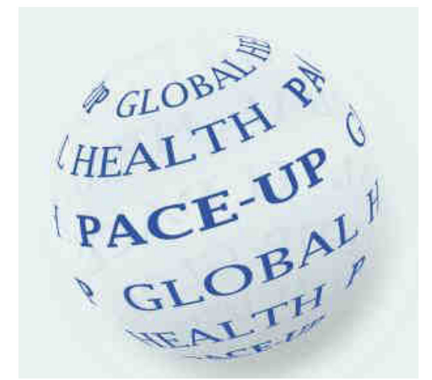


Workshop
Molecular Diagnostic Techniques
(Genomics and Bioinformatics – Infectious diseases)
Module-1

Dr. Srinivas Reddy Pallerla, Mr. Tung Tran and Ms. Linh Le
Prof. Dr. Thirumalaisamy P. Velavan, University Hospital, Tübingen, Germany
Prof. Dr. Daniel Olusola Ojurongbe, Nigeria
Prof. Dr Mohamed Osman, University of Khartoum, Sudan

25 July 2022

Sponsor: *Alexander von Humboldt Foundation*

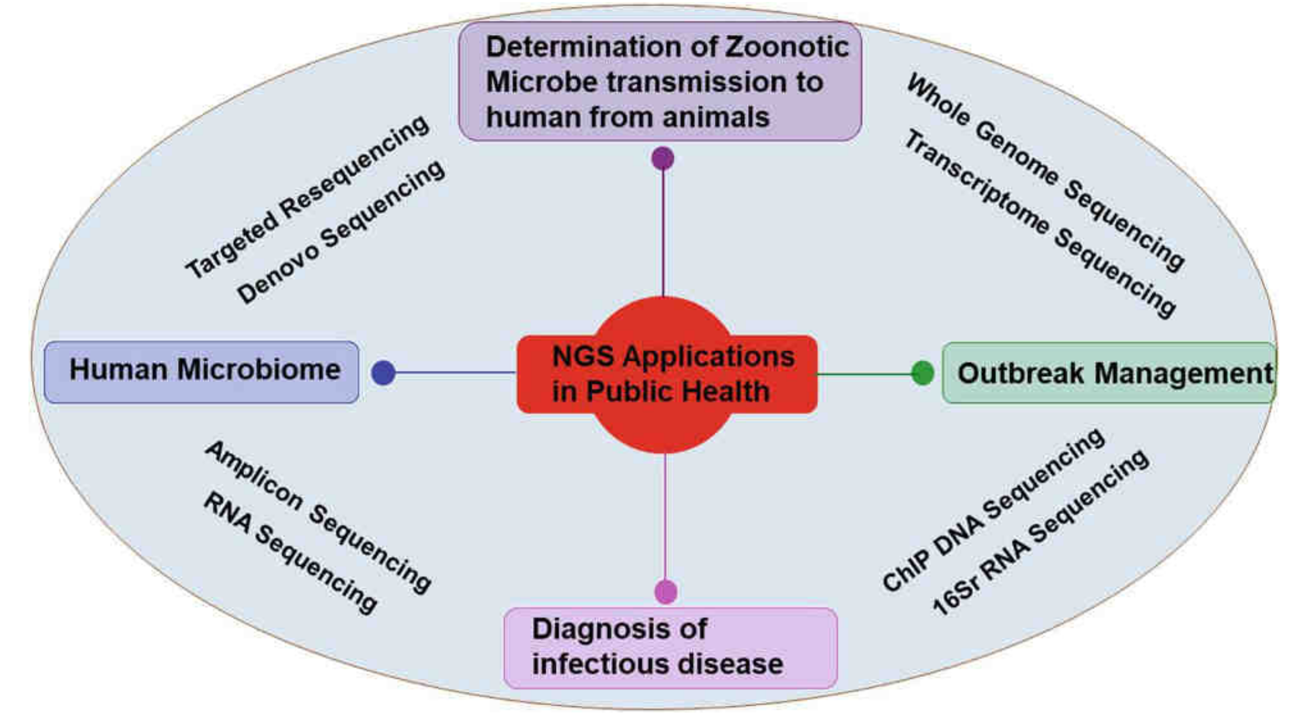
Training schedule

Time/Day	Monday 25th	Tuesday 26th	Wednesday 27th	Thursday 28th	Friday 29th
9-10 am	Overview of CERID Workshop rules – Prof Ojurongbe Objectives and expectations; Workshop Pre-test Prof Olowe	**Wet Lab** Intro. to Laboratory equipment Pipetting skills DNA extraction DBS – Plasmodium Crude - Bacteria isolate – Dr. Ayoola	RNA Extraction Nasal Swab for Sars-Cov-2 detection Tung/Dolapo/Salma	Oxford Nanopore Seq MODULE -3 Demonstration of Nanopore Seq Barcoding, Flow cell loading, Sequencing Flow cell washing Tung/Dolapo/Salma	Bioinformatics: MODULE -4 Nanopore data analysis Linux (Ubuntu) OS. Linux commands Open terminal and practice commands Dr. Pallerla/Tung
10-11 am	Principle and Types of PCR - Prof Opaleye				
11-11.15 am	**Coffee/Tea Break**				
11.15-11 am	Real-Time PCR Principles and Types – Prof Ojurongbe	**Hands on PCR** Master Mix prep PCR amplification Gel electrophoresis -Drs. Funwei / Ayoola	▪ RT-cDNA Synthesis Real-time PCR Sars-Cov-2 and Monkey Pox detection Tung/Dolapo/Salma	Oxford Nanopore Seq Demonstration of Nanopore Seq Barcoding, Flow cell loading, Sequencing Flow cell washing Tung/Dolapo/Salma	Bioinformatics: Nanopore data analysis Linux (Ubuntu) OS. Linux commands Open terminal and practice commands Dr. Pallerla/Tung
11-12 pm					
12-1 pm	Application of RT-PCR COVID-19 and other pathogen detection Dr. Pallerla				
1-2 pm	**LUNCH BREAK**				
2-3 pm	Introduction to NGS and Bioinformatics – Dr. Pallerla MODULE -1	**Hands on PCR** Master Mix prep PCR amplification Gel electrophoresis - Drs. Funwei / Ayoola	▪ RT-cDNA Synthesis Real-time PCR Sars-Cov-2 and Monkey Pox detection Tung/Dolapo/Salma	Oxford Nanopore Seq Demonstration of Nanopore Seq Barcoding, Flow cell loading, Sequencing Flow cell washing Tung/Dolapo/Salma	Other Bioinformatic tools MODULE - 5 MEGA Phylogenetic Analysis Dr. Ayoola/Tung
3-4 pm					
4-4.15 pm	**Coffee/Tea Break**				
4.15-5 pm	Introduction to Nanopore Sequencing - Dr. Pallerla MODULE -2	**Hands on PCR** Master Mix prep PCR amplification Gel electrophoresis Drs. Funwei / Ayoola	▪ RT-cDNA Synthesis Real-time PCR Sars-Cov-2 and Monkey Pox detection Tung/Dolapo/Salma	Oxford Nanopore Seq Demonstration of Nanopore Seq Barcoding, Flow cell loading, Sequencing Flow cell washing Tung/Dolapo/Salma	Other Bioinformatic tools MEGA Phylogenetic Analysis Dr. Ayoola/Tung
5-5.45pm					Workshop Post test

Learning objectives of this modul

- Introduction NGS and bioinformatics
- First, second and third generation sequencing technologies
- Genomics and bioinformatics
- Genomics in Infectious disease research and public health
- File systems, terminology and softwares/repositories
- Oxford Nanopore Sequencing and data analysis
- Programming languages used commonly (Python, R and Bash)
- Linux (Ubuntu) operating system
- Linux commands usage: Open terminal and practice to use these commands

Genomics in Infectious disease research and public health



Diagnostics,

monitoring for viral evolution,

transmissibility and virulence,

transmission dynamics,

antibiotic resistance,

nosocomial outbreaks,

vaccine development, and

the overall assessment of infection prevention and control measures

COVID-19 pandemic has highlighted the need to incorporate pathogen genomics for enhanced disease surveillance and outbreak management.

First, second and third generation sequencing technologies

First Generation

Second Generation
(Next Generation Sequencing)

Third Generation

Sanger Sequencing
Maxam and Gilbert
Sanger Chain-termination

- Infer nucleotide identity using dNTPs then visualize with electrophoresis
- 500-1000 bp fragments

454, Solexa,
Ion Torrent
Illumina

- High throughput from the parallelization of sequencing reactions
- ~50-500 bp fragments

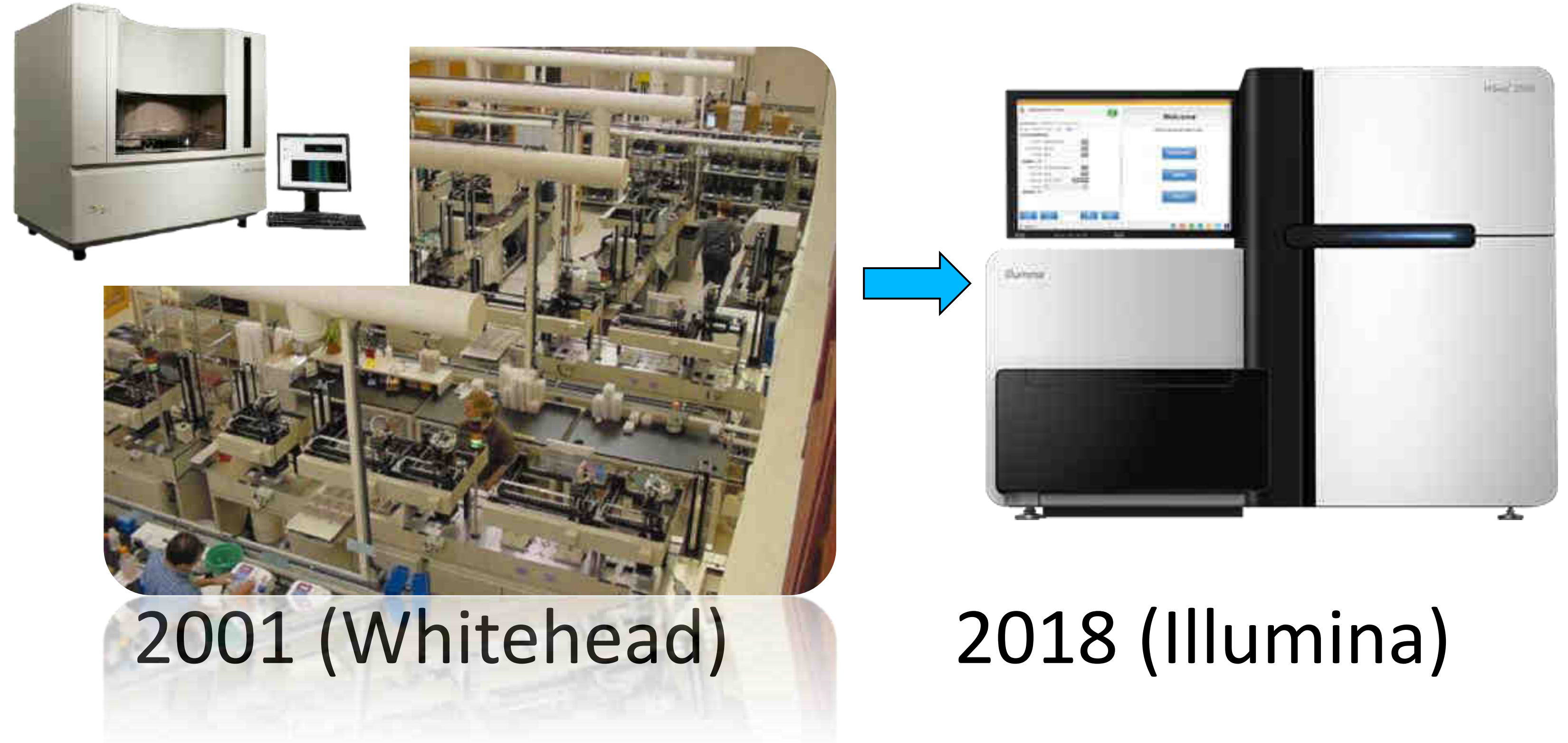
PacBio
Oxford Nanopore

- Sequence native DNA in real time with single-molecule resolution
- Tens of kb fragments, on average

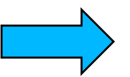
Short-read sequencing (First and Second Generation)

Long-read sequencing (Third Generation)

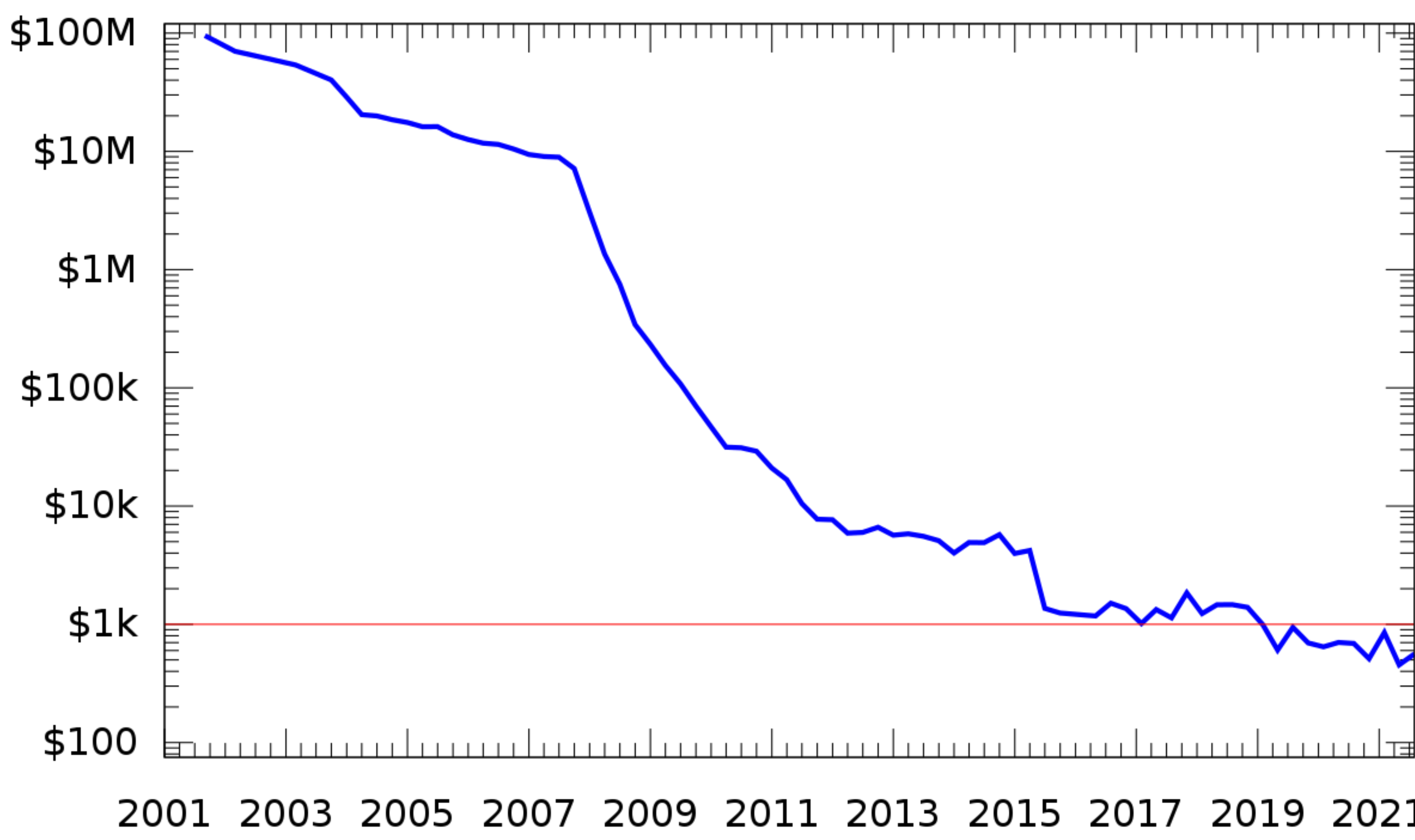
Advances in Sequencing technologies



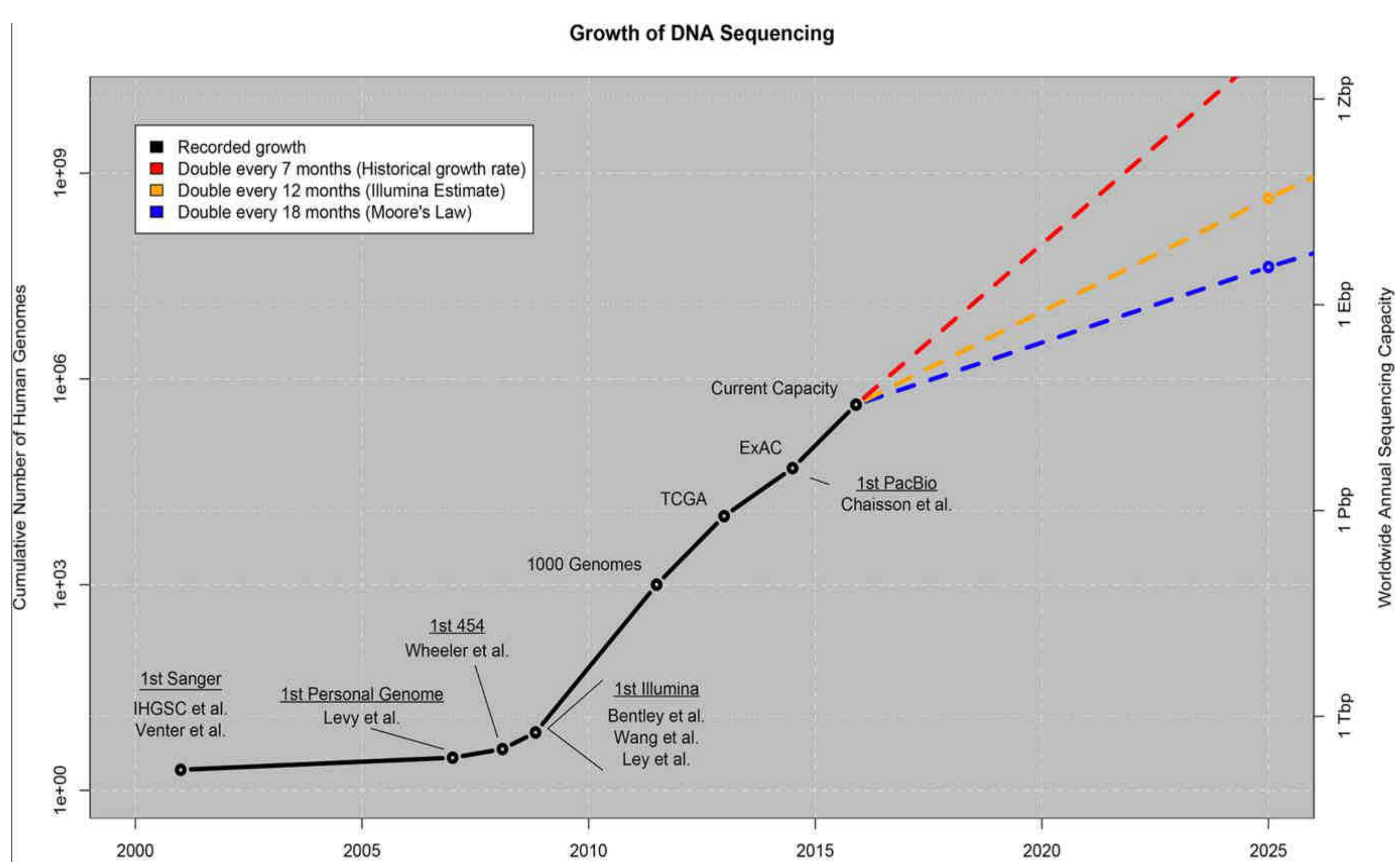
13 Years & 3.8 billion $



<12 hours & 1000$



https://en.wikipedia.org/wiki/National_Human_Genome_Research_Institute



https://goo.gl/3RhFkH

Genomics, Big Data, and Medicine Seminar Series -- Eric D. Green
The Largest Current Bottleneck in Genomics…

Advantages and disadvantages for short vs long read sequencing

Short-read sequencing

- Higher sequence fidelity
- Cheap
- Can sequence fragmented DNA
- Not able to resolve structural variants or distinguish highly homologous genomic regions
- Unable to provide coverage of some repetitive regions

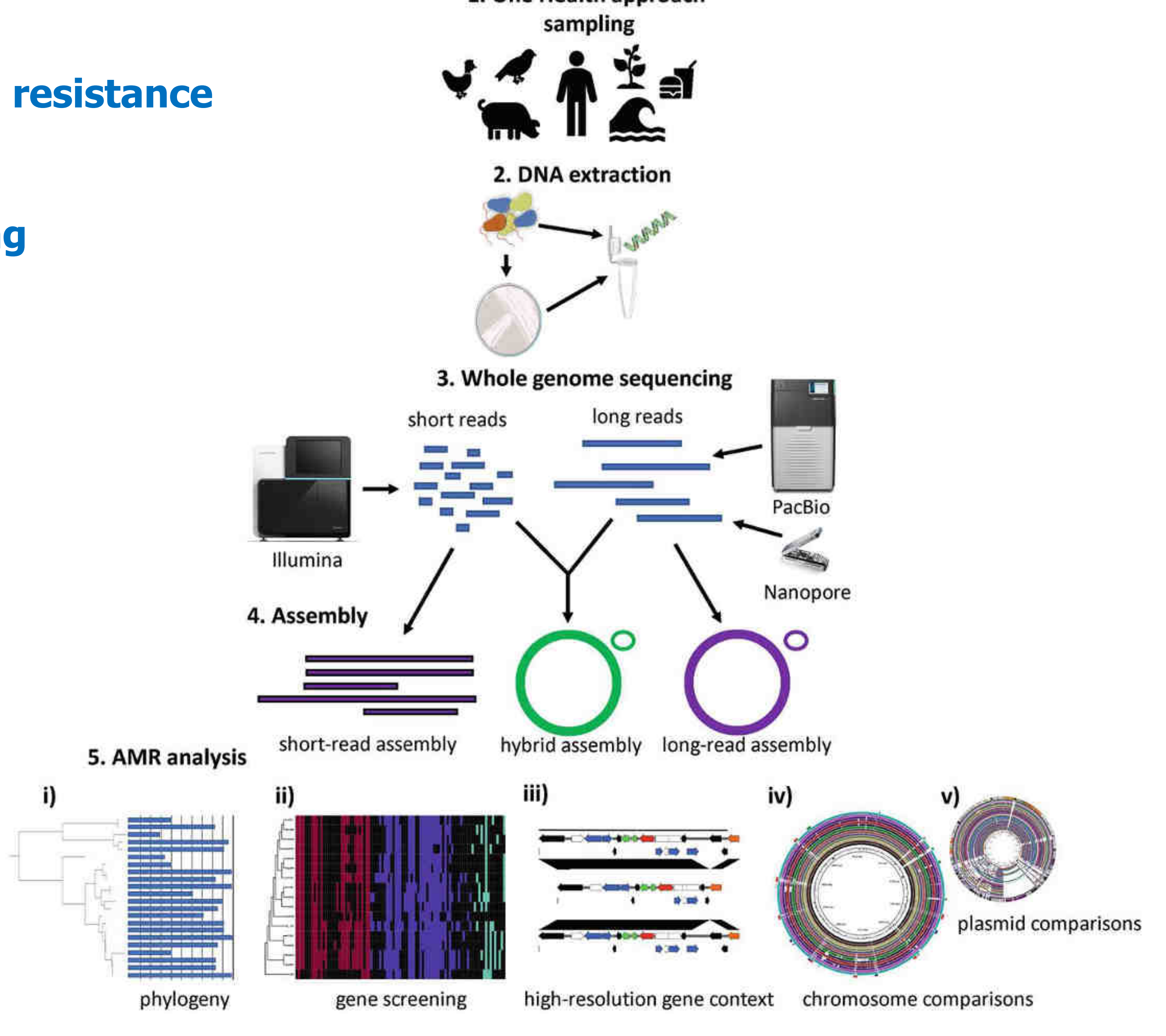
Long-read sequencing

- Able to sequence genetic regions that are difficult to characterize with short-read seq due to repeat sequences
- Able to resolve structural rearrangements or homologous regions
- Able to read through an entire RNA transcript to determine the specific isoform
- Assists de novo genome assembly
- Lower per read accuracy
- Bioinformatic challenges, caused by coverage biases, high error rates in base allocation, scalability and limited availability of appropriate pipelines

Hybrid Bioinformatic assembly: Data generated from the same sample using long and short read sequencers

Whole genome sequencing – Bacterial genomes - Antimicrobial resistance (AMR)

Both short vs long read sequencing



https://link.springer.com/chapter/10.1007/698_2020_626

Various steps showing SARS-CoV-2 sequencing using NGS platforms

https://www.beckman.de/liquid-handlers/genomic-automated-workstations

Oxford Nanopore sequencing technology

Research Areas:

- Microbiology
- Microbiome
- Environmental
- Animal
- Infectious disease
- Human genomics
- Clinical research
- Cancer
- Transcriptome
- Populations genomics

Applications:
Nanopore sequencing offers advantages in all areas of research; includes **DNA**, **RNA** sequencing and gene expression.

https://www.youtube.com/c/OxfordNanoporeTechnologies/playlists

Key considerations - library preparation and sequencing platform

(a) Research question ? being asked

(b) Sample type

(c) Short-read or long-read sequencing

(d) DNA or RNA sequencing – do you need to look at the genome or transcriptome?

(e) Is the whole genome required or only specific regions?

(f) Read depth (coverage) needed – experiment-specific

(g) Extraction method

(h) Sample concentration

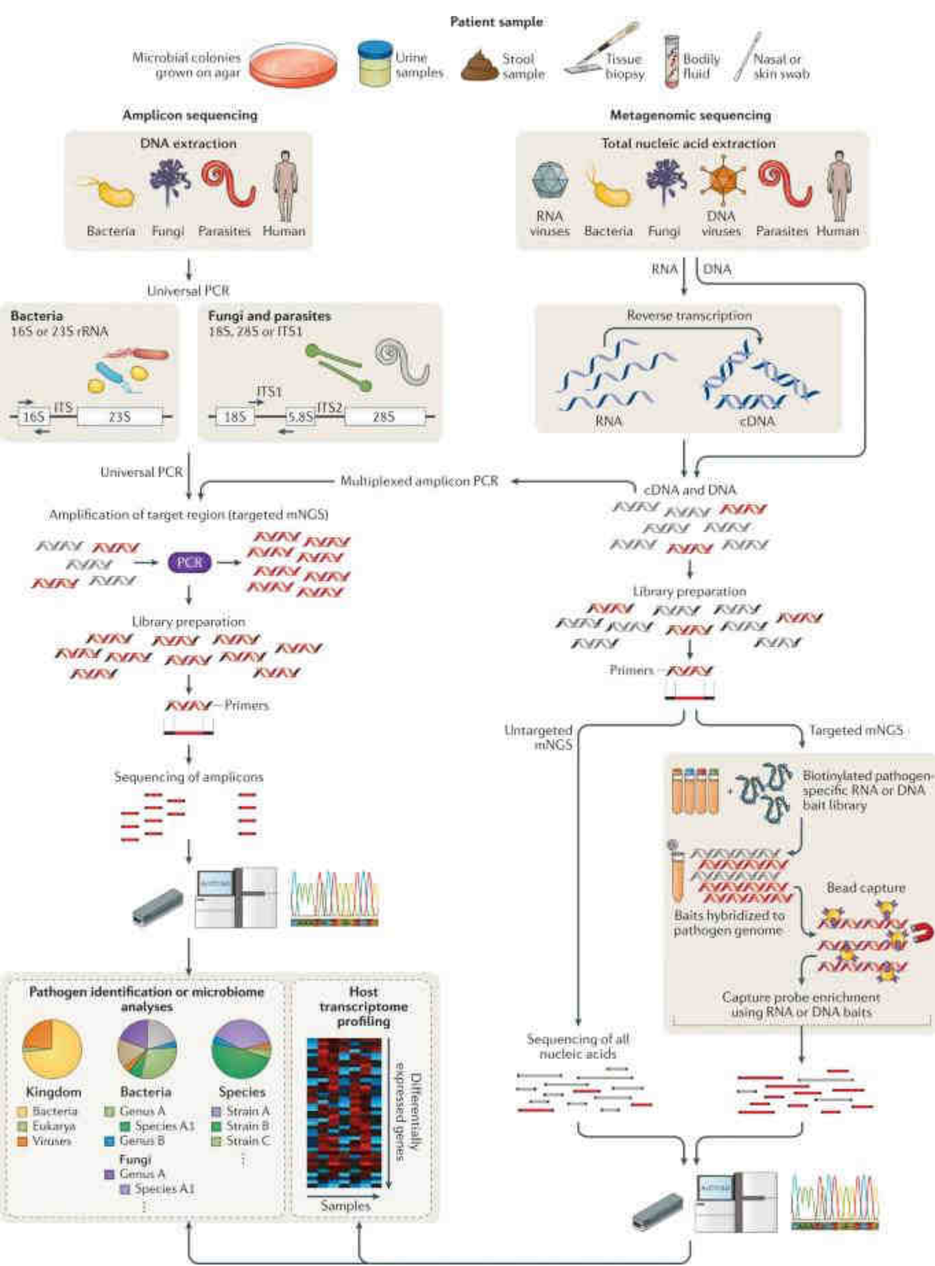
(i) Single end, paired end or mate pair reads

(j) Specific read length required

(K) Could samples be multiplexed ?

(l) Bioinformatic tools – experiment dependent. Depending on the sample and the biological question, the entire process of sequence analysis can be adapted

Infectious Disease Diagnostics



Sequencing Methods

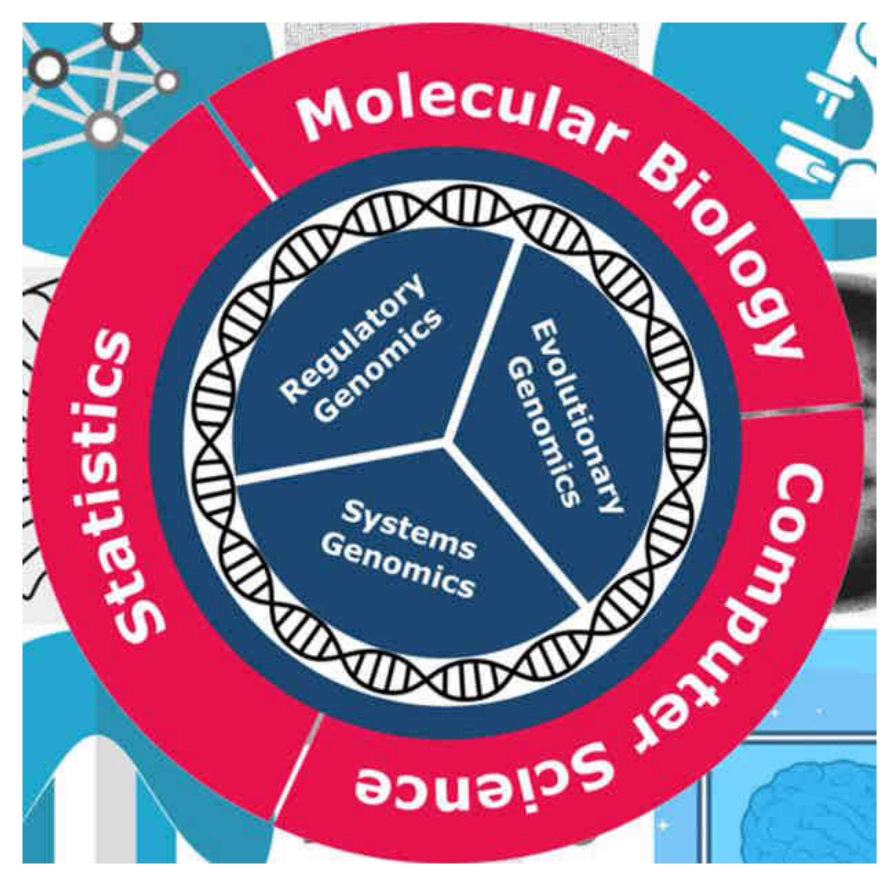
Whole genome sequencing
Targeted sequencing
Metagenomics
Rna sequencing
Methylation sequencing

Chiu and Miller. Clinical metagenomics. Nat Rev Genet. 2019.

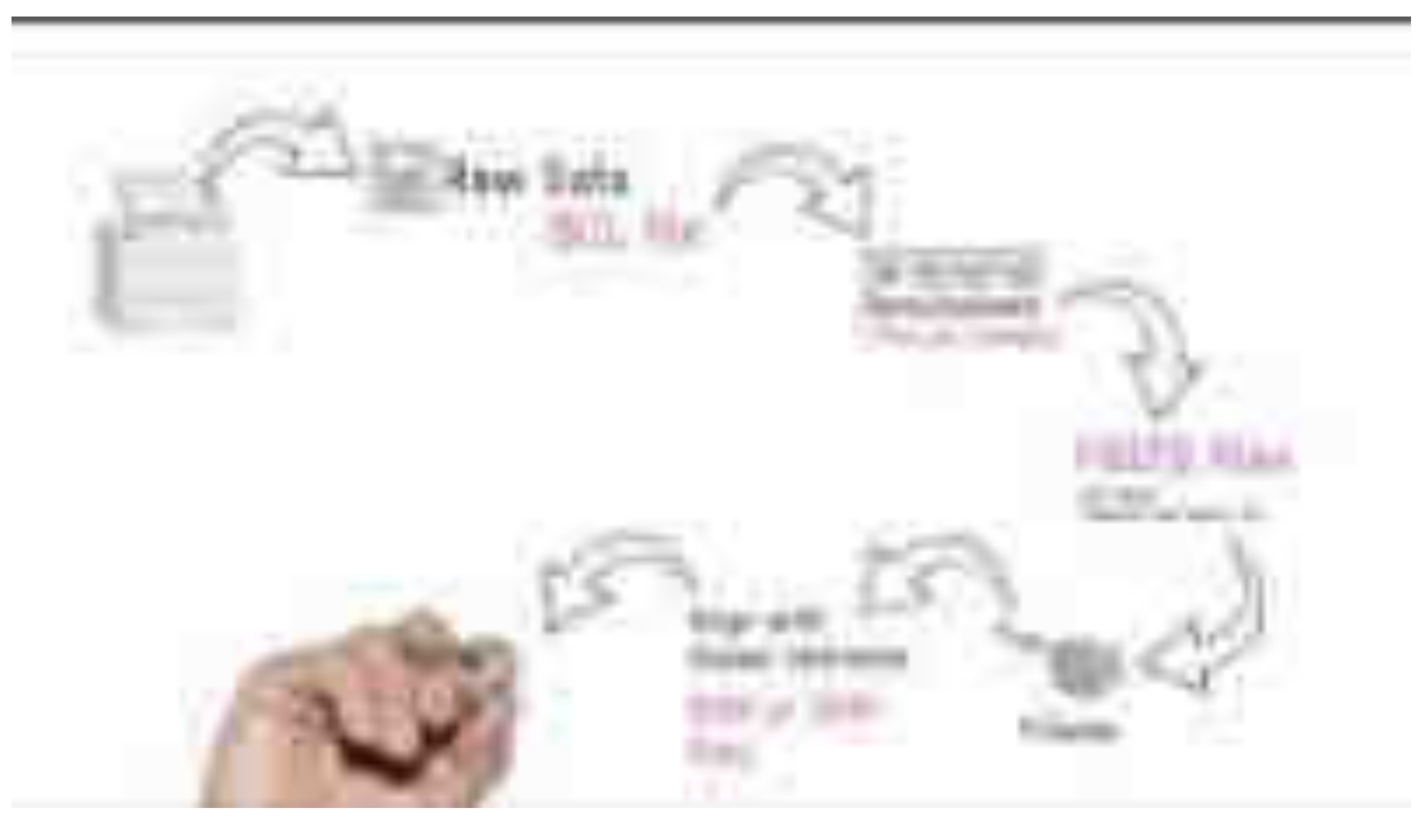
Bioinformatics

- Bioinformatics and sequencing technologies have flourished in recent years.

- Bioinformatic deals with, how to mine, interpret, as well as utilize sequencing and experimental data, and convert them into useful information.



NGS data analysis, Pipelines and Bioinformatics



https://www.youtube.com/watch?v=RkttaYc8hfw

2021 Canadian Bioinformatics Workshop (CBW) series: Bioinformatics DotCa
https://www.youtube.com/channel/UCKbkfKk65PZyRCzUwXOJung/featured

Any one can become a Bioinformatician

Category 1: "Biologists that use tools“- Biologists view the details of quantitative tools as largely unimportant. Rather the only thing that is important is 'biological questions’. Bioinformatics is about using computational approaches to study biological questions that would be impossible or infeasible to approach with other methods.

