Hands-on SARS-CoV-2 genome analysis workshop

Module1
Sequencing, short reads, and cleaning

Sara Javornik Cregeen & Anil Surathu

29 March 2021

Baylor College of Medicine

ALKEK CENTER FOR METAGENOMICS AND MICROBIOME RESEARCH

SARS-CoV2 genome

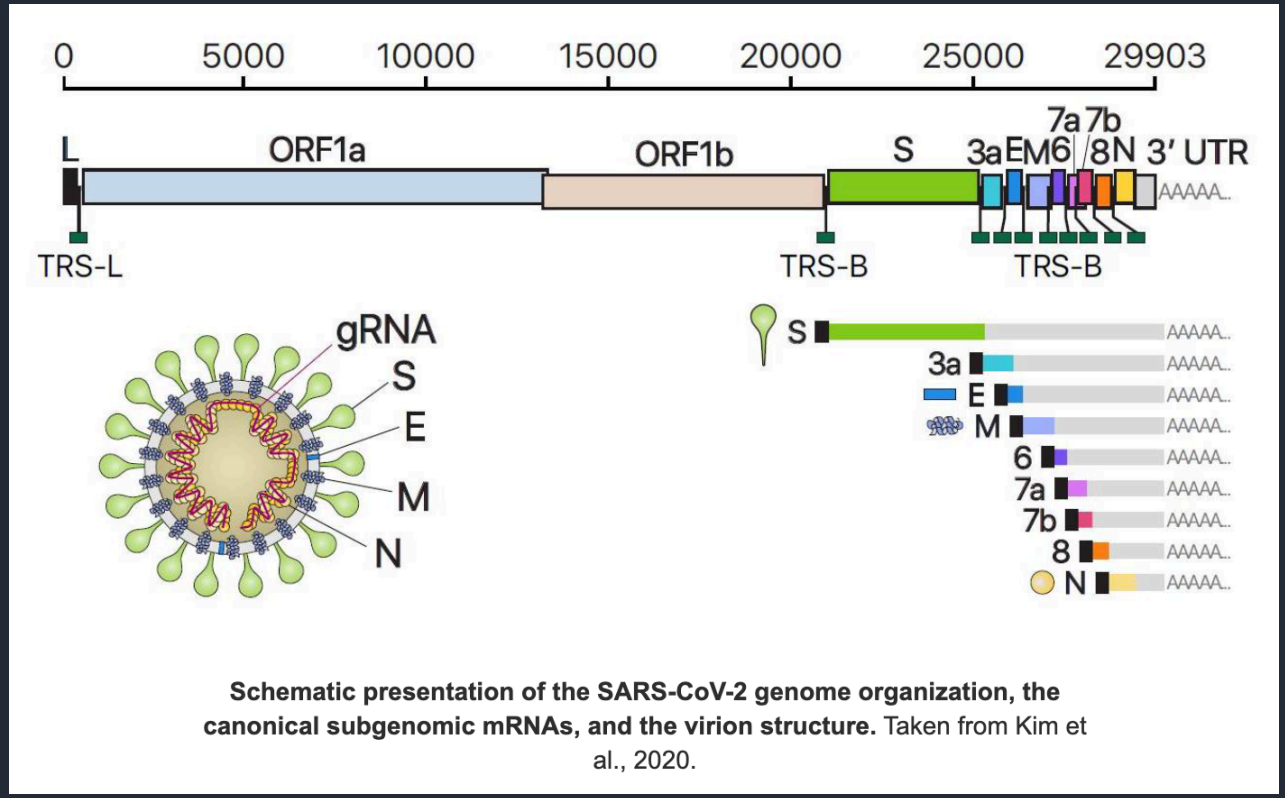
Positive-sense single-stranded RNA

Length: ~30kb

Contains:

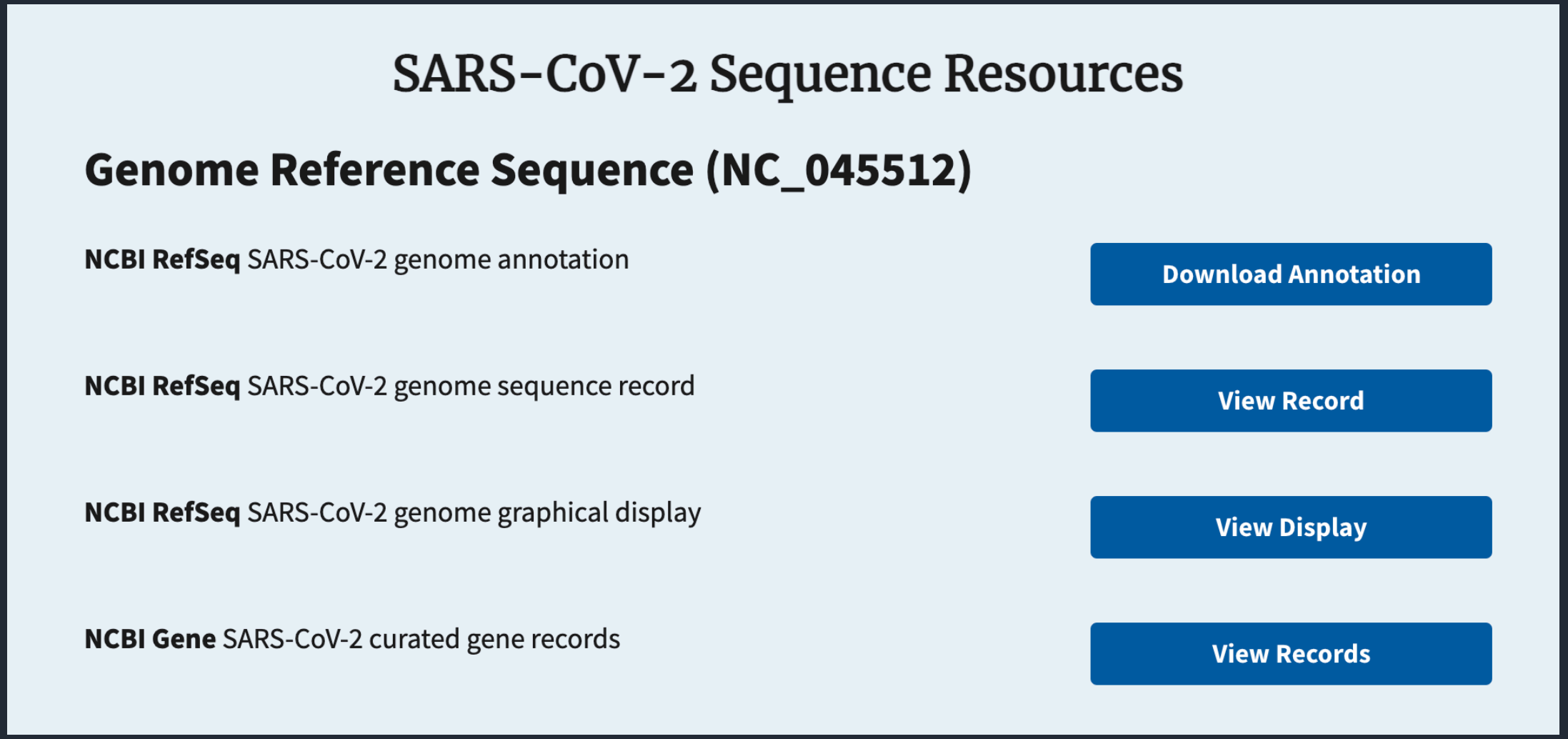
- viral structural elements (S, E, M, and N proteins)
- accessory genes (ORF 1a, 1b, 3a, 6, 7a, 7b, 8, and 10)

Of particular interest are mutations in the Spike (S) protein - enables viral infection via ACE-2 receptor recognition and membrane fusion.



Schematic presentation of the SARS-CoV-2 genome organization, the canonical subgenomic mRNAs, and the virion structure. Taken from Kim et al., 2020.

