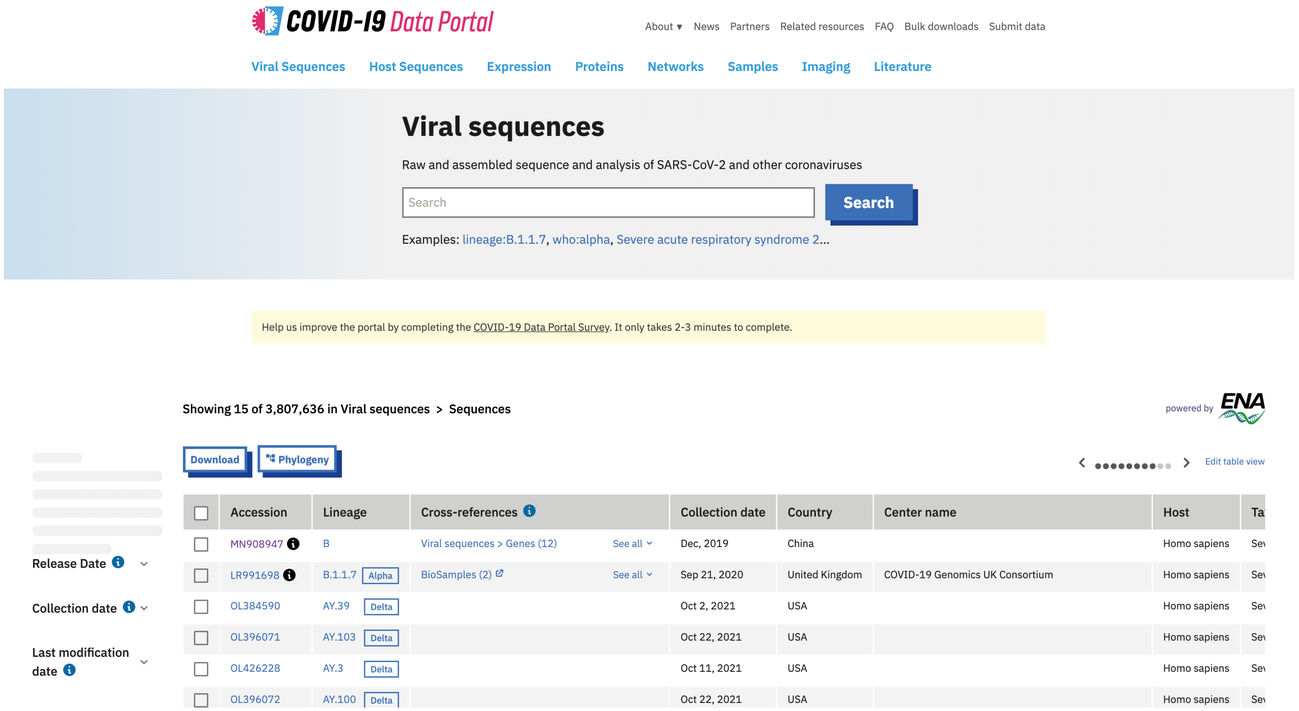
European Nucleotide Archive (ENA) : EMBL

NIH / NCBI : GenBank

General Feature File (GFF)