Category 2: Professions who develops methods and software tools for understanding biological data, in particular when the data sets are large and complex.

https://www.youtube.com/c/SimonCockell/videos

Any one can become a Bioinformatician

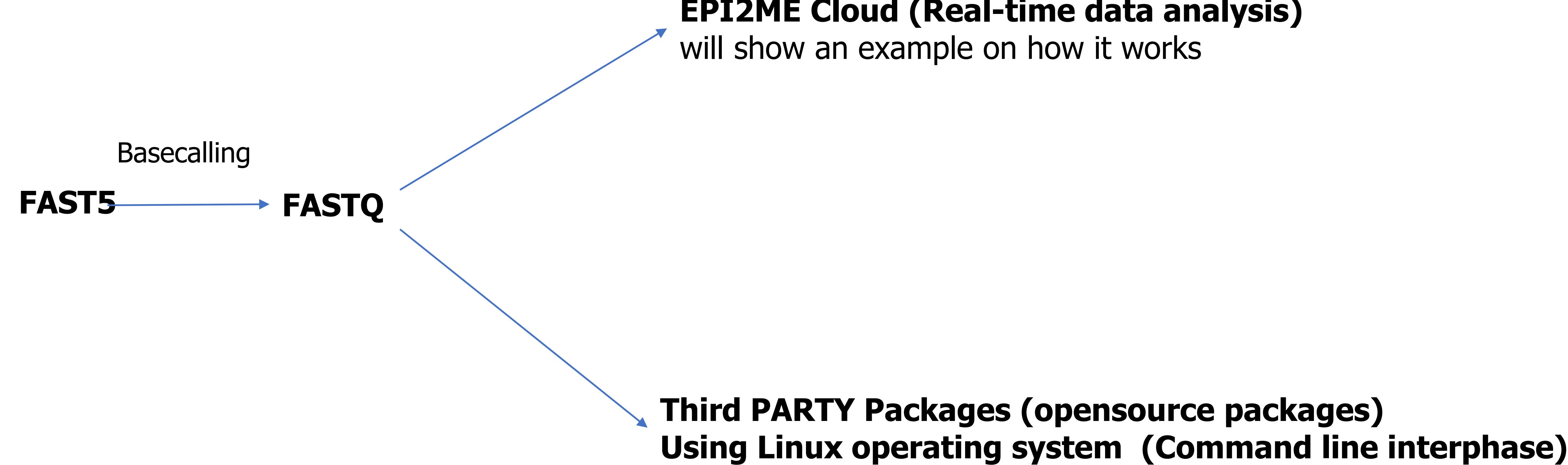
- Most bioinformatics (and most experts) are self-taught, and you can do the same – learn from online resources (Google and Youtube) (you cannot believe where you will reach….)

- Require curiosity and an interest in bioinformatics

Oxford Nanopore Sequencing and data analysis

- Oxford Nanopore produces FAST5 (HDF5) files and/or FASTQ files.
- FAST5 files contain raw signal data that can be used for basecalling.
- FASTQ format that contain nucleotide sequence data (obtained from FAST5 Raw data)

Data Analysis



Common file systems, terminology and softwares/repositories

File Systems

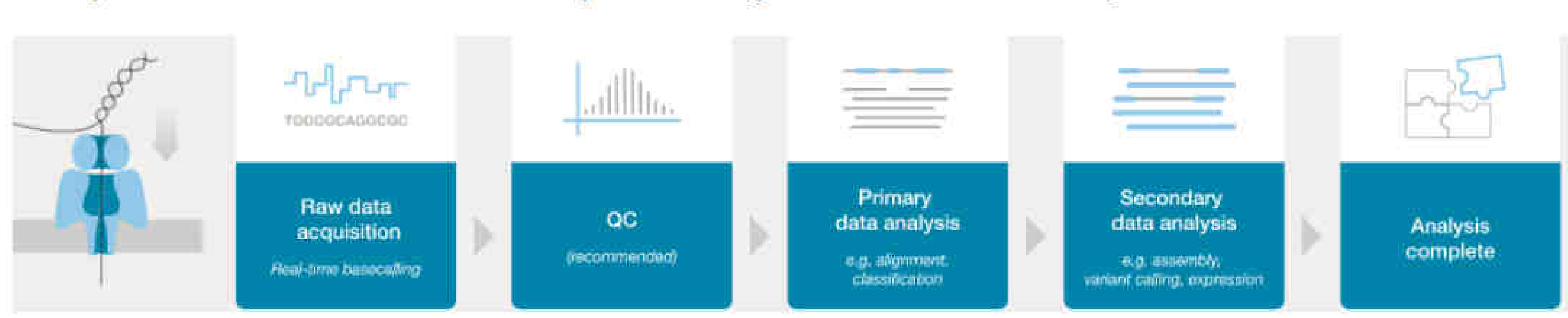
FAST5 (Nanopore)
FASTQ
FASTA
BAM/SAM
GTF/GFF/BED
VCF

Terminology

Pipeline
MultiQC
Reference sequence
Consensus sequence
De novo assembly
Alignment to reference

Softwares/repositories

Github
Conda (Anaconda, Miniconda, Bioconda)
Medaka
Minimap2



File Systems

FASTQ: Sequence (Fasta) with quality score

Identifier	`@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1`
Sequence	`TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGA`
+ sign & identifier	`+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1`
Quality scores	``efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd` ``

Base T
phred Quality] = 29

FASTA: Text-based format for representing nucleotide sequences

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

File Systems

FAST5 (Nanopore) : Raw reads from Oxford Nanopore sequencer (only for Nanopore)

SAM/BAM : Sequence Alignment Map (SAM) is a text-based format originally for storing biological sequences aligned to a reference sequence; The binary equivalent of a SAM file is a Binary Alignment Map (BAM) file

GTF/GFF : General Feature Format (GFF) file is a simple tab-delimited text file for describing genomic features.

VCF : Variant Call Format is text file used in storing gene sequence variations

Terminology

Pipeline : Composed of a wide array of software algorithms to process raw sequencing data and generate a consensus and list of annotated sequence variants

Reference sequences: The Reference Sequence (RefSeq) database is an open access, annotated and curated collection of publicly available nucleotide sequences

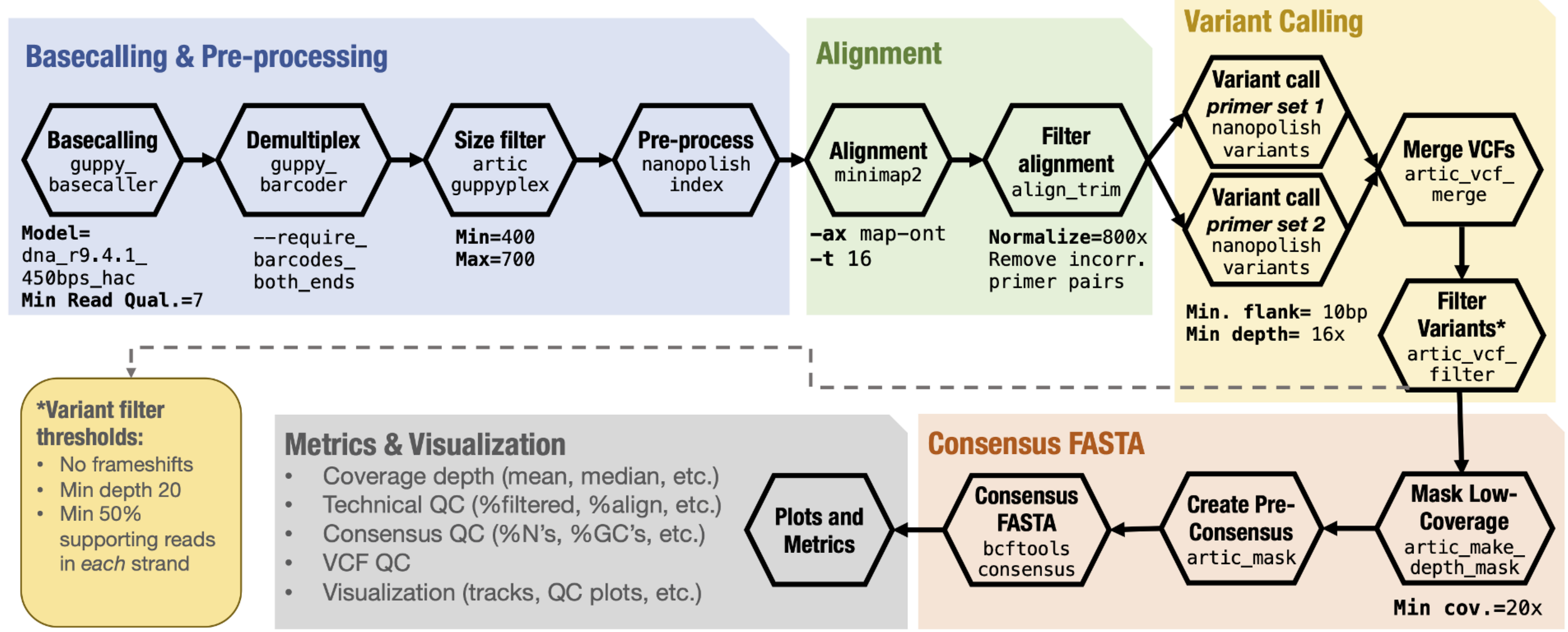
MultiQC : Aggregate results from bioinformatics analyses across many samples into a single report

Sequence alignment : Aligns the reads/sequences based in the similarities

Consensus sequence: A consensus sequence is a sequence of DNA represents aligned, related sequences.

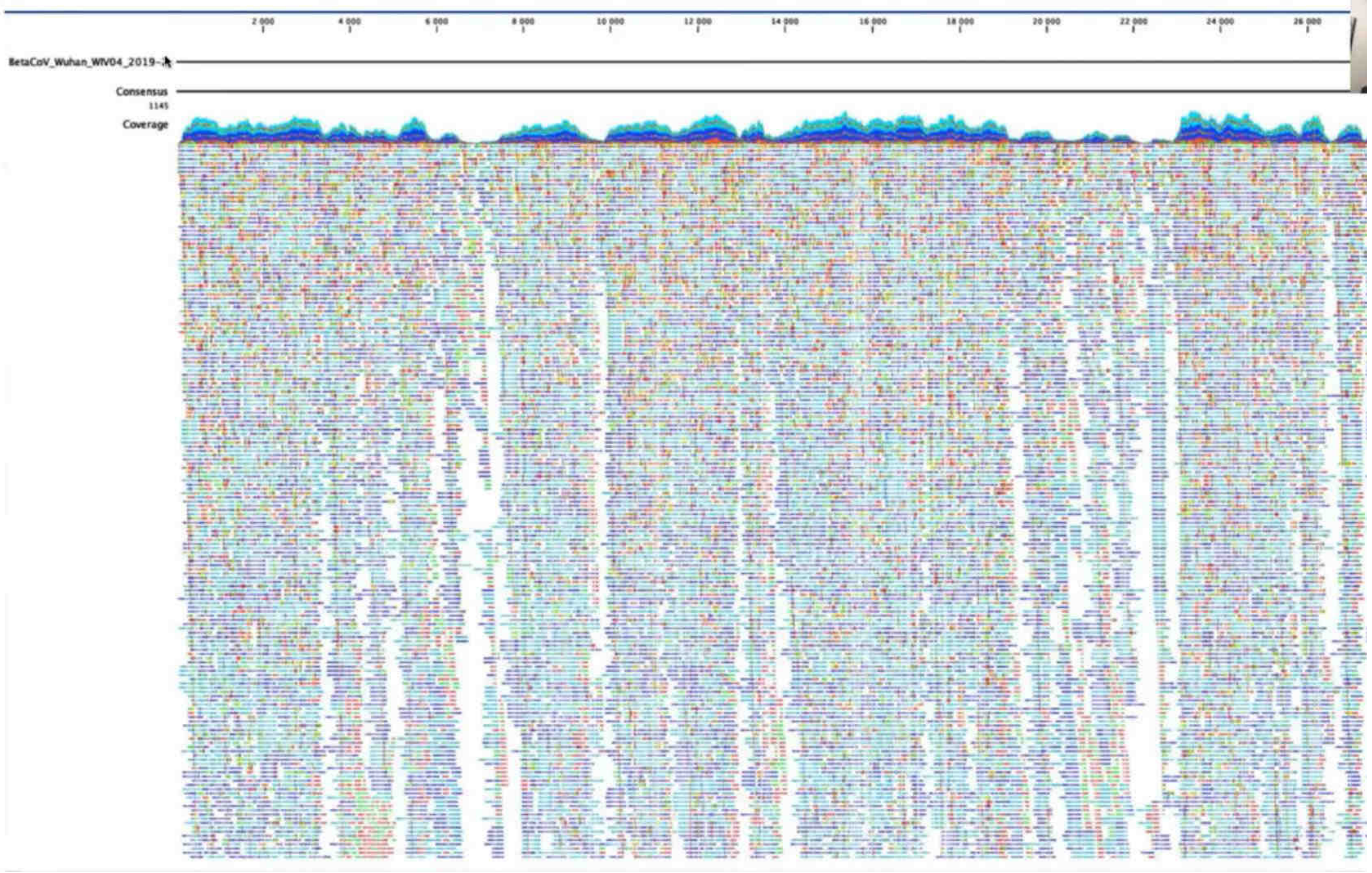
De novo assembly: De novo sequence assemblers are a type of program that assembles short nucleotide sequences into longer ones without the use of a reference genome.

SARS-CoV2 **ARTIC** Analysis Pipeline
(Nanopolish)



Reference Genome: Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1 (GenBank MN908947.3)
Software versions: `guppy-GPU v3.4.4; minimap2 v2.17; samtools v1.9; bcftools v1.9; bedtools v2.27.0; python v3.6; nanopolish v0.13.1; muscle v3.8.31`
ARTIC pipeline cloned from: https://github.com/artic-network/fieldbioinformatics

Read alignment (Bowtie 2) : Wuhan reference sequence to the raw reads to get the final consensus seq.



Softwares and repositories

Github: Millions of developers and companies build, ship, and maintain their software on GitHub—the largest and most advanced development platform in the world.

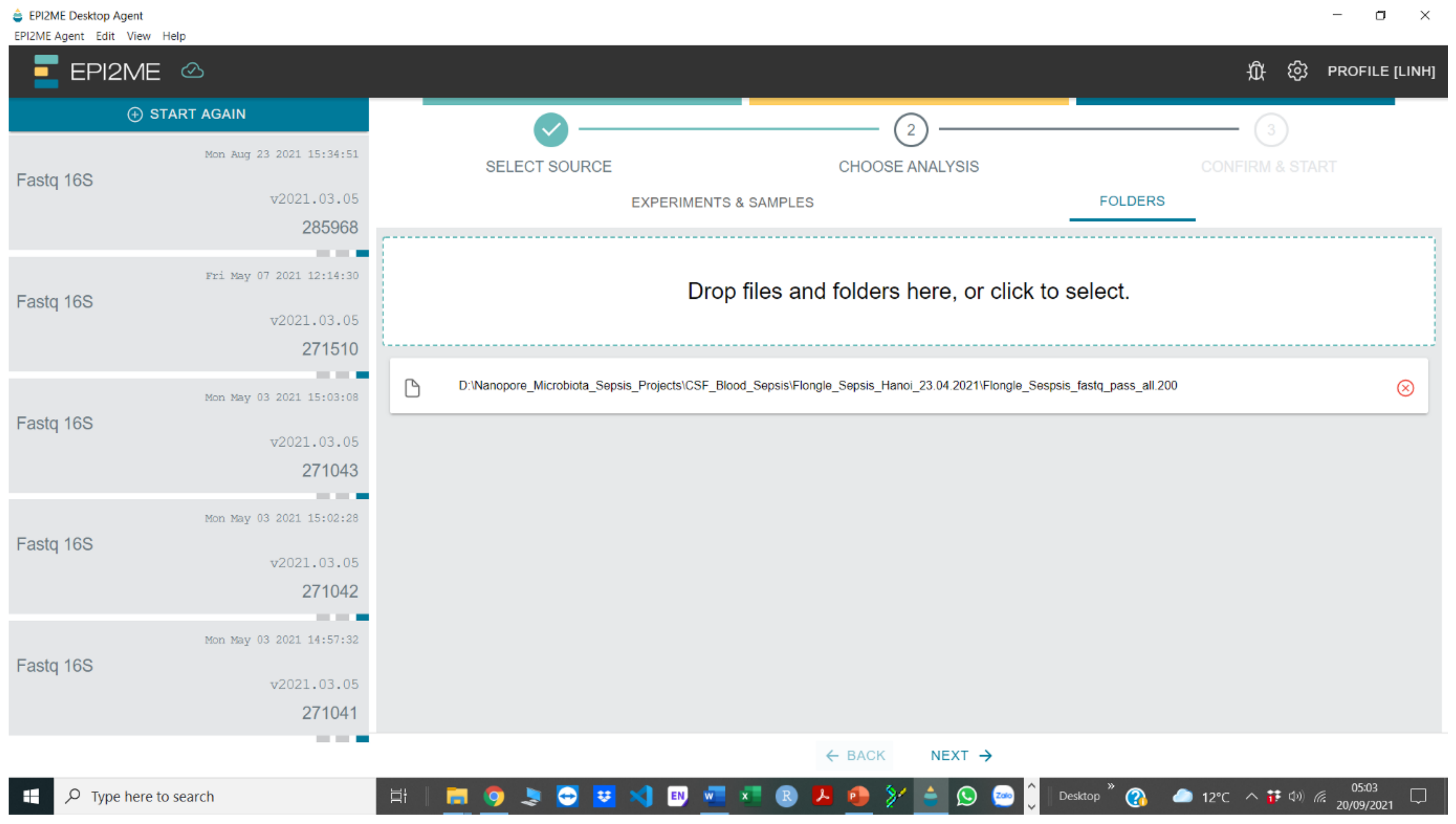
Conda (Anaconda, Miniconda, Bioconda): Conda is an open-source package management system and environment management system that runs on Windows, macOS, and Linux. Conda quickly installs, runs, and updates packages and their dependencies. Conda easily creates, saves, loads, and switches between environments on your local computer. It was created for Python programs but it can package and distribute software for any language.

Minimap2: is a versatile sequence alignment program

Medaka: is a tool to create consensus sequences and variant calls from nanopore sequencing data

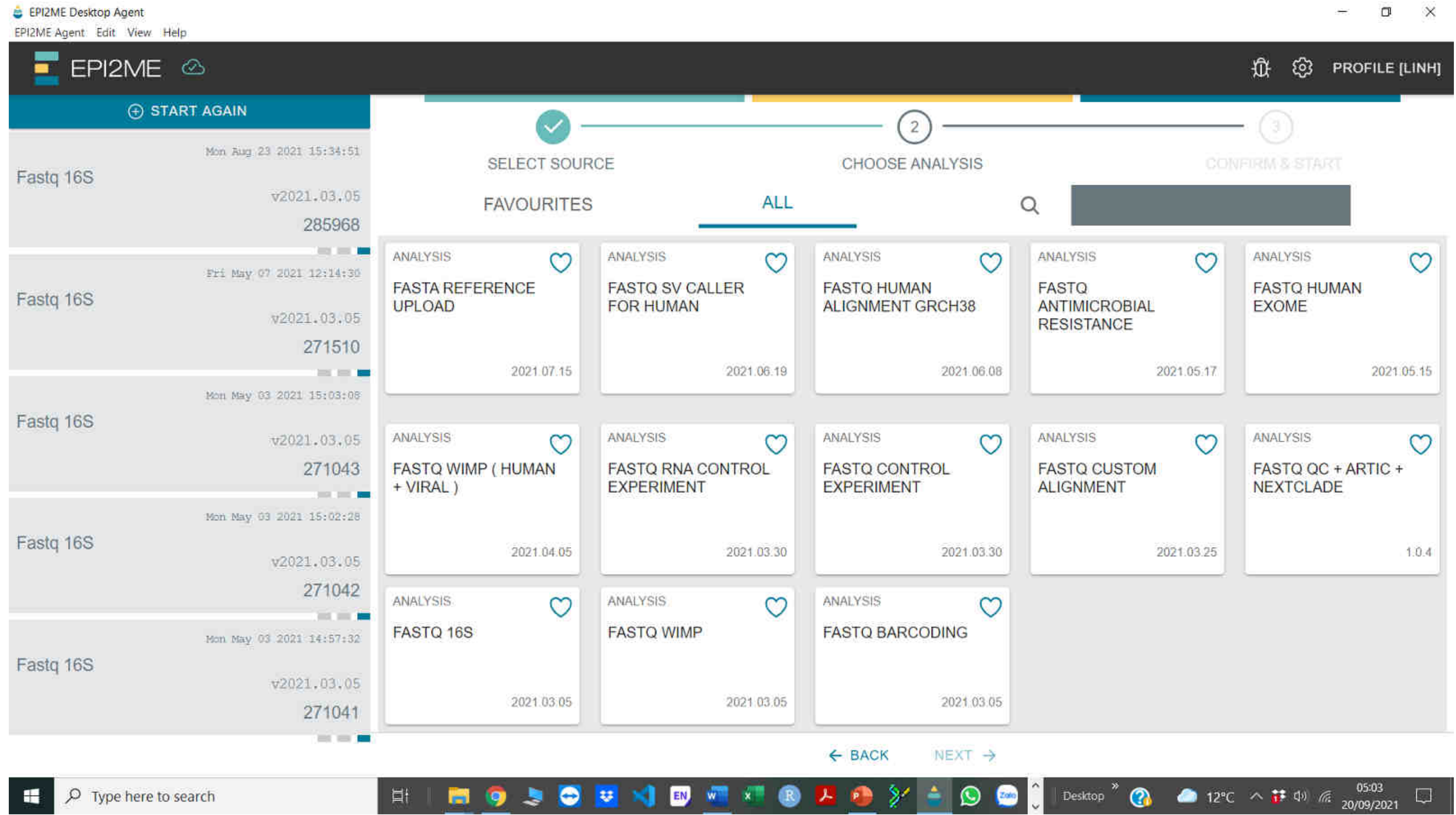
EPI2ME Cloud (Real-time data analysis)- No need bioinformatics

Bacterial sepsis/meningitis detection using real-time data analysis on EPI2ME cloud (16S amplicon sequencing)

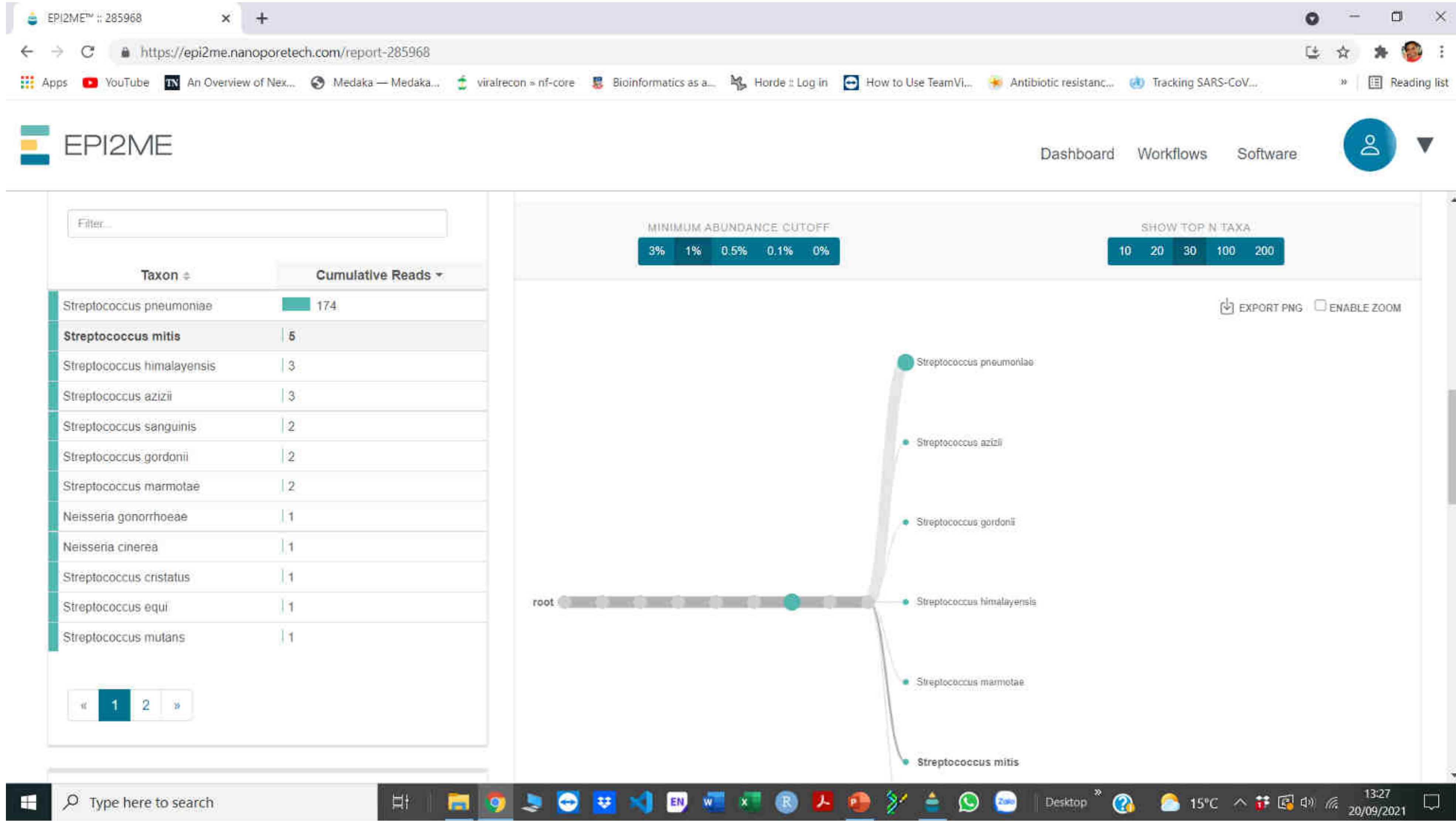


Epi2ME Cloud example live demo

EPI2ME Cloud (Real-time data analysis)

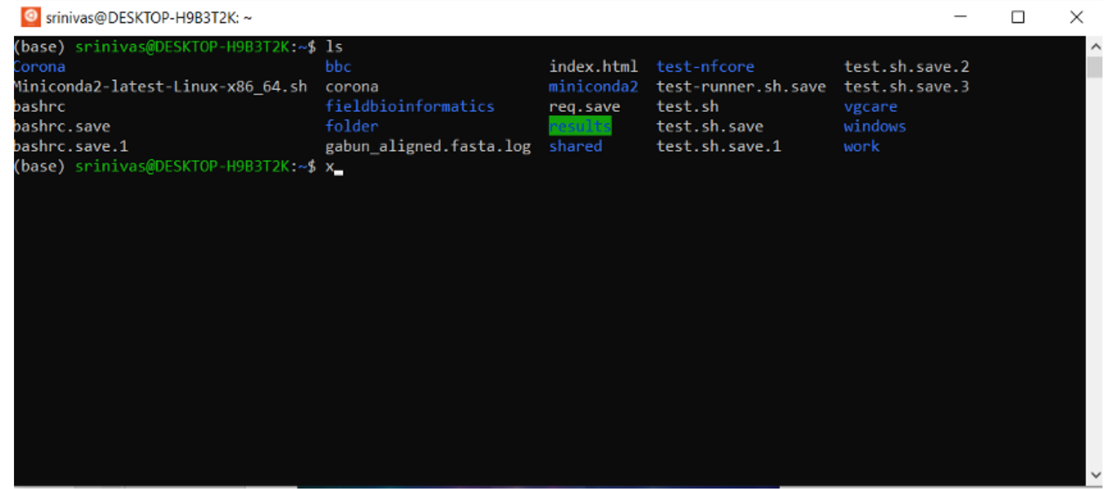


EPI2ME Cloud (Real-time data analysis)



Third Party packages (open source packages) on Linux operating system (Command line interphase)

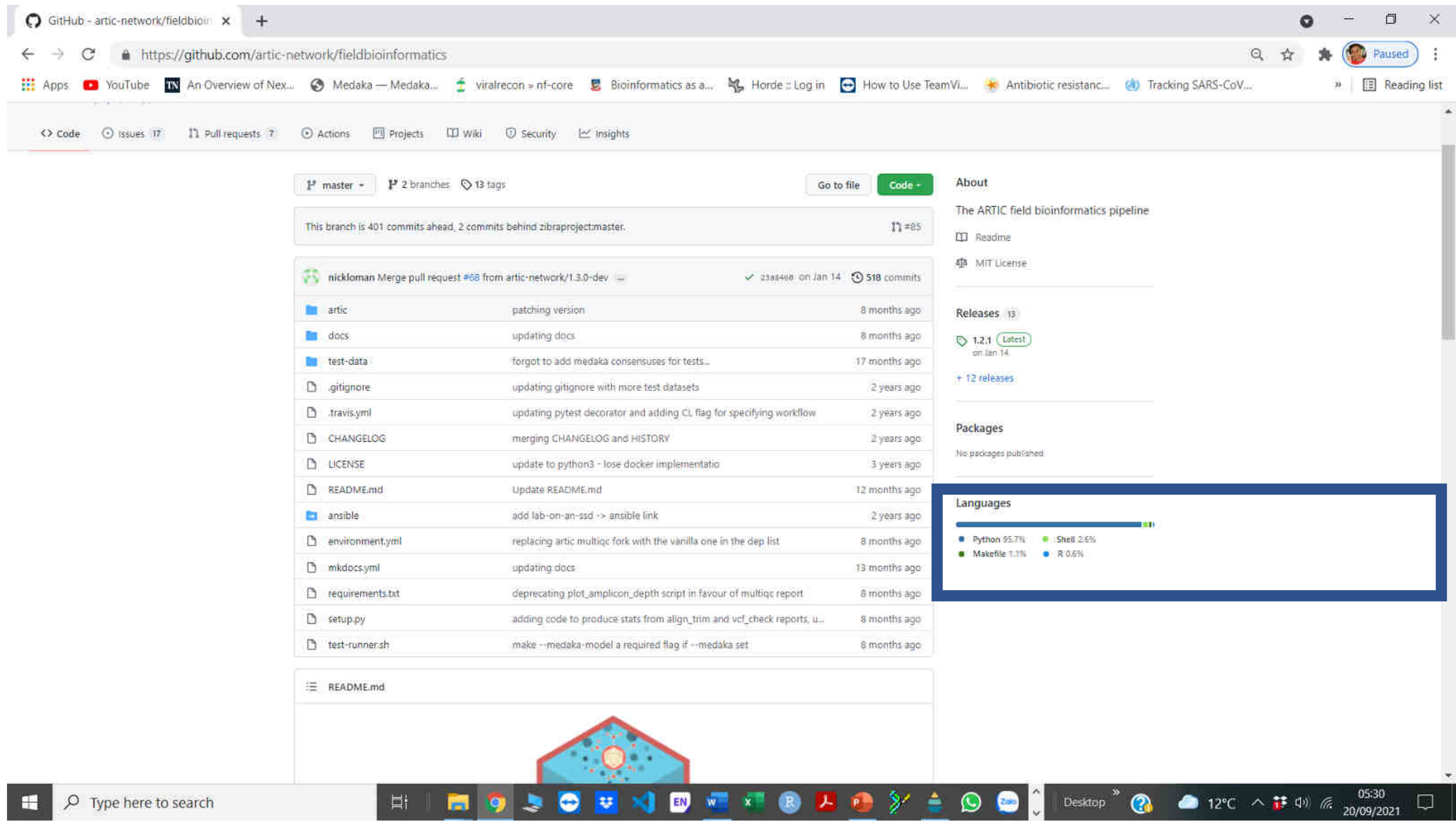
Linux Operating System (Ubuntu)



Most common languages used in the Third PARTY Packages (opensource packages)

.r .sh .py

Fieldbioinformatics package on Github (SARS-CoV-2 data analysis)



https://github.com/artic-network/fieldbioinformatics

Difference between Linux and Windows

S.NO	Linux	Windows
1.	Linux is a open source operating system.	While windows are the not the open source operating system.
2.	Linux is free of cost.	While it is costly.
3.	It's file name case-sensitive.	While it's file name is case-insensitive.
4.	In linux, monolithic kernel is used.	While in this, micro kernel is used.
5.	Linux is more efficient in comparison of windows.	While windows are less efficient.
6.	There is forward slash is used for Separating the directories.	While there is back slash is used for Separating the directories.
7.	Linux provides more security than windows.	While it provides less security than linux.
8.	Linux is widely used in hacking purpose based systems.	While windows does not provide much efficiency in hacking.

