INTRODUCTION TO NGS DATA



DARYL DOMMAN, PHD DARRELL DINWIDDIE, PHD DDOMMAN@GMAIL.COM









..................

Intro talk on NGS data formats

Bioinformatics Module 1 (~30-40 min)

Session wrap up & VM install guide

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



- Geneious Prime

Free "Point and Click" - UGENE

- MEGA

Web-based pipelines



- Galaxy Server

- Terra (Google cloud)
- Illumina Basespace

Command line Thousands of individual programs

existrain bash-3,25 is RMCC_ncv_ID_root-sequence.json RMCC_ncv_ID_root-sequence.json ncv_ID_focus_json										
<pre>Softer 2017-04/2019/ 2017-07 2017-02-11 2012_02_11/ WRCL remains.cmcL.sh existrain bash-3.25 bis/.i/remains.cmcL.sh Existrain bash-3.25 bis//remains.cmcL.sh Existrain bash-3.25 bis</pre>	extstrain) bash-3.2\$ ls CC_ncov_AR_json CC_ncov_AR_focus.json CC_ncov_AR_focus_root-seq CC_ncov_AR_focus_tip-freq CC_ncov_AR_tip-frequencie CC_ncov_IR_tip-frequencie CC_ncov_ID_json extstrain) bash-3.2\$ bash	uence.json uencies.json json s.json //	RMCC_ncov_ID_root-sequ RMCC_ncov_ID_tip-frequ RMCC_ncov_MT.json RMCC_ncov_MT_root-sequ RMCC_ncov_MT_tip-frequ RMCC_ncov_NM_json RMCC_ncov_NM_root-sequ	uence.json uencies.json uence.json uencies.json uence.json	RMCC_ncov_NM_tip-frequencie RMCC_ncov_WY.json RMCC_ncov_WY_root-sequence. RMCC_ncov_WY_tip-frequencie ncov_AR_focus_json ncov_AR_focus_top-trequenci ncov_AR_focus_tip-frequenci	es.json .json es.json e.json ies.json	ncov_ID_focus.json ncov_ID_focus_root- ncov_ID_focus_tjo-f ncov_MT_focus_json ncov_MT_focus_root- ncov_MT_focus_tjo-f ncov_NM_focus.json	sequence.json requencies.json sequence.json requencies.json	ncov_NM_focus_ro ncov_NM_focus_ti ncov_MY_focus.js ncov_WY_focus_ro ncov_WY_focus_ti	ot-sequence.json p-frequencies.json on ot-sequence.json p-frequencies.json
ADME.md auspice images narratives #2rebp giftim giftod2 git git-credential-osxkeychain git-receive-pack git-upload-pack gitsi git-credential-osxkeychain git-shell git-shell git-upload-archive extstrain) bash-3.25 git t git-cvsserver git-shell git-upload-pack git-credential-osxkeychain git-receive-pack git-upload-archive gitk extstrain) bash-3.25 conda activate bare	5_Store 2022-02-097 extstrain) bash-3.25 bash extstrain) bash-3.25 ls CC_ncov_AR.json CC_ncov_AR_focus.json CC_ncov_AR_focus_root-sec CC_ncov_AR_focus_tip-free extstrain) bash-3.25 ls	2022-02 //renamu uence.json uencies.json ./	-15/ 2022-02-21/ e_RMCC.sh RMCC_ncov_AR_root-seq RMCC_ncov_AR_tip-freq RMCC_ncov_ID_json RMCC_ncov_ID_root-sequ	2022_02_21/ uence.json pencies.json uence.json	RMCC/ rename_RMCC RMCC_ncov_ID_tip-frequencie RMCC_ncov_MT.json RMCC_ncov_MT_root-sequence. RMCC_ncov_MT_tip-frequencie	C.sh es.json .json es.json	RMCC_ncov_NM.json RMCC_ncov_NM_root-s RMCC_ncov_NM_tip-fr RMCC_ncov_WY.json	equence.json equencies.json	RMCC_ncov_WY_roo RMCC_ncov_WY_tip	t-sequence.json -frequencies.json
extstrain) bash-3.25 git t git-cvsserver git-shell git-upload-pack t-credential-osxkeychain git-receive-pack git-upload-archive gitk extstrain) bash-3 15 conda activate base	ADME.md auspice extstrain) bash-3.2\$ gi f2rgb f2webp fbuild	images gifclrmp giffix giftext	narratives giftog giftoc ginsi	gd2 Dl	git git-credential-osxkeych git-cvsserver	git-rec hain git-she git-upl	eive-pack ll oad-archive	git-upload-pack gitk		
navietesia) hach 3 16 conda activata haca	nextstrain) bash-3.2\$ git it it-credential-osxkevchain	git-cvsserv	er git-sl	nell pload-archive	git-upload-pack gitk					
	(nevtetrain) hash-3 2% cond	a activate b	200		6 i u					



RAW SEQUENCE DATA FILES

Illumina











Nanopore

Ion Torrent















Fastq format

- 1 @SEQ_ID
- 3 +

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).