SARS-CoV2 genome



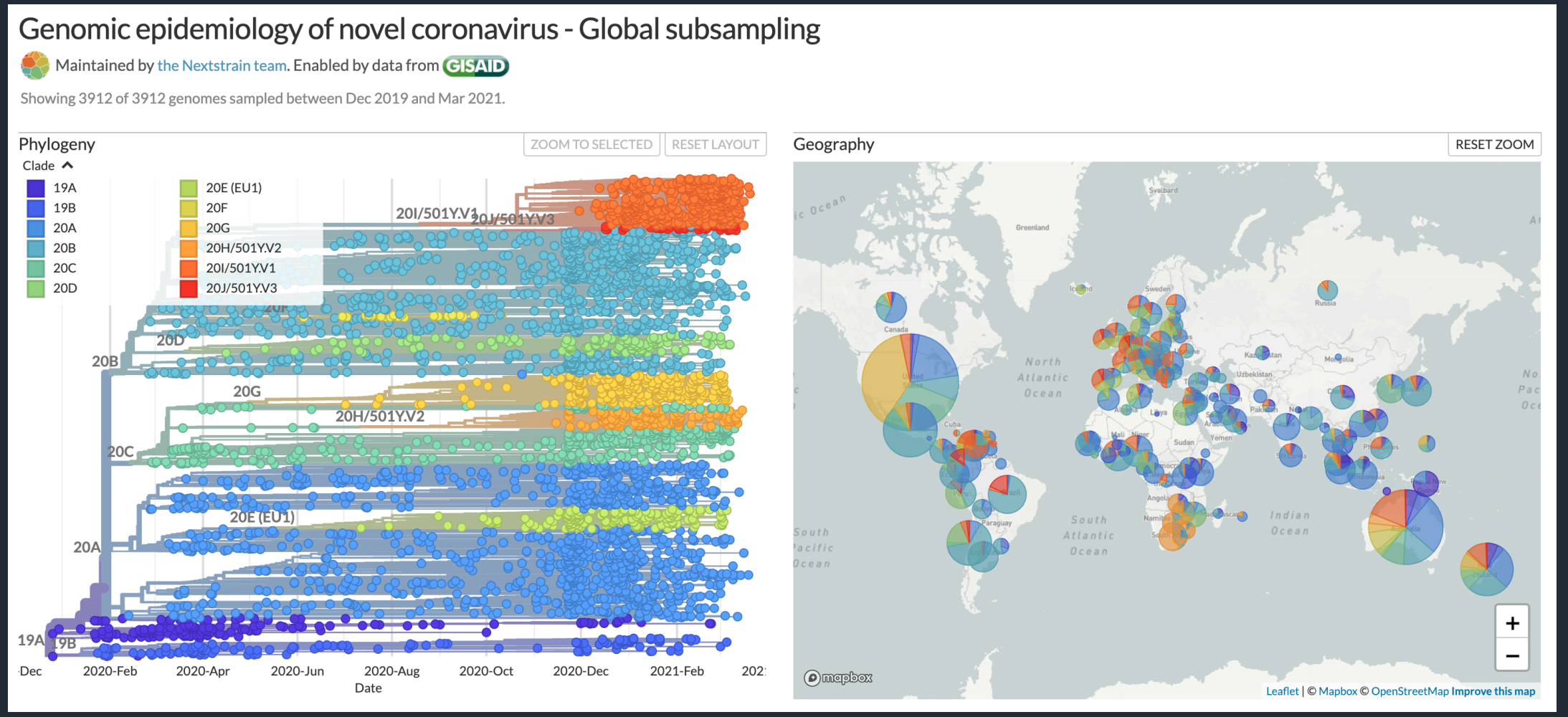
https://www.ncbi.nlm.nih.gov/sars-cov-2/

Why are we interested in SARS-CoV-2 genome sequencing?

- It is important to understand various aspects of new and emerging pathogens, such as SARS-CoV-2.
- Full-length genome sequencing gives us insight into:
  - Relatedness to other viruses
  - Rates of evolution
  - Geographical spread and adaptation to human hosts.
- When genomic information is combined with collected metadata, it can be particularly powerful in assisting in epidemiological investigations.
- The more data is collected and made publicly available the better our overall understanding can become.

Important to continuously upload genomes to public repositories

- GISAID (https://www.gisaid.org/)
- GenBank NCBI (https://www.ncbi.nlm.nih.gov/genbank/)
- Can be used in global analysis (e.g https://nextstrain.org/ncov/global) and lineage assignment (e.g. https://pangolin.cog-uk.io/).