What after performing the Bioinformatic analysis

- Deposit the RAW data and the final sequences to the public repositories.
- If metagenomics is there any pathogen detected – is it novel ?
- Combine the phenotype and genotype data
- Perform statistical (in case there are groups)/phylogenetic analysis (Viruses)
- Check the results
- Write a manuscript…. edit…edit…..Final draft
- Submit and publish

Next-generation sequencing data repositories (to upload and download raw data)

GISAID - https://www.gisaid.org/ (Repository)

https://www.ncbi.nlm.nih.gov/sra (NCBI SRA Toolkit to download the data directly)

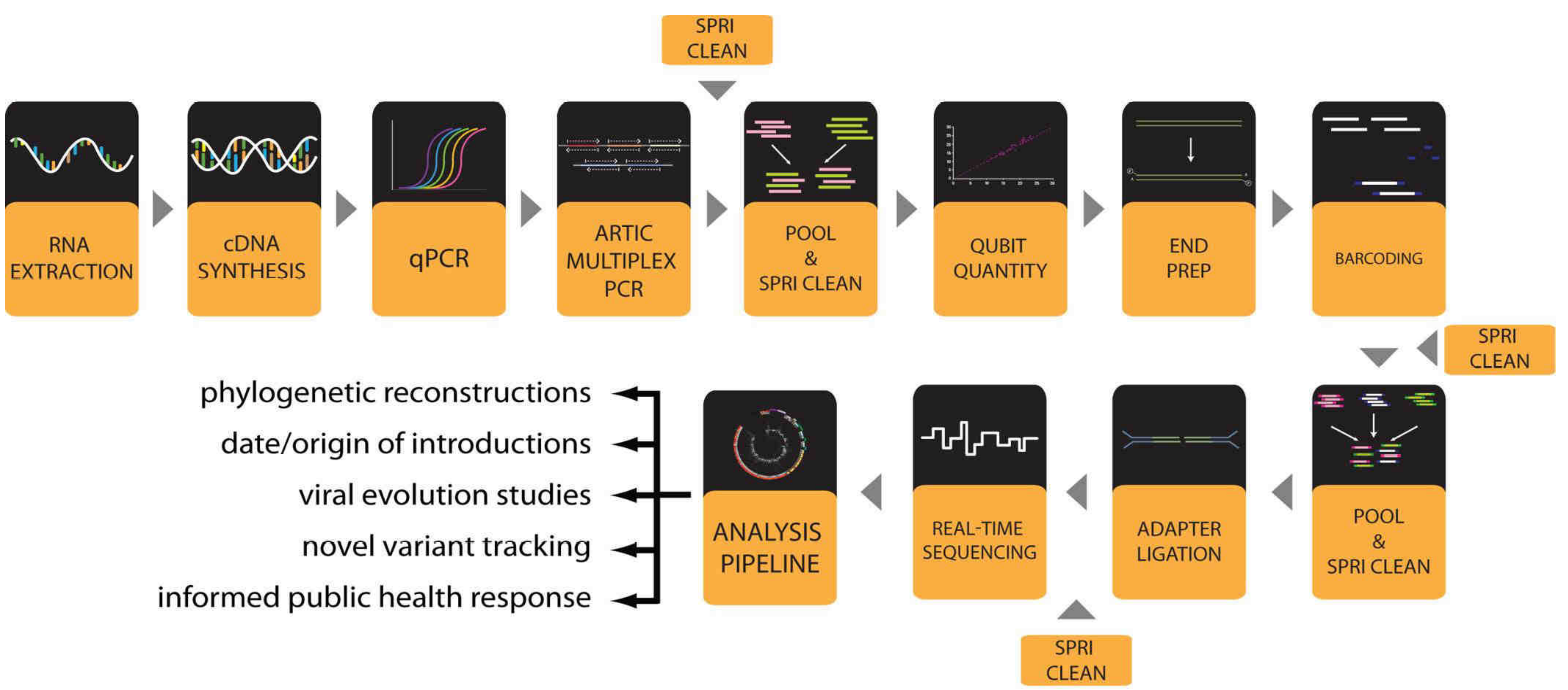
https://www.ncbi.nlm.nih.gov/genbank

https://www.insdc.org/

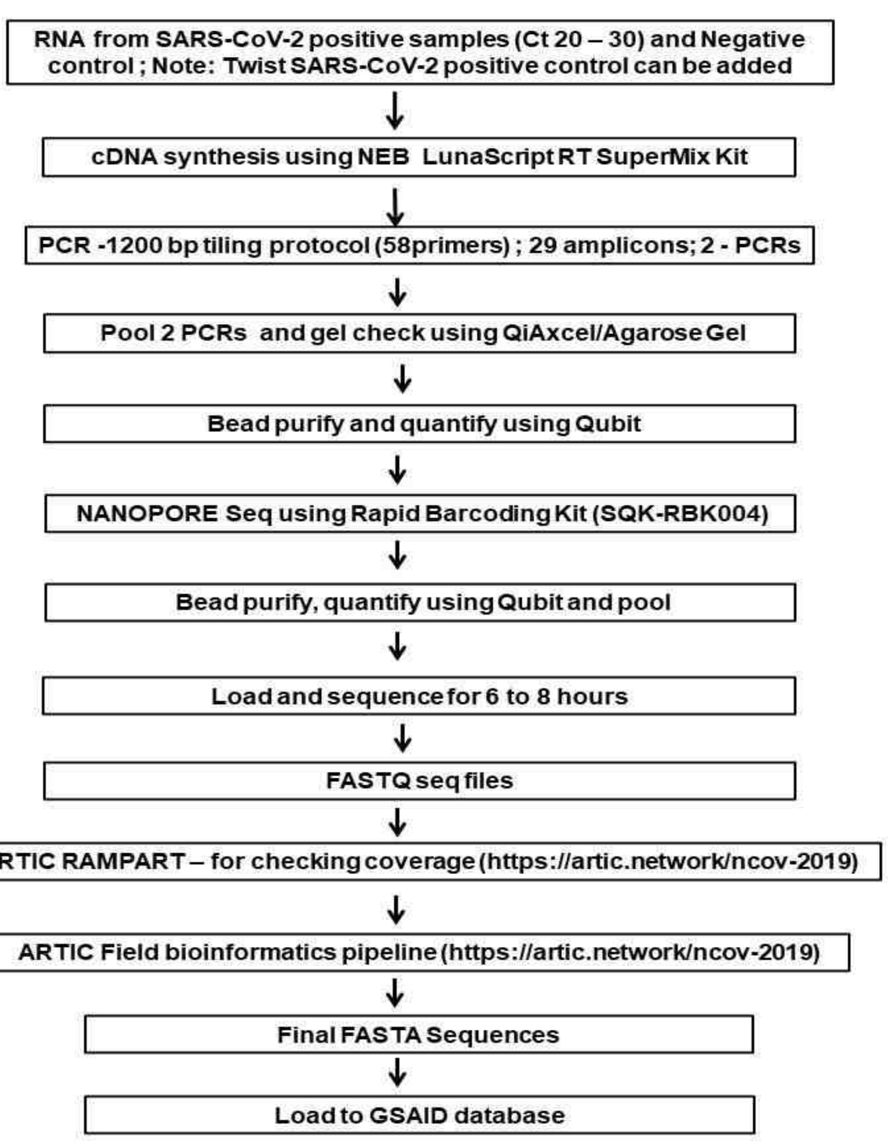
https://www.ebi.ac.uk/ena/browser/home

https://www.ebi.ac.uk/submission/

If you don’t have the your own raw data and want to work on the pipelines: Data can be downloaded from NCBI or other databases ……


SPRI CLEAN
RNA EXTRACTION
cDNA SYNTHESIS
qPCR
ARTIC MULTIPLEX PCR
POOL & SPRI CLEAN
QUBIT QUANTITY
END PREP
BARCODING
SPRI CLEAN
POOL & SPRI CLEAN
ADAPTER LIGATION
REAL-TIME SEQUENCING
ANALYSIS PIPELINE
SPRI CLEAN
phylogenetic reconstructions
date/origin of introductions
viral evolution studies
novel variant tracking
informed public health response

Flow diagram of the SARS-CoV-2 Whole-genome sequencing using Oxford Nanopore



THANK YOU

Linux : important links for using linux operating system

https://www.youtube.com/c/SimonCockell/videos

https://www.youtube.com/watch?v=v_1zB2WNN14

https://www.youtube.com/watch?v=V1y-mbWM3B8

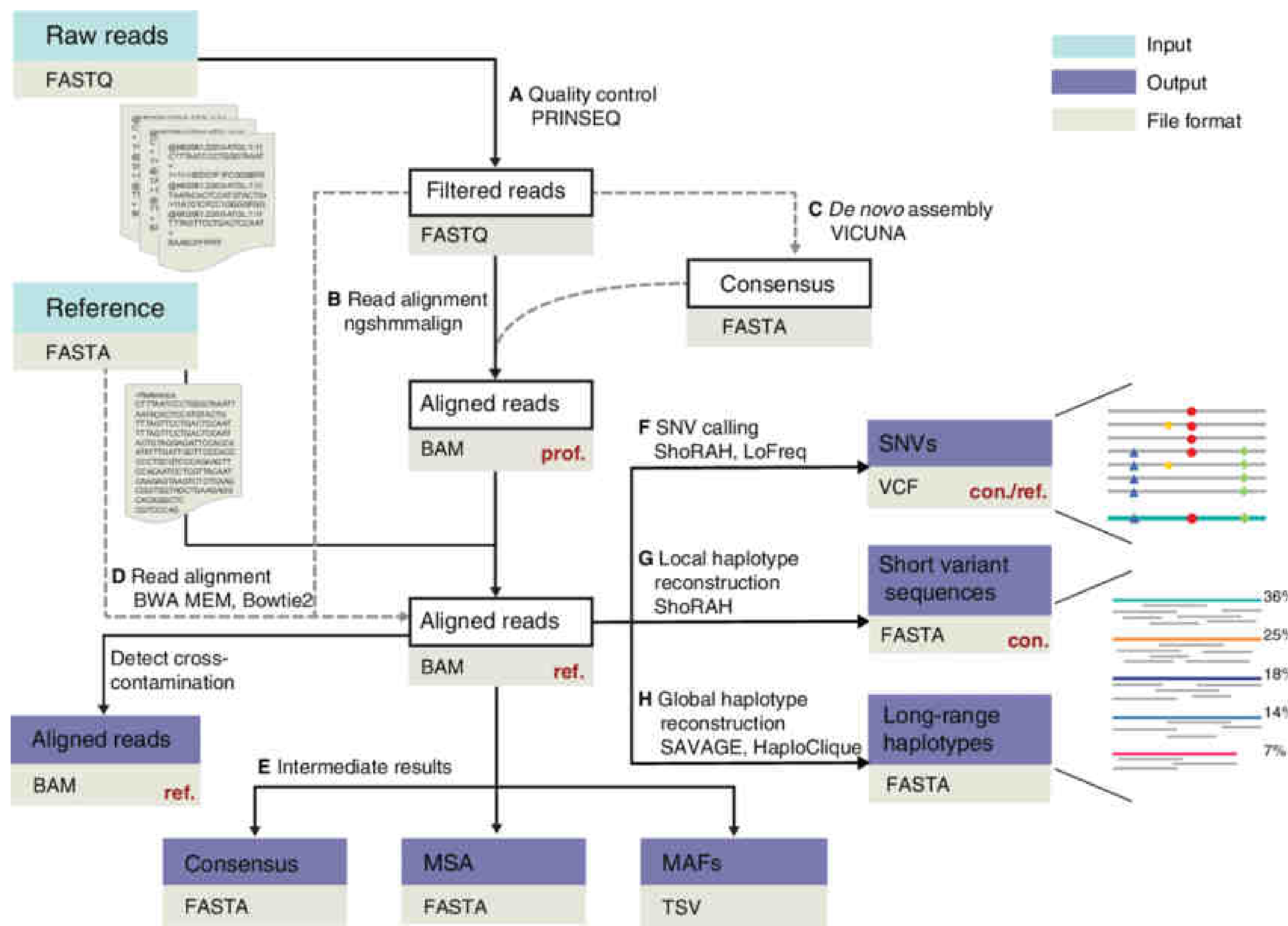
https://www.youtube.com/watch?v=x5MhydijWmc

https://www.youtube.com/watch?v=G36I1iqDZig

V-Pipe: Viral Quasipecies analysis pipeline (example)

Illumina
Nanopore

FAST5 reads
Nanopore



- Pipeline
- Softwares used
- File formats

https://github.com/cbg-ethz/V-pipe

Posada et al., 2021 Bioinformatics

Flow chart showing Sars-CoV-2 sequencing using both illumina and Nanopore

RNA from 4 SARS-CoV-2 positive samples (Ct 20 – 30) and Neg control

Note: Twist positive control can be added

↓

cDNA synthesis – (4 samples and negative)

↓

1200 bp tiling protocol (58primers) ; 29 amplicons ; 2 - PCRs

↓

Pool 2 PCRs

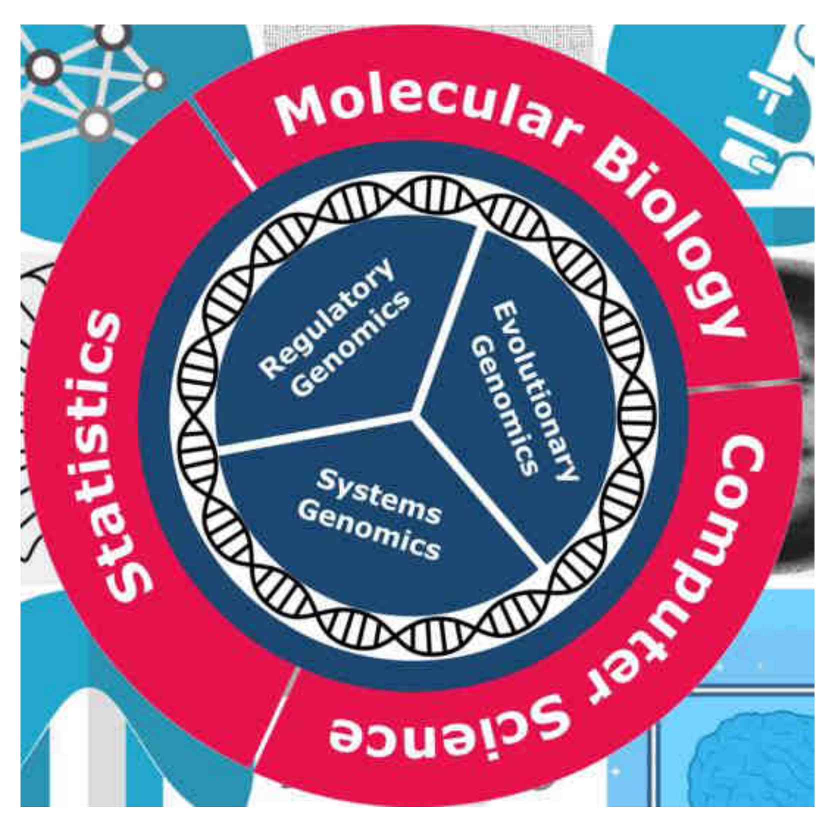
↓

QiAxcel – check

↓

Bead purify and quantify using Qubit

NANOPORE	MISEQ
NANOPORE Seq	MISEQ Seq
Rapid Barcoding Kit (SQK-RBK004)	Illumina Nextera XT Kit
Bead purify, quantify using Qubit and pool	Bead purify, quantify using Qubit and pool
Load and sequence for 4 to 6 hours (But base calling will be longer)	Load and sequence for 36 to 48 hours
FASTQ seq files	FASTQ seq files
ARTIC Field bioinformatics pipeline https://artic.network/ncov-2019	nf-core bioinformatics pipeline https://nf-co.re/viralrecon


Molecular Biology
Computer Science
Statistics
Regulatory Genomics
Evolutionary Genomics
Systems Genomics

Workshop
Molecular Diagnostic Techniques
(Genomics and Bioinformatics – Infectious diseases)
module-2

Dr. Srinivas Reddy Pallerla, Mr. Tung Tran and Ms. Linh Le
Prof. Dr. Thirumalaisamy P. Velavan, University Hospital, Tübingen, Germany
Prof. Dr. Daniel Olusola Ojurongbe, Nigeria
Prof. Dr Mohamed Osman, University of Khartoum, Sudan

26 July 2022

Training Sponsor: *Alexander von Humboldt Foundation*

Overview

- Introduction
  - Who is who? Who knows what?

- How will the training be organized?
  - What is in the “additional information”?

- Overview of NGS-nanopore work-flow as described in the protocol

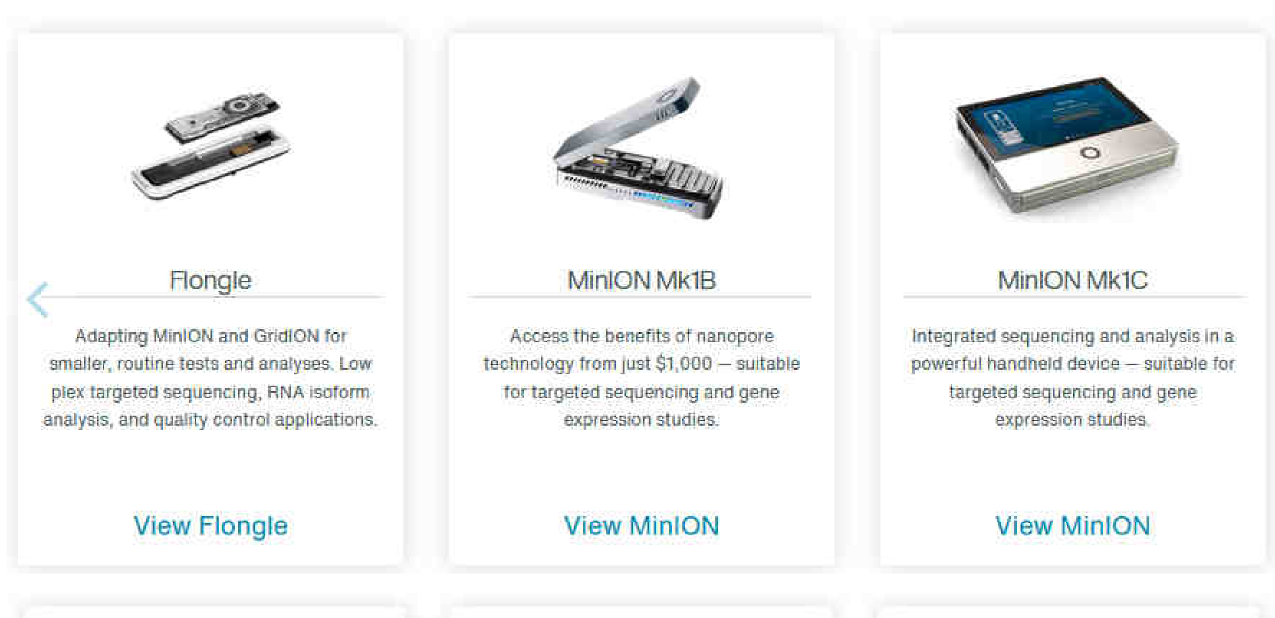
- Theory of nanopore-sequencing
  - Material + devices

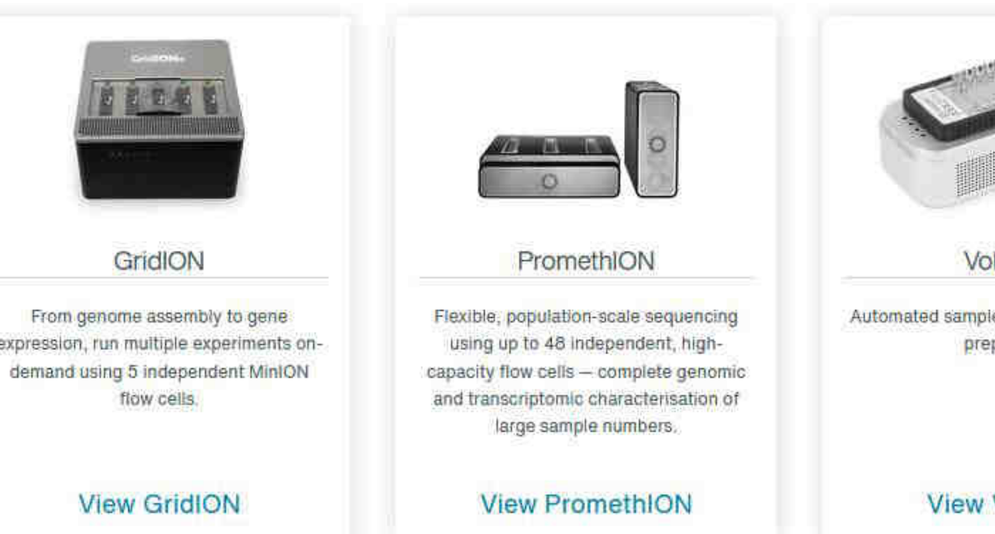
“Our goal is to enable the analysis of anything, by anyone, anywhere”.
Fom Nanopore

https://www.youtube.com/c/OxfordNanoporeTechnologies/featured

https://nanoporetech.com/

Nanopore technique – Devices available





https://nanoporetech.com/how-nanopore-sequencing-works

Nanopore sequencing, the only technology that offers scientific researchers:

- Sequence any DNA/RNA fragment length from short to ultra-long

Characterise more genetic variation, versatile to broad applications

- Direct sequencing of native DNA/RNA

Generate content-rich data, including methylation

- Data available in real time

Rapid insights, and analyses that can respond to results in real time

- Scalable from portable devices to ultra-high throughput desktop devices

Sequence anything, anywhere

- No capital investment required

Accessible and cost effective

- Simple & rapid, or automated, library prep

Easy to use and versatile

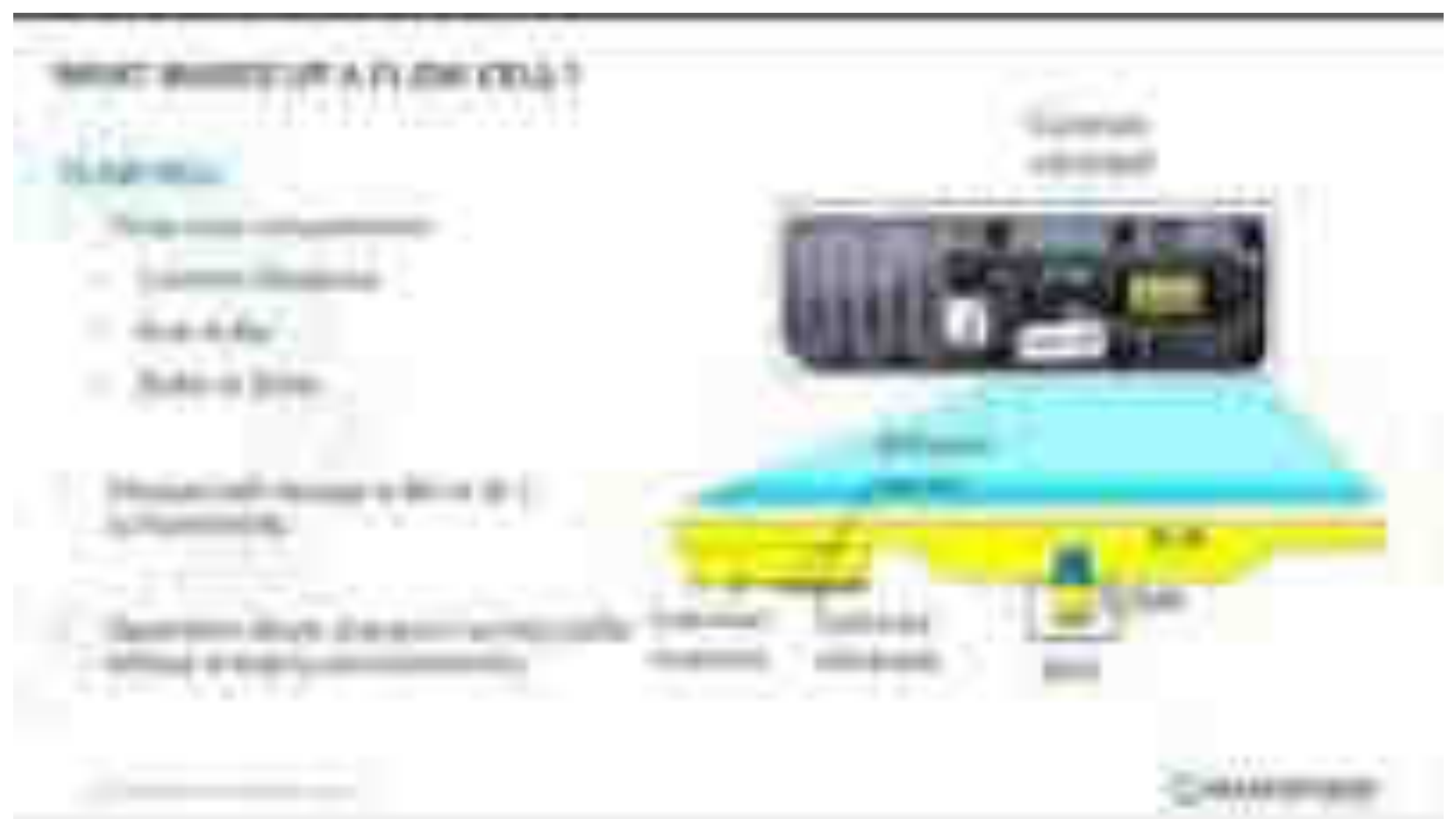
https://nanoporetech.com/

Overview

https://www.youtube.com/c/OxfordNanoporeTechnologies/featured

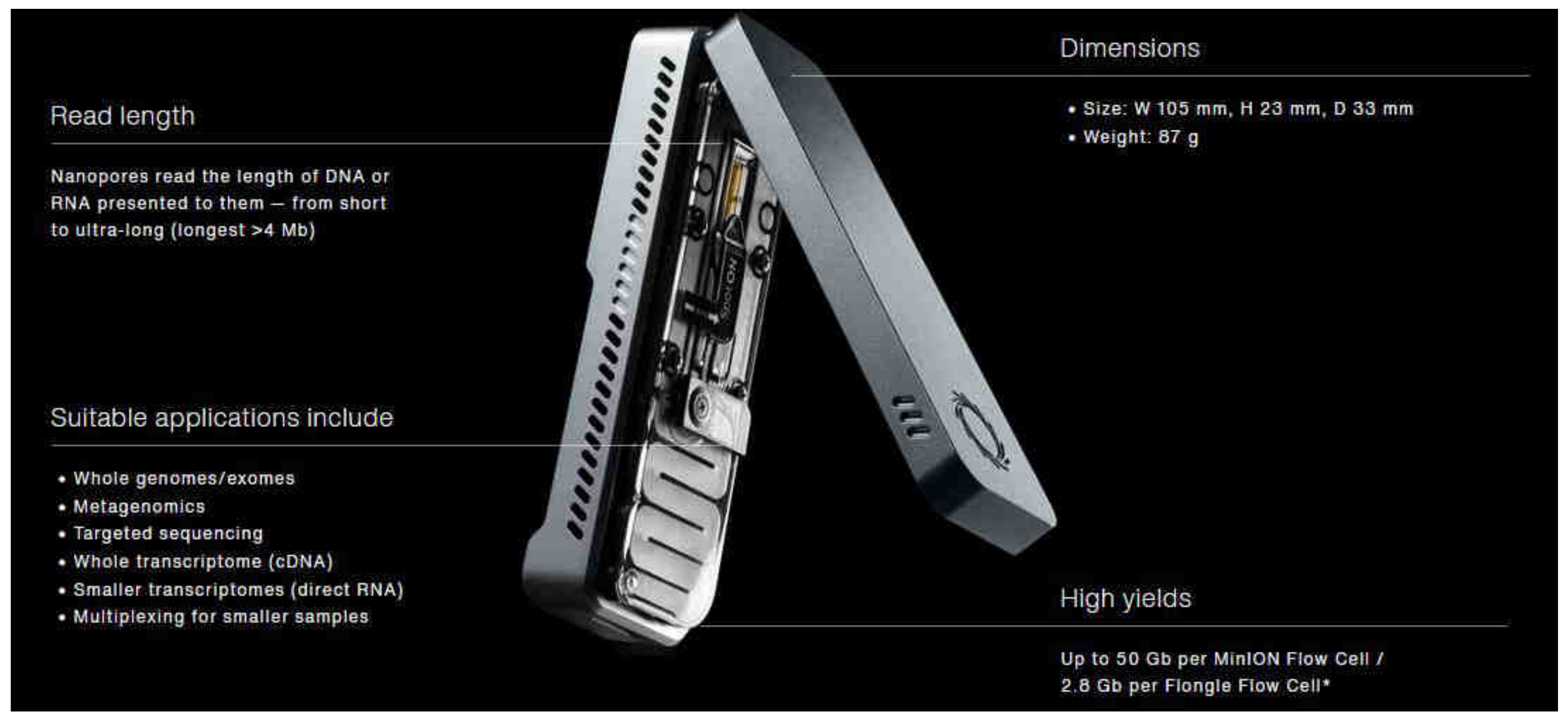
All the updates on the Nanopore sequencing advancements, kits, improvements, ……. please follow to adapt to the updates….. that will reduce the costs, new kits, new products…….

VolTRAX – a small device designed to perform library preparation automatically



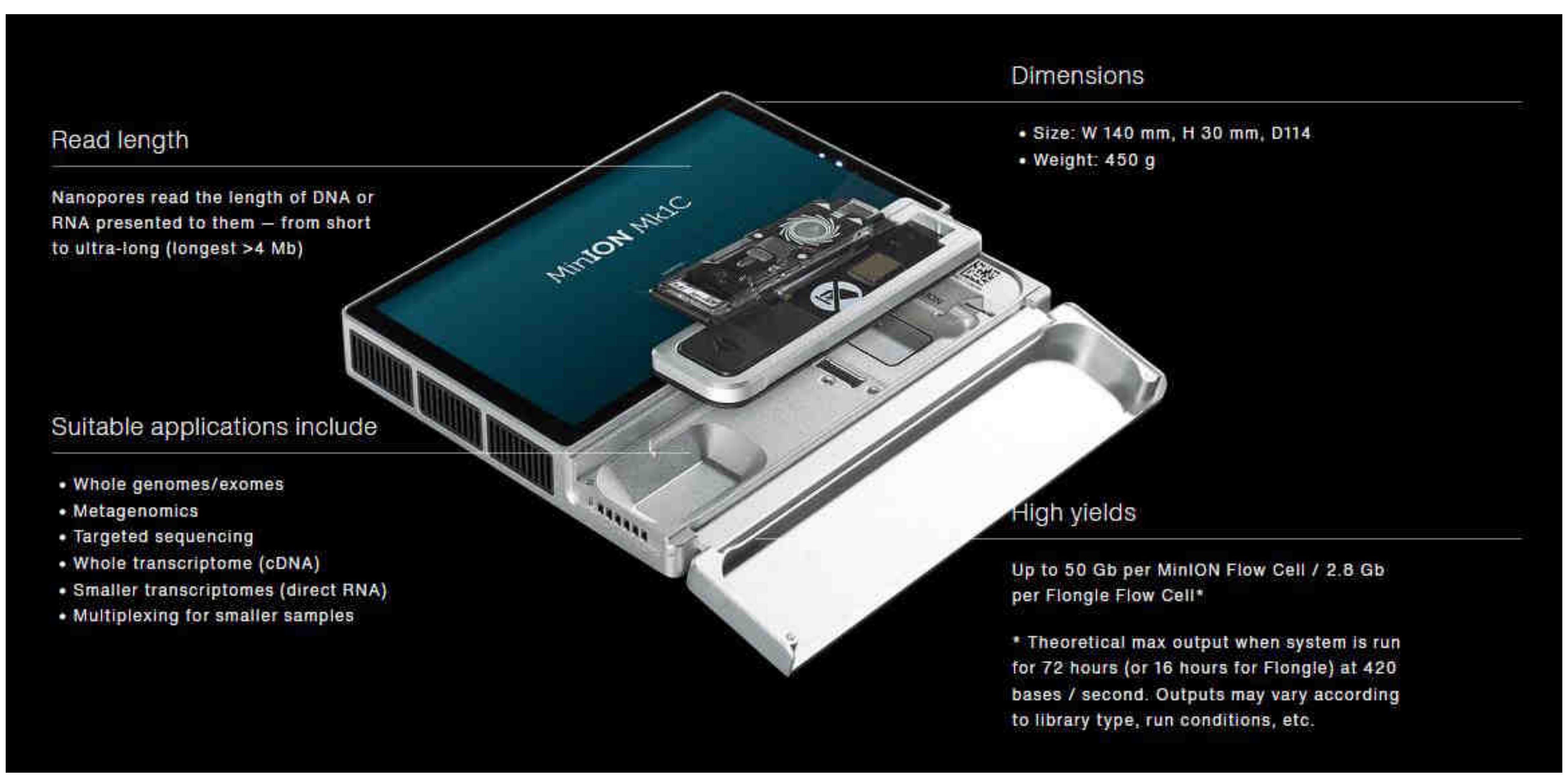
https://nanoporetech.com/how-nanopore-sequencing-works

Nanopore technique – MinION Mk1B



https://nanoporetech.com/products/minion#tabs-0=mk1b

Nanopore technique – MinION Mk1C



https://nanoporetech.com/products/minion#tabs-0=mk1c

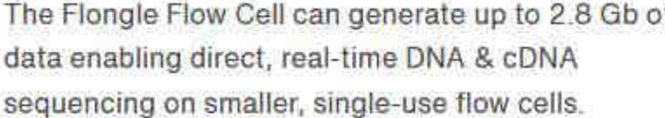
Nanopore technique – Flowcell

https://community.nanoporetech.com/protocols/flow-cell-wash-kit-exp-wsh004/v/wfc_9120_v1_revb_08dec2020/flushing-a-minion-gridion-flow-cell?devices=minion

Nanopore technique – Flongle (adapter for MinION for smaller, single-use flowcells)

https://nanoporetech.com/products/flongle

Nanopore technique – Flowcell



The Flongle Flow Cell can generate up to 2.8 Gb of data enabling direct, real-time DNA & cDNA sequencing on smaller, single-use flow cells.

The MinION Flow Cell can generate up to 50 Gb of data for sequencing DNA, cDNA or native RNA in real-time.

The PromethION Flow Cell can generate up to 290 Gb for sequencing DNA, cDNA or native RNA in real-time.

https://nanoporetech.com/flow-cells-and-nanopores

What is in the additional information ?

1. Theory

- 1.5 Links to get familiar with technology and experiments that you may plan in future using NANOPORE

  -> https://nanoporetech.com/

  -> https://nanoporetech.com/nanopore-sequencing-data-analysis

  -> https://www.youtube.com/c/OxfordNanoporeTechnologies/playlists

What is in the additional information ?

1. Theory

- 1.5 Links to get familiar with technology and experiments that you may plan in future using NANOPORE

  -> https://nanoporetech.com/

  -> https://nanoporetech.com/nanopore-sequencing-data-analysis

  -> https://www.youtube.com/c/OxfordNanoporeTechnologies/playlists

Nanopore technique

- Video : Introduction to nanopore sequencing

https://www.youtube.com/watch?v=qzusVw4Dp8w&list=PLxpxXZj-IgXyR0it4QjoBETC5JeEanW-o&index=2

Nanopore technique

- Oxford Nanopore Website :

https://nanoporetech.com/

Nanopore technique – Kits

Different kits used in the workflow:

- cDNA-synthesis – LunaScript RT SuperMix Kit by NEB
- Multiplex PCR – Q5 Hot Start HF Polymerase by NEB
  dNTP mix, 10mM by Thermo Fisher Scientific
  (+ in-house Primer Pool)
- Purification – Agencourt Ampure XP Beads by Beckman Coulter
- Quantification – Qubit dsDNA BR Kit by Thermo Fisher Scientific
- Library preparation – Rapid Barcoding Kit by Oxford Nanopore
  Flow Cell Priming Kit by Oxford Nanopore
- Flowcell washing – Flowcell Wash Kit by Oxford Nanopore

Nanopore technique – Oxford Nanopore Kits

- Many different kits available, for an overview check: https://store.nanoporetech.com/sample-prep.html

- ONT-Kits used in the workflow:
  - Library preparation – Rapid Barcoding Kit by Oxford Nanopore
Flow Cell Priming Kit by Oxford Nanopore
  - Flowcell washing – Flowcell Wash Kit by Oxford Nanopore

Nanopore technique – Rapid Barcoding Kit

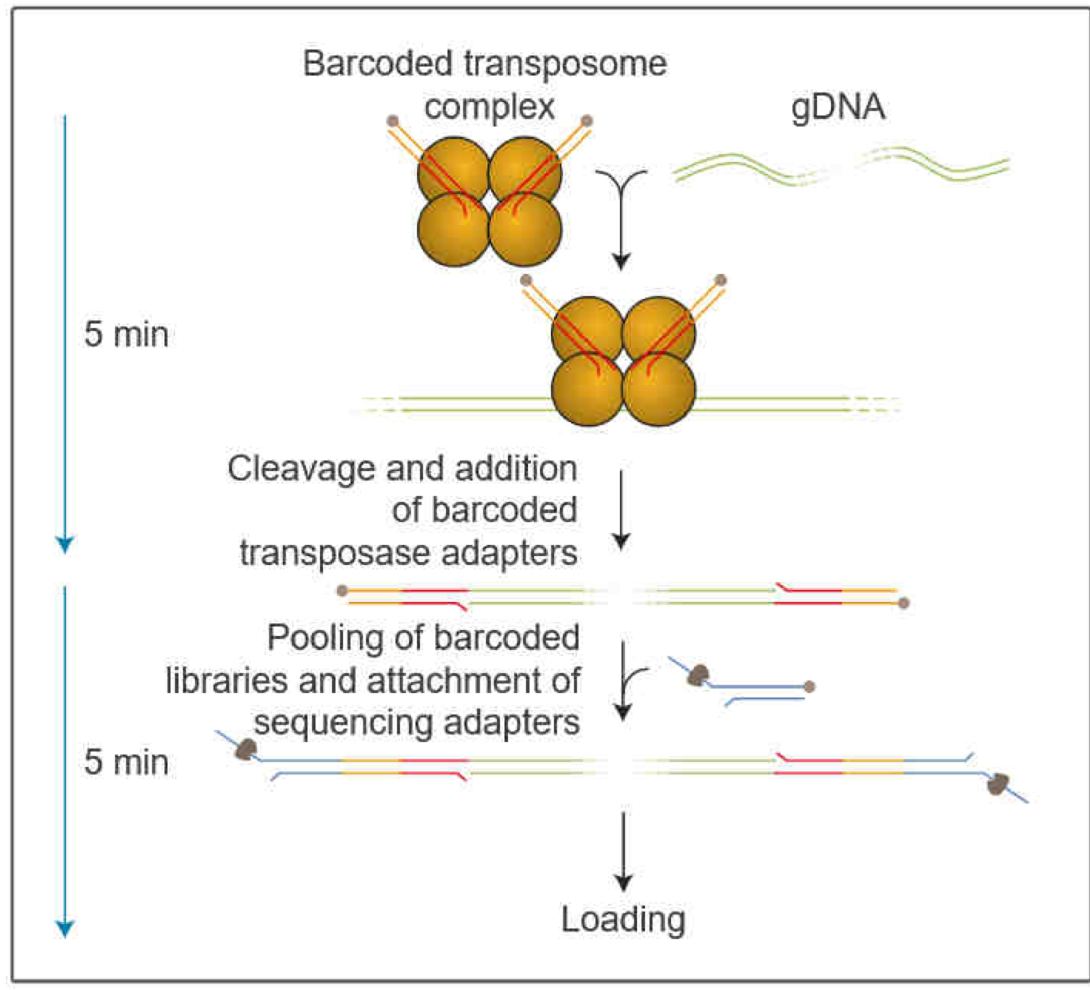
Rapid Barcoding Kit

SQK-RBK004

- Simple and rapid library preparation, with barcoding for up to 12 gDNA samples
- Shipped at 2–8°C
- Long-term storage -20°C

https://store.nanoporetech.com/rapid-barcoding-kit.html

Nanopore technique – Rapid Barcoding Kit



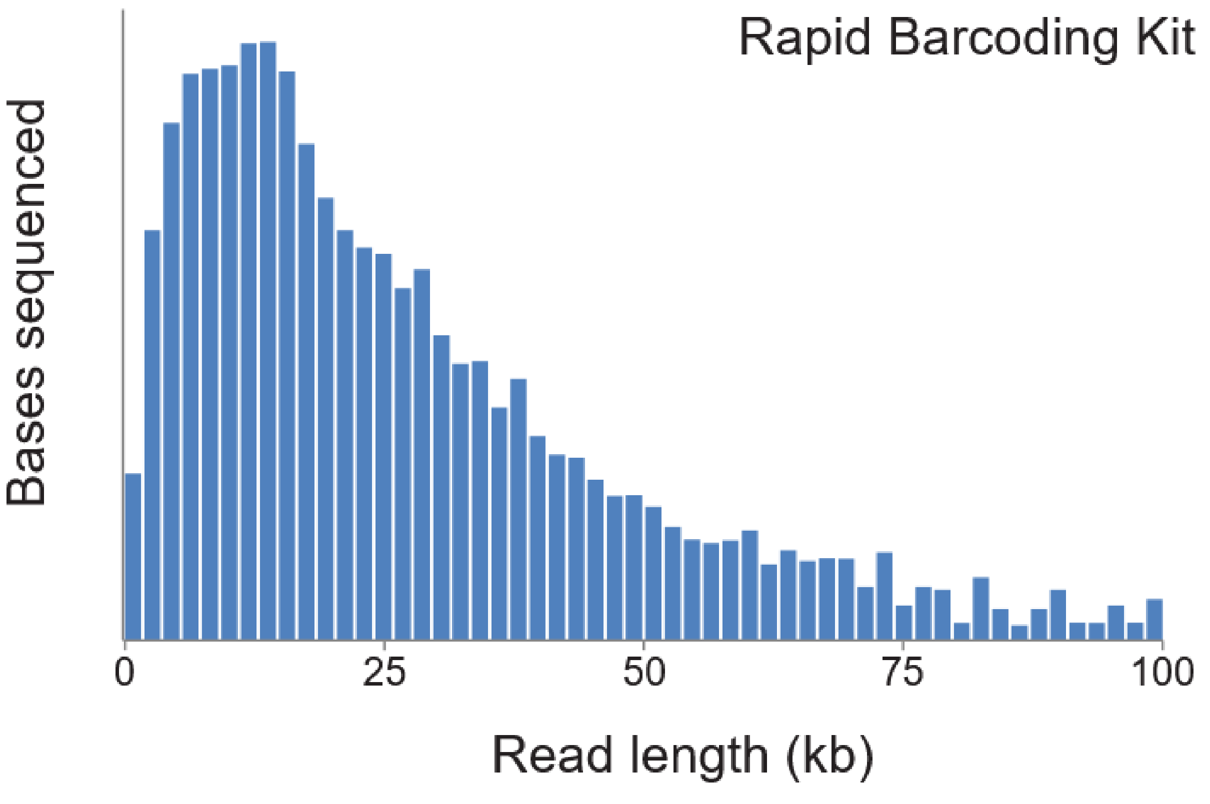
https://store.nanoporetech.com/rapid-barcoding-kit.html

The Rapid Barcoding Kit generates barcoded sequencing libraries from extracted gDNA in 10 minutes using a simple 2-step protocol.