Line 2 is the raw sequence letters.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

4 !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>CCCCCC65

Fastq format



a. unique instrument name

- b. run id
- c. flowcell id
- d. flowcell lane
- e. tile number within the flowcell lane
- f. x-coordinate of the cluster within the tile
- g. y-coordinate of the cluster within the tile

- i. Y if the read fails filter (read is bad), N otherwise
- 0 when no control bits are on
- k. index sequence

fastq header format (version > 1.8)

+Sequence ID d e f k g hi @HWI-ST486:166:C06K9ACXX:7:1101:1443:1995 1:N:0:ACAGTG

h. the member of a pair, 1 or 2 (paired-end or mate-pair reads only)

https://www.gdc-docs.ethz.ch/MDA/site/getdata/



Fastq format



Known as **Phred score**

@A00178:71:HGT77DSXX:1:1507:30291:23422 1:N:0:ACAGCAAC+GTTGCTGT ACATAGAGCTTGATGTTGTTGGCCTTCTTCCTGGTGTCGAAGAGGTCAAAGGGGGGGCCTCTTGGGGACAAAAGGACAGCCTTGAACTCAAGCT

@A00178:71:HGT77DSXX:1:1507:30291:23422 2:N:0:ACAGCAAC+GTTGCTGT

CTGGATGAGGAAGCCTGAGGAGATCACCAAGGAGGAGGAGTATGCTGCTTTCTATAAAAGCTTGACAAATGACTGGGAAGAGCATCTGGCTGTCAAG

@A00178:71:HGT77DSXX:1:2413:22806:35790 1:N:0:ACAGCAAC+GTTGCTGT GCTTGATGTTGTTGGCCTTCTTCCTGGTGTCGAAGAGGTCAAAGGGGGGCCTCTTGGGGACAAAAGGACAGCCTTGAACTCAAGCTGCCCCTC

@A00178:71:HGT77DSXX:1:2413:22806:35790 2:N:0:ACAGCAAC+GTTGCTGT

GAGAAGAAAAAAGAAGAAGACGATCAAGGAGGTTTCTCATGAATGGTCCTTGATCAACAAGCAGAAACCTATCTGGATGAGGAAGCCTGAGGAGATCA

@A00178:71:HGT77DSXX:1:2354:5620:8876 1:N:0:ACAGCAAC+GTTGCTGT ATGTTGTTGGCCTTCTTCCTGGTGTCGAAGAGGTCAAAGGGGGGGCCTCTTGGGGACAAAAGGACAGCCTTGAACTCAAGCTGCCCCTCTACAG

@A00178:71:HGT77DSXX:1:2354:5620:8876 2:N:0:ACAGCAAC+GTTGCTGT

@A00178:71:HGT77DSXX:1:1560:6741:9815 1:N:0:ACAGCAAC+GTTGCTGT GCAGGATTTTACCATGATCGACTACTTTTTGTCATGCCCAGAGAAGCTAGATTTTGCCAATGATGTTTATAGACCATTTAACGTTTCGCCAAGC

TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA

Quality score interpretation

$Q = -10 \log_{10} P$ \longrightarrow $P = 10^{-1}$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

The quality (Q), also called Phred score, is the probability (P) that the corresponding basecall is incorrect.



$$P = 10^{\frac{-Q}{10}}$$



fast5 format

Binary file (not human readable)

Contains:

- Sequence of a read
- Raw signal data from pore
- Additional log files

Typically convert fast5 to fastq for downstream analyses



BAM format for read data

<u>Binary Alignment Map format</u>

Binary conversion of the <u>Sequence</u> <u>Alignment</u> <u>Map</u> (SAM) file

Typically convert bam to fastq for downstream analyses

File format: SAM / BAM (each line: one aligned sequence read)

The SAM/BAM file format is very powerful. It is unlikely that you will need to work with the contents of a SAM/BAM file directly, but it is very informative to visualize it in a viewer and it is a great format to do further analysis with. The format specifications are at http://samtools.sourceforge.net/SAM1.pdf. Below is a brief overview of the information contained in such files.