Full genomes give us insight into the emergence of variants.

Lineage Reports

sequences classified as a particular PANGO lineage

https://outbreak.info/situation-reports

How to interpret these reports

B.1.1.7

VARIANT OF CONCERN

first identified in United Kingdom

a.k.a. **Variant of Concern 202012/01, VOC-202012/01, 20B/501Y.V1, 20I/501Y.V1**

B.1.351

VARIANT OF CONCERN

first identified in South Africa

a.k.a. **20H/501Y.V2**

B.1.427

VARIANT OF CONCERN

first identified in California

a.k.a. **CA VUI1, CAL.20C**

B.1.429

VARIANT OF CONCERN

first identified in California (United States)

a.k.a. **CA VUI1, CAL.20C**

P.1

VARIANT OF CONCERN

first identified in Brazil

a.k.a. **B.1.1.28.1, 20J/501Y.V3**

B.1.526

VARIANT OF INTEREST

first identified in New York

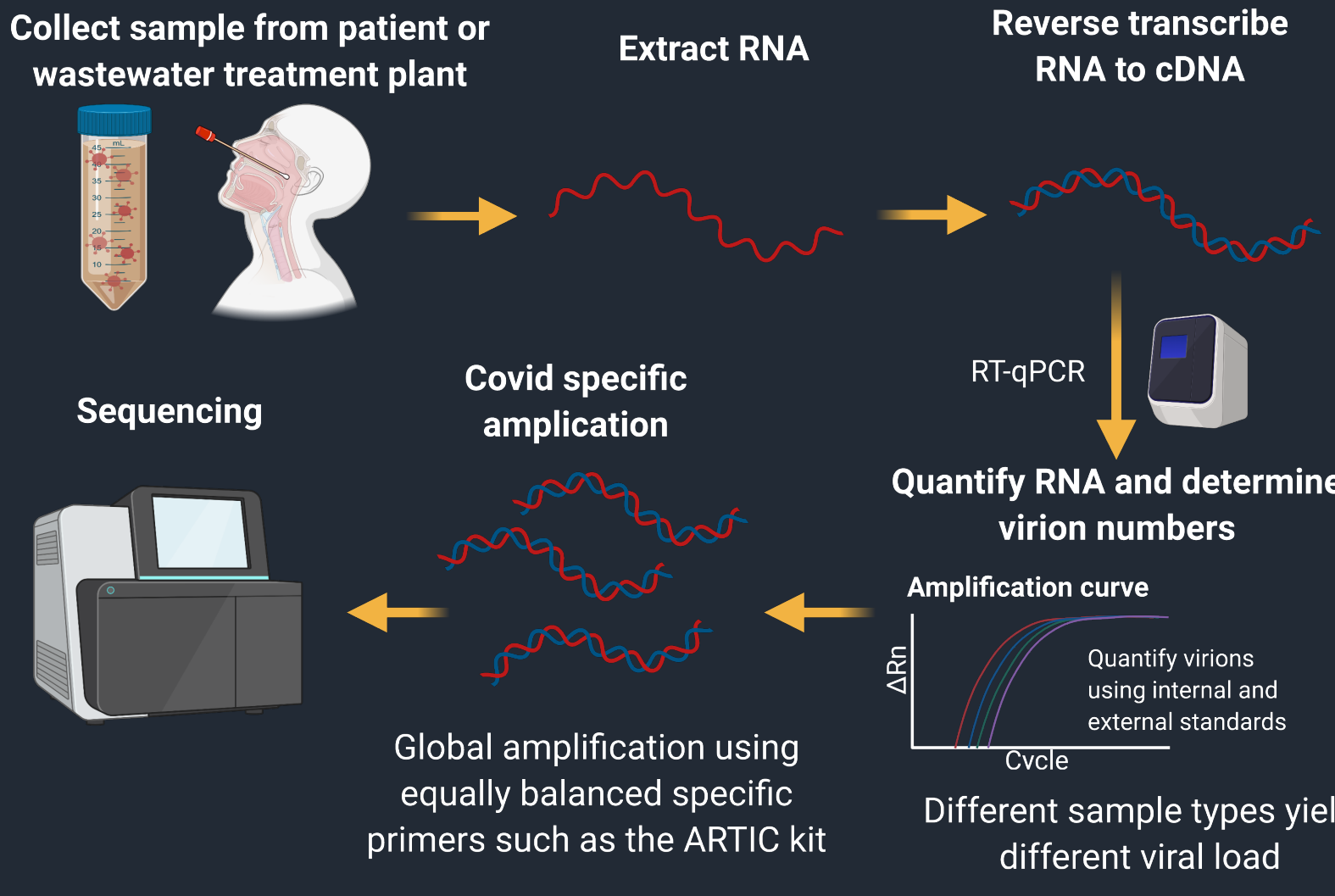
Sample collection and sequencing

Basic sample collection to sequencing workflow.

Different sample types will present different challenges for detection and amplification of full-length genomes.

Low vs. high viral load.
Presence of non target sequences.
Choice of method based on cost and scalability.

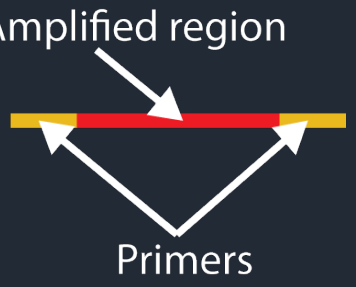
Short-read sequencing (Illumina) is usually 2x250 bp or 2x300 bp.



Created with BioRender.com

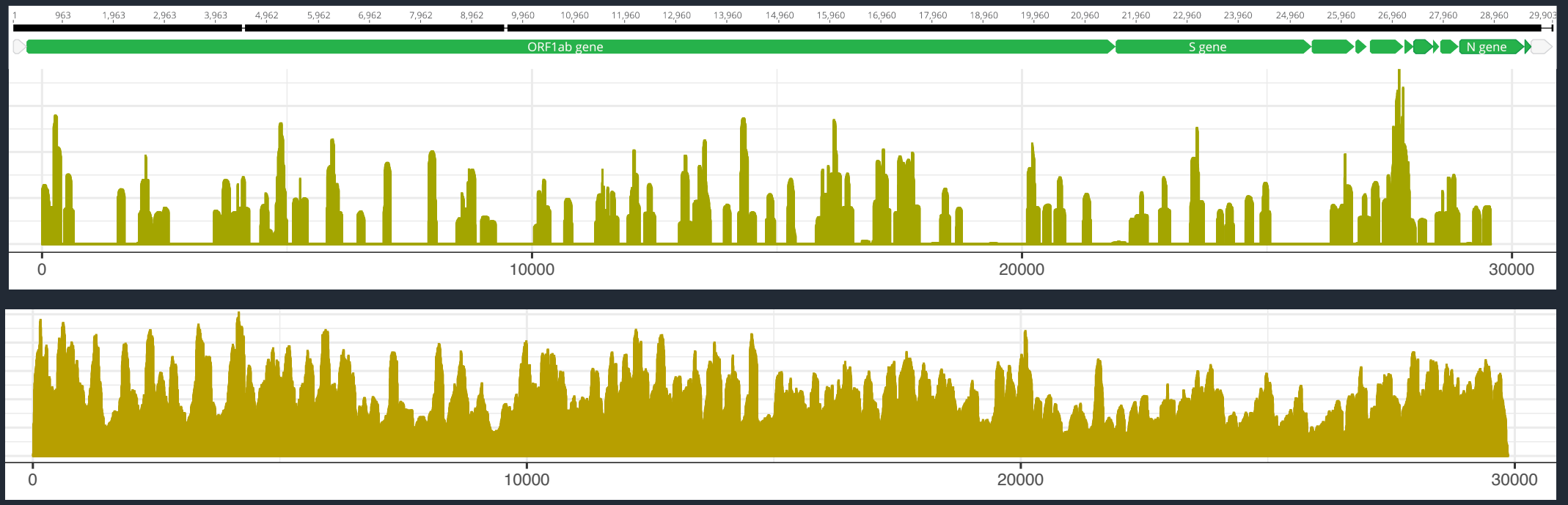
Targeted amplification-based approaches

Amplicon:



Tiled Amplification

Depth of sequencing



Targeted amplification-based approaches

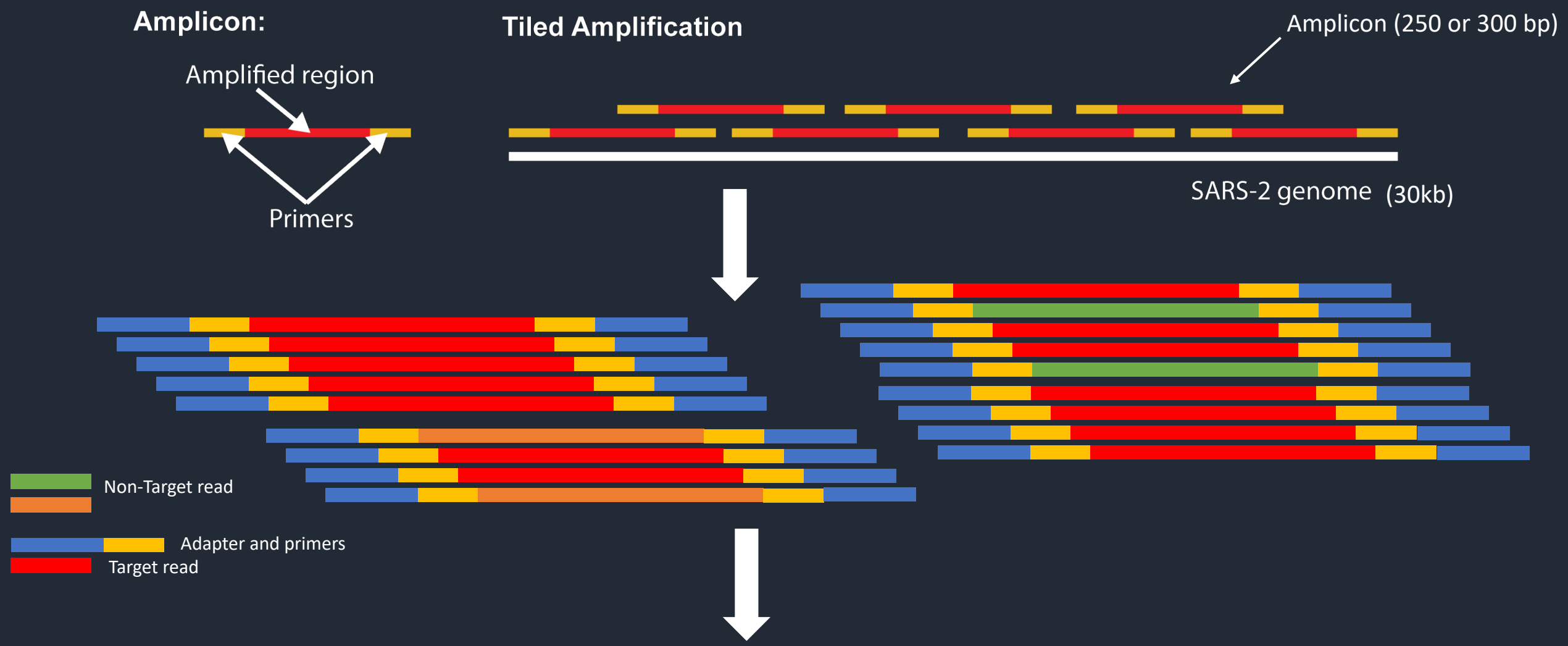
Aims of a good sequencing protocol:

- High specificity to target
- Ensures evenness of coverage along the entire genome length
- Few amplicon dropouts and little amplification bias
- Scalability to large sample size
- More recently: does it capture emerging variants.

Some available protocols:

- ARTIC V3 (https://artic.network/ncov-2019)
- SWIFT (https://swiftbiosci.com/swift-amplicon-sars-cov-2-panel/)
- TWIST (https://www.twistbioscience.com/products/ngs/fixed-panels/sars-cov-2-research-panel)

Sequencing completed, what are we working with?



Bioinformatics pipelines

Sequence quality control

- We want to ensure that only good quality reads are included in downstream analysis.
- All sequencing platforms and protocols require addition of non-biological sequences, such as sequencing adapters and primers, that will need to be removed.
- No sequencing is perfect. All wet lab preparation steps introduce their own biases and different sequencing platforms are prone to specific types of error.