At the heart of the kit is a transposase which simultaneously cleaves template molecules and attaches barcoded tags to the cleaved ends: there are 12 unique barcoded tags in the kit.

Barcoded samples are pooled and Rapid Sequencing Adapters are then added to the tagged ends.

Nanopore technique – Rapid Barcoding Kit



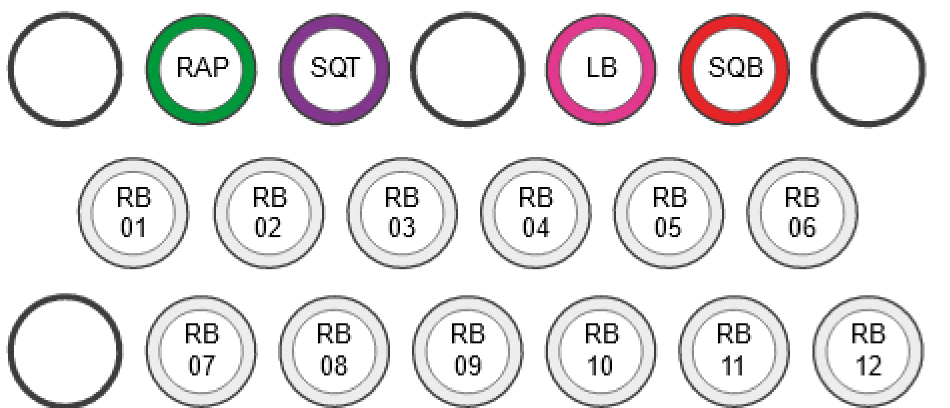
Typical read length histogram when preparing an E.coli library with the Rapid Barcoding Kit.

https://store.nanoporetech.com/rapid-barcoding-kit.html

Due to the simple nature of the workflow and the fact that little sample manipulation is required (e.g. minimal pipetting steps and no clean-ups), some very long reads can be achieved with this kit, despite the required transposase fragmentation. However, in order for long reads to be observed in sequencing, long fragments need to be present in the sample in the first place.

Nanopore technique – Rapid Barcoding Kit

The Rapid Barcoding Kit contains twelve unique barcodes and sufficient reagents to generate six sequencing libraries.



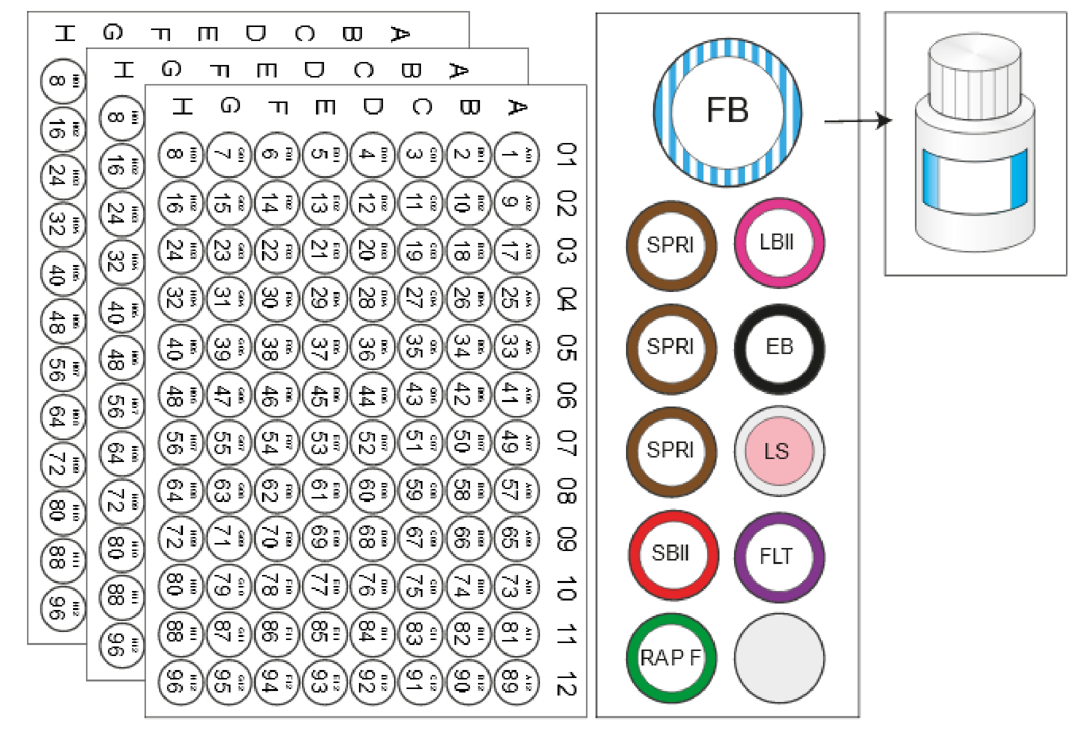
RAP : Rapid adapter
SQT : Sequencing tether
LB : Loading beads
SQB : Sequencing buffer
RB01 : Fragmentation Mix RB 1
RB02 : Fragmentation Mix RB 2
RB03 : Fragmentation Mix RB 3
RB04 : Fragmentation Mix RB 4
RB05 : Fragmentation Mix RB 5
RB06 : Fragmentation Mix RB 6
RB07 : Fragmentation Mix RB 7
RB08 : Fragmentation Mix RB 8
RB09 : Fragmentation Mix RB 9
RB10 : Fragmentation Mix RB 10
RB11 : Fragmentation Mix RB 11
RB12 : Fragmentation Mix RB 12

https://store.nanoporetech.com/rapid-barcoding-kit.html

Nanopore technique – Rapid Barcoding Kit 96

Rapid Barcoding Kit 96

SQK-RBK110.96



FB : Flush Buffer
SPRI : SPRI Beads
SBII : Sequencing Buffer II
RAP F : Rapid Adapter F

LBII : Loading Beads II
EB : Elution Buffer
LS : Loading Solution
FLT : Flush Tether

https://store.nanoporetech.com/rapid-barcoding-kit-1.html

Nanopore technique – Flow Cell Wash Kit

Flow Cell Wash Kit

EXP-WSH004

- Shipped at 2–8°C
- Long-term storage -20°C

https://store.nanoporetech.com/flow-cell-wash-kit-r9.html

Nanopore technique – Flowcell

https://store.nanoporetech.com/flow-cell-r9-4-1.html

- Never freeze!!
- Room temperature: flow cells can be stored, unopened, at room temperature for one month.
- 2-8°C: flow cells can be stored, unopened, at 2-8°C for 12 weeks.

Nanopore technique – Flowcell

How to load the flowcell:

https://www.youtube.com/watch?v=Pt-iaemrM88

Nanopore technique – MinION Mk1B

Presentation of MinION Video:

https://www.youtube.com/watch?v=1_mER5qmaVk&list=PLxpxXZj-IgXyR0it4QjoBETC5JeEanW-o&index=12

Nanopore technique – MinION Mk1C

Presentation of MinION-Mk1C Video:

https://www.youtube.com/watch?v=A1NLE0Jbvo8&list=PLxpxXZj-IgXyR0it4QjoBETC5JeEanW-o&index=15

Nanopore technique – Flongle

Presentation of Flongle Video:

https://www.youtube.com/watch?v=Ov889BEkh5I&list=PLxpxXZj-IgXyR0it4QjoBETC5JeEanW-o&index=16

SARS-CoV-2: Nanopore MinION sequencing

Aim: Sustainable capacities to be built place at CERID

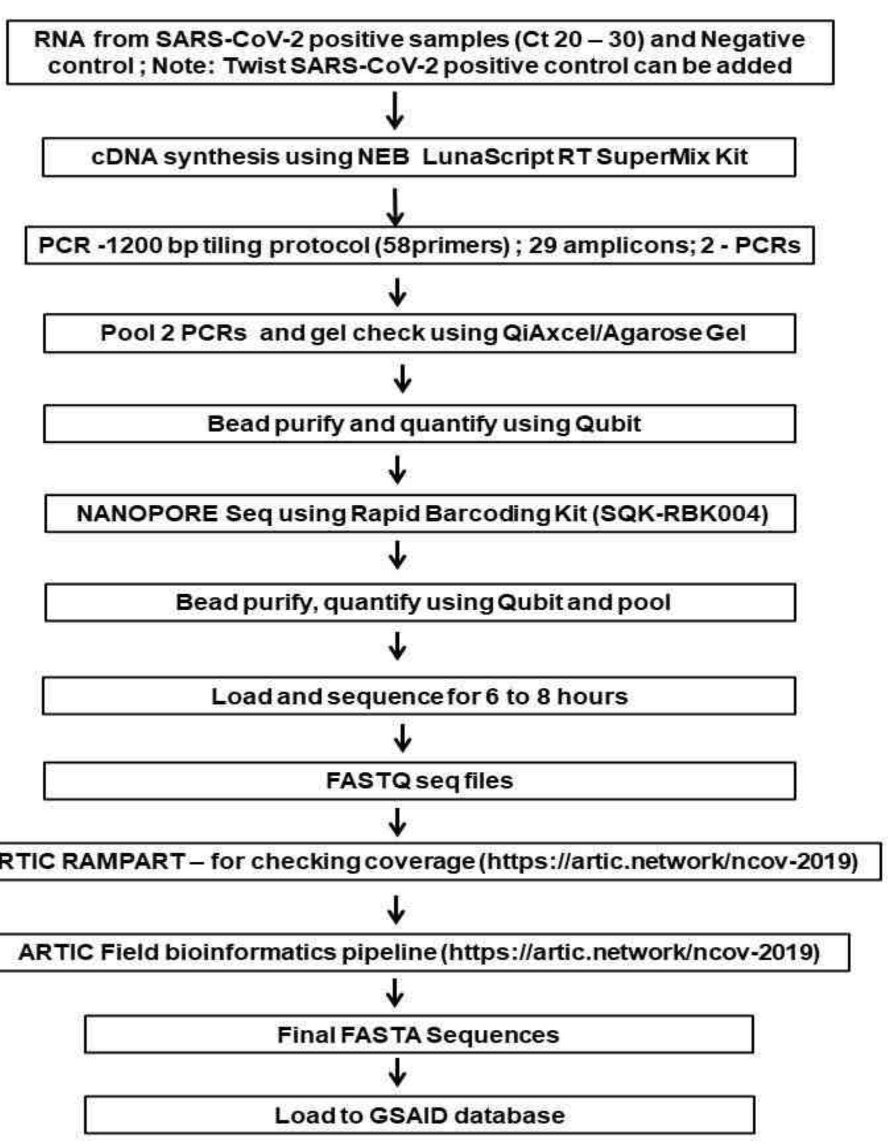
Training module:

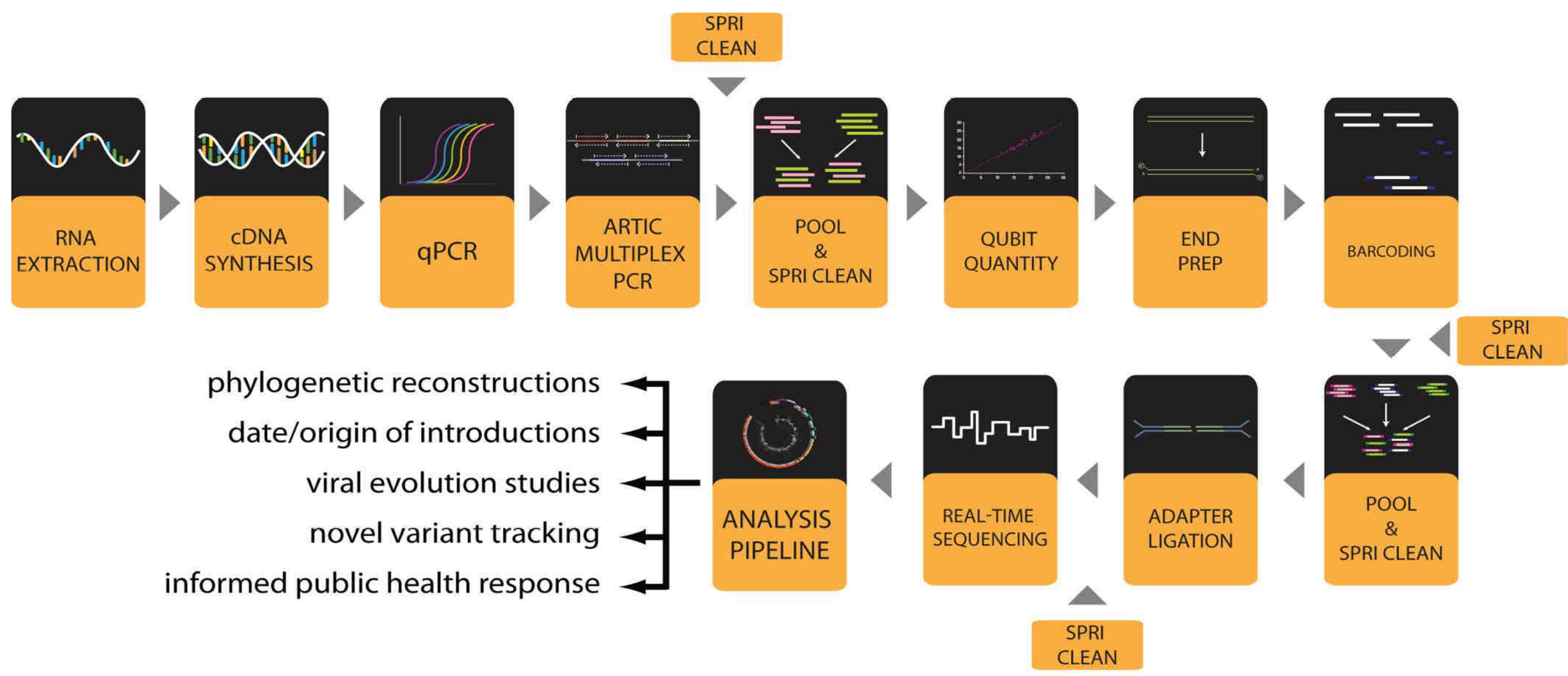
Theory : NGS and Oxford nanopore methodology

Wet-lab : SARS-CoV-2 sequencing

Dry-lab : LINUX and Bioinformatics pipelines

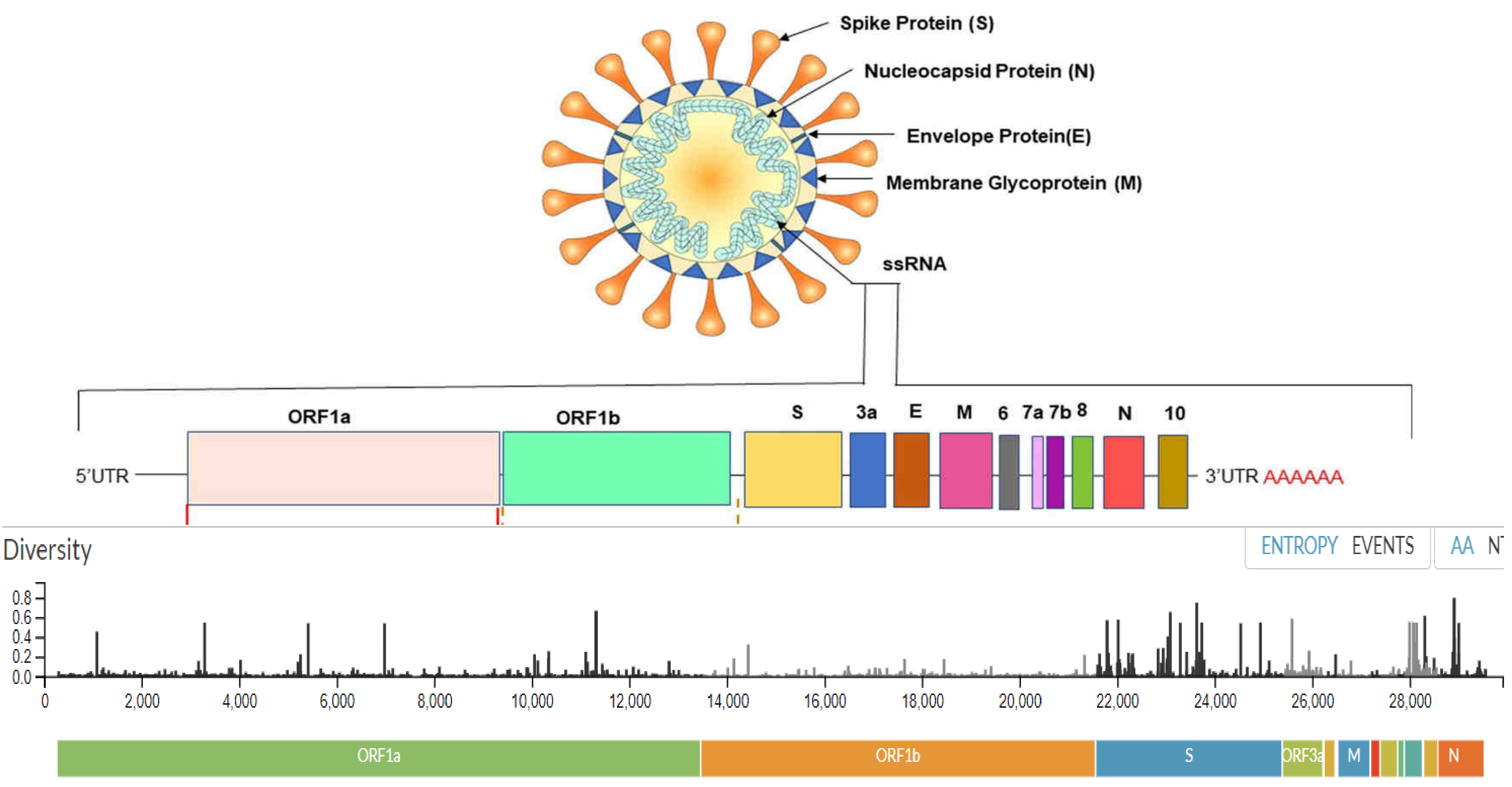
Flow diagram of the SARS-CoV-2 Whole-genome sequencing using Oxford Nanopore





SPRI CLEAN
RNA EXTRACTION
cDNA SYNTHESIS
qPCR
ARTIC MULTIPLEX PCR
POOL & SPRI CLEAN
QUBIT QUANTITY
END PREP
BARCODING
SPRI CLEAN
POOL & SPRI CLEAN
ADAPTER LIGATION
REAL-TIME SEQUENCING
ANALYSIS PIPELINE
SPRI CLEAN
phylogenetic reconstructions
date/origin of introductions
viral evolution studies
novel variant tracking
informed public health response

SARS-CoV-2 genome – 29.9 kb

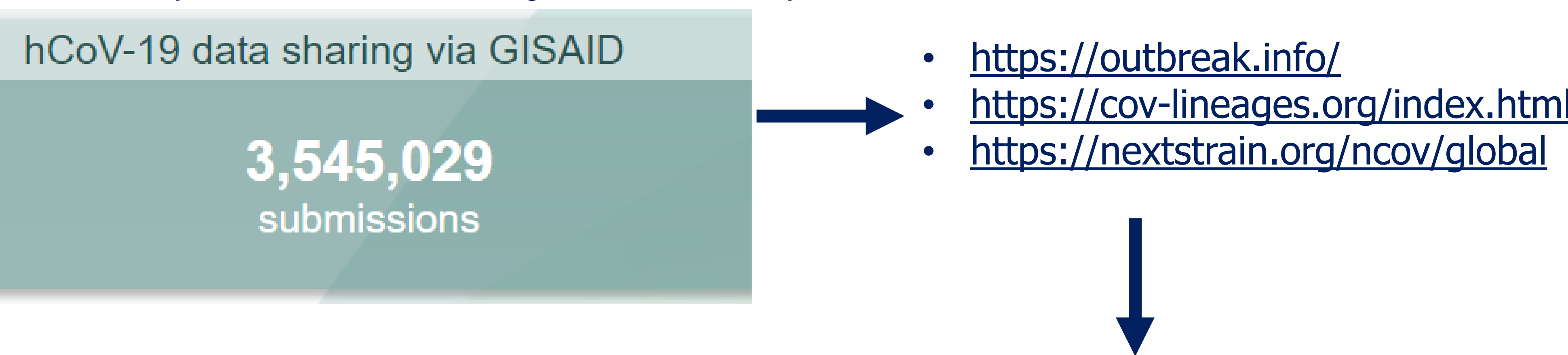


Rastogi et al., Respir Res. 2020

How concerning is the SARS-CoV-2 variants of concern ?

- Increased transmissibility;
- Increased morbidity and mortality;
- Evade detection by diagnostic tests;
- Decreased neutralizing to antibodies (e.g., convalescent plasma or monoclonal antibodies);
- Evade natural immunity (e.g., causing reinfections);
- Infect vaccinated individuals;
- Increased risk - multisystem inflammatory syndrome or long COVID;
- Increased affinity for particular demographic or clinical groups, such as children or immunocompromised individuals.

Global - genomic surveillance of SARS-CoV-2



- Global/Country/State/City to local
- Circulating variants
- New viral lineages
- Variants of interest, under investigation, of concern
- Distribution over time and space
- No of mutation accumulating
- Much more…….

What is in the additional information ?

1. Theory

- 1.1 SARS-CoV-2
  -> Presentation given on 15.09.2021 (pdf)

- 1.2 COVID-19 Genomic Epidemiology Toolkit by CDC
  -> https://www.cdc.gov/amd/training/covid-19-gen-epi-toolkit.html

- 1.3 Details about Nanopore Technology -> workshop-presentations (pdf)
  -> https://nanoporetech.com/how-nanopore-sequencing-works
  -> https://nanoporetech.com/nanopore-sequencing-data-analysis

- 1.4 Links SARS-CoV-2 sequencing principles videos Oxford Nanopore:
  -> https://nanoporetech.com/resource-centre/bioinformatics-workflows-sars-cov-2-raw-nanopore-reads-consensus-genomes-using
  -> https://www.youtube.com/watch?v=rYaFcDE-Ewg

What is in the additional information ?

2. Wet-lab

- 2.1 Protocol! (pdf)

- 2.2 Video links on Oxford Nanopore Flow cell introduction, Priming and loading flow cell

    -> https://www.youtube.com/watch?v=zC6lAtzqi_k

    -> https://www.youtube.com/watch?v=Pt-iaemrM88

    -> https://www.youtube.com/watch?v=IknVaEnuDz0&t=604s

- 2.3. Flow cell wash protocol (pdf)

- 2.4 Methods manuscript, literature protocol (pdf)

What is in the additional information ?

3. Dry-lab

3.1 Protocol! (pdf)

Protocol NGS_WGS_SARS-CoV-2_1200bp amplicons using Nanopore_V1.2.pdf

3.2 Introduction to bioinformatics

-> https://www.youtube.com/c/SimonCockell/videos

3.3 Linux

-> linuxcommands (pdf)

-> Links on Linux operating system

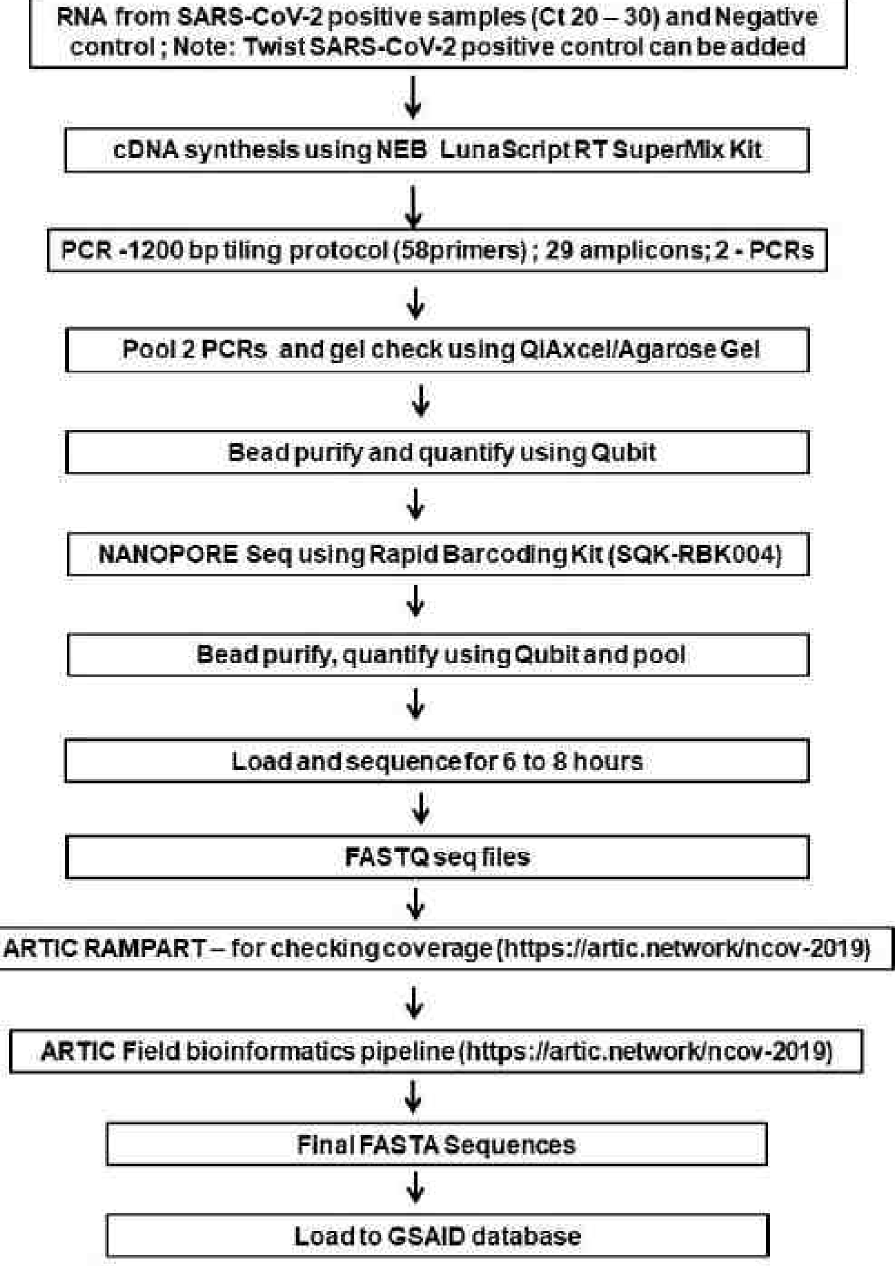
https://www.youtube.com/watch?v=v_1zB2WNN14

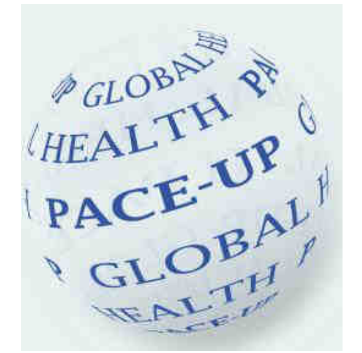
https://www.youtube.com/watch?v=V1y-mbWM3B8Ubuntu

https://www.youtube.com/watch?v=x5MhydijWmc

https://www.youtube.com/watch?v=G36I1iqDZig

NGS-nanopore WORK-FLOW Overview





Workshop
Molecular Diagnostic Techniques
(Genomics and Bioinformatics – Infectious diseases)
Module-3

Dr. Srinivas Reddy Pallerla, Mr. Tung Tran and Ms. Linh Le
Prof. Dr. Thirumalaisamy P. Velavan, University Hospital, Tübingen, Germany
Prof. Dr. Daniel Olusola Ojurongbe, Nigeria
Prof. Dr Mohamed Osman, University of Khartoum, Sudan

28 July 2022

Sponsor: *Alexander von Humboldt Foundation*

SARS-CoV-2 - Nanopore sequencing training module

1. Theory module

- Introduction into NGS technology and the library preparation
- Experimental design
- Standard operating procedures
- Introduction to Oxford Nanopore technology and its application

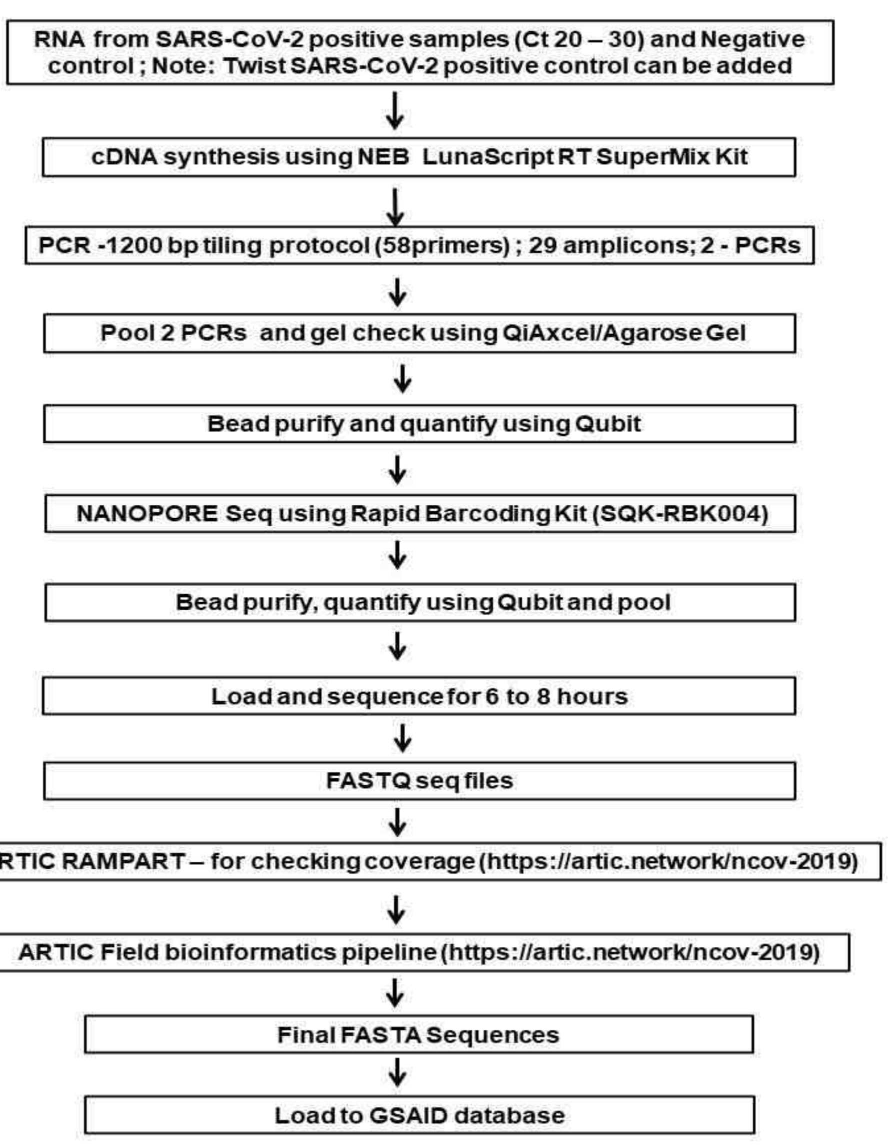
2. Wet-lab module specific for SARS-CoV-2

- NGS library preparation and validation.
- Multiplexing and barcoding
- Quantify and quality control of the prepared libraries

3. Dry-lab module

- Introduction to the bioinformatics pipelines
- NGS data analysis using Linux operating system
- Basic Linux commands for beginners
- SARS-CoV-2 data analysis using ARTIC bioinformatic pipelines (https://artic.network/ncov-2019

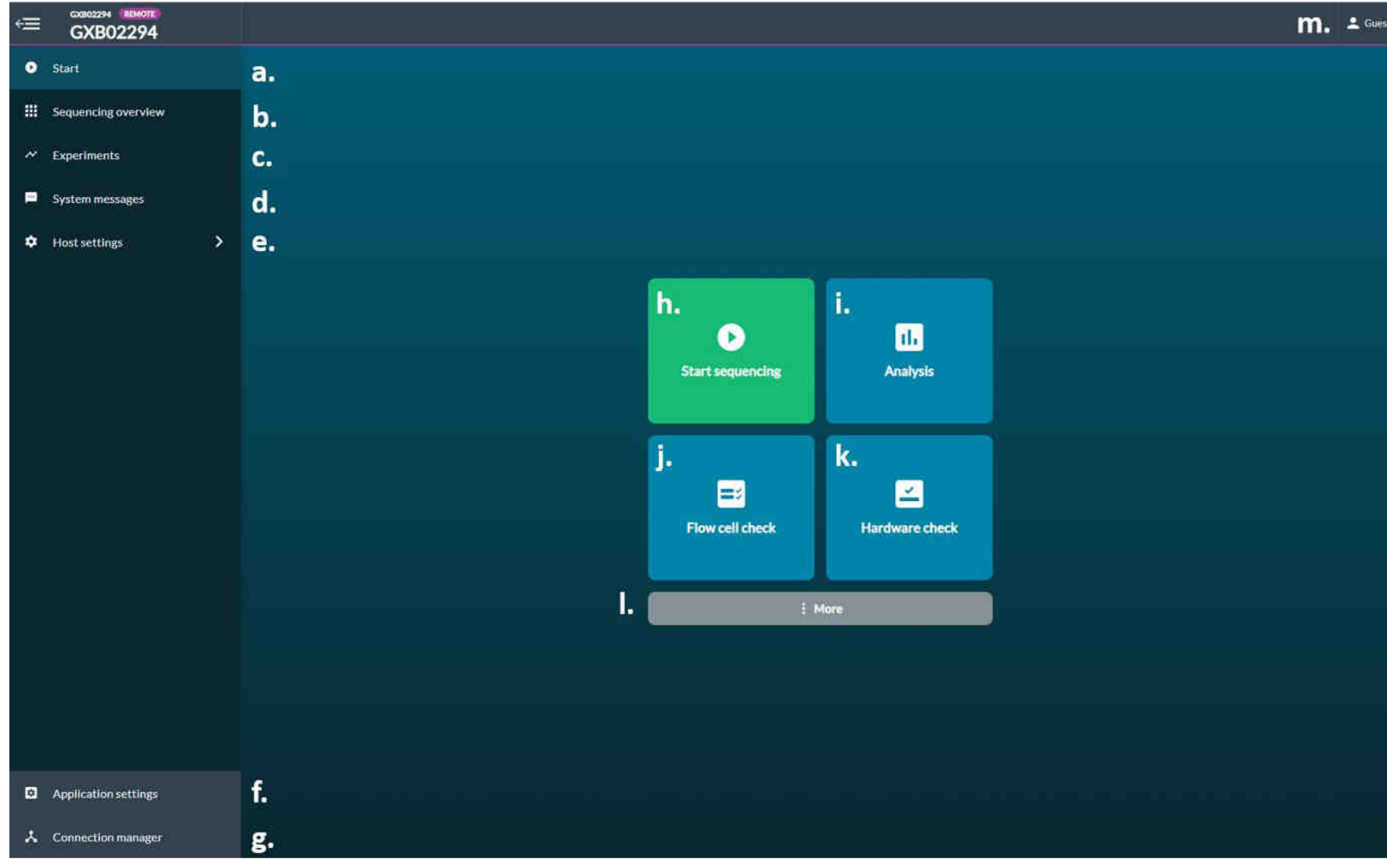
Flow diagram of the SARS-CoV-2 Whole-genome sequencing using Oxford Nanopore



Overview

- MinKNOW-Software
- Data formats
- Programs / work-flow used for data procession and analysis
    - Installation
    - Execution