Fasta format



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

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

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

File name extensions:

4. Sequence (nucleotides or AA)

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



Two common paths to generating genome fasta files







GENOME ANNOTATION FILES



Genome annotation files







	SDC	International Nuc	leotide Sequence D	atabase Collabora	tion
ABOUT	INSDC	POLICY	ADVISORS	DOCUMENT	rs
ENA S NCBI	Internation • The line operations asser config	al Nucleotide Sequence International Nucleotide Se ates between DDBJ, EMB Imblies to functional annota gurations.	equence Database Collaboration <u>L-EBI</u> and <u>NCBI</u> . INSDC co ation, enriched with context	on ration (INSDC) is a long- vers the spectrum of dat ual information relating to	standing foundational initiative a raw reads, through alignmer o samples and experimental
Ö DDDI	2	Data type	DDBJ	EMBL-EBI	NCBI
& DDBJ		Next generation reads	Sequence Read Archive		Sequence Read Archive
		Capillary reads	Trace Archive	European Nucleotide	Trace Archive
		Annotated sequences	DDBJ		GenBank
		Samples	<u>BioSample</u>		BioSample
		Studies	<u>BioProject</u>		<u>BioProject</u>



COVID-19 Data Portal – ENA

Note
News
Partners
Related resource
Robit works
Submit data

Viral Sequences
Host Sequences
Kecelerating research through data sharing
Read and sign our letter in support of open COVID-19 data

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

Viral sequences 🕤

 $\leftarrow \rightarrow$ C \triangleq covid19dataportal.org

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

8,257,541 records >

Expression 🕤

Gene and protein expression data of human genes implicated in the virus infection of the host cells. Identifying cell types and genes with highest expression in SARS-CoV-2 infections.

Host sequences 🕑

Raw and assembled sequence and analysis of human and other hosts.

21,146 records >

Proteins 🕤

Curated functional and classification data on the SARS-CoV-2 protein entries and associated protein receptors.

3,121 records >



] COVID-19 Da	About 🔻 News	Partners Rel	ated resources	FAQ Bulk downloads	Submit data		
Vira	al Sequences Host Se	quences Expression	Proteins	Networks	Samples	Imaging	Literature	
	Ri	Viral sequer w and assembled sequer earch amples: lineage:B.1.1.7,	ICES nce and analysis who:alpha, Seve	of SARS-CoV-2	2 and other co ratory syndroi	oronaviruses	ırch	
Не	lelp us improve the portal by co	npleting the <u>COVID-19 Data F</u>	ortal Survey. It only	y takes 2-3 minute	es to complete.			

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

	Download										
		Accession	Lineage	Cross-references		Collection date	Country	Center name			
		MN908947	В	Viral sequences > Genes (12)	See all ~	Dec, 2019	China				
Release Date 👽 🗸		LR991698 🕄	B.1.1.7 Alpha	BioSamples (2)	See all 🛩	Sep 21, 2020	United Kingdom	COVID-19 Genomics UK Consortium			
Collection date \bigcirc \checkmark		OL384590	AY.39 Delta			Oct 2, 2021	USA				
Last modification		OL396071	AY.103 Delta			Oct 22, 2021	USA				
date 1		OL426228	AY.3 Delta			Oct 11, 2021	USA				
		OL396072	AY.100 Delta			Oct 22, 2021	USA				



Edit table viewHostTaHomo sapiensSevHomo sapiensSevHomo sapiensSevHomo sapiensSevHomo sapiensSevHomo sapiensSevHomo sapiensSevHomo sapiensSev

EMBL format from EBI

	AC	MN908947;	
	XX DT	13-JAN-2020 (R	el. 143, Create
	DT	19-MAR-2020 (R	el. 144, Last u
	XX DE	Severe acute r	espiratory synd
	DE	complete genom	le.
	КW		
	XX		
	0S	Severe acute r	espiratory synd
	OC	Orthocoronavir	inae; Betacoron
	XX		
	RN RP	[1] 1-29903	
	RX	PUBMED; 320155	08.
	RA	Wu F., Zhao S.	, Yu B., Chen Y
	RA	Zheng J.J., Xu	L., Holmes E.C
	RT	"A new coronav	irus associated
	RL XX	Nature 579(779	8):265-269(2020
	RN	[2]	
	RP	1-29903 WU E Zhao S	Vu B Chen V
	RA	Tian JH., Pe	i YY., Yuan M
Two-character line code	RA	Wang QM., Zh	eng JJ., Xu L
	RL	; Submitted (05-	JAN-2020) to th
indicates the type of information	RL	Shanghai Publi	C Health Clinic
contained in the line	RL XX	University, Sh	anghai, China
	DR	MD5; 105c82802	b67521950854a85
	XX	On Jan 17, 202	0 this sequence
	CC	##Assembly-Dat	a-START##
	CC	Assembly Metho	d :: Mega
	CC	##Assembly-Dat	a-END##
	XX	Wass	Tanating (Oval
	FH FH	кеу	Location/Quai
	FT	source	129903
Feature	FT FT		/organism="Se /host="Homo s
Kay	\mathbf{FT}		/isolate="Wuh
кеу	FT FT		/mol_type="ge /country="Chi
	FT		/collection_d
	FT		/db_xref="tax
	FT	gene	26621555
	FT		/gene="orflab
	FT FT	CDS	<pre>join(266134 /codon start=</pre>
	\mathbf{FT}		/ribosomal_sl
	FT FT		/gene="orflab /product="orf
	FT		/note="pplab;
	FT		<pre>/protein_id="</pre>
	FT		LKDGTCGLVEVEK
	FT		GVLVPHVGEIPVA
	FT VV		KHSSGVTRELMRE
	SO	Sequence 29	903 BP: 8954
	- Z	attaaaggtt	tataccttcc ca
		gttctctaaa	cgaactttaa aa
		cacgcagtat a	aattaataac ta
		ttctgcaggc	tgcttacggt tt
			cyaccgaaag gt
		agactccoto	gaggaggtet ta

ID MN908947; SV 3; linear; genomic RNA; STD; VRL; 29903 BP.

updated, Version 6)

XX

drome coronavirus 2 isolate Wuhan-Hu-1,

drome coronavirus 2 les; Cornidovirineae; Coronaviridae; navirus; Sarbecovirus.