Sequence quality control

- Most of these artefacts and errors can be accounted for and minimized during the bioinformatics processing steps.
- A good way of evaluating the quality and reproducibility of your sequencing efforts is to include controls in your experiment.
- Check that our sequences represent what we intended to sequence – contamination.

Sequence quality control

- We want to ensure that only good quality reads are included in downstream analysis.
- All sequencing platforms and protocols require addition of non-biological sequences, such as sequencing adapters and primers, that will need to be removed.
- No sequencing is perfect. All wet lab preparation steps introduce their own biases and different sequencing platforms are prone to specific types of error.
- These can be accounted for and minimized during the bioinformatics processing steps.
- A good way of evaluating the quality and reproducibility of your sequencing efforts is to include controls in your experiment.

Bad data in, bad data out!

FASTQ file format

Text-based format for storing both a **biological sequence** and its corresponding **quality scor**es.

Always four lines per sequence read:

@Label	* *Sequence identifier (starts with @)*
TATCGGAAGAGCACACGTCTGAACTCCAGTCACGCAA	* Sequence (ATGC sequence)
+	* A separator (usually +)
--ABCC@;@9<E9DCFCFFGGGGCEGGGFGGFFGGE7	* Base call quality scores (Phred +33 encoded)

FASTQ file format

Text-based format for storing both a **biological sequence** and its corresponding **quality scor**es.

Always four lines per sequence read:

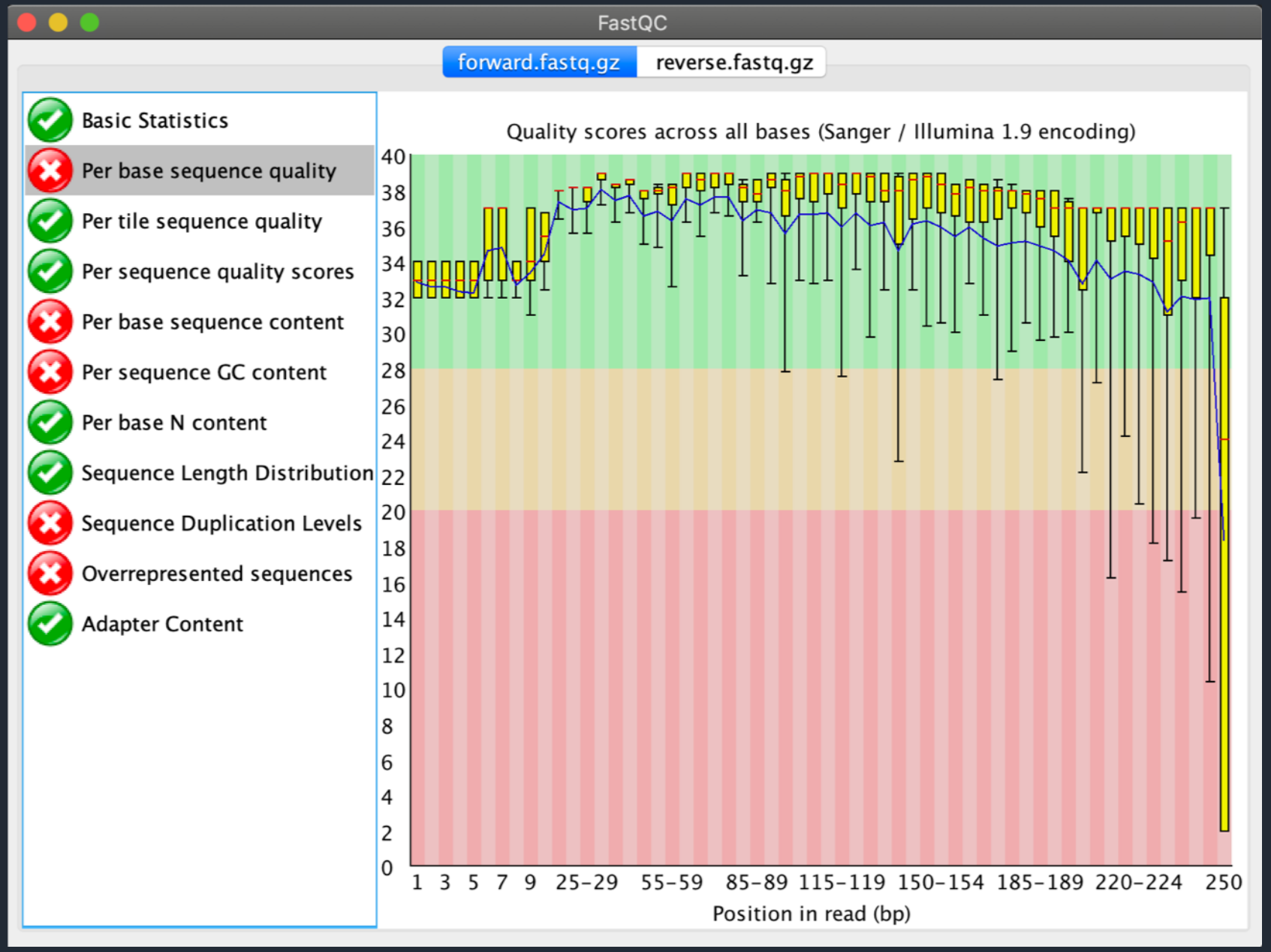
@Label	* *Sequence identifier (starts with @)*
TATCGGAAGAGCACACGTCTGAACTCCAGTCACGCAA	* Sequence (ATGC sequence)
+	* A separator (usually +)
--ABCC@;@9<E9DCFCFFGGGGCEGGGFGGFFGGE7	* Base call quality scores (Phred +33 encoded)