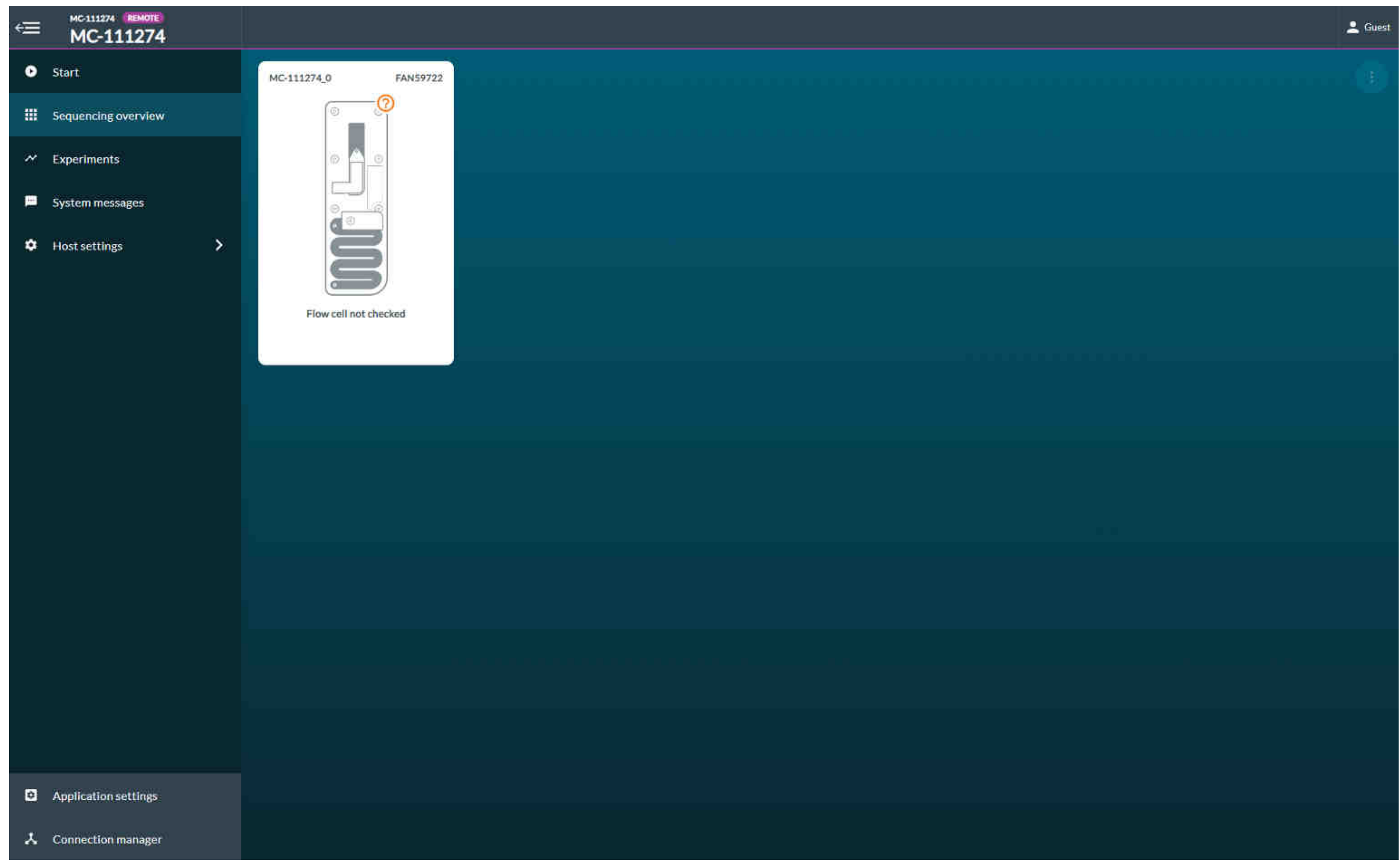
MinKNOW-Software



https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/homepage

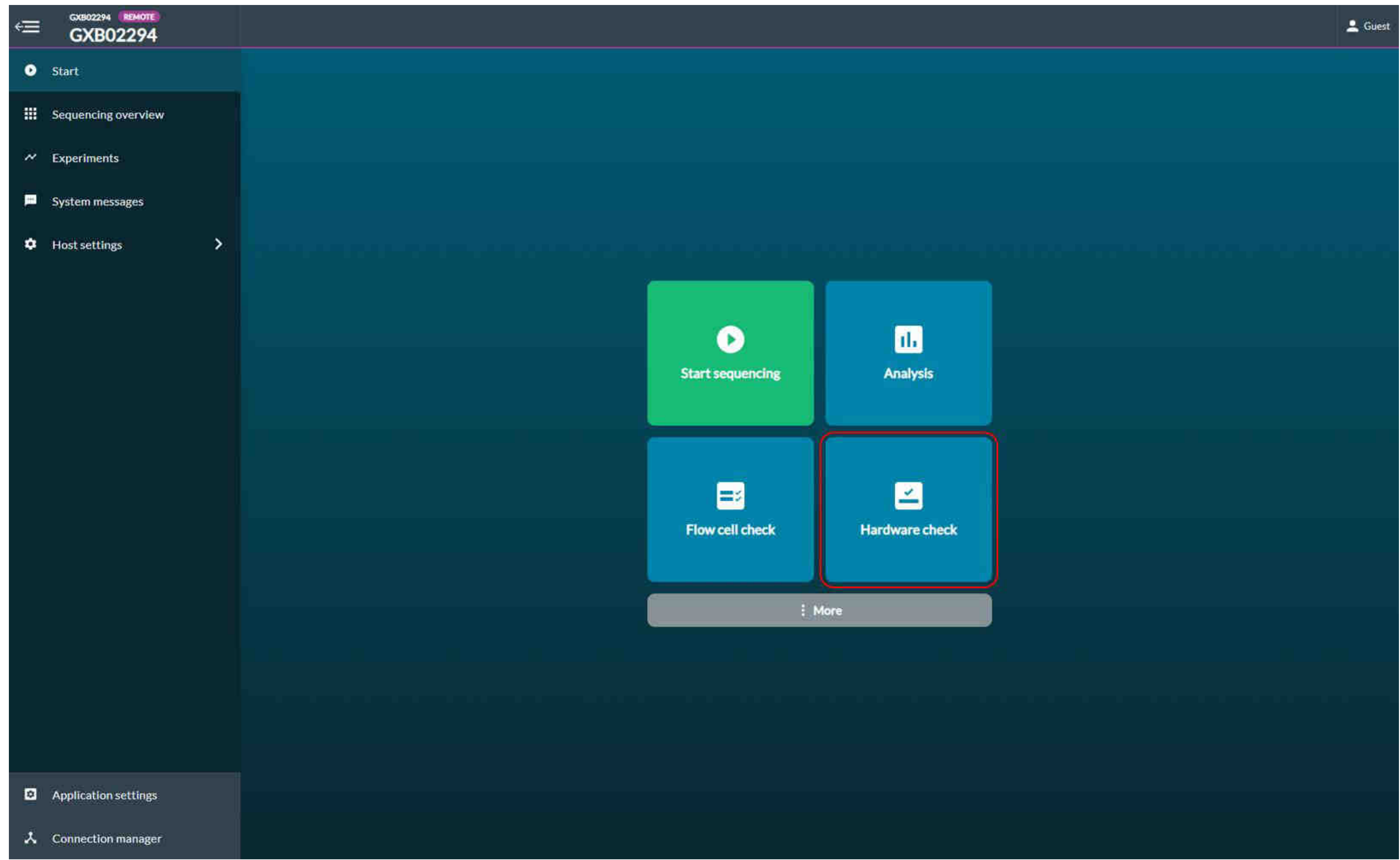
a. **Start** homepage
b. **Sequencing Overview** of connected flow cells
c. Recent and current **Experiments**
d. **System Messages**
e. **Host Settings**
f. **Application Settings** contain tutorials, account/login settings and MinKNOW GUI information
g. **Connection Manager** to connect with other available devices
h. **Start Sequencing** experiment
i. Post-run **Analysis**
j. **Flow Cell Check**
k. **Hardware Check**
l. **More** includes option to generate .mmi from .fasta file or to import a sample sheet
m. **Guest/initials** to logout

MinKNOW-Software



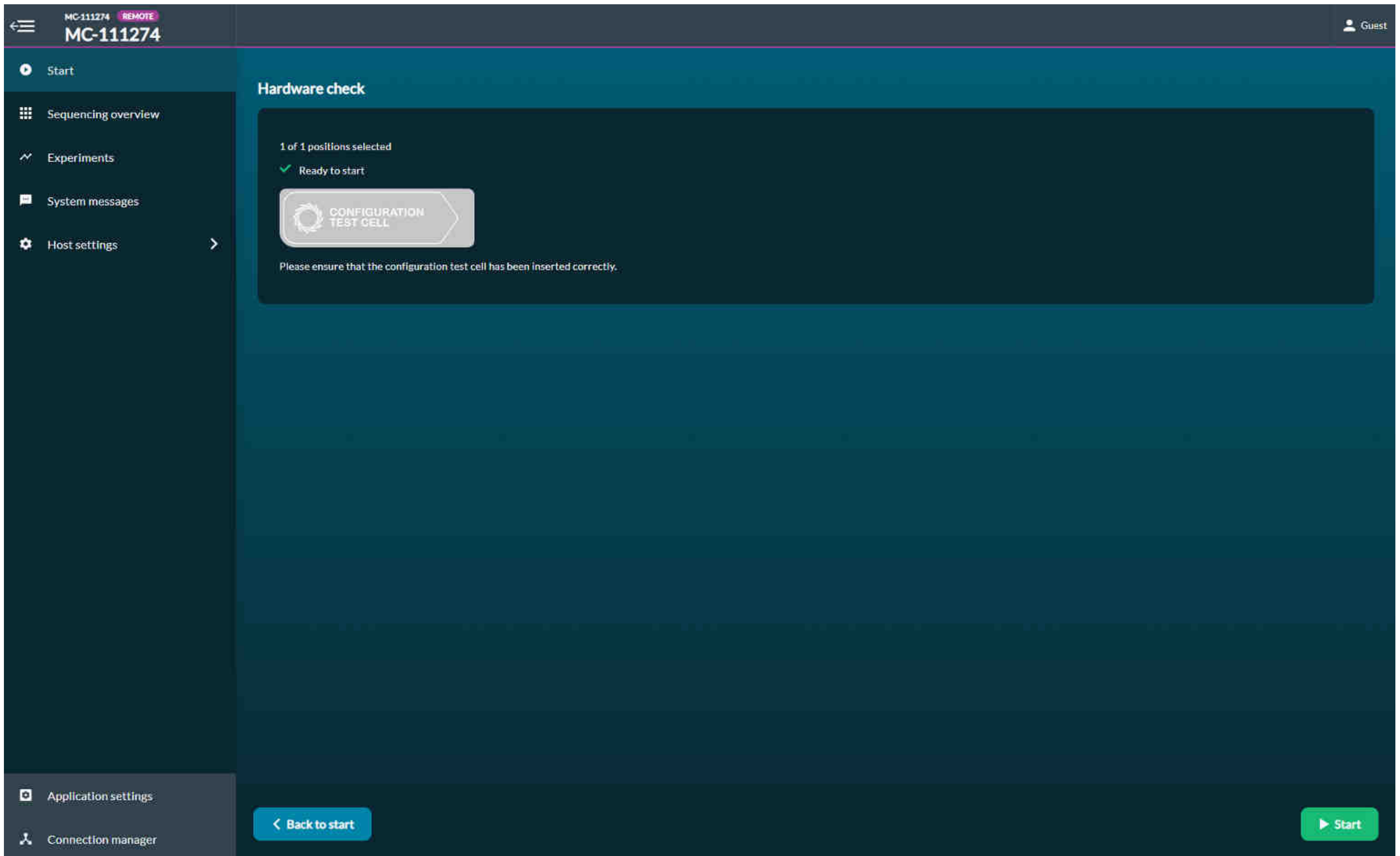
https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/homepage

MinKNOW-Software



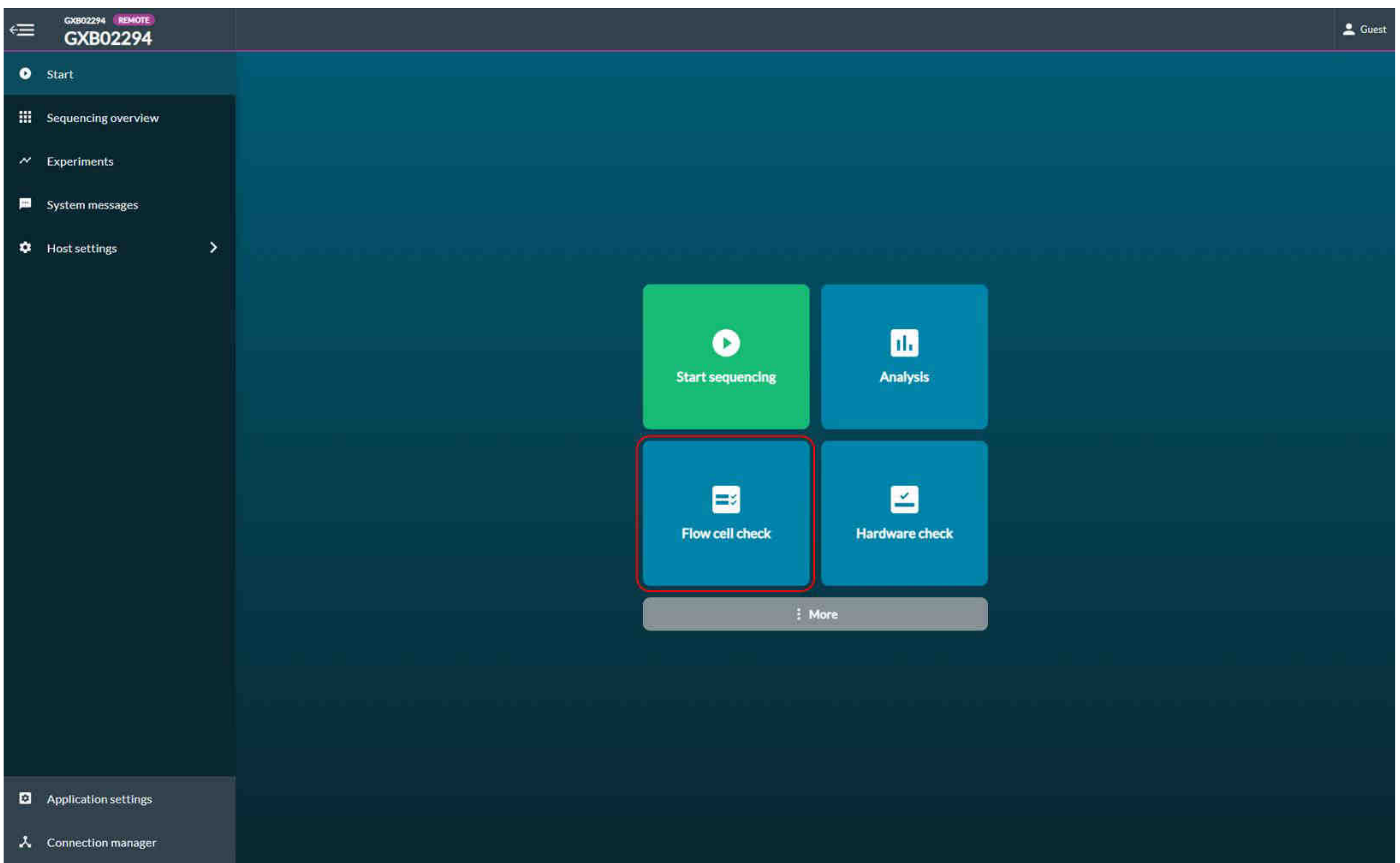
https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/homepage

MinKNOW-Software



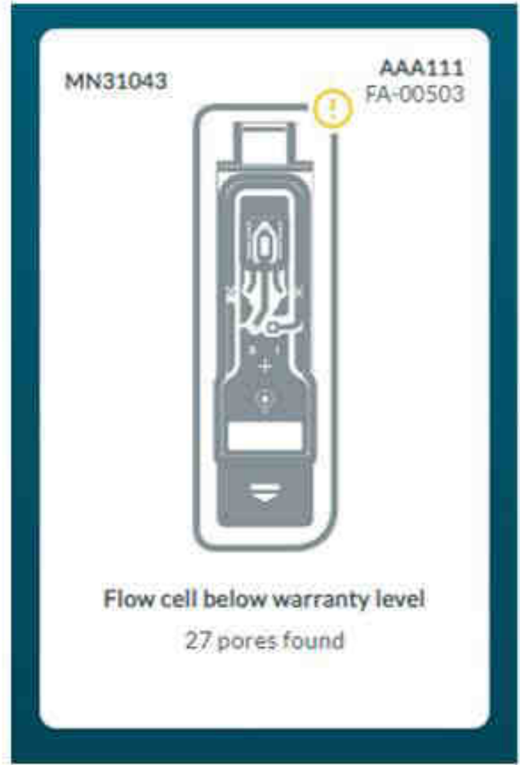
https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/homepage

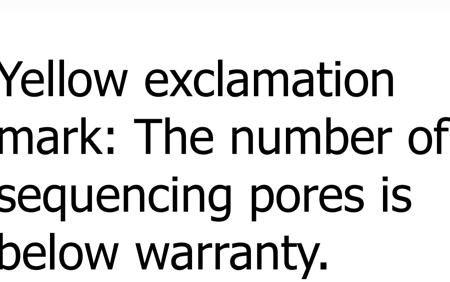
MinKNOW-Software

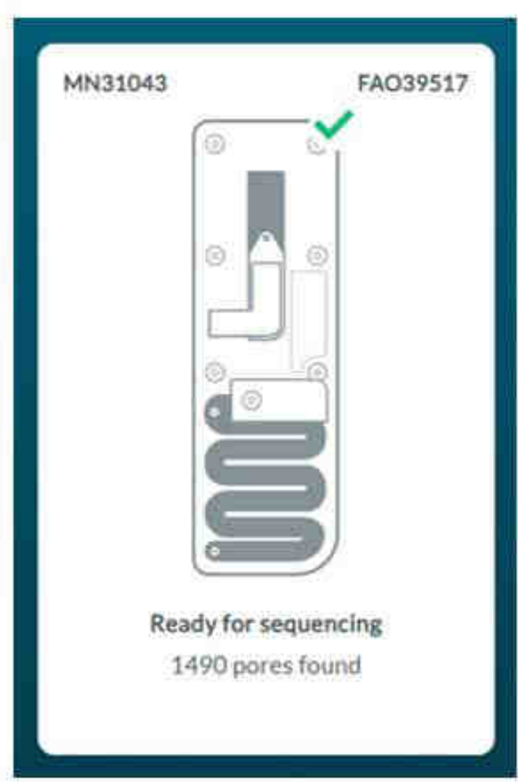


https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/homepage

MinKNOW-Software




Yellow exclamation mark: The number of sequencing pores is below warranty.

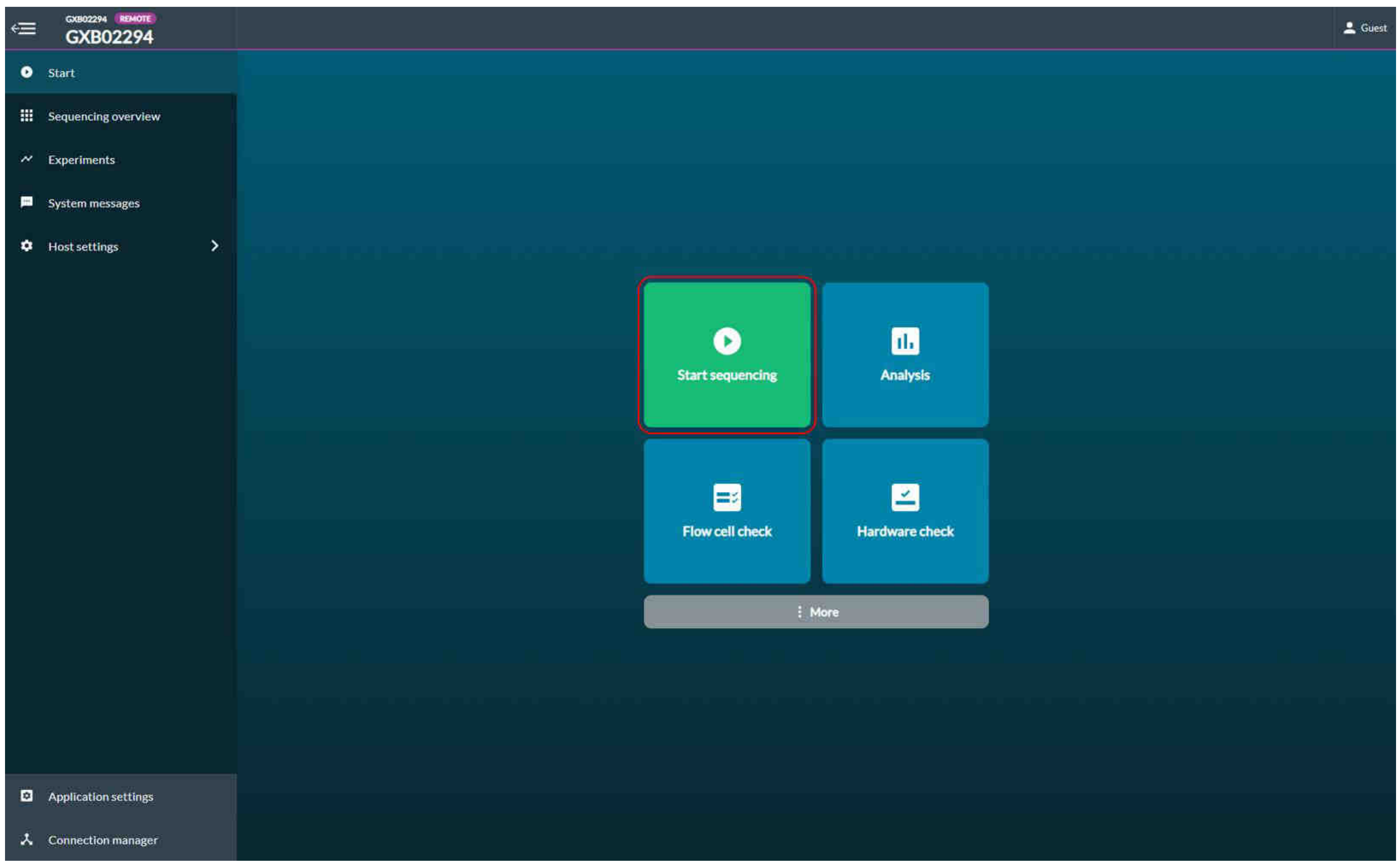


Green tick: The number of sequencing pores is above warranty and ready for sequencing.

- warranty for flow cells -> 800 nanopores or above checked within 5 days of receipt
- Sequencing can also be run with less than 800 nanopores.
- The lower the number of pores, the longer will it need to sequence.

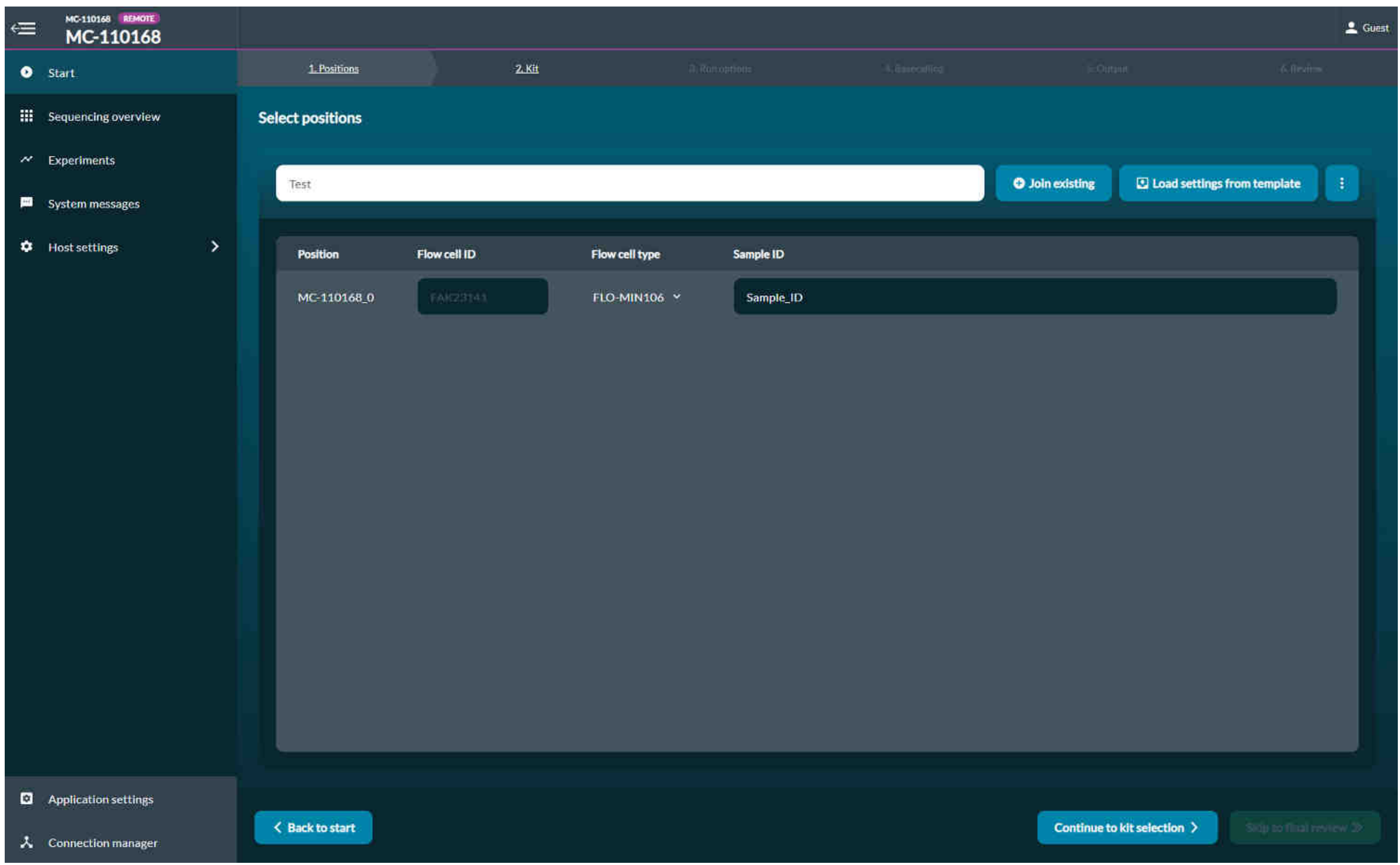
https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/homepage

MinKNOW-Software



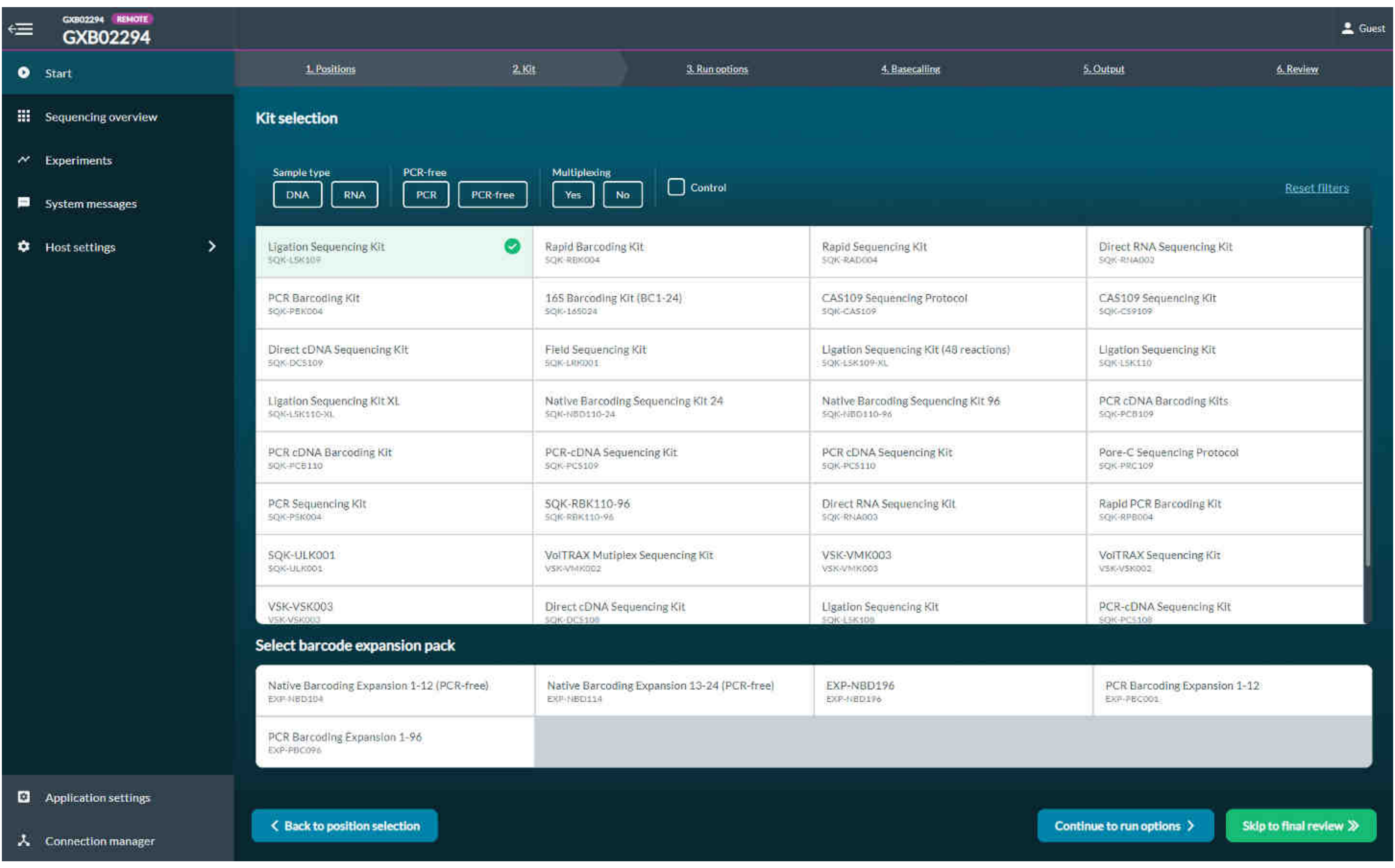
https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/starting-a-sequencing-run-on-minion

MinKNOW-Software



https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/starting-a-sequencing-run-on-minion

MinKNOW-Software

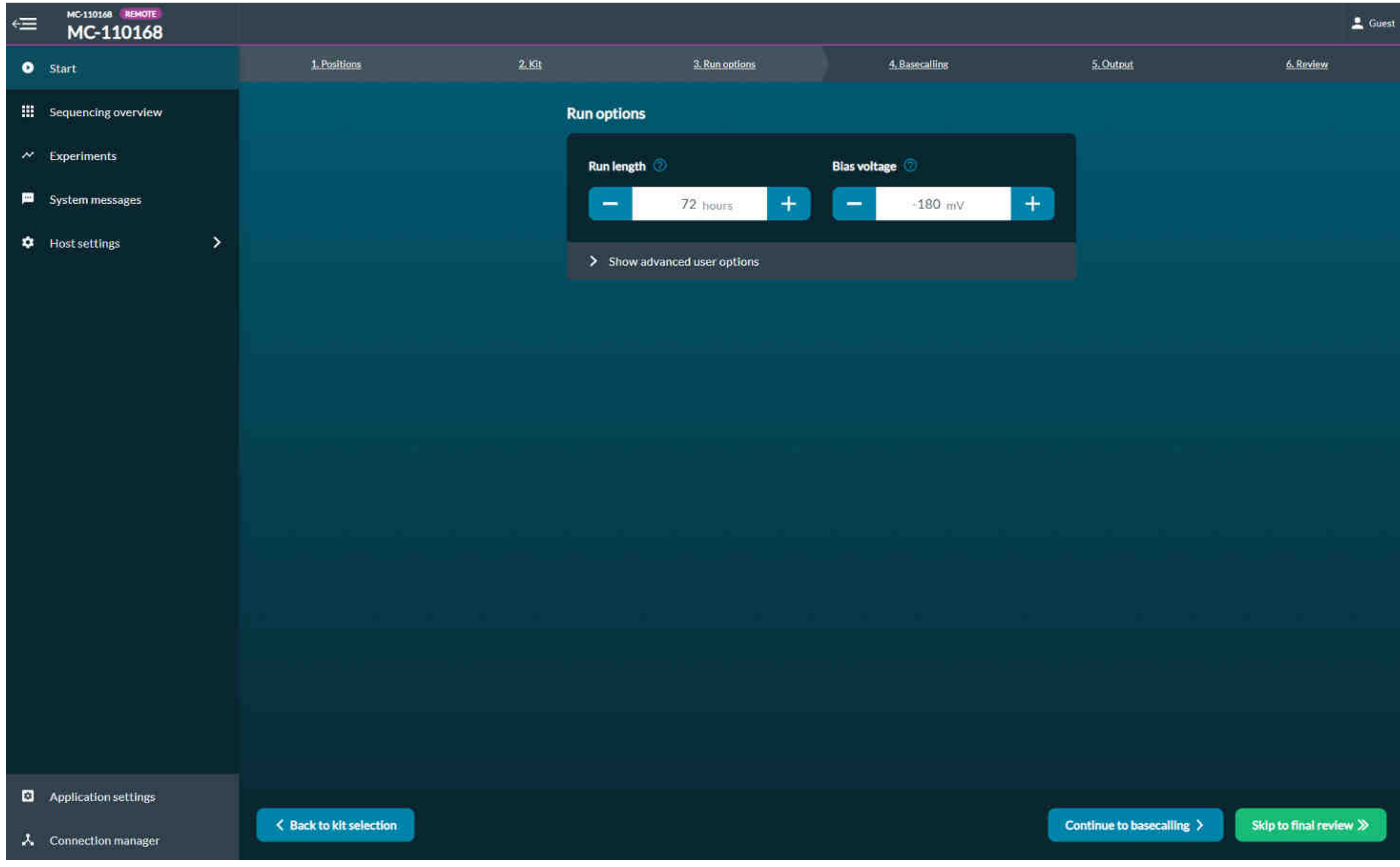


Select the kit!