M., Wang W., Song Z.G., Hu Y., Tao Z.W., L., Zhang Y.L., Dai F.H., Liu Y., Wang Q.M., C., Zhang Y.Z.; with human respiratory disease in China";).

.-M., Wang W., Hu Y., Song Z.-G., Tao Z.-W., 1.L., Zhang Y.-L., Dai F.-H., Liu Y., ., Holmes E.C., Zhang Y.-Z.;

ne INSDC. cal Center & School of Public Health, Fudan

51fc6eefd.

version replaced MN908947.2. ahit v. V1.1.3 ımina

lifiers

evere acute respiratory syndrome coronavirus 2" sapiens" nan-Hu-1" enomic RNA" lna" late="Dec-2019" con:2697049"

468,13468..21555) Lippage lab polyprotein" translated by -1 ribosomal frameshift" 'QHD43415.1" ="MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQH (GVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGETL YRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTD ELNGGAYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFI

A; 5492 C; 5863 G; 9594 T; 0 other; aggtaacaa accaaccaac tttcgatctc ttgtagatct atctgtgtg gctgtcactc ggctgcatgc ttagtgcact aattactgt cgttgacagg acacgagtaa ctcgtctatc tcgtccgtg ttgcagccga tcatcagcac atctaggttt taagatgga gagccttgtc cctggtttca acgagaaaac tgttttaca ggttcgcgac gtgctcgtac gtggctttgg atcagaggc acgtcaacat cttaaagatg gcacttgtgg

Header

Annotation

Sequence

60 120

180

240 300

360

420

NIH / NCBI Virus

COVID-19 Information

Public health information (CI

NCBI Virus Sequences for discovery

SARS-CoV-2 Data Hub

🖽 Tabular View

Refine Results

Dashboard V

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

			NIH	National Lib	Drary of Medic Biotechnology Informa	cine ^{tion}					Log in			
<u>C)</u> <u>Res</u>	search ir	nformation (NIH)	<u>SARS-CoV-2 data (NCB</u>	I) Prevention and	l treatment informatio	on (HHS) Español								×
										About Us ∽	Find Data ∽ Help ∽	How to Participate	 Submit Sequences 	<u>Contact Us</u>
Down	load ~]				Quick Links	Betacoronavirus BLAST CDC Outbreak Information		SARS-CoV-2 A SRA Data	rticles in PubMed		NCBI SARS-CoV-2 Resource Datasets command line	25	
sualizati	ions	Mutations in	SRA ⁰ 📲 Complete	Tree ⁰								Se	elected Results: 0 Align B	uild Phylogenetic Tree
	Nuc	leotide (4,083,187	') Protein (24,	012,480) R	RefSeq Genome (1)									Select Columns
ble		Accession 🗢	Submitters 🗘	Release Date ≑	Pangolin 🗢	Isolate 🗢	Species 🗢	Molecule type ≑	Length 🌲	Geo Location 🗢	USA 🗘	Host 🗘	Isolation Source 🗢	Collection Date 4
kpand Ta		NC 045512 RefSeq	Wu,F., et al.	2020-01-13	В	Wuhan-Hu-1	Severe acute respiratory syndrome-r	ssRNA(+)	29903	China		Homo sapiens		2019-12
<u> </u>		<u>OM840138</u>	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-UI-IIDS-U0847	Severe acute respiratory syndrome-r	ssRNA(+)	29781	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
		<u>OM840139</u>	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-UI-IIDS-U0850	Severe acute respiratory syndrome-r	ssRNA(+)	29808	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
Ī		<u>OM840140</u>	Andrews,K.R., et al.	2022-02-27	B.1	ID-UI-IIDS-U0852	Severe acute respiratory syndrome-r	ssRNA(+)	29780	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
		<u>OM840141</u>	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-UI-IIDS-U0853	Severe acute respiratory syndrome-r	ssRNA(+)	29717	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
Ī		<u>OM840142</u>	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-UI-IIDS-U0856	Severe acute respiratory syndrome-r	ssRNA(+)	29775	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
		<u>OM840143</u>	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-UI-IIDS-U0857	Severe acute respiratory syndrome-r	ssRNA(+)	29717	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
		<u>OM840144</u>	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-UI-IIDS-U0862	Severe acute respiratory syndrome-r	ssRNA(+)	29779	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
		<u>OM840145</u>	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-UI-IIDS-U0863	Severe acute respiratory syndrome-r	ssRNA(+)	29808	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
		<u>OM840146</u>	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-UI-IIDS-U0864	Severe acute respiratory syndrome-r	ssRNA(+)	29808	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
		<u>OM840147</u>	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-UI-IIDS-U0866	Severe acute respiratory syndrome-r	ssRNA(+)	29780	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
		<u>OM840148</u>	Andrews,K.R., et al.	2022-02-27	B.1	ID-UI-IIDS-U0867	Severe acute respiratory syndrome-r	ssRNA(+)	29780	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
		<u>OM840149</u>	Andrews,K.R., et al.	2022-02-27	B.1	ID-UI-IIDS-U0870	Severe acute respiratory syndrome-r	ssRNA(+)	29779	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
		<u>OM840150</u>	Andrews,K.R., et al.	2022-02-27	B.1.2	ID-UI-IIDS-U0872	Severe acute respiratory syndrome-r	ssRNA(+)	29779	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
		<u>OM840151</u>	Andrews,K.R., et al.	2022-02-27	B.1	ID-UI-IIDS-U0877	Severe acute respiratory syndrome-r	ssRNA(+)	29808	USA: Moscow, Idah	o ID	Homo sapiens	oronasopharynx	2020-11-11
	\square	01/0/01/50			D 1 0		<u> </u>				ID.		1	