Phred score is the probability that the called base is incorrect. The symbols in the fastq translate to ASCII characters representing quality scores. Higher Q-score is better.

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

FASTQC – a sequence quality visualization tool



Basic sequence pre-processing steps

There are many different tools available to complete these QC steps.

In the Module1 practical we will be doing the following:

- Perform quality filtering and adapter trimming using bbduk.
- Remove human reads from trimmed fastqs using bbmap.
- Remove ARTIC amplicon primers from processed reads using iVar.

Basic sequence pre-processing steps

There are many different tools available to complete these QC steps.

In the Module1 practical we will be doing the following:

- Perform quality filtering and adapter trimming using bbduk.
- Remove human reads from trimmed fastqs using bbmap.
- Remove ARTIC amplicon primers from processed reads using iVar.

Optional step (not covered today):

- Using a metagenomic classifier to profile your data.
- Kraken2 (http://ccb.jhu.edu/software/kraken2/)
- Centrifuge (https://ccb.jhu.edu/software/centrifuge/manual.shtml#what-is-centrifuge)

Basic sequence pre-processing steps

There are many tools available to complete these QC steps.

In the Module1 practical we will be doing the following:

- Perform quality filtering and adapter trimming using bbduk.
- Remove human reads from trimmed fastqs using bbmap.
- Remove ARTIC amplicon primers from processed reads using iVar.

Links and more detailed descriptions of all these steps is available in the Module1 walkthrough that you will be following in the practical session.

Workshop dataset

SRX8941978: ARTIC Illumina Sequencing of SARS-CoV-2
1 ILLUMINA (Illumina MiSeq) run: 265,305 spots, 153.7M bases, 92.2Mb downloads

Design: ARTIC v3 Protocol Sequencing of RNA extracted from NP Swab samples, following typical protocol with amplicons fed into NEB Next Library prep and sequenced on MiSeq 600v3 cartridge

Submitted by: JHU SARS-CoV-2 Genome Sequencing (JHU_SARS-CoV-2)

Study: Johns Hopkins Viral Genomics of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
PRJNA650037 • SRP277377 • All experiments • All runs
show Abstract

Sample:
SAMN15691633 • SRS7197758 • All experiments • All runs
Organism: Severe acute respiratory syndrome coronavirus 2

Library:
Name: MDHP-00059_Illumina
Instrument: Illumina MiSeq
Strategy: AMPLICON
Source: VIRAL RNA
Selection: PCR
Layout: PAIRED

Runs: 1 run, 265,305 spots, 153.7M bases, 92.2Mb

Run	# of Spots	# of Bases	Size	Published
SRR12447392	265,305	153.7M	92.2Mb	2020-08-13

Questions?