For the training we use: „Rapid Barcoding-Kit“ SQK-RB004

https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/starting-a-sequencing-run-on-minion

MinKNOW-Software

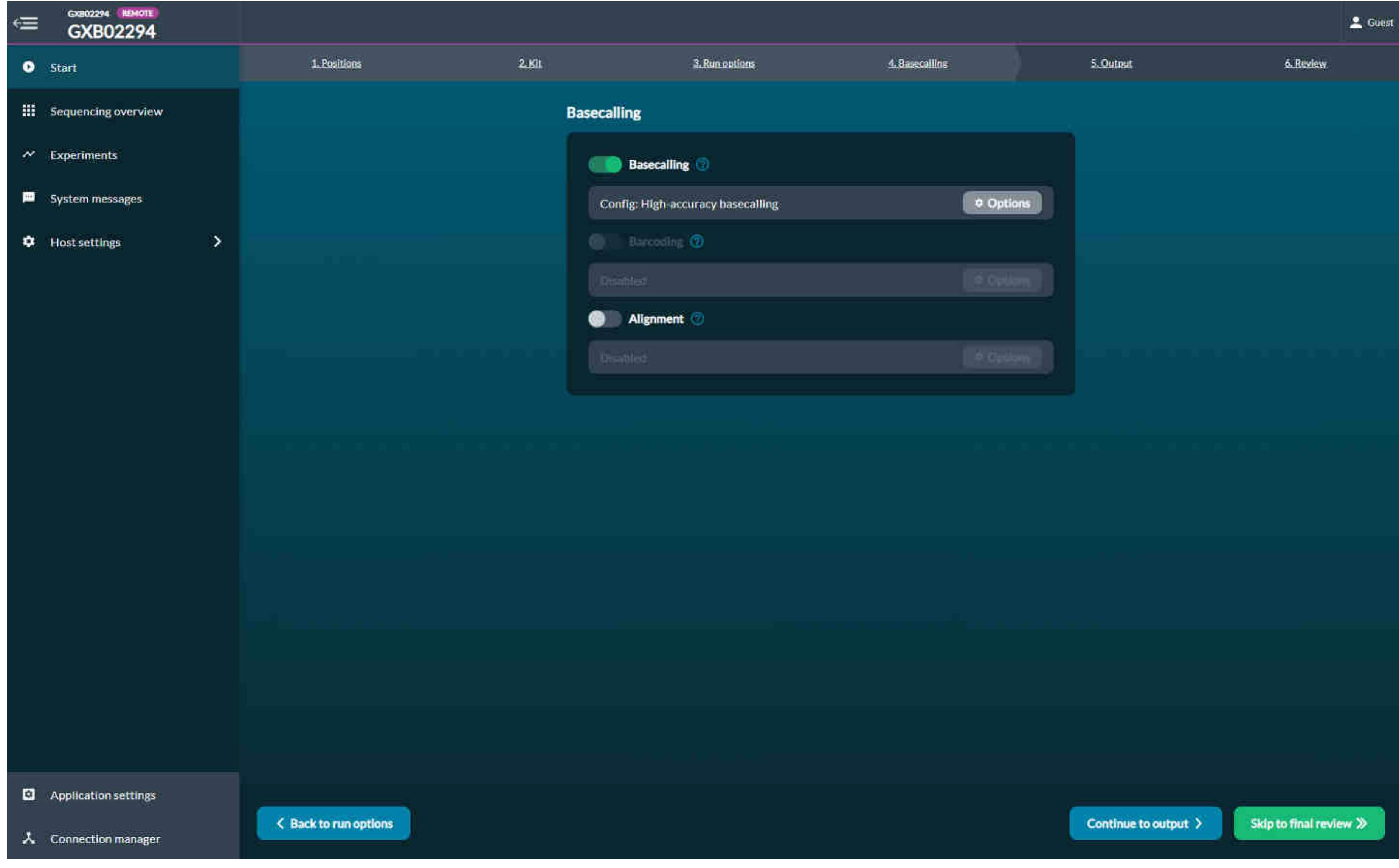


Set run length to 6 hours

(it can be stopped earlier)

https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/starting-a-sequencing-run-on-minion

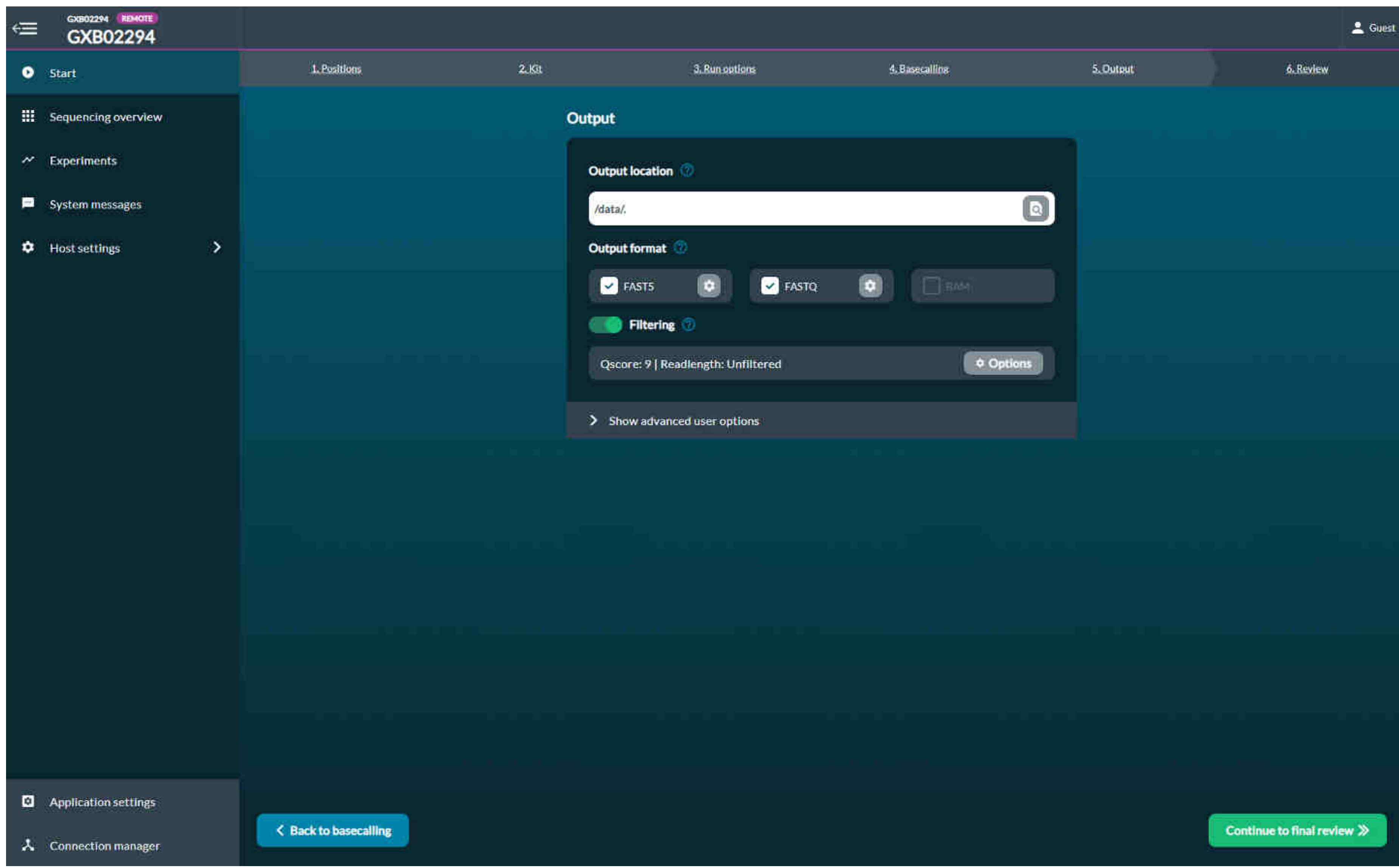
MinKNOW-Software



Select basecalling and barcoding

https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/starting-a-sequencing-run-on-minion

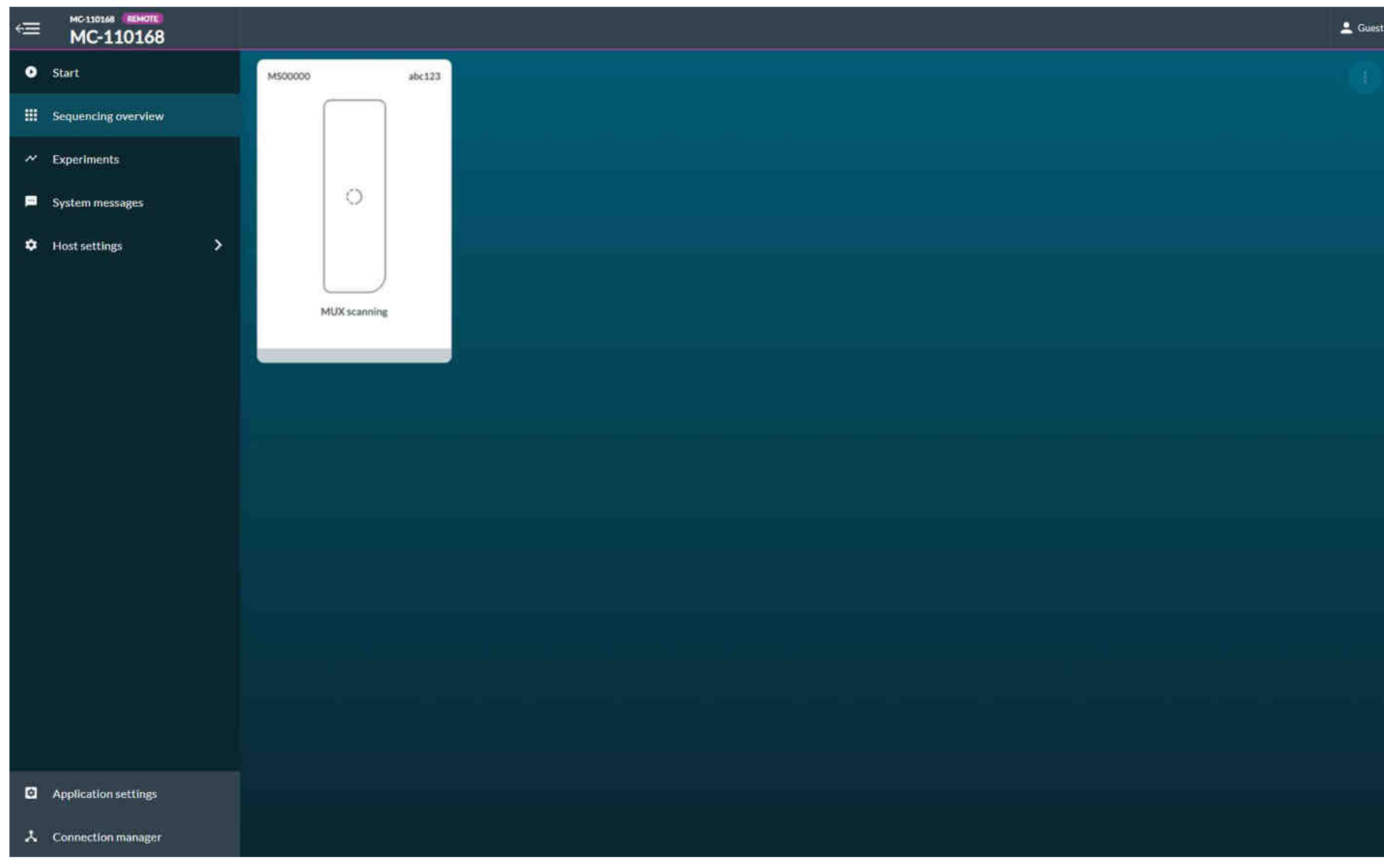
MinKNOW-Software



Switch off fastq-compression

https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/starting-a-sequencing-run-on-minion

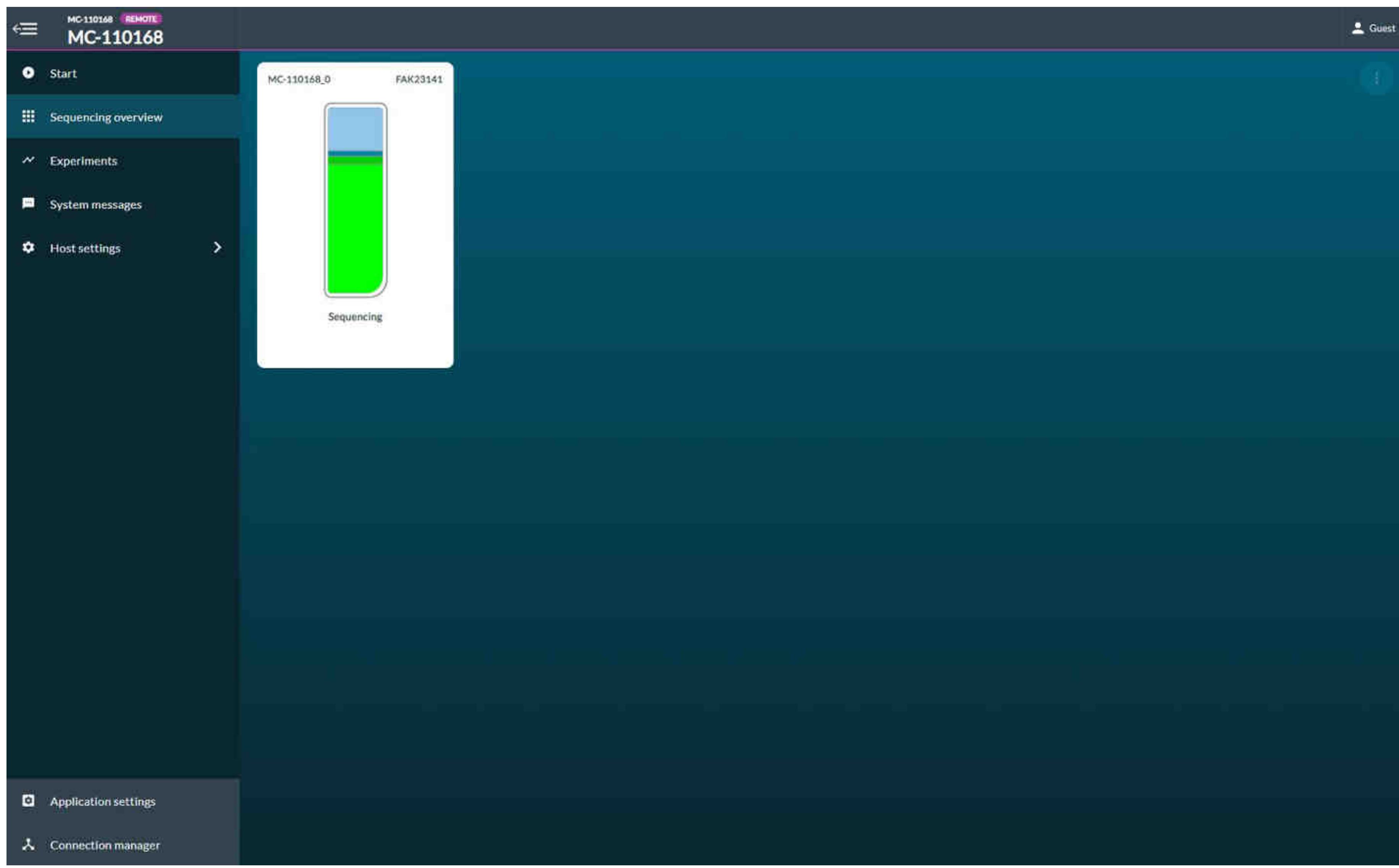
MinKNOW-Software



MUX-Scan: multiplex scan, allows MinKNOW to prioritise the order in which the nanopores are used, maximising the data output in the initial stages of the run.

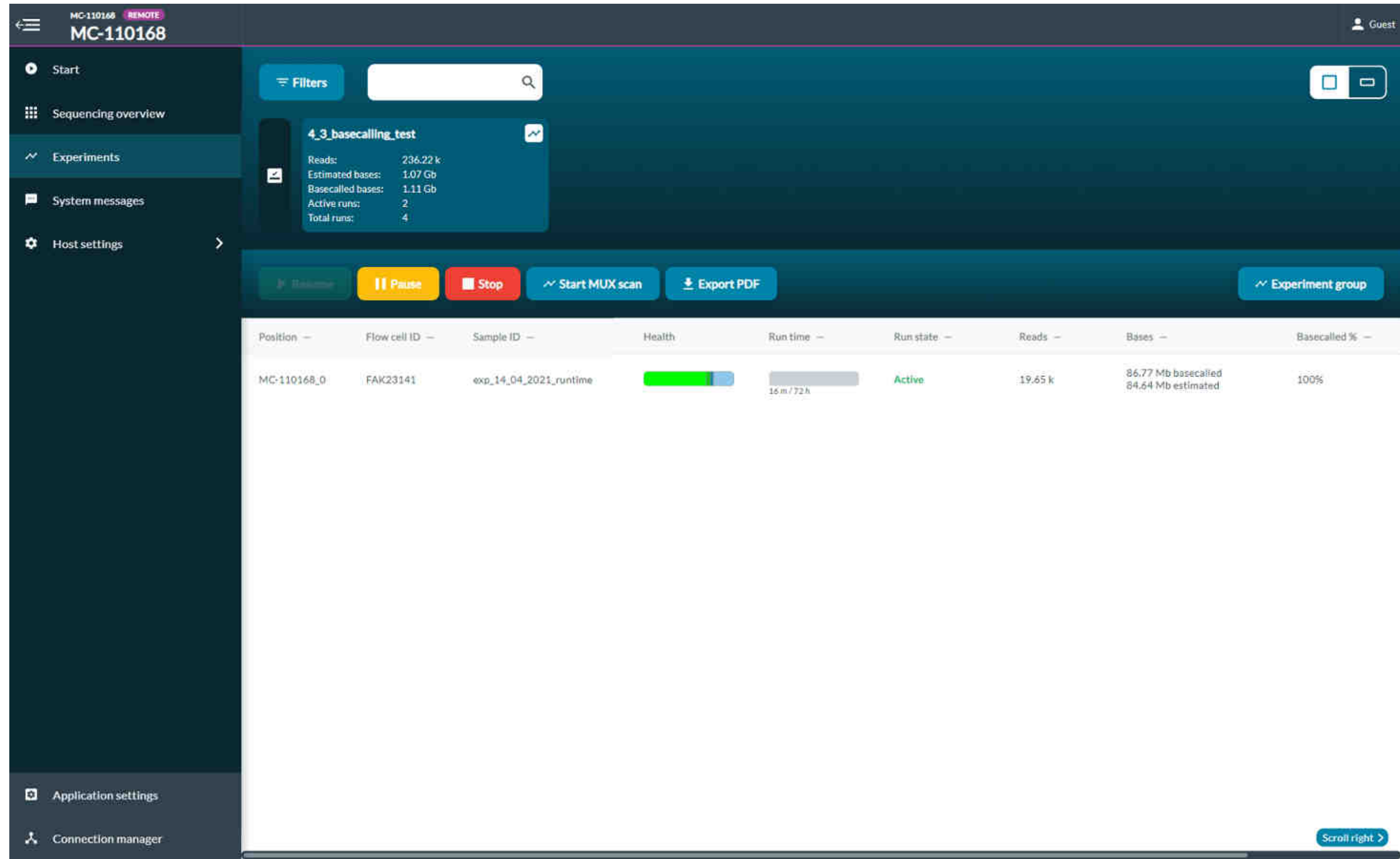
https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/homepage

MinKNOW-Software



https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/homepage

MinKNOW-Software

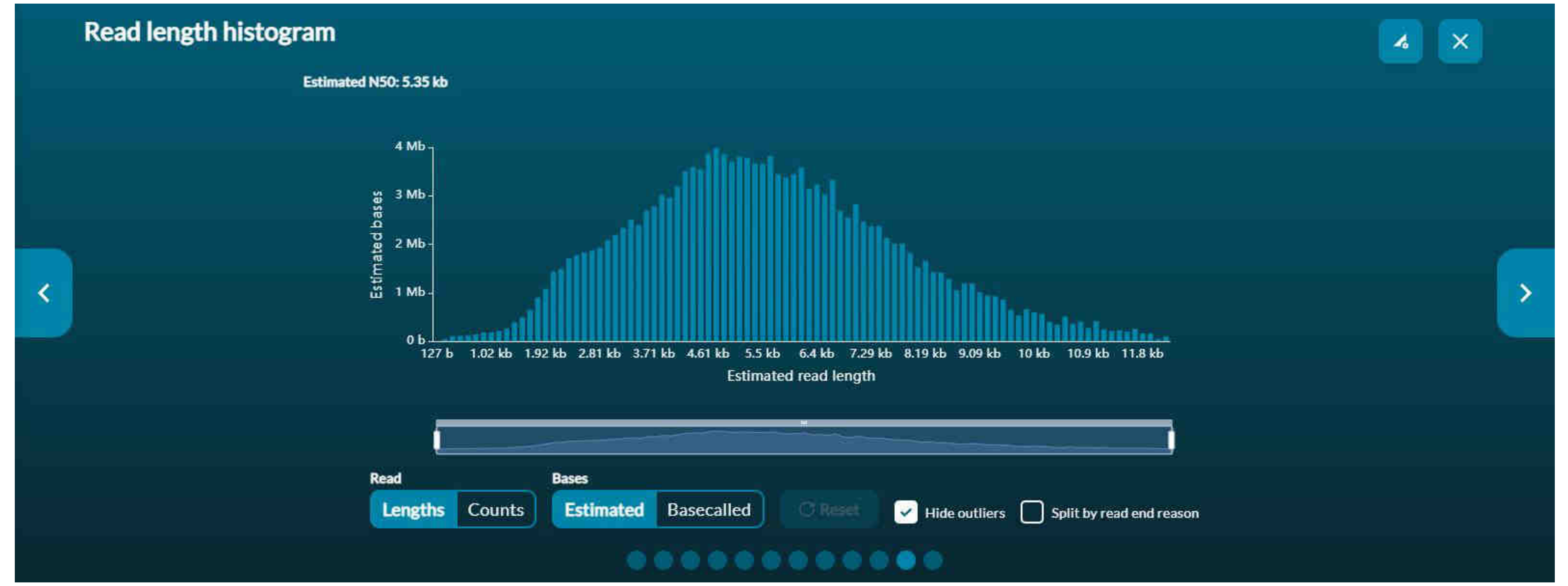


https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/homepage

MinKNOW-Software

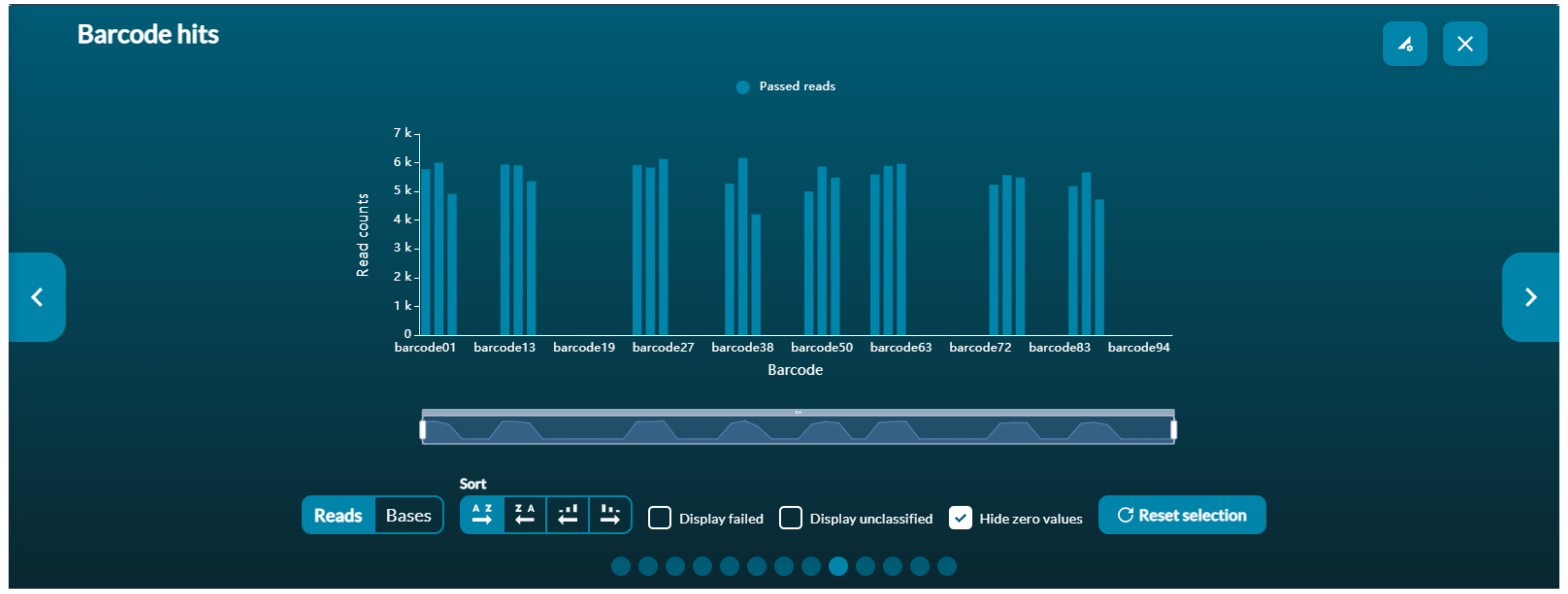
https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/progression-of-minknow-pro

MinKNOW-Software



https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/progression-of-minknow-pro

MinKNOW-Software



https://community.nanoporetech.com/protocols/experiment-companion-minknow/v/mke_1013_v1_revbp_11apr2016/progression-of-minknow-pro

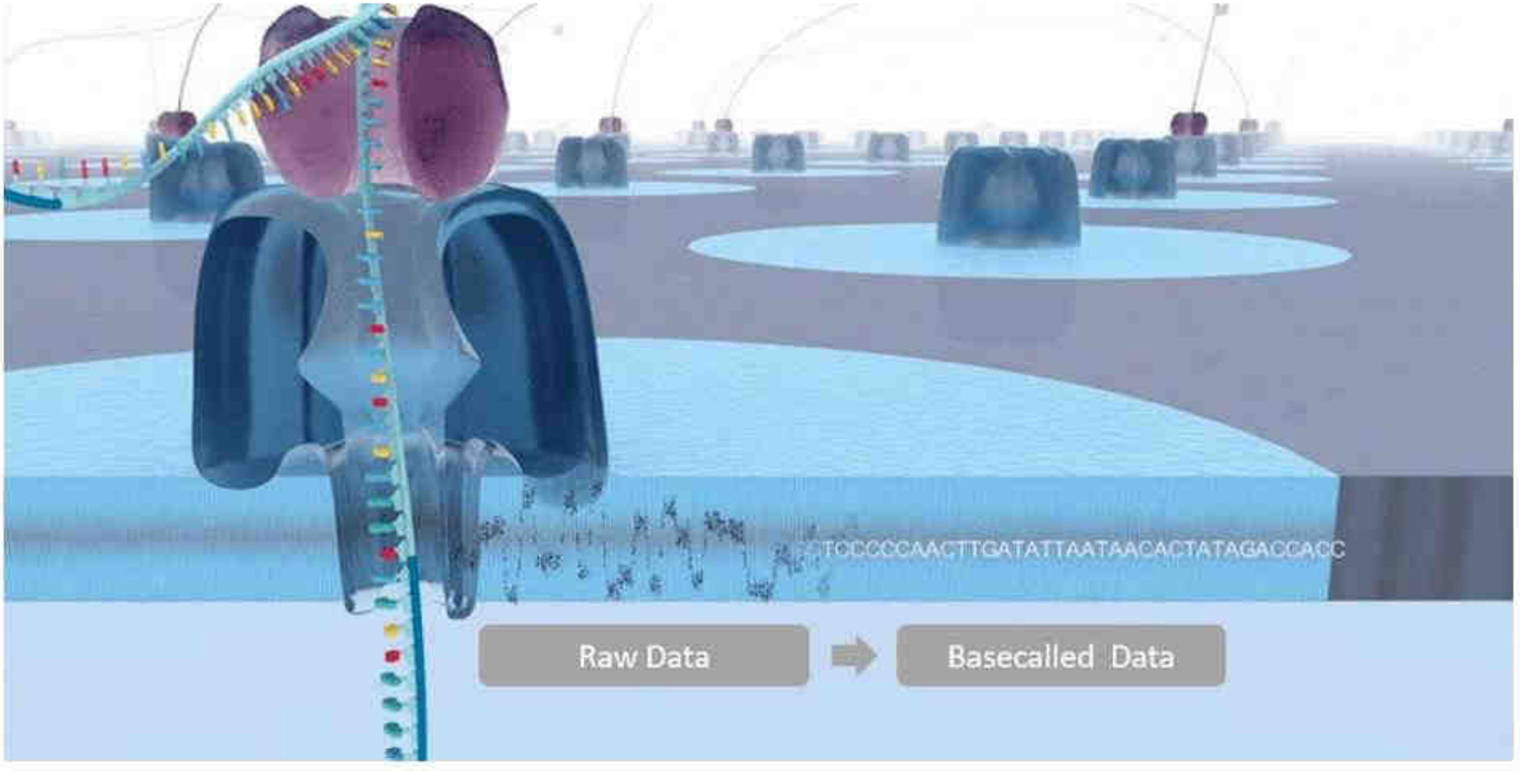
Data Formats – fast5

- raw-data
- contain all information that was generated while sequencing
- big size
- not human-readable

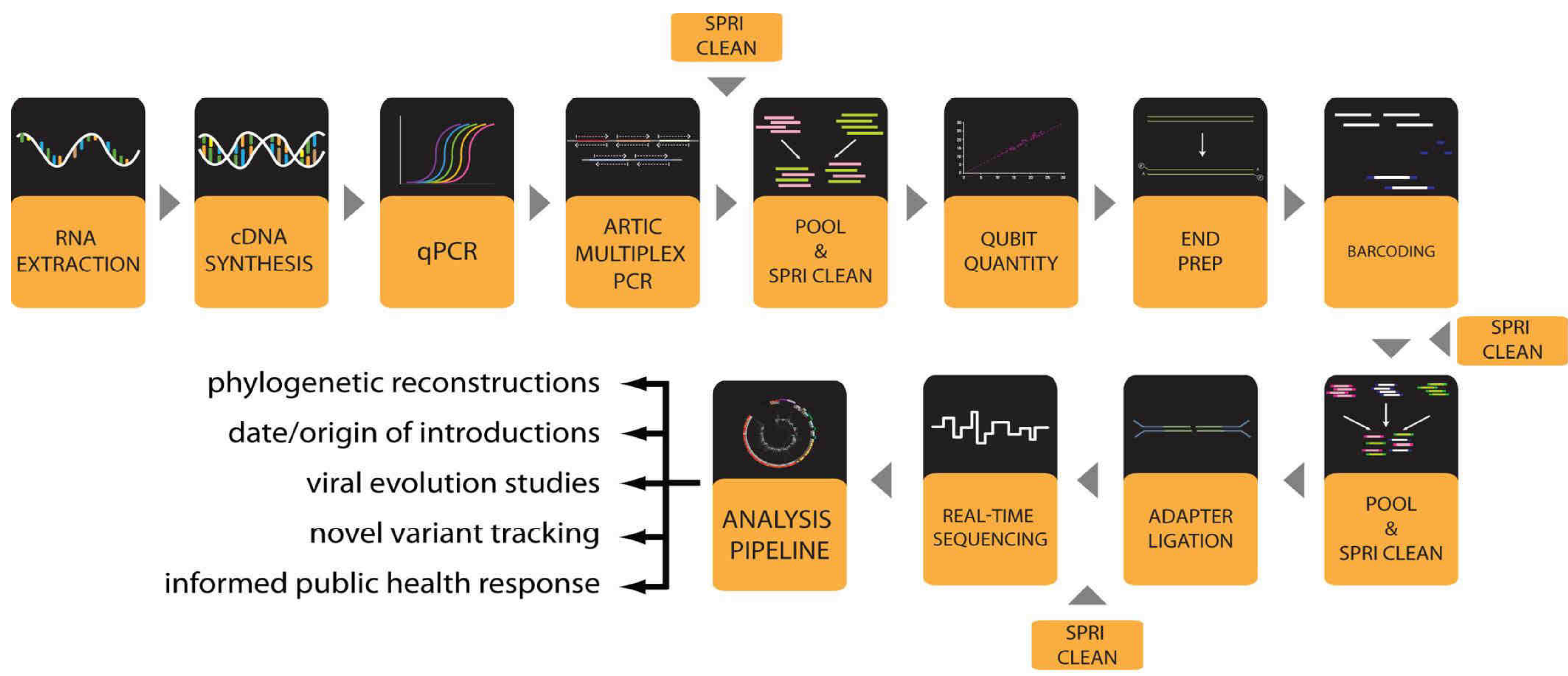
High accuracy basecalling with Guppy

- Toolkit by Oxford Nanopore, integrated in MinKNOW
- Basecalling algorithm based on a Recurrent Neural Network (RNN)

Basecalling is the process of converting the electrical signals generated by a DNA or RNA strand passing through the nanopore into the corresponding base sequence of the strand. The general data flow in a nanopore sequencing experiment is shown below.



https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revz_14dec2018/guppy-software-overview



RNA EXTRACTION
cDNA SYNTHESIS
qPCR
ARTIC MULTIPLEX PCR
SPRI CLEAN
POOL & SPRI CLEAN
QUBIT QUANTITY
END PREP
BARCODING
SPRI CLEAN
POOL & SPRI CLEAN
ADAPTER LIGATION
SPRI CLEAN
REAL-TIME SEQUENCING
ANALYSIS PIPELINE
phylogenetic reconstructions
date/origin of introductions
viral evolution studies
novel variant tracking
informed public health response

Overview

- Introduction
  - Who is who? Who knows what?

- How will the training be organized?
  - Detailed schedule
  - What is in the "additional information"?

- Overview of NGS-nanopore work-flow as described in the protocol

- Theory of nanopore-sequencing
  - Material + devices

What is in the additional information ?

1. Theory

- 1.1 SARS-CoV-2

  -> Presentation given on 15.09.2021 (pdf)

- 1.2 COVID-19 Genomic Epidemiology Toolkit by CDC

  -> https://www.cdc.gov/amd/training/covid-19-gen-epi-toolkit.html

- 1.3 Details about Nanopore Technology -> workshop-presentations (pdf)

  -> https://nanoporetech.com/how-nanopore-sequencing-works

  -> https://nanoporetech.com/nanopore-sequencing-data-analysis

- 1.4 Links SARS-CoV-2 sequencing principles videos Oxford Nanopore:

  -> https://nanoporetech.com/resource-centre/bioinformatics-workflows-sars-cov-2-raw-nanopore-reads-consensus-genomes-using

  -> https://www.youtube.com/watch?v=rYaFcDE-Ewg

What is in the additional information ?

1. Theory

- 1.5 Links to get familiar with technology and experiments that you may plan in future using NANOPORE

  -> https://nanoporetech.com/

  -> https://nanoporetech.com/nanopore-sequencing-data-analysis

  -> https://www.youtube.com/c/OxfordNanoporeTechnologies/playlists

What is in the additional information ?

2. Wet-lab

- 2.1 Protocol! (pdf)

- 2.2 Video links on Oxford Nanopore Flow cell introduction, Priming and loading flow cell

  -> https://www.youtube.com/watch?v=zC6lAtzqi_k

  -> https://www.youtube.com/watch?v=Pt-iaemrM88

  -> https://www.youtube.com/watch?v=IknVaEnuDz0&t=604s

- 2.3. Flow cell wash protocol (pdf)

- 2.4 Methods manuscript, literatur protocol (pdf)

What is in the additional information ?

3. Dry-lab

3.1 Protocol! (pdf)

Protocol NGS_WGS_SARS-CoV-2_1200bp amplicons using Nanopore_V1.2.pdf

3.2 Introduction to bioinformatics

-> https://www.youtube.com/c/SimonCockell/videos

3.3 Linux

-> linuxcommands (pdf)

-> Links on Linux operating system

https://www.youtube.com/watch?v=v_1zB2WNN14

https://www.youtube.com/watch?v=V1y-mbWM3B8Ubuntu

https://www.youtube.com/watch?v=x5MhydijWmc

https://www.youtube.com/watch?v=G36I1iqDZig

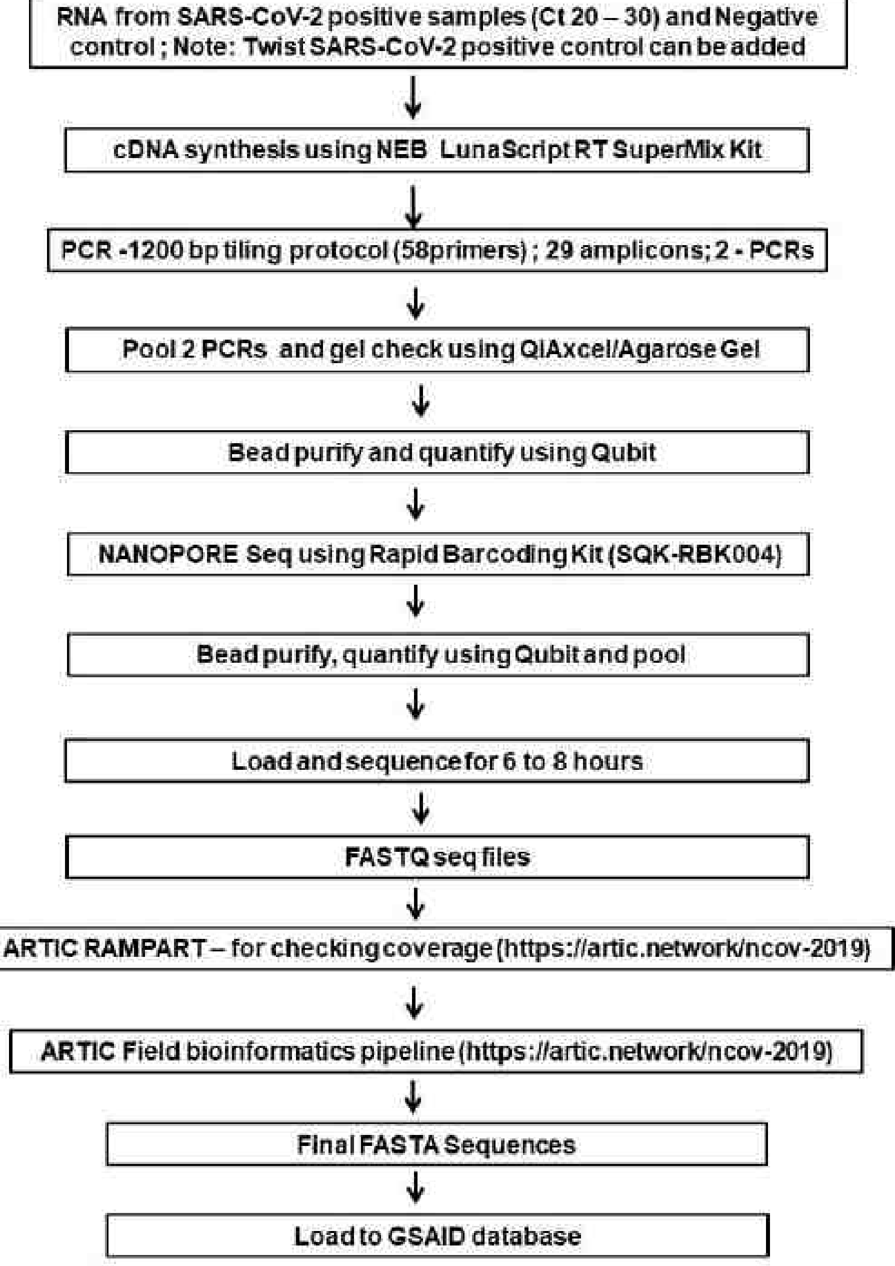
Workshop
Molecular Diagnostic Techniques
(Genomics and Bioinformatics – Infectious diseases)
Module-4

Dr. Srinivas Reddy Pallerla, Mr. Tung Tran and Ms. Linh Le
Prof. Dr. Thirumalaisamy P. Velavan, University Hospital, Tübingen, Germany
Prof. Dr. Daniel Olusola Ojurongbe, Nigeria
Prof. Dr Mohamed Osman, University of Khartoum, Sudan

29 July 2022

Training Sponsor: *Alexander von Humboldt Foundation*

NGS-nanopore WORK-FLOW Overview



What is in the additional information ?