< Page 1 of 20416 >

GenBank format from NCBI

	NC_045512 29903 bp ss-RNA linear VRL 18-JUL-2020
DELIMITION	complete genome.
ACCESSION	NC_045512
DBLINK	BioProject: PRJNA485481
KEYWORDS	RefSeq.
SOURCE	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
ONGANTSH	Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
	Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae;
REFERENCE	Betacoronavirus; Sarbecovirus. 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.,
TITLE	A new coronavirus associated with human respiratory disease in
	China
JOURNAL	Nature 579 (7798), 265-269 (2020)
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.,
TITLE	Programmed ribosomal frameshifting in decoding the SARS-CoV genome
JOURNAL	Virology 332 (2), 498-510 (2005)
PUBMED	15680415 3 (bases 29728 to 29768)
AUTHORS	Robertson, M.P., Igel, H., Baertsch, R., Haussler, D., Ares, M. Jr. and
TTT	Scott,W.G.
IIILE	The structure of a rigorously conserved KNA element within the SARS virus genome
JOURNAL	PLoS Biol. 3 (1), e5 (2005)
PUBMED	15630477
AUTHORS	4 (Dases 29609 to 29657) 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
PUBMED	J. VIFOL. 73 (10), 8349-8355 (1999) 10482585
REFERENCE	5 (bases 1 to 29903)
CONSRTM	NCBI Genome Project
JOURNAL	Submitted (17–JAN–2020) National Center for Biotechnology
	Information, NIH, Bethesda, MD 20894, USA
REFERENCE	6 (bases 1 to 29903) Wu E Zhao S Yu B Chen Y M Wang W Hu Y Song Z G
Admons	Tao,ZW., Tian,JH., Pei,YY., Yuan,M.L., Zhang,YL.,
	Dai,FH., Liu,Y., Wang,QM., Zheng,JJ., Xu,L., Holmes,E.C. and
TTTLE	Zhang,YZ. Direct Submission
JOURNAL	Submitted (05-JAN-2020) Shanghai Public Health Clinical Center &
COMMENT	School of Public Health, Fudan University, Shanghai, China
COMMENT	REVIEWED REFSEU: 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. ###
	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
	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_FND##
	COMPLETENESS: full length.
FEATURES	Location/Qualifiers
source	/organism="Severe acute respiratory syndrome coronavirus
	2"
	/mol_type="genomic RNA" (isolate="Wwwhap Hu 1"
	/host="Homo sapiens"
	/db_xref="taxon:2697049"
	/country="China" (collection_date="Dec_2010"
5'UTR	1265
gene	26621555
	/gene="ORF1ab" /locus_tag="GU280_gp01"
	/db_xref="GeneID:43740578"
CDS	join(26613468,1346821555)
	/gene="UKFlab" /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"
	/ LTAINS LA LION="MESLVPGENEKTHVQLSLPVLQVKDVLVKGFGDSVEEVLSEARQ HLKDGTCGLVEVEKGVLPOLEOPYVFTKRSDARTAPHGHVMVFT VAFT FGT0YGRSGF
	TLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQEN
	WNTKHSSGVTRELMRELNGGAYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQ
ORIGIN	
1 att	aaaggtt tataccttcc caggtaacaa accaaccaac tttcqatctc ttqtaqa
61 gtt	ctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgc
121 cac	gcagtat aattaataac taattactgt cgttgacagg acacgagtaa ctcgtct