COVID-19 Data Portal — ENA





EMBL format from EBI

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

Feature Key

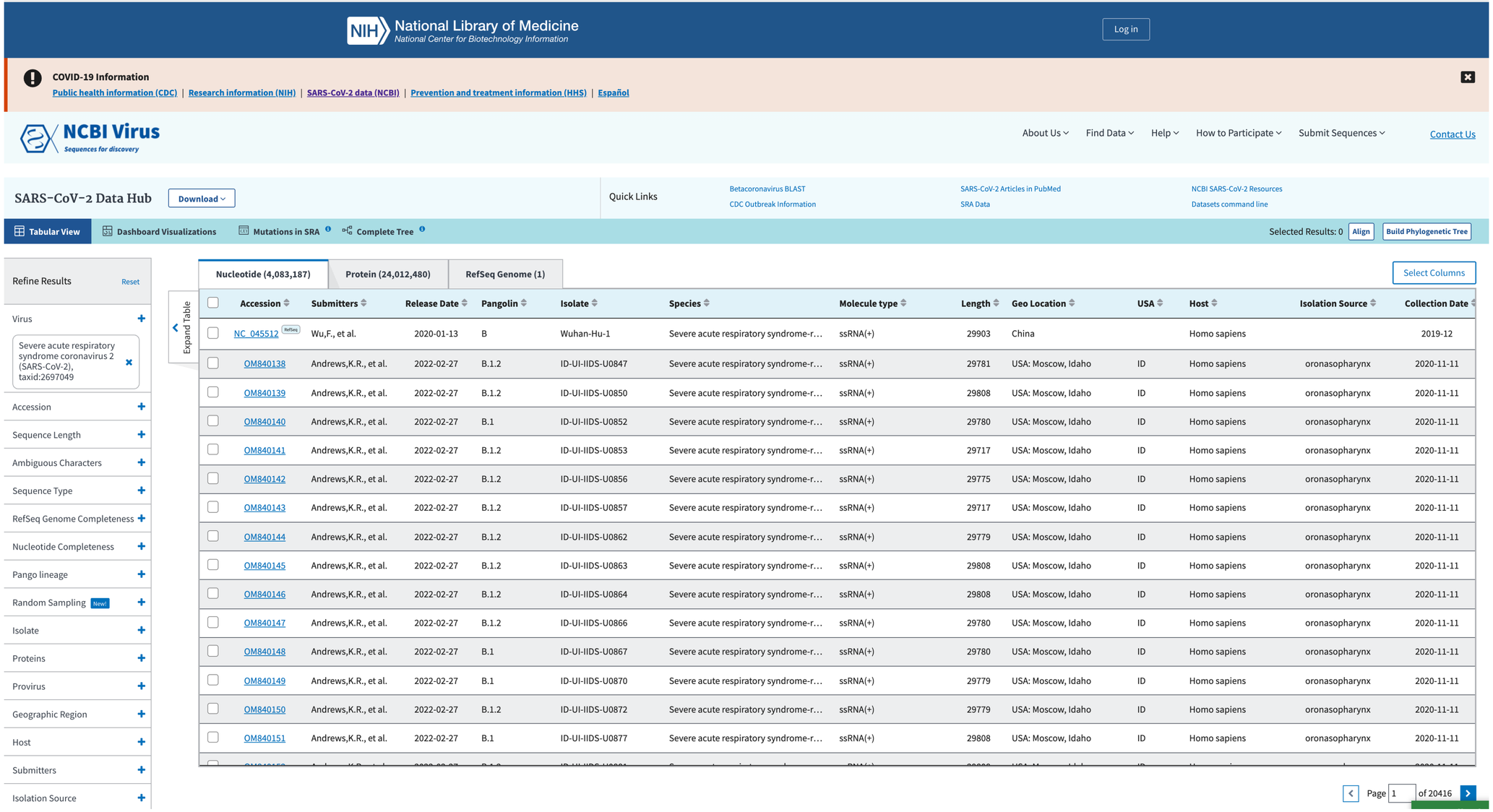
```
ID   MN908947; SV 3; linear; genomic RNA; STD; VRL; 29903 BP.
XX
AC   MN908947;
XX
DT   13-JAN-2020 (Rel. 143, Created)
DT   19-MAR-2020 (Rel. 144, Last updated, Version 6)
XX
DE   Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,
DE   complete genome.
XX
KW   .
XX
OS   Severe acute respiratory syndrome coronavirus 2
OC   Viruses; Riboviria; Nidovirales; Cornidovirineae; 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
RL   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="pp1ab; translated by -1 ribosomal frameshift"
FT                   /protein_id="QHD43415.1"
FT                   /translation="MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQH
FT                   LKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGETL
FT                   GVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQENWNT
FT                   KHSSGVTRELMRELNGGAYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFI
XX
SQ   Sequence 29903 BP; 8954 A; 5492 C; 5863 G; 9594 T; 0 other;
     attaaaggtt tataccttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct        60
     gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact       120
     cacgcagtat aattaataac taattactgt cgttgacagg acacgagtaa ctcgtctatc       180
     ttctgcaggc tgcttacggt ttcgtccgtg ttgcagccga tcatcagcac atctaggttt       240
     cgtccgggtg tgaccgaaag gtaagatgga gagccttgtc cctggtttca acgagaaaac       300
     acacgtccaa ctcagtttgc ctgttttaca ggttcgcgac gtgctcgtac gtggctttgg       360
     agactccgtg gaggaggtct tatcagaggc acgtcaacat cttaaagatg gcacttgtgg       420
```

Header

Annotation

Sequence

NIH / NCBI Virus



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; Pisoniviricetes;
            Nidovirales; Cornidovirineae; 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.
  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)
  CONSRTM   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 SARSr-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             Location/Qualifiers
     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           1..265
     gene            266..21555
                     /gene="ORF1ab"
                     /locus_tag="GU280_gp01"
                     /db_xref="GeneID:43740578"
     CDS             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="MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQ
                     HLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGE
                     TLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQEN
                     WNTKHSSGVTRELMRELNGGAYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQ

ORIGIN
        1 attaaaggtt tataccttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct
       61 gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact
      121 cacgcagtat aattaataac taattactgt cgttgacagg acacgagtaa ctcgtctatc
      181 ttctgcaggc tgcttacggt ttcgtccgtg ttgcagccga tcatcagcac atctaggttt
      241 cgtccgggtg tgaccgaaag gtaagatgga gagccttgtc cctggtttca acgagaaaac
```

Header

Annotation

Sequence

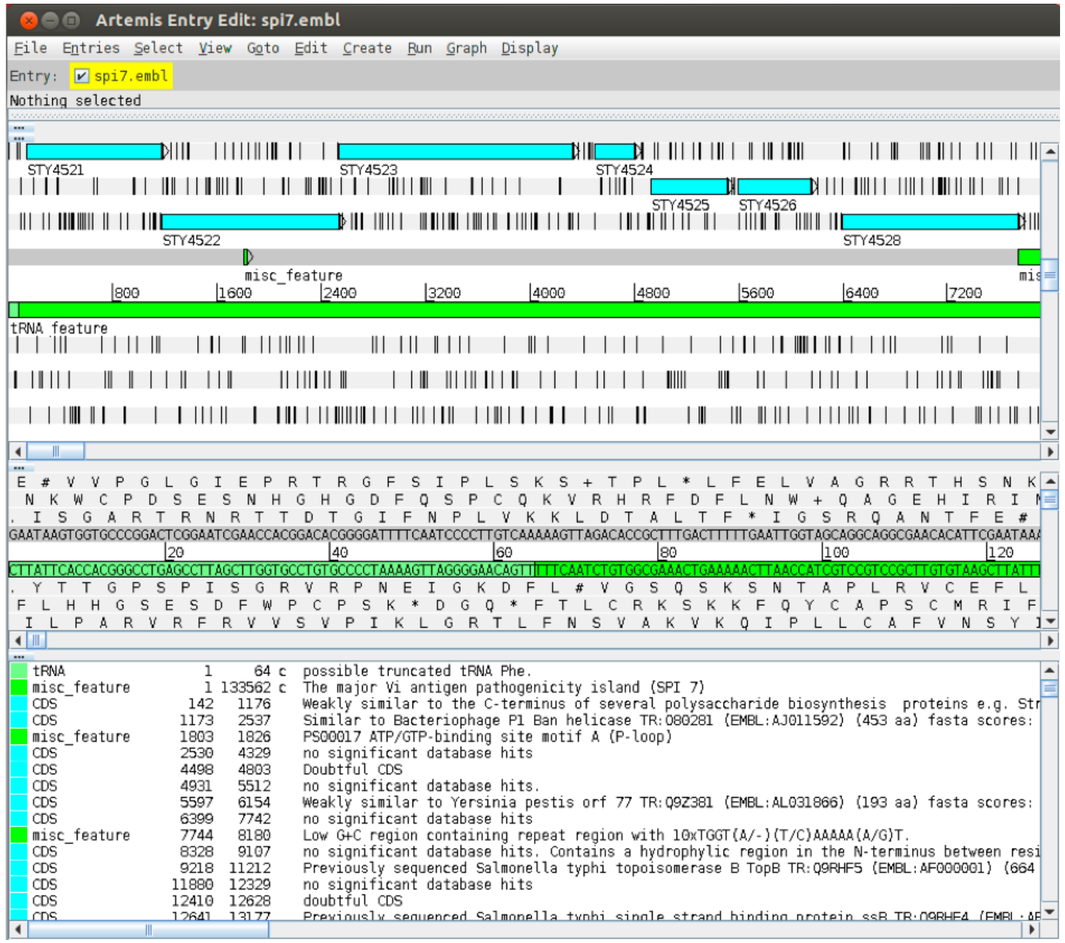
General Feature File (GFF)

```
##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_009724389.1:181..818;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_009724389.1:819..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
NC_045512.2	RefSeq	mature_protein_region_of_CDS	8555	10054	.	+	.	ID=id-YP_009724389.1:2764..3263;Note=nsp4B_TM%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 cleavages downstream of nsp4. 3D structure of the SARSr-CoV homolog has been determined (Yang et al.%2C 2003)%3B produced by both pp1a and pp1ab;Parent=cds-YP_009724
NC_045512.2	RefSeq	mature_protein_region_of_CDS	10973	11842	.	+	.	ID=id-YP_009724389.1:3570..3859;Note=nsp6_TM%3B putative transmembrane domain%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	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 nsp13_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_ExoN 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_OMT%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=ORF1a 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..818;Note=produced by both pp1a and pp1ab;Parent=cds-YP_009725295.1;gbkey=Prot;product=nsp2;protein_id=YP_009742609.1
```

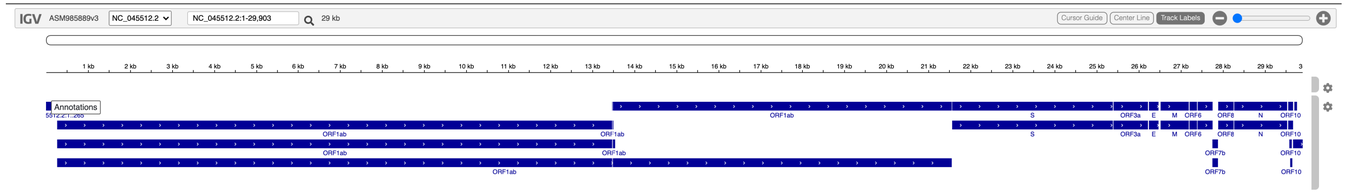
General GFF3 structure

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. Augustus or RepeatMasker) or an organization (like TAIR).
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 standards released by the Sequence Ontology Project.
4	start	Genomic start of the feature, with a **1-base offset**. This is in contrast with other 0-offset half-open sequence formats, like BED.
5	end	Genomic end of the feature, with a **1-base offset**. This is the same end coordinate as it is in 0-offset half-open sequence formats, like BED.[*citation needed*]
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 strand 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



Artemis



Integrated Genome Viewer (IGV)

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

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