3. Dry-lab

3.1 Protocol! (pdf)

Protocol NGS_WGS_SARS-CoV-2_1200bp amplicons using Nanopore_V1.2.pdf

3.2 Introduction to bioinformatics

-> https://www.youtube.com/c/SimonCockell/videos

3.3 Linux

-> linuxcommands (pdf)

-> Links on Linux operating system

https://www.youtube.com/watch?v=v_1zB2WNN14

https://www.youtube.com/watch?v=V1y-mbWM3B8Ubuntu

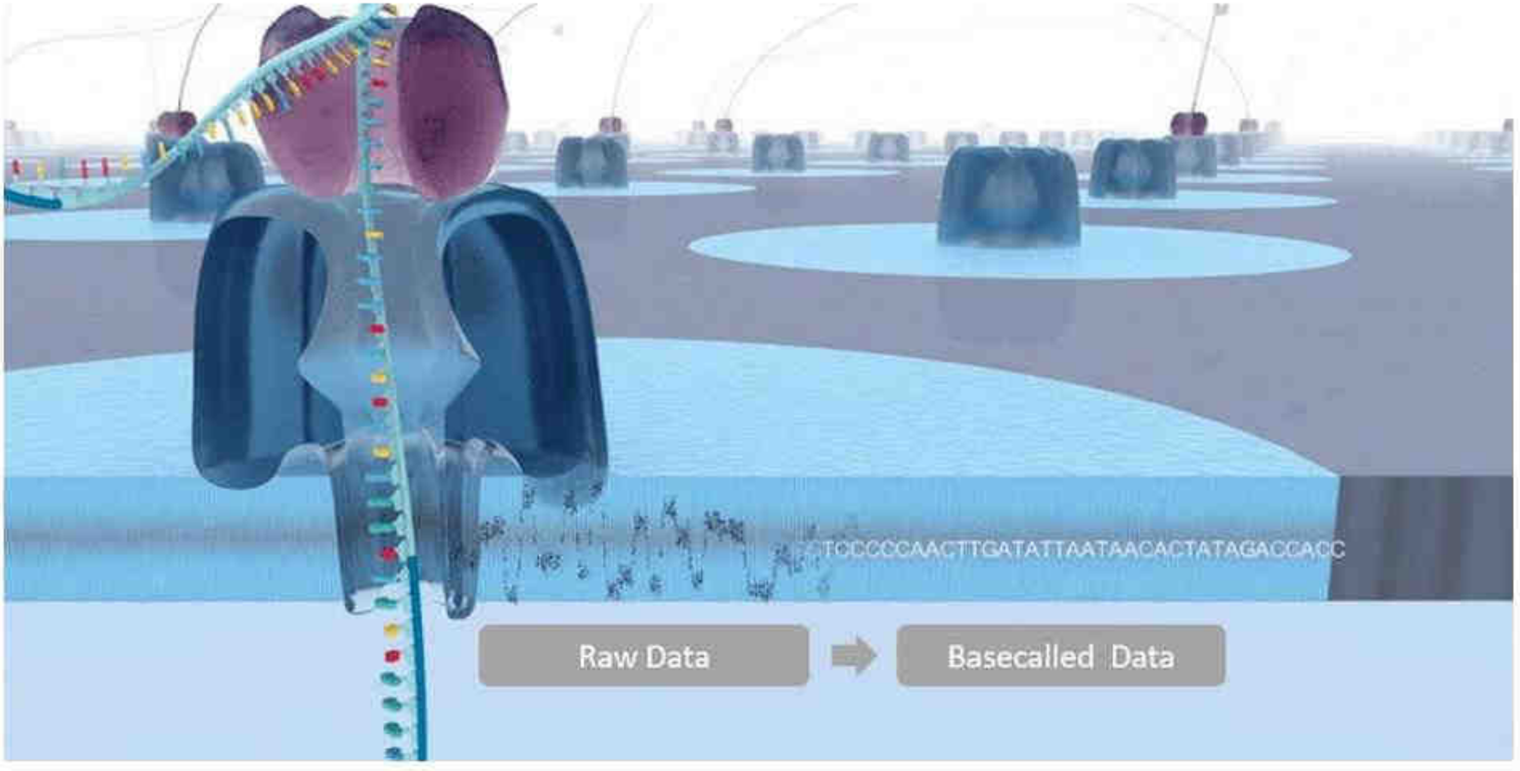
https://www.youtube.com/watch?v=x5MhydijWmc

https://www.youtube.com/watch?v=G36I1iqDZig

High accuracy basecalling with Guppy

- Toolkit by Oxford Nanopore, integrated in MinKNOW
- Basecalling algorithm based on a Recurrent Neural Network (RNN)

Basecalling is the process of converting the electrical signals generated by a DNA or RNA strand passing through the nanopore into the corresponding base sequence of the strand. The general data flow in a nanopore sequencing experiment is shown below.



https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revz_14dec2018/guppy-software-overview

Programs / work-flow – fieldbioinformatics

For further analysis, you have to merge all fasta-files into one, this can either be done manually or with the script below.

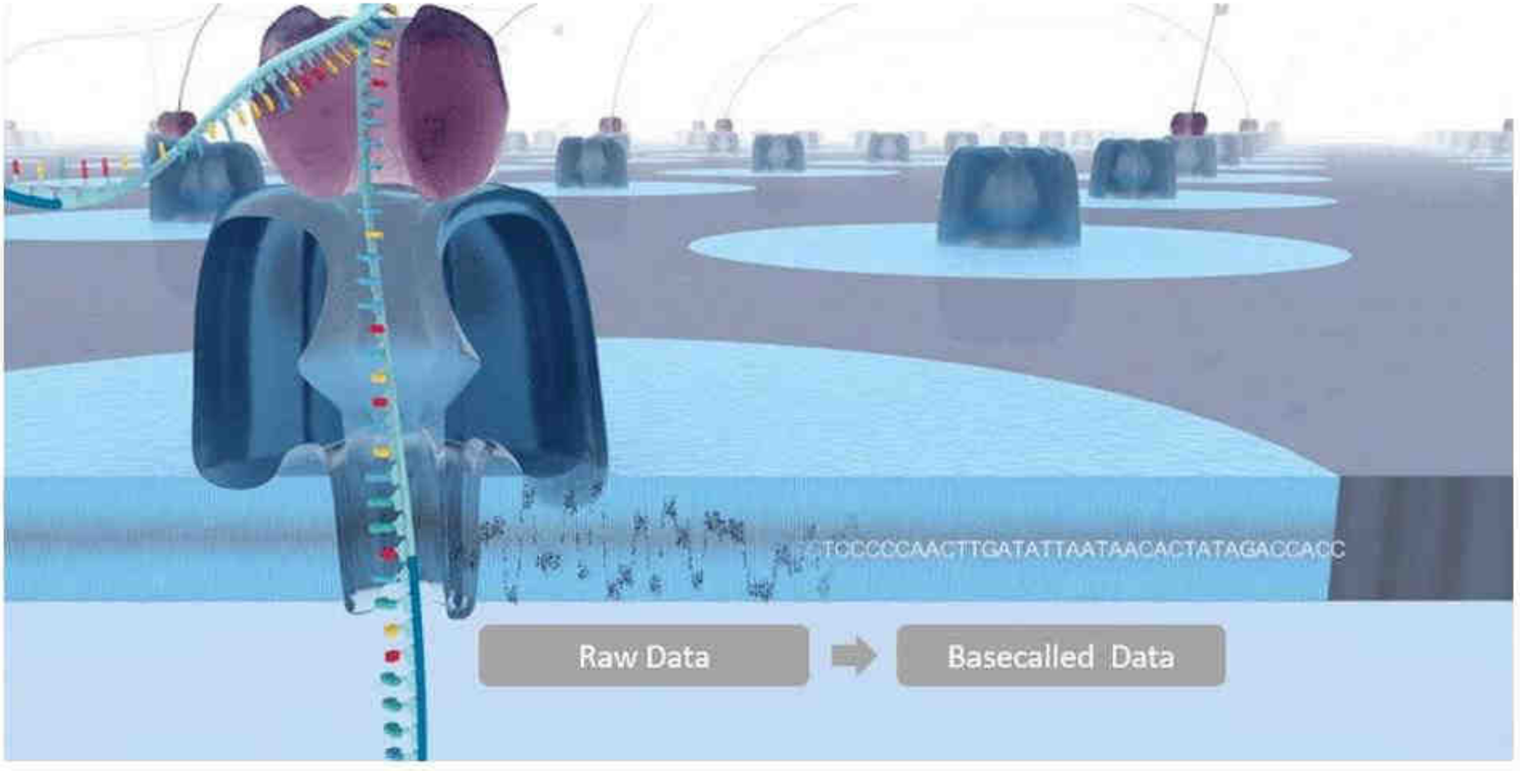
(the script is also in the dropbox, you have to edit in input-path)

```
GNU nano 4.8                                        merge_consensus_files.sh
#!/bin/bash

# create file

touch all_barcodes_merged.fasta
echo "file created"

# add consensus files

for N in {07..11};
        do
        cat ~/data/data_210820/fastq_pass_high_accuracy/barcode$N/SARS-CoV_high_accuracy_Barcode$N/SARS-CoV_high_accuracy_Barcode$N.consensus.fasta \
         >> all_barcodes_merged.fasta
        echo "done adding Barcode$N"
        done
```

High accuracy basecalling with Guppy

- Toolkit by Oxford Nanopore, integrated in MinKNOW
- Basecalling algorithm based on a Recurrent Neural Network (RNN)

Basecalling is the process of converting the electrical signals generated by a DNA or RNA strand passing through the nanopore into the corresponding base sequence of the strand. The general data flow in a nanopore sequencing experiment is shown below.



https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revz_14dec2018/guppy-software-overview

High accuracy basecalling with Guppy

Toolkit:

- Basecaller
    - Calibration strand detection
    - Adapter trimming
- Barcoding/demultiplexing
- Alignment
- Modified basecalling

https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revz_14dec2018/guppy-software-overview

High accuracy basecalling with Guppy

System-requirements:

- 4 GB RAM plus 1 GB per thread for 1D basecalling
- 4 GB RAM plus 2 GB per thread for 1D2 basecalling
- Administrator access for .deb or .msi installers
- 2 GB of drive space for installation, minimum 512 GB storage space for basecalled read files (1 TB recommended)

- CUDA 10 for Linux 4 Tegra running Ubuntu 18

More details can be checked at the Oxford Nanopore website, GPU is recommended

https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revz_14dec2018/guppy-software-overview

High accuracy basecalling with Guppy

Different models:

“The Fast model is designed to keep up with data generation on Oxford Nanopore devices (MinION Mk1C, GridION, PromethION). The HAC model provides a higher raw read accuracy than the Fast model and is currently 5-8 times more computationally-intensive. The Super accurate model has an even higher raw read accuracy, and is ~3 times more intensive than the HAC model.”

Model	R9.4.1 modal accuracy	R10.3 modal accuracy
Fast	95.8	91.4
HAC	97.8	95.7
sup	98.3	97.5

https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revz_14dec2018/guppy-software-overview

High accuracy basecalling with Guppy

Different models:

Device	Basecalling speed in Gbases per hour			Keep-up, number of flow cells
	Fast model	HAC model	sup model	Fast model
PromethION P24	82	19	5.2	23
GridION	34	5	2	51
MinION Mk1C	0.8	0.07	-	1.25
MinION CPU, high-spec laptop	0.13	0.014	0.003	0.21
MinION GPU, RTX2070 laptop	14	1.6	0.06	22

"Keep up" is defined as 80% of the theoretical flow cell output.

e.g. MinION = 4000 kHz x 512 channels x 0.8 = 1.6 M samples/s = 160 kbases/s at 400 b/s

https://community.nanoporetech.com/protocols/Guppy-protocol/v/gpb_2003_v1_revz_14dec2018/guppy-software-overview

High accuracy basecalling with Guppy

Command line – program

```
vgcare@vgcare: ~
File  Edit  View  Search  Terminal  Help
(base) vgcare@vgcare:~$ guppy_basecaller --input_path ~/test_data/fast5_barcode07 --save_path ~/test_data/fast5_barcode07
 --flowcel FLO-MIN106 --kit SQK-RBK004 -x "cuda:0"
```

"guppy_basecaller" – program command
"- - input_path" – argument, path to input data
"- - save_path" – argument, where to store output data
"- - flowcel" – argument, flowcell model
"- - kit" – argument, barcoding kit used
"- x" – argument, use GPU

High accuracy basecalling with Guppy

Command line – program

```
vgcare@vgcare: ~
File  Edit  View  Search  Terminal  Help
(base) vgcare@vgcare:~$ guppy_basecaller --input_path ~/test_data/fast5_barcode07
 --save_path ~/test_data/fast5_barcode07 --flowcel FLO-MIN106 --kit SQK-RBK004 -x
 "cuda:0"
ONT Guppy basecalling software version 5.0.14+8f53ee955
config file:        /opt/ont/guppy/data/dna_r9.4.1_450bps_hac.cfg
model file:         /opt/ont/guppy/data/template_r9.4.1_450bps_hac.jsn
input path:         /home/vgcare/test_data/fast5_barcode07
save path:          /home/vgcare/test_data/fast5_barcode07
chunk size:         2000
chunks per runner:  256
minimum qscore:     9
records per file:   4000
num basecallers:    4
gpu device:         cuda:0
kernel path:
runners per device: 4

Found 8 fast5 files to process.
Init time: 584 ms

0%   10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
```

Data Formats – fast5

- raw-data
- contain all information that was generated while sequencing
- big size
- not human-readable

Data Formats – fastq

- q = quality
- Information about sequencing run and quality score
- Human-readable but complicated to read

```
@42b099dc-b20b-4f3d-af60-d859ebc8d976 runid=a55ad8caa1c516614cc04cc21c9dbdb356b66c41 read=101 ch=58 start_time=2021-08-20T13:47:46Z flow_cell_id=FAP81457 protocol_group_id=1234 sample_id=no_sa
GTTGTACTTCGTTCAGTTTGAATTTGGGTGTTCTACGTGTTTGGTGCGTGGGAATGAATCCTTGTTTTCGCATTTATCGTGAAACGCTTTCGCGTTTTTGGTGCGCCGCTTCATCTTGTGCAGAATGAATTCTCCCGTAACTACGTAGCACAAGTAGATGTAGTTAACTTTAATCTCCTGGCAATCTTTAAT
+
&'(.0,,-2.&).4''',+&$&$&+**3.]D8,%%&(,//-*'&&&$%(*//39/+%&%.22>;<>FC@+,,,,65582/--8<;44564617336810(')(*)*+)&+004:;9;//155(787.+,+,/+3'&+232..04+++*+*+,&&0599:;81/2442422100>==],-/,...669>?=89
@129269c0-d564-4052-8359-9c2a91f80a43 runid=a55ad8caa1c516614cc04cc21c9dbdb356b66c41 read=65 ch=316 start_time=2021-08-20T13:47:46Z flow_cell_id=FAP81457 protocol_group_id=1234 sample_id=no_sa
GTACTTCACGTTCAATCAGTTGGGTGTTTAACATTTACCGTGGGAATGGAATCCTTGATGGGGTGAAGCAACTTTCGCGTTTTTCGTGCGCCCTTCAGCATTTGCAGTACAAAAGACATACTGTTCTAATGTTGAATTCACTTTGAATTTATCAAAACACTCTACACGAGCACGTGCAGGTATAATTCTACT
+
$'&&&0,&%'+00+)))+''(+))+]63,(&&$$%0''''(+1864<0:427('&&&(''))+,.03**(((-,+&)*,/9:961368-,,-&--+-00<=9'&&+,..,-?=@;9<5,+**;>A+./:><=@:=<7=:9;:65232348<?4>?=;+,.86//*)'+-+.045-.-044:]]+297624
@aa9e903f-69cf-415b-8d4e-8af728fb8cc9 runid=a55ad8caa1c516614cc04cc21c9dbdb356b66c41 read=52 ch=390 start_time=2021-08-20T13:47:46Z flow_cell_id=FAP81457 protocol_group_id=1234 sample_id=no_sa
GTTGTACTTCGTTCGATTTCAATTTGGGTGTTTATGTGTTACCGTGGGAATGGATGTTTCATGTTATCGTGAAACGCTTTCGCGTTTTCGTGCGCCAACTTCACGAACGCTGTATTACAAGTGGGAGCTTCGCAGCACGTGTAGCAGAGTGACTCAGGTTTTACTGCATACAGTCGCTACAGGATTAGCAAC
+
$,,//..010+.33%$$%/'$$%&&&%+-];64*&'(.95***)%&'&2:10((+&'+--(*%$%%%.)%&$(('%),334))*58AB?.-*++**'%(*30-,./8843/(((,,&&'0,%%')+')),,04346&&05]70110-,,)*4304)).,9??+++1446692112 8:8542074.-'%%&&'
```

Data Formats – fastq

A single read sequence in a FASTQ file is described in four lines:

1. Line 1 begins with a '@' and is followed by a header containing information about the sequencing run.
2. Line 2 is the basecalled sequence (using A, C, T, G and N).
3. Line 3 contains a '+'.
4. Line 4 encodes the per-base quality scores for the sequence in Line 2.

An example of a FASTQ file generated by MinKNOW or Guppy is shown below:

```
@75be78f7-bd62-4972-92d2-aba16f465b0d runid=ff83cfafb0cb3bfc28ac370b841f59798ab3d63a
sampleid=RB02_lambda_ovn1 read=19343 ch=53 start_time=2019-12-23T13:44:31Z
CGGTATTACTTCGTTCAGTTTCGGACAGGTGTTTTAACC[...]TCGTACCTAT
+
'%+-($&&&&'(':+7)-%(&$$.%##))868;;87/9;[...]68(*(2)/%$
```

https://community.nanoporetech.com/technical_documents
/data-analysis/v/datd_5000_v1_revo_22aug2016/fastq-files

Data Formats – fastq

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                  3.....9..............................41
0.2......................26...31........41
0..................20........30........40........50..........................................93
```

```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

MinKNOW writes in Sanger / Illumina 1.9

https://en.wikipedia.org/wiki/FASTQ_format

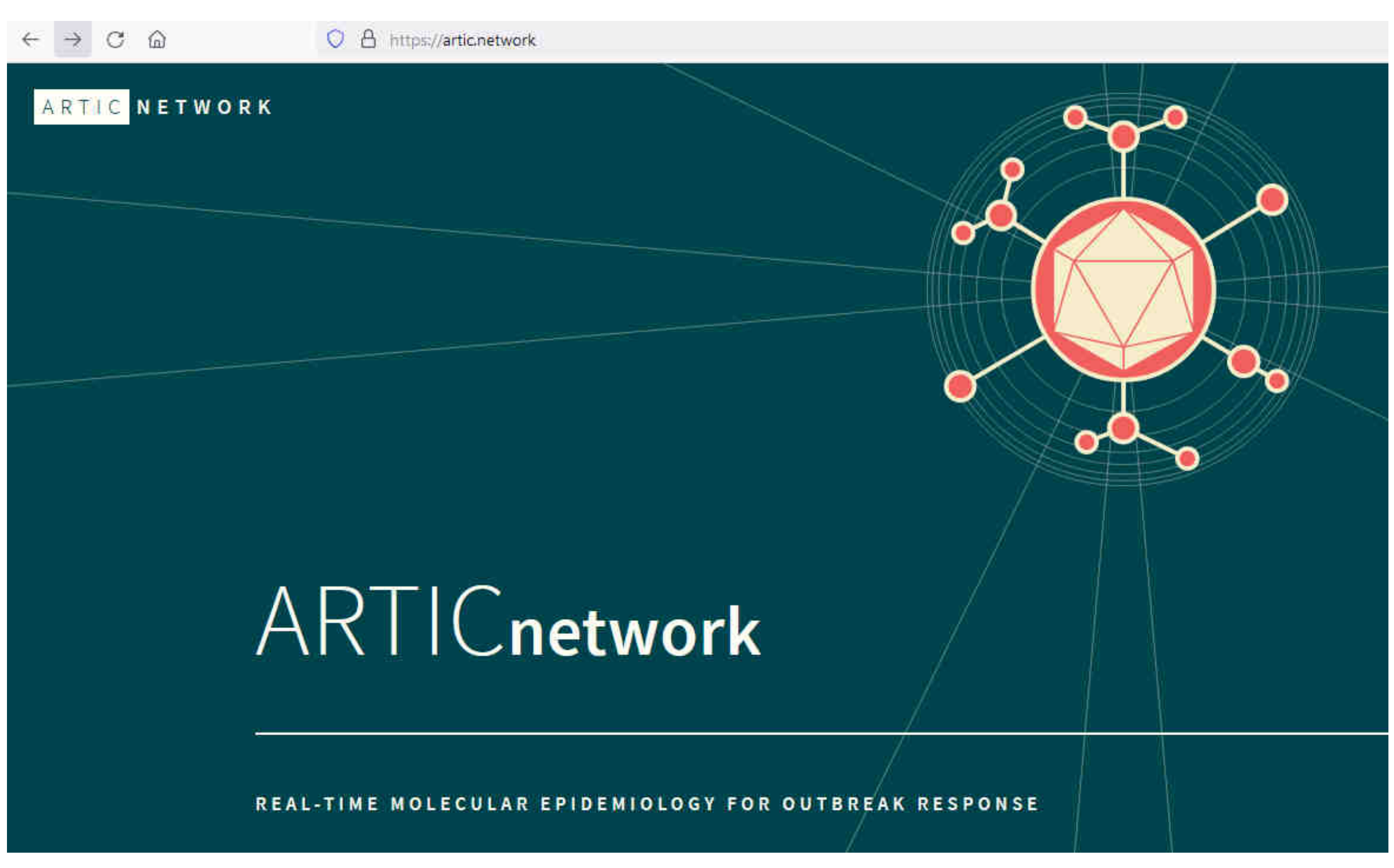
Data Formats – fasta

- text-based format
- Plain sequence
- New sequences always begin with “>”-Symbol
- standard format for alignments

```
>SARS-CoV_high_accuracy_Barcode07/ARTIC/medaka MN908947.3
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATCT
GTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACT
CACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATC
TTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
TGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAAC
ACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGG
AGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACT
CGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGG
CGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGG
TGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGA
TCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGA
ACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTC
```

Programs / work-flow – fieldbioinformatics

- Bioinformatics pipeline for virus sequencing, developed by the ARTIC-Network



Programs / work-flow – fieldbioinformatics

Features include:

- read filtering
- primer trimming
- amplicon coverage normalization
- variant calling
- consensus building

Information, download and installation-instructions:
https://github.com/artic-network/fieldbioinformatics

To install the fieldbioinformatics-pipeline you have to have conda installed!

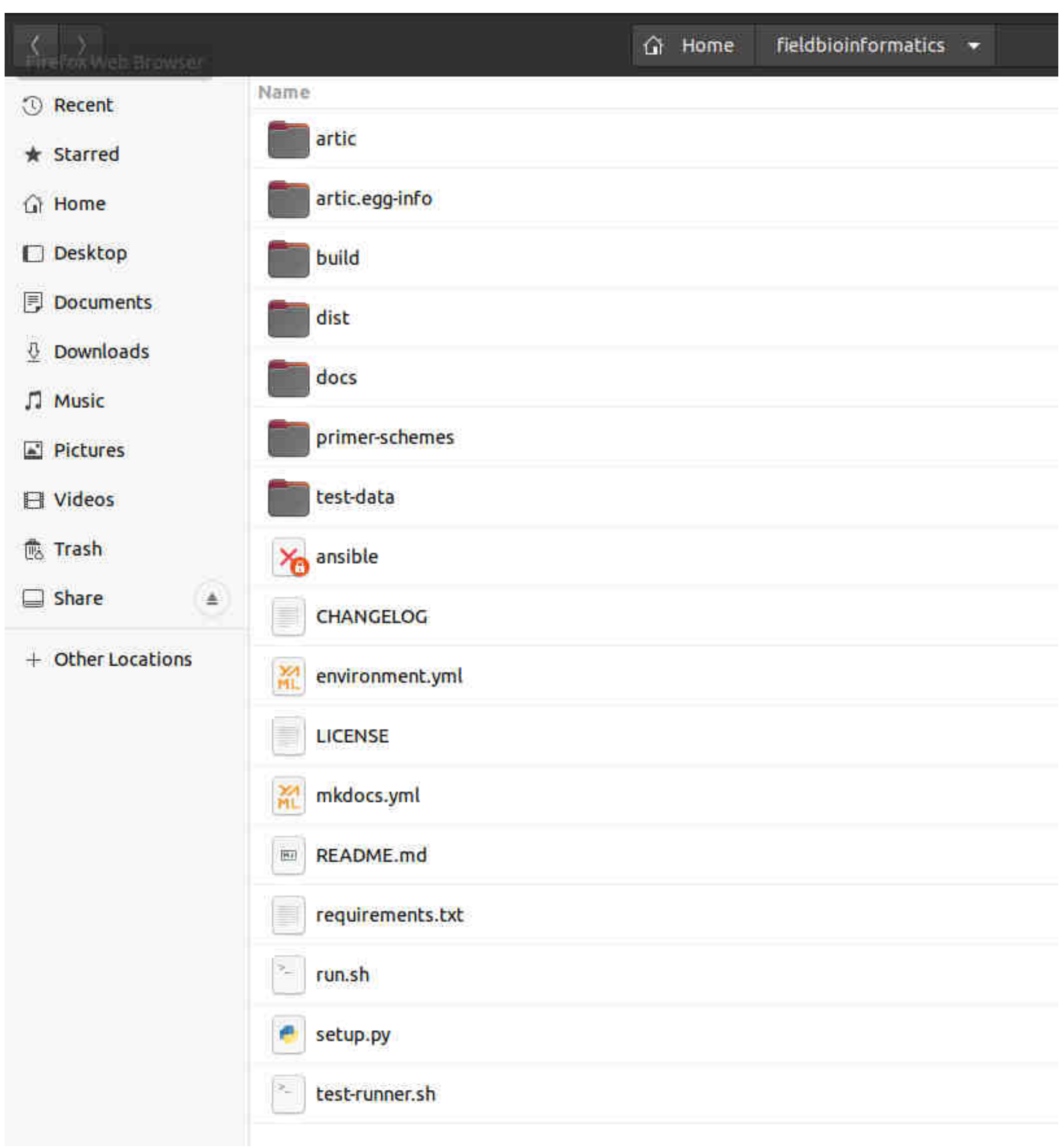
Programs / work-flow – fieldbioinformatics

- After downloading and installing fieldbioinformatics:

  -> exchange run.sh file

  -> add our primer schemes into the fieldbioinformatics folder

  (both are available in the dropbox-folder shared with you)



Programs / work-flow – fieldbioinformatics

hannah@hannah-VirtualBox: ~/fieldbioinformatics

```
(base) hannah@hannah-VirtualBox:~$ cd fieldbioinformatics
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$ ls -lh
total 72K
lrwxrwxrwx 1 hannah hannah   13 Aug 10 00:08 ansible -> lab-on-an-ssd
drwxrwxr-x 3 hannah hannah 4,0K Aug 10 00:08 artic
drwxrwxr-x 2 hannah hannah 4,0K Aug 10 11:52 artic.egg-info
drwxrwxr-x 4 hannah hannah 4,0K Aug 10 11:52 build
-rw-rw-r-- 1 hannah hannah 2,7K Aug 10 00:08 CHANGELOG
drwxrwxr-x 2 hannah hannah 4,0K Aug 10 11:52 dist
drwxrwxr-x 2 hannah hannah 4,0K Aug 10 00:08 docs
-rw-rw-r-- 1 hannah hannah  445 Aug 10 00:08 environment.yml
-rw-rw-r-- 1 hannah hannah 1,1K Aug 10 00:08 LICENSE
-rw-rw-r-- 1 hannah hannah  541 Aug 10 00:08 mkdocs.yml
drwxrwx--- 3 hannah hannah 4,0K Aug 10 12:02 primer-schemes
-rw-rw-r-- 1 hannah hannah 2,8K Aug 10 00:08 README.md
-rw-rw-r-- 1 hannah hannah   55 Aug 10 00:08 requirements.txt
-rwxrwx--- 1 hannah hannah 5,1K Sep 19 20:32 run.sh
-rw-rw-r-- 1 hannah hannah 1,9K Aug 10 00:08 setup.py
drwxrwxr-x 5 hannah hannah 4,0K Aug 10 00:08 test-data
-rwxrwxr-x 1 hannah hannah 4,6K Aug 10 00:08 test-runner.sh
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$ nano run.sh
```

Programs / work-flow – fieldbioinformatics

```
hannah@hannah-VirtualBox: ~/fieldbioinformatics
  GNU nano 4.8                              run.sh
#!/usr/bin/env bash
set -e
#
# test-runner.sh runs a the entire ARTIC field bioinformatics pipeline using a small set
# of data (the Mayinga barcode from an Ebola amplicon library sequenced on a flongle).
#
# full data available: http://artic.s3.climb.ac.uk/run-folders/EBOV_Amplicons_flongle.tar.gz
#
# usage:
#       ./test-runner.sh [medaka|nanopolish]
#
#   specify either medaka or nanopolish to run the respective workflow of the pipeline
#
###########################################################################################
# Setup the data, commands and the testing function.

for N in {07..12};
do
    inputData="/home/hannah/Share/data_210820/fastq_pass210820/barcode$N"
    primerSchemes="/home/hannah/fieldbioinformatics/primer-schemes"
    primerScheme="nCoV-2019/V1200"
    barcode=$N
    threads=5
    #samplesheet="/home//NGS-data/SARS_tr1/fastq_pass/samplesheet.csv"
    #OLDIFS=$IFS
    #while IFS=';' read -r field1 field2
    #do
    #    echo "$field1 : $field2"

    prefix="SARS-CoV_Barcode$N"

# pipeline commands

^G Get Help   ^O Write Out  ^W Where Is   ^K Cut Text   ^J Justify    ^C Cur Pos    M-U Undo     M-A Mark Text
^X Exit       ^R Read File  ^\ Replace    ^U Paste Text ^T To Spell   ^_ Go To Line M-E Redo     M-6 Copy Text
```

Modify the script for each analysis (input-path, barcodes, prefix..).

Programs / work-flow – fieldbioinformatics

```
hannah@hannah-VirtualBox: ~/fieldbioinformatics
(base) hannah@hannah-VirtualBox:~$ cd fieldbioinformatics
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$ ls -lh
total 72K
lrwxrwxrwx 1 hannah hannah   13 Aug 10 00:08 ansible -> lab-on-an-ssd
drwxrwxr-x 3 hannah hannah 4,0K Aug 10 00:08 artic
drwxrwxr-x 2 hannah hannah 4,0K Aug 10 11:52 artic.egg-info
drwxrwxr-x 4 hannah hannah 4,0K Aug 10 11:52 build
-rw-rw-r-- 1 hannah hannah 2,7K Aug 10 00:08 CHANGELOG
drwxrwxr-x 2 hannah hannah 4,0K Aug 10 11:52 dist
drwxrwxr-x 2 hannah hannah 4,0K Aug 10 00:08 docs
-rw-rw-r-- 1 hannah hannah  445 Aug 10 00:08 environment.yml
-rw-rw-r-- 1 hannah hannah 1,1K Aug 10 00:08 LICENSE
-rw-rw-r-- 1 hannah hannah  541 Aug 10 00:08 mkdocs.yml
drwxrwx--- 3 hannah hannah 4,0K Aug 10 12:02 primer-schemes
-rw-rw-r-- 1 hannah hannah 2,8K Aug 10 00:08 README.md
-rw-rw-r-- 1 hannah hannah   55 Aug 10 00:08 requirements.txt
-rwxrwx--- 1 hannah hannah 5,1K Sep 19 20:32 run.sh
-rw-rw-r-- 1 hannah hannah 1,9K Aug 10 00:08 setup.py
drwxrwxr-x 5 hannah hannah 4,0K Aug 10 00:08 test-data
-rwxrwxr-x 1 hannah hannah 4,6K Aug 10 00:08 test-runner.sh
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$ nano run.sh
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$ chmod u+x run.sh
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$
```

When running the first time, you may have to add execution-permissions to the script.

Programs / work-flow – fieldbioinformatics

```
hannah@hannah-VirtualBox: ~/fieldbioinformatics
(base) hannah@hannah-VirtualBox:~$ cd fieldbioinformatics/
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$ ls -lh
total 72K
lrwxrwxrwx 1 hannah hannah   13 Aug 10 00:08 ansible -> lab-on-an-ssd
drwxrwxr-x 3 hannah hannah 4,0K Aug 10 00:08 artic
drwxrwxr-x 2 hannah hannah 4,0K Aug 10 11:52 artic.egg-info
drwxrwxr-x 4 hannah hannah 4,0K Aug 10 11:52 build
-rw-rw-r-- 1 hannah hannah 2,7K Aug 10 00:08 CHANGELOG
drwxrwxr-x 2 hannah hannah 4,0K Aug 10 11:52 dist
drwxrwxr-x 2 hannah hannah 4,0K Aug 10 00:08 docs
-rw-rw-r-- 1 hannah hannah  445 Aug 10 00:08 environment.yml
-rw-rw-r-- 1 hannah hannah 1,1K Aug 10 00:08 LICENSE
-rw-rw-r-- 1 hannah hannah  541 Aug 10 00:08 mkdocs.yml
drwxrwx--- 3 hannah hannah 4,0K Aug 10 12:02 primer-schemes
-rw-rw-r-- 1 hannah hannah 2,8K Aug 10 00:08 README.md
-rw-rw-r-- 1 hannah hannah   55 Aug 10 00:08 requirements.txt
-rwxrwx--- 1 hannah hannah 5,1K Sep 19 20:32 run.sh
-rw-rw-r-- 1 hannah hannah 1,9K Aug 10 00:08 setup.py
drwxrwxr-x 5 hannah hannah 4,0K Aug 10 00:08 test-data
-rwxrwxr-x 1 hannah hannah 4,6K Aug 10 00:08 test-runner.sh
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$ nano run.sh
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$ chmod u+x run.sh
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$ conda activate artic
(artic) hannah@hannah-VirtualBox:~/fieldbioinformatics$
```

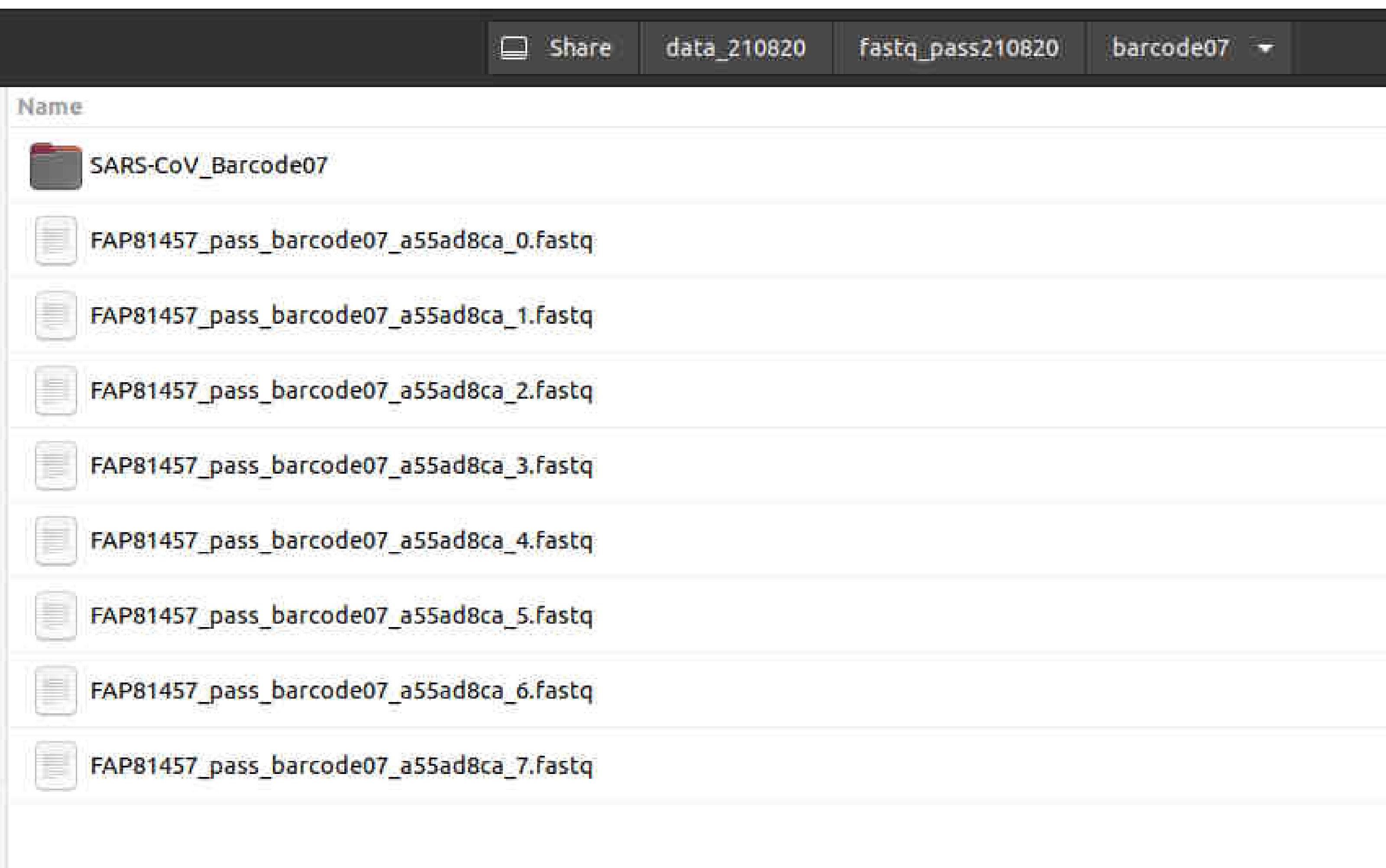
Activate the correct conda environnement.