spiratory syndrome coronavirus 2 (SARS-CoV-2) spiratory syndrome coronavirus 2 ria; Orthornavirae; Pisuviricota; Pisoniviricetes; rnidovirineae; Coronaviridae; Orthocoronavirinae; Sarbecovirus. 29903) Yu,B., Chen,Y.M., Wang,W., Song,Z.G., Hu,Y., J.H., Pei,Y.Y., Yuan,M.L., Zhang,Y.L., Dai,F.H., M., Zheng,J.J., Xu,L., Holmes,E.C. and Zhang,Y.Z. us associated with human respiratory disease in 8), 265-269 (2020) 2020 Apr;580(7803):E7. PMID: 32296181] to 13503) enderson,C.M., Anderson,C.B., Gesteland,R.F., Howard.M.T. somal frameshifting in decoding the SARS-CoV genome), 498-510 (2005) to 29768) Igel,H., Baertsch,R., Haussler,D., Ares,M. Jr. and ^a a rigorously conserved RNA element within the SARS), e5 (2005) to 29657) Chang,R.Y. and Brian,D.A. lly conserved hairpin-type 3' untranslated region tions in coronavirus RNA replication 0), 8349-8355 (1999) 29903) ject AN-2020) National Center for Biotechnology H, Bethesda, MD 20894, USA 29903) Yu,B., Chen,Y.-M., Wang,W., Hu,Y., Song,Z.-G., ,J.-H., Pei,Y.-Y., Yuan,M.L., Zhang,Y.-L., Y., Wang,Q.-M., Zheng,J.-J., Xu,L., Holmes,E.C. and AN-2020) Shanghai Public Health Clinical Center & Health, Fudan University, Shanghai, China This record has been curated by NCBI staff. The nce is identical to MN908947. this sequence version replaced NC_045512.1. added using homology to SARSr-CoV NC_004718.3. ### 'Wuhan seafood market pneumonia virus.' If you have ggestions, please email us at info@ncbi.nlm.nih.gov accession number NC_045512.### Protein structures i.nlm.nih.gov/structure/?term=sars-cov-2.### Find e acute respiratory syndrome coronavirus 2 quences at i.nlm.nih.gov/genbank/sars-cov-2-seqs/ -START## :: Megahit v. V1.1.3 nology :: Illumina -END## ull length. on/Qualifiers ism="Severe acute respiratory syndrome coronavirus ype="genomic RNA" te="Wuhan-Hu-1" "Homo sapiens" ef="taxon:2697049" ry="China" ction_date="Dec-2019" 1555 "ORF1ab" _tag="GU280_gp01" ef="GeneID:43740578" 6..13468,13468..21555) s_tag="GU280_gp01" comal_slippage "pplab; translated by -1 ribosomal frameshift" _start=1 ct="ORF1ab polyprotein" n_id="YP_009724389.1" ef="GeneID:43740578" ation="MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQ CGLVEVEKGVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGE

ttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct tttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact ataac taattactgt cgttgacagg acacgagtaa ctcgtctatc 181 ttctgcaggc tgcttacggt ttcgtccgtg ttgcagccga tcatcagcac atctaggttt 241 cgtccgggtg tgaccgaaag gtaagatgga gagccttgtc cctggtttca acgagaaaac

Header

Sequence

Annotation

General Feature File (GFF)