Programs / work-flow – fieldbioinformatics

```
hannah@hannah-VirtualBox: ~/fieldbioinformatics
total 72K
lrwxrwxrwx 1 hannah hannah   13 Aug 10 00:08 ansible -> lab-on-an-ssd
drwxrwxr-x 3 hannah hannah 4,0K Aug 10 00:08 artic
drwxrwxr-x 2 hannah hannah 4,0K Aug 10 11:52 artic.egg-info
drwxrwxr-x 4 hannah hannah 4,0K Aug 10 11:52 build
-rw-rw-r-- 1 hannah hannah 2,7K Aug 10 00:08 CHANGELOG
drwxrwxr-x 2 hannah hannah 4,0K Aug 10 11:52 dist
drwxrwxr-x 2 hannah hannah 4,0K Aug 10 00:08 docs
-rw-rw-r-- 1 hannah hannah  445 Aug 10 00:08 environment.yml
-rw-rw-r-- 1 hannah hannah 1,1K Aug 10 00:08 LICENSE
-rw-rw-r-- 1 hannah hannah  541 Aug 10 00:08 mkdocs.yml
drwxrwx--- 3 hannah hannah 4,0K Aug 10 12:02 primer-schemes
-rw-rw-r-- 1 hannah hannah 2,8K Aug 10 00:08 README.md
-rw-rw-r-- 1 hannah hannah   55 Aug 10 00:08 requirements.txt
-rwxrwx--- 1 hannah hannah 5,1K Sep 19 20:32 run.sh
-rw-rw-r-- 1 hannah hannah 1,9K Aug 10 00:08 setup.py
drwxrwxr-x 5 hannah hannah 4,0K Aug 10 00:08 test-data
-rwxrwxr-x 1 hannah hannah 4,6K Aug 10 00:08 test-runner.sh
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$ nano run.sh
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$ chmod u+x run.sh
(base) hannah@hannah-VirtualBox:~/fieldbioinformatics$ conda activate artic
(artic) hannah@hannah-VirtualBox:~/fieldbioinformatics$ ./run.sh medaka
Starting tests...
 - using the medaka workflow

running the pipeline...
###########################################################################################
Running: artic gather --min-length 400 --max-length 1200 --prefix SARS-CoV_Barcode07 --directory /home/hannah/Share/data_21
0820/fastq_pass210820/barcode07 --no-fast5s

Processing 8 files in barcode07
SARS-CoV_Barcode07_barcode07.fastq      23426   23426
Processing 1 files in SARS-CoV_Barcode07
```

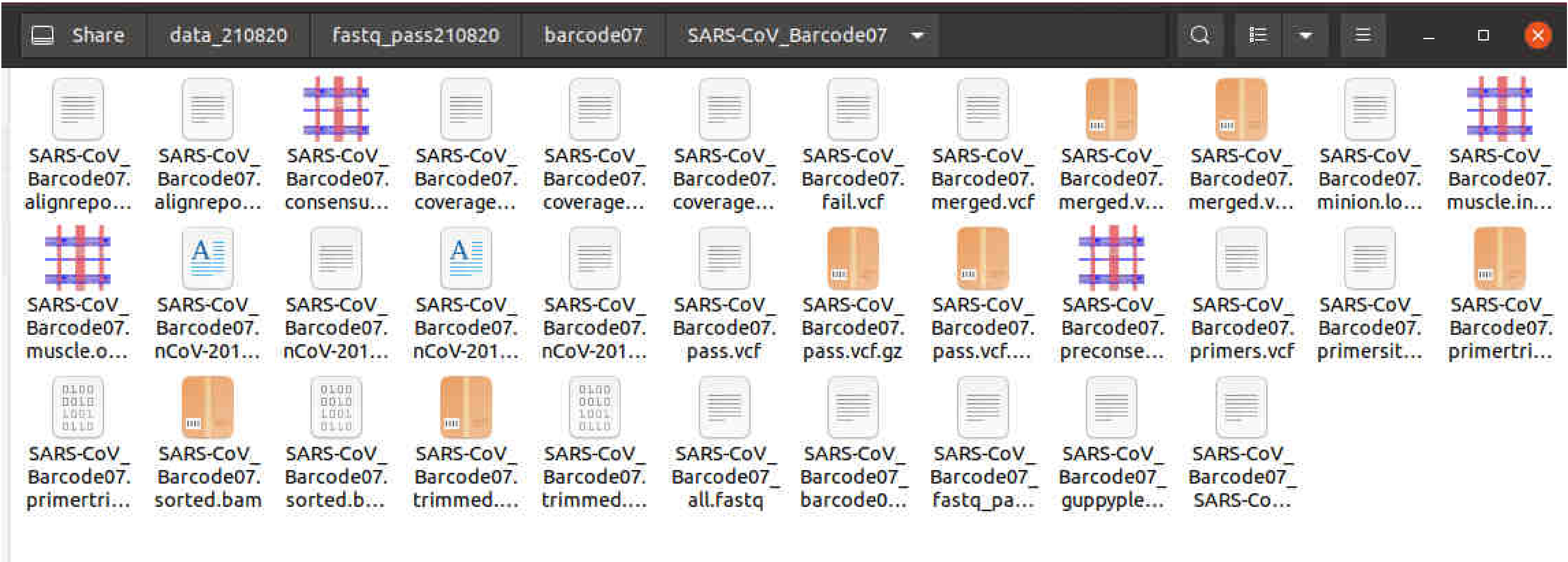
Run the pipeline.

Programs / work-flow – fieldbioinformatics



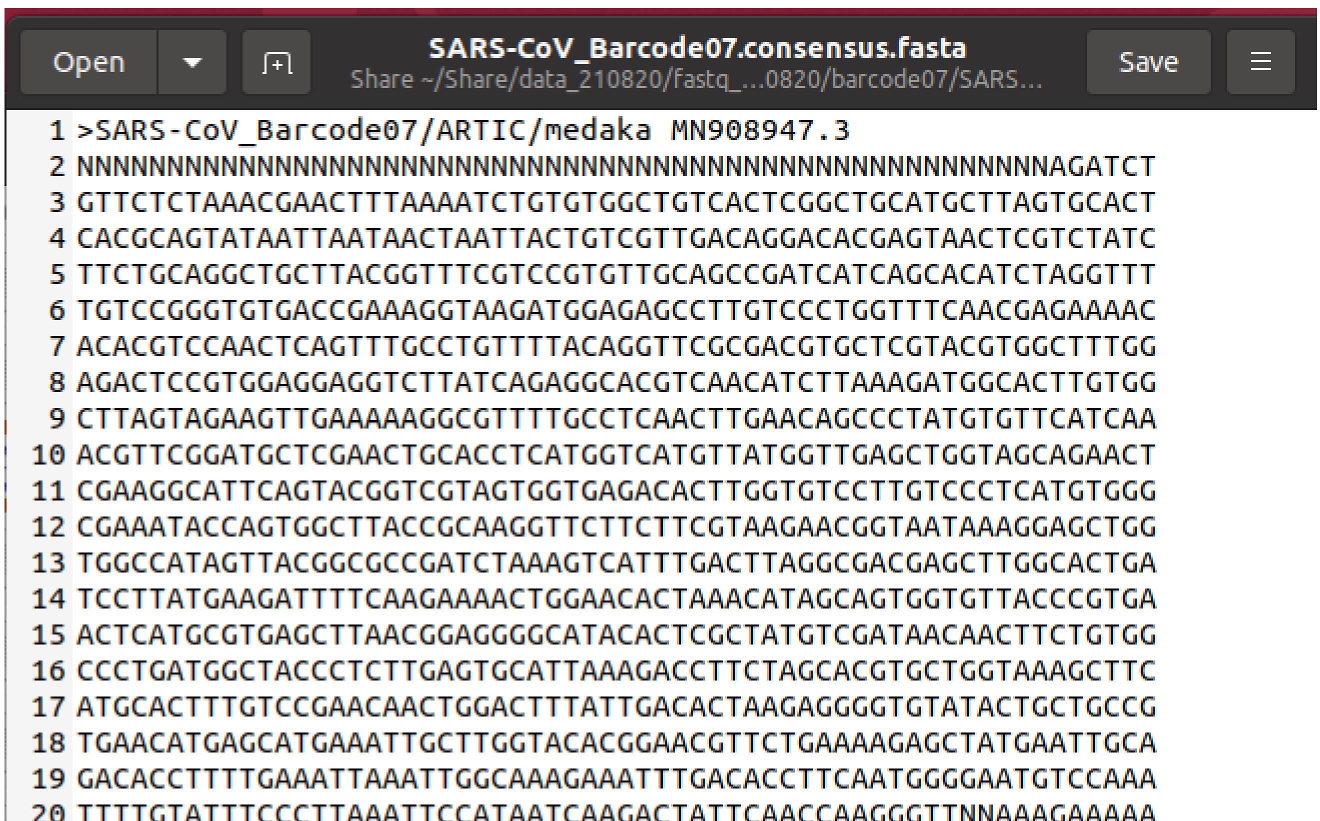
The output-folder will be in the same place as the input-data.

Programs / work-flow – fieldbioinformatics



The content of the output-folder.

Programs / work-flow – fieldbioinformatics



The ".consensus.fasta" file is the one, which interests us most.

Programs / work-flow – fieldbioinformatics

Open

*SARS-CoV_Barcode07.nCoV-2019_1.vcf
Share ~/Share/data_210820/fastq_...0820/barcode07/SARS-CoV_Barcode07

```
1 ##fileformat=VCFv4.1
2 ##medaka_version=1.0.3
3 ##contig=<ID=MN908947.3>
4 ##FORMAT=<ID=GT,Number=1,Type=String,Description="Medaka genotype.">
5 ##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Medaka genotype quality score">
6 #CHROM       POS    ID   REF   ALT   QUAL    FILTER  INFO   FORMAT  SAMPLE
7 MN908947.3   40     .    C     T     36.972  PASS           GT:GQ   1:37
8 MN908947.3   241    .    C     T     55.435  PASS           GT:GQ   1:55
9 MN908947.3   1130   .    GA    G     0.817   PASS           GT:GQ   1:1
10 MN908947.3  3037   .    C     T     23.787  PASS           GT:GQ   1:24
11 MN908947.3  4457   .    A     G     42.077  PASS           GT:GQ   1:42
12 MN908947.3  4891   .    C     T     42.832  PASS           GT:GQ   1:43
13 MN908947.3  5184   .    C     T     35.099  PASS           GT:GQ   1:35
14 MN908947.3  9204   .    A     G     32.587  PASS           GT:GQ   1:33
15 MN908947.3  11074  .    CT    C     8.197   PASS           GT:GQ   1:8
16 MN908947.3  11430  .    A     G     31.292  PASS           GT:GQ   1:31
17 MN908947.3  15187  .    A     G     40.072  PASS           GT:GQ   1:40
18 MN908947.3  15324  .    C     T     48.63   PASS           GT:GQ   1:49
19 MN908947.3  22802  .    C     A     40.773  PASS           GT:GQ   1:41
20 MN908947.3  22912  .    T     G     51.175  PASS           GT:GQ   1:51
21 MN908947.3  23403  .    A     G     35.788  PASS           GT:GQ   1:36
22 MN908947.3  25445  .    GTGAAATCAAGGATGCTACTCCTTCAGATTT      G       1029.872
23 MN908947.3  27576  .    T     C     36.103  PASS           GT:GQ   1:36
24 MN908947.3  28846  .    C     T     46.757  PASS           GT:GQ   1:47
25 MN908947.3  28887  .    C     T     31.681  PASS           GT:GQ   1:32
26 MN908947.3  29730  .    C     T     32.8    PASS           GT:GQ   1:33
```

The ".vcf" file lists all variations from the original WUHAN-sequence.

Analysis of the sequences for variants/clades/lineages/mutations

https://pangolin.cog-uk.io/

Help

Drag and drop fasta file

Select fasta file to upload

Pangolin COVID-19 Lineage Assigner

Phylogenetic Assignment of Named Global Outbreak LINeages

You can upload one or more sequences by dragging and dropping a (multi)fasta file or clicking "Select fasta file to upload" and selecting a (multi)fasta file.

This Web Application assigns lineages to COVID-19 sequenes based on the methodology described in this article

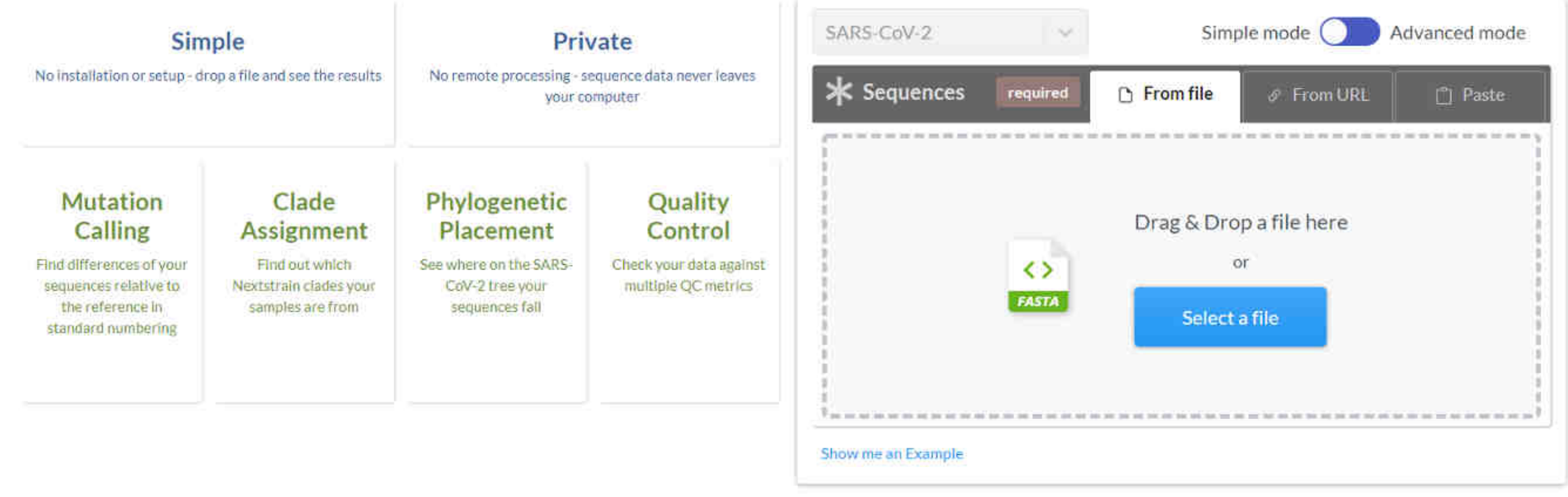
The software to assign lineages based on the algorithm that was developed by Áine O'Toole, Verity Hill, JT McCrone, Emily Scher and Andrew Rambaut.
The source code can be found here

Recommended browsers or

Analysis of the sequences for variants/clades/lineages/mutations

https://clades.nextstrain.org/

Clade assignment, mutation calling, and sequence quality checks



High accuracy basecalling with Guppy

Back

Done. Total sequences: 5. Succeeded: 5

ID	Sequence name	QC	Clade	Mut.	non-ACGTN	Ns	Gaps	Ins.
0	SARS-CoV_high_accuracy_Barcode07/ARTIC/me	N M P C F S	20A	25	0	220	2	9
1	SARS-CoV_high_accuracy_Barcode08/ARTIC/me	N M P C F S	20A	26	0	220	2	9
2	SARS-CoV_high_accuracy_Barcode09/ARTIC/me	N M P C F S	20A	30	0	220	11	9
3	SARS-CoV_high_accuracy_Barcode10/ARTIC/me	N M P C F S	20A	29	0	220	2	9
4	SARS-CoV_high_accuracy_Barcode11/ARTIC/me	N M P C F S	19B	3	0	189	0	0

THANK YOU

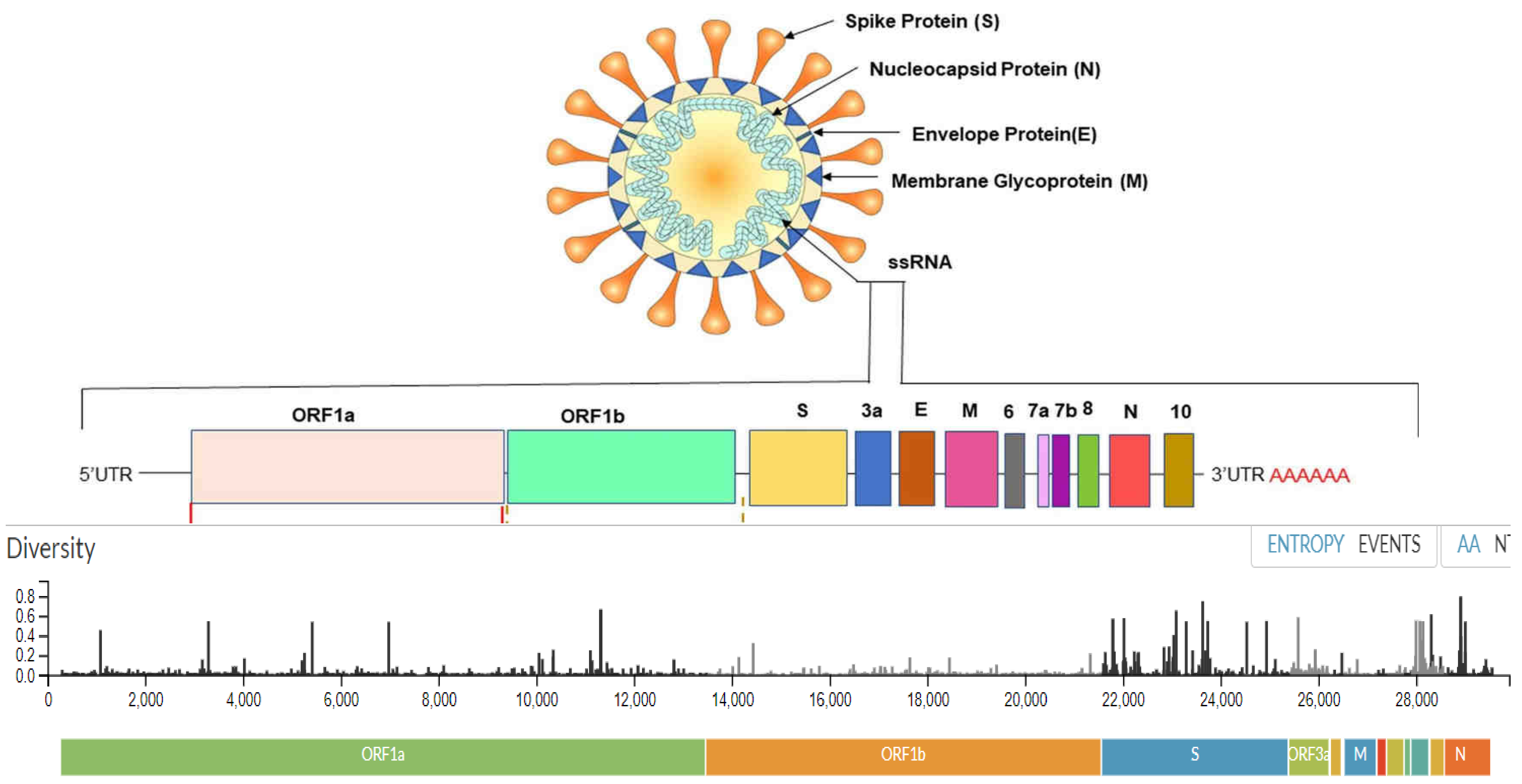
Workshop
Molecular Diagnostic Techniques
(Genomics and Bioinformatics – Infectious diseases)
module-5

Dr. Srinivas Reddy Pallerla, Mr. Tran Thanh Tung and Ms. Linh Le
Prof. Dr. Thirumalaisamy P. Velavan, University Hospital, Tübingen, Germany
Prof. Dr. Daniel Olusola Ojurongbe, Nigeria
Prof. Dr Mohamed Osman, University of Khartoum, Sudan

29 July 2022

Training Sponsor: *Alexander von Humboldt Foundation*

SARS-CoV-2 genome – 29.9 kb



Rastogi et al., Respir Res. 2020

Rationale for sequencing of SARS-CoV-2

- Evolution and epidemiology of the virus
- Understand transmissibility, morbidity and mortality
- Evade detection by diagnostic tests
- Decreased neutralizing to antibodies (e.g., convalescent plasma or monoclonal antibodies)
- Evade natural immunity (e.g., causing reinfections)
- Infect vaccinated individuals (outbreak infections)
- Monitor variants of concern (VOC) and variants of interest (VOI)
- Monitor trends at the national level
- Monitor emergence of important new strains
- Monitor trends after interventions such as vaccination
- Better understand epidemiology at the local level

Global - genomic surveillance of SARS-CoV-2

GISAID (Global Initiative on Sharing All Influenza Data)

hCoV-19 data sharing via GISAID

3,545,029

submissions

- https://outbreak.info/
- https://cov-lineages.org/index.html
- https://nextstrain.org/ncov/global

- Global/Country/State/City to local
- Circulating variants
- New viral lineages
- Variants of interest, under investigation, of concern
- Distribution over time and space
- No of mutation accumulating
- Much more…….

Nomenclatures for SARS-CoV-2 variants

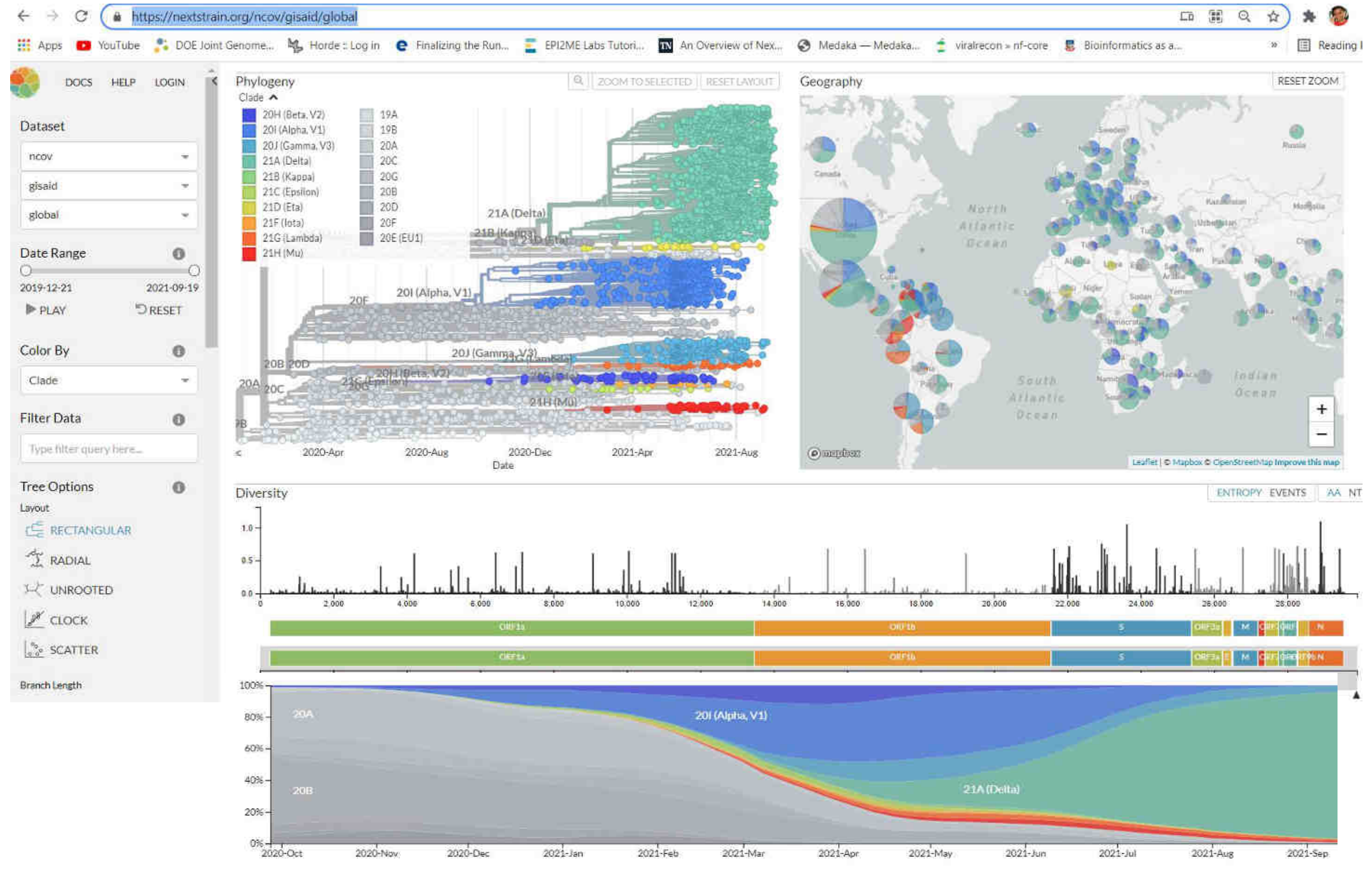
GISAID Clades: (e.g. L, V, S, O, G, GH, GR, GV, GRY) ; https://www.gisaid.org/

NeXTstrain Clades: Major clades: 19A, 19B, and 20A–20I ; https://nextstrain.org/ncov/gisaid/global

PANGOLIN Lineages: Phylogenetic Assignment of Named Global Outbreak Lineages ; https://cov-lineages.org/ Predominant - B lineage: e.g. B.1.1.7 (Alpha), B.1.351 (Beta), B.1.617.2 (Delta) etc.

Accessed on 15 Sep 2021 Source: GSAID

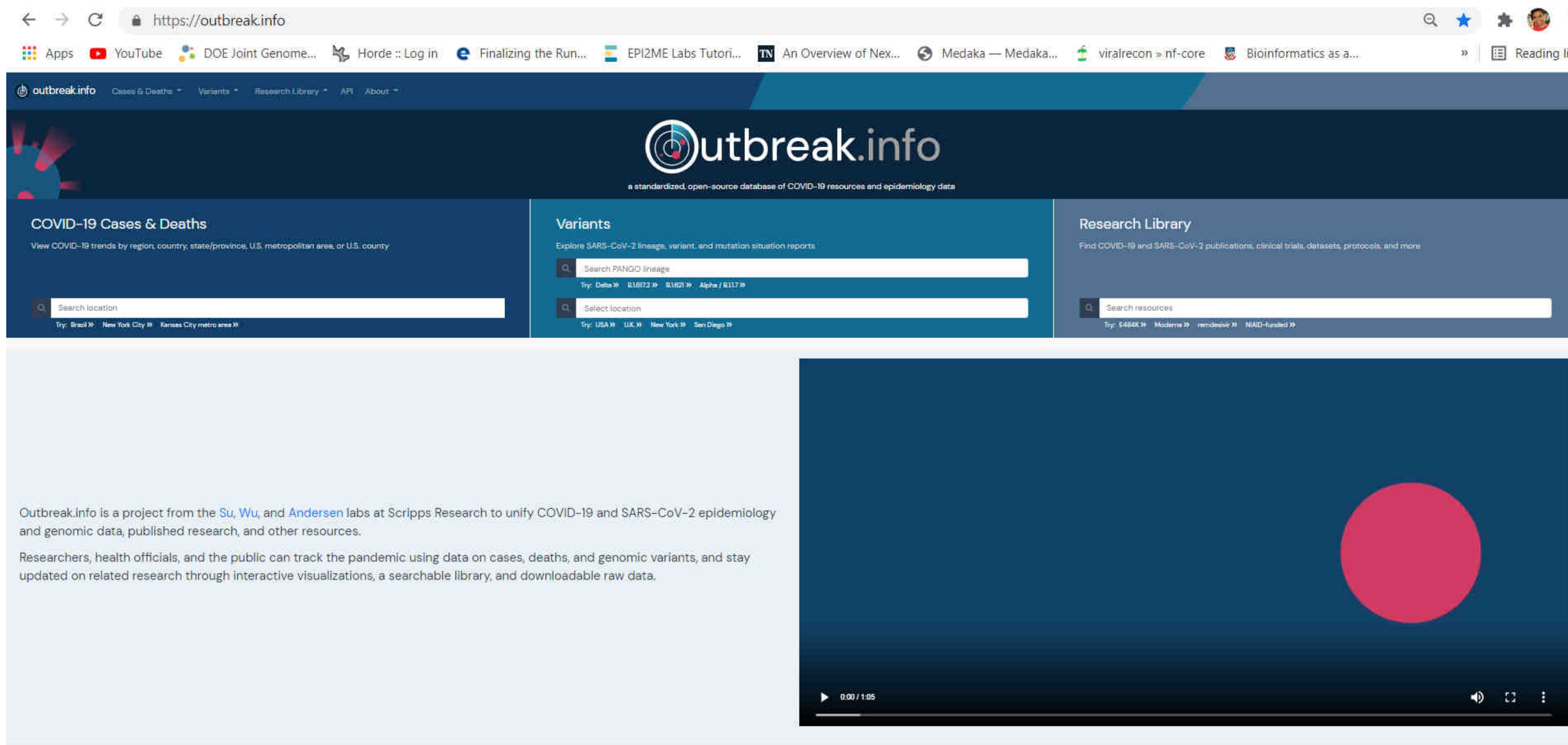
Genomic epidemiology of novel coronavirus - Global subsampling (nextstrain.org)



https://nextstrain.org/ncov/gisaid/global

Details will be showed in the browser

Genomic epidemiology of novel coronavirus (outbreak.info)



https://outbreak.info/

Details will be showed in the browser

Phylogenetic analysis of SARS-CoV-2 genomes

Short video on the phylogenetics

https://www.cdc.gov/amd/training/covid-toolkit/module1-3.html

https://www.youtube.com/watch?v=Ny38Aa2edk8

https://www.cdc.gov/amd/training/covid-19-gen-epi-toolkit.html

Load MAFFT and IQTREE programmes using following commands in Linux terminal

12.5 Phylogeny

For phylogeny you can use a program called **MegaX** https://megasoftware.net/ which comes with a graphical user interface. There is a free version which allows all important steps, for more features you can buy a full version. But MegaX is rather slow, especially for large numbers of samples.

Alternatively there are some command line tools.

MAFFT https://mafft.cbrc.jp/alignment/software/ is a multiple sequence alignment program. It can be installed via the command line

$ sudo apt install mafft

To run mafft type:

$ mafft all_barcodes_merged.fasta > all_barcodes_alinged.fasta (input > output)

IQTREE http://www.iqtree.org/ is a phylogeny program. It can be installed via the command line.

$ sudo apt install iqtree

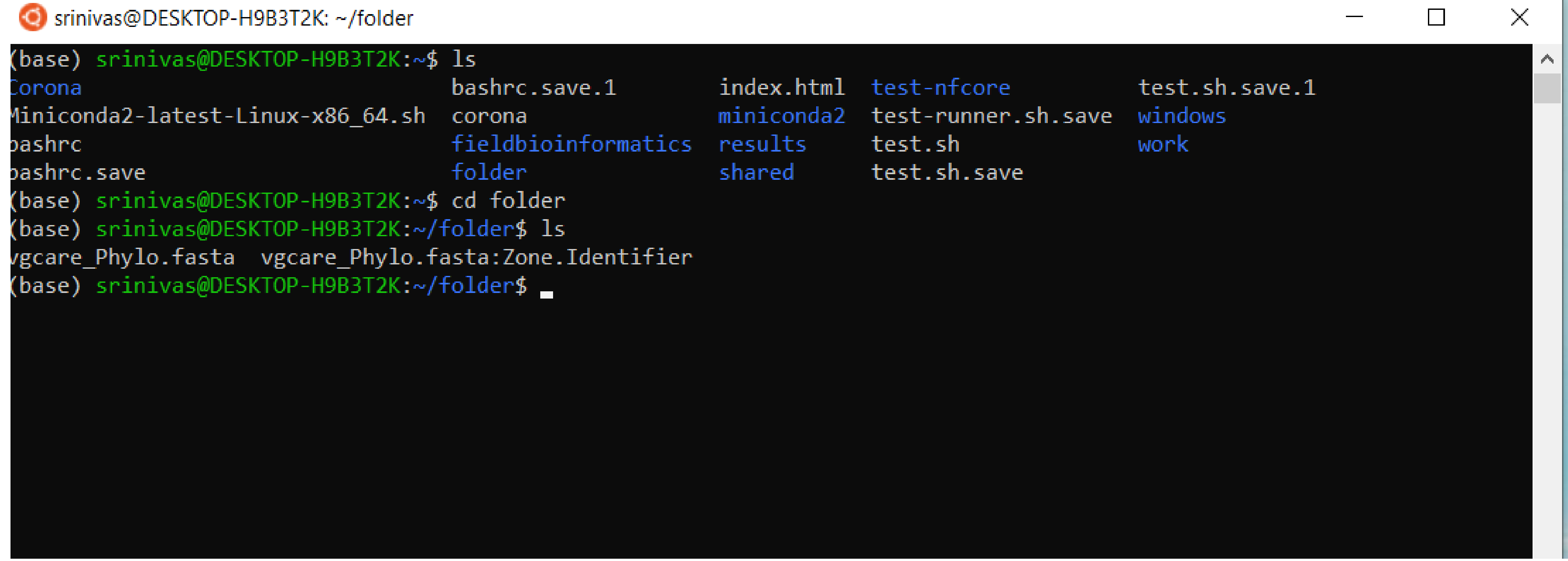
To run type:

$ iqtree -s all_barcodes_alinged.fasta -bb 1000

"-bb 100" is to set the ultrafast bootstrap (UFBoot) feature to 1000 replicates which will give more unbiased values.

To visualize your phylogenetic analysis you can load the treefile of the iqtree-output to the **ITOL-website**: https://itol.embl.de/upload.cgi

1. Create a folder (using command mkdir xxx) and copy vgcare_phylo.fasta file from dropbox



2. $ mafft input > output files as shown below

srinivas@DESKTOP-H9B3T2K: ~/folder

```
(base) srinivas@DESKTOP-H9B3T2K:~$ ls
Corona                              bashrc.save.1        index.html  test-nfcore          test.sh.save.1
Miniconda2-latest-Linux-x86_64.sh  corona               miniconda2  test-runner.sh.save  windows
bashrc                              fieldbioinformatics  results     test.sh              work
bashrc.save                         folder               shared      test.sh.save
(base) srinivas@DESKTOP-H9B3T2K:~$ cd folder
(base) srinivas@DESKTOP-H9B3T2K:~/folder$ ls
vgcare_Phylo.fasta  vgcare_Phylo.fasta:Zone.Identifier
(base) srinivas@DESKTOP-H9B3T2K:~/folder$ mafft vgcare_Phylo.fasta > vgcare_aligned.fasta
```

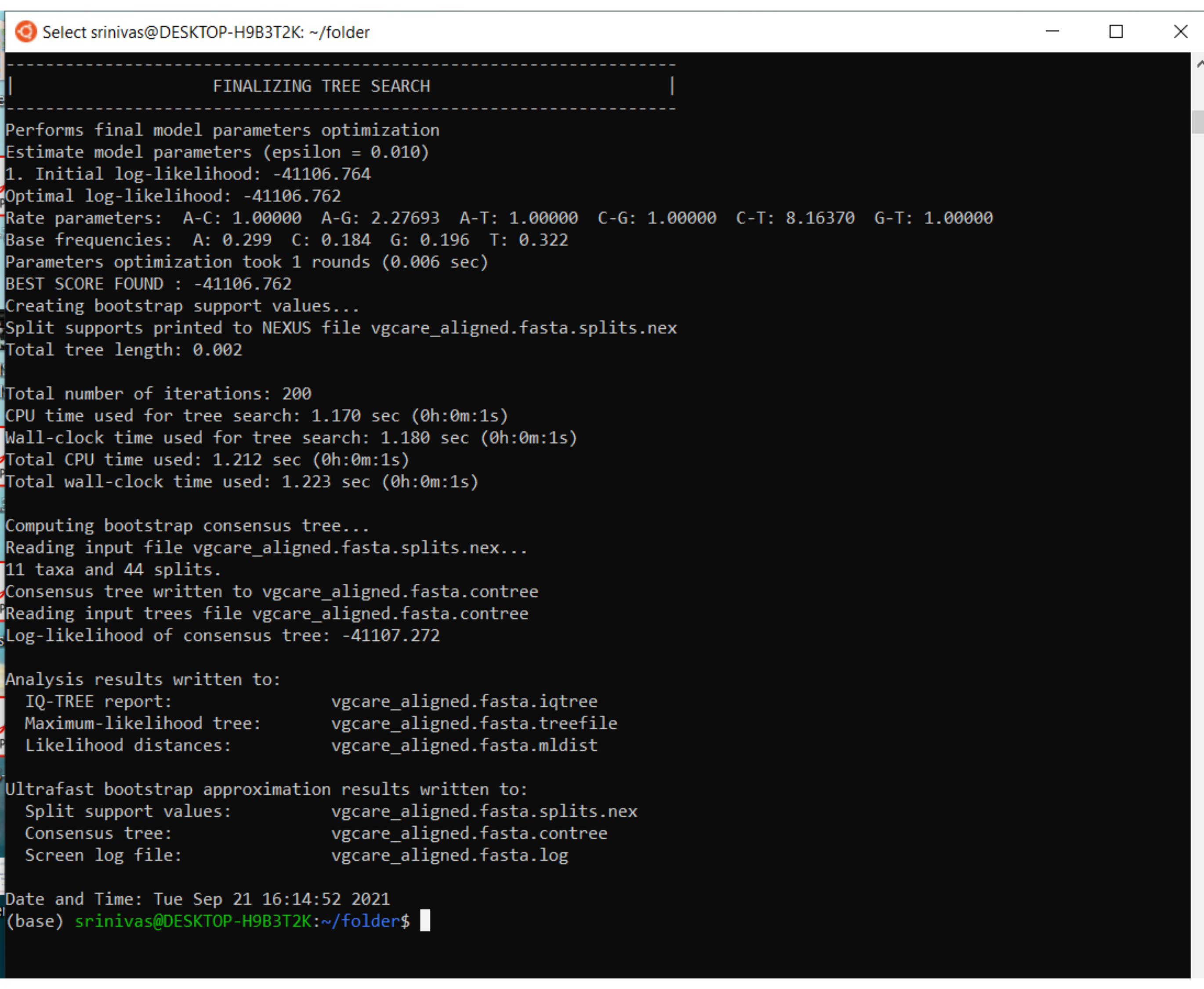
Load MAFFT and IQTREE programmes using following commands in Linux terminal

3. $ "iqtree –s vgcare_aligned.fasta –bb 1000" command