##sequence-regio	on NC_045	512.2 1 29903			utov ogi	2:4-26070	240			
##species nucles	://www.nc		ахопошу/ Б	rowser/w	wtax.cgi	10=20976	949			
NC_045512.2	RefSeq	region 1	29903	•	+	•	ID=NC_	045512.	2:129903	;Dbxref=taxon:2697049;collection-date=Dec-2019;country=China;gb-acronym=S
NC_045512.2	RefSeq	five_prime_UTR	1	265		+		ID=ic	1-NC_04551	2.2:1265;gbkey=5'UTR
NC_045512.2	RefSeq	gene 266	21555		+		ID=gen	e-GU280	_gp01;Dbxı	ref=GeneID:43740578;Name=ORF1ab;gbkey=Gene;gene=ORF1ab;gene_biotype=protei
NC_045512.2	RefSeq	CDS 266	13468		+	0	ID=cds	-YP_009	724389.1;F	<pre>Parent=gene-GU280_gp01;Dbxref=Genbank:YP_009724389.1,GeneID:43740578;Name=</pre>
NC_045512.2	RefSeq	CDS 13468	21555		+	0	ID=cds	-YP_009	724389.1;	<pre>Parent=gene-GU280_gp01;Dbxref=Genbank:YP_009724389.1,GeneID:43740578;Name=</pre>
NC_045512.2	RefSeq	<pre>mature_protein_</pre>	region_of	_CDS	266	805		+		ID=id-YP_009724389.1:1180;Note=nsp1%3B produced by both pp1a and pp1a
NC_045512.2	RefSeq	<pre>mature_protein_</pre>	region_of	_CDS	806	2719		+		ID=id-YP_009724389.1:181818;Note=produced by both pp1a and pp1ab;Pare
NC_045512.2	RefSeq	mature_protein_	region_of	_CDS	2720	8554		+		ID=id-YP_009724389.1:8192763;Note=former nsp1%3B conserved domains ar
NC_045512.2	RefSeq	<pre>mature_protein_</pre>	region_of	_CDS	8555	10054		+		ID=id-YP_009724389.1:27643263;Note=nsp4B_TM%3B contains transmembrane
NC_045512.2	RefSeq	<pre>mature_protein_</pre>	region_of	_CDS	10055	10972		+		<pre>ID=id-YP_009724389.1:32643569;Note=nsp5A_3CLpro and nsp5B_3CLpro%3B n</pre>
NC_045512.2	RefSeq	mature_protein_	region_of	_CDS	10973	11842		+		ID=id-YP_009724389.1:35703859;Note=nsp6_TM%3B putative transmembrane
NC_045512.2	RefSeq	mature_protein_	region_of	_CDS	11843	12091		+		ID=id-YP_009724389.1:38603942;Note=produced by both pp1a and pp1ab;Pa
NC_045512.2	RefSeq	mature_protein_	region_of	_CDS	12092	12685		+		ID=id-YP_009724389.1:39434140;Note=produced by both pp1a and pp1ab;Pa
NC_045512.2	RefSeq	<pre>mature_protein_</pre>	region_of	_CDS	12686	13024		+		ID=id-YP_009724389.1:41414253;Note=ssRNA-binding protein%3B produced
NC_045512.2	RefSeq	mature_protein_	region_of	_CDS	13025	13441		+		ID=id-YP_009724389.1:42544392;Note=nsp10_CysHis%3B formerly known as
NC_045512.2	RefSeq	mature_protein_	region_of	_CDS	13442	13468		+		ID=id-YP_009724389.1:43935324;Note=nsp12%3B NiRAN and RdRp%3B produce
NC_045512.2	RefSeq	<pre>mature_protein_</pre>	region_of	_CDS	13468	16236		+		ID=id-YP_009724389.1:43935324;Note=nsp12%3B NiRAN and RdRp%3B produce
NC_045512.2	RefSeq	<pre>mature_protein_</pre>	region_of	_CDS	16237	18039		+		ID=id-YP_009724389.1:53255925;Note=nsp13_ZBD%2C nsp13_TB%2C and nsp_H
NC_045512.2	RefSeq	mature_protein_	region_of	_CDS	18040	19620		+		ID=id-YP_009724389.1:59266452;Note=nsp14A2_ExoN and nsp14B_NMT%3B pro
NC_045512.2	RefSeq	mature_protein_	region_of	_CDS	19621	20658		+		ID=id-YP_009724389.1:64536798;Note=nsp15-A1 and nsp15B-NendoU%3B proc
NC_045512.2	RefSeq	<pre>mature_protein_</pre>	region_of	_CDS	20659	21552		+		ID=id-YP_009724389.1:67997096;Note=nsp16_0MT%3B 2'-o-MT%3B produced b
NC_045512.2	RefSeq	CDS 266	13483		+	0	ID=cds	-YP_009	725295.1;F	<pre>Parent=gene-GU280_gp01;Dbxref=Genbank:YP_009725295.1,GeneID:43740578;Name=</pre>
NC_045512.2	RefSeq	<pre>mature_protein_</pre>	region_of	_CDS	266	805		+		ID=id-YP_009725295.1:1180;Note=nsp1%3B produced by both pp1a and pp1a
NC_045512.2	RefSeq	<pre>mature_protein_</pre>	region_of	_CDS	806	2719		+		ID=id-YP_009725295.1:181818;Note=produced by both pp1a and pp1ab;Pare

Position index	Position name	
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
3	type	The feature type name, like "gene" or "exon". In a well structured GFF file, all the children features always follow features and their relationships should be compatible with the standards released by the Sequence Ontology Pro-
4	start	Genomic start of the feature, with a 1-base offset. This is in contrast with other 0-offset half-open sequence form
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 se
6	score	Numeric value that generally indicates the confidence of the source in the annotated feature. A value of "." (a dot)
7	strand	Single character that indicates the strand of the feature; it can assume the values of "+" (positive, or 5'->3'), "-", (r
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
9	attributes	All the other information pertaining to this feature. The format, structure and content of this field is the one which

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

n coding;locus tag=GU280 gp01

=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 =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 ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=leader protein;protein_id=YP_009725297.1

ent=cds-YP_009724389.1;gbkey=Prot;product=nsp2;protein_id=YP_009725298.1 re: 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 domain 2 (TM2)%3B produced by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=nsp4;protein_id=YP_009725300.1

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 domain%3B produced by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=nsp6;protein_id=YP_009725302.1 arent=cds-YP_009724389.1;gbkey=Prot;product=nsp7;protein_id=YP_009725303.1

arent=cds-YP_009724389.1;gbkey=Prot;product=nsp8;protein_id=YP_009725304.1

d by both pp1a and pp1ab;Parent=cds-YP_009724389.1;gbkey=Prot;product=nsp9;protein_id=YP_009725305.1 s 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 ced by pp1ab only;Parent=cds-YP_009724389.1;gbkey=Prot;product=RNA-dependent RNA polymerase;protein_id=YP_009725307.1 ced by pp1ab only;Parent=cds-YP_009724389.1;gbkey=Prot;product=RNA-dependent RNA polymerase;protein_id=YP_009725307.1

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 oduced by pp1ab only;Parent=cds-YP_009724389.1;gbkey=Prot;product=3'-to-5' exonuclease;protein_id=YP_009725309.1

duced by pp1ab only;Parent=cds-YP_009724389.1;gbkey=Prot;product=endoRNAse;protein_id=YP_009725310.1 by pp1ab only;Parent=cds-YP_009724389.1;gbkey=Prot;product=2'-0-ribose methyltransferase;protein_id=YP_009725311.1

:YP_009725295.1;Note=pp1a;gbkey=CDS;gene=ORF1ab;locus_tag=GU280_gp01;product=ORF1a_polyprotein;protein_id=YP_009725295.1

ab;Parent=cds-YP_009725295.1;gbkey=Prot;product=leader protein;protein_id=YP_009742608.1

ent=cds-YP_009725295.1;gbkey=Prot;product=nsp2;protein_id=YP_009742609.1

General GFF3 structure

Description

n (like TAIR).

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 oject ៤.

nats, like BED.

equence formats, like BED.[citation needed]

is used to define a null value.

negative, or 3'->5'), "." (undetermined).

on below for a detailed explanation.

varies the most between the three competing file formats.





Annotation files can be visualized and explored

😕 🗇 🗊 Artemis Entry Edit: spi7.embl
<u>File Entries Select View Goto Edit Create Run Graph D</u> isplay
Entry: 🔽 spi7.embl
Nothing selected
STY4521 STY4523 STY4524
STY4522 STY4528
misc_feature mise
800 1600 2400 3200 4000 4800 5600 6400 7200
tRNA feature
 E # V V P G L G I E P R T R G F S I P L S K S + T P L * L F E L V A G R R T H S N K
N K W C P D S E S N H G H G D F Q S P C Q K V R H R F D F L N W + Q A G E H I R I M
, I S G A K T K N K T T D T G I F N P L V K K L D T A L T F * I G S R Q A N T F E # GAATAAGTGGTGCCCGGACTCGGAATCGAACCACGGACACGGGGATTTCAATCCCCTTGTCAAAAAGTTAGACACCGCTTTGACTTTGAATTGGTAGCAGGCAG
20 40 60 80 100 120
FLHHGSESDFWPCPSK*DGQ*FTLCRKSKKFQYCAPSCMRIF
I L P A R V R F R V V S V P I K L G R T L F N S V A K V K Q I P L L C A F V N S Y 🖛
tRNA 1 64 c possible truncated tRNA Phe.
misc_reature = 1 133562 c The major VI antigen pathogenicity island (SPI 7) CDS = 142 1176 Weakly similar to the C-terminus of several polysarcharide biosynthesis proteins e.g. Str
CDS 1173 2537 Similar to Bacteriophage Pl Ban helicase TR:080281 (EMBL:AJ011592) (453 aa) fasta scores:
misc_feature 1803 1826 PS00017 ATP/GTP-binding site motif A (P-loop)
CDS 2530 4329 no significant database hits
CDS 4498 4803 Doubtful CDS
CDS 4931 0012 NO SIGNIFICANT DATADASE NITS. CDS 5597 6154 Weekly similar to Yersinia pestis orf 77 TR-007001 (EMPL-AL001066) (102 as) facts courses
CDS 6399 7742 no significant database hits
misc feature 7744 8180 Low G+C region containing repeat region with l0xTGGT(A/-)(T/C)AAAAA(A/G)T.
CDS 8328 9107 no significant database hits. Contains a hydrophylic region in the N-terminus between resi
CDS 9218 11212 Previously sequenced Salmonella typhi topoisomerase B TopB TR:09RHF5 (EMBL:AF000001) (664
CDS 11880 12329 no significant database hits
CDS 12410 12020 UUUDITUI CDS 12641 13177 Previously sequenced Salmonella typhi single strand hinding protein SSB TR-OORHE4 (EMBL-OF

Artemis



Integrated Genome Viewer (IGV)











..................

Intro talk on NGS data formats

Bioinformatics Module 1 (~30-40 min)

Session wrap up & VM install guide

Bioinformatics Module 1

https://domman-genomics.github.io/Oman_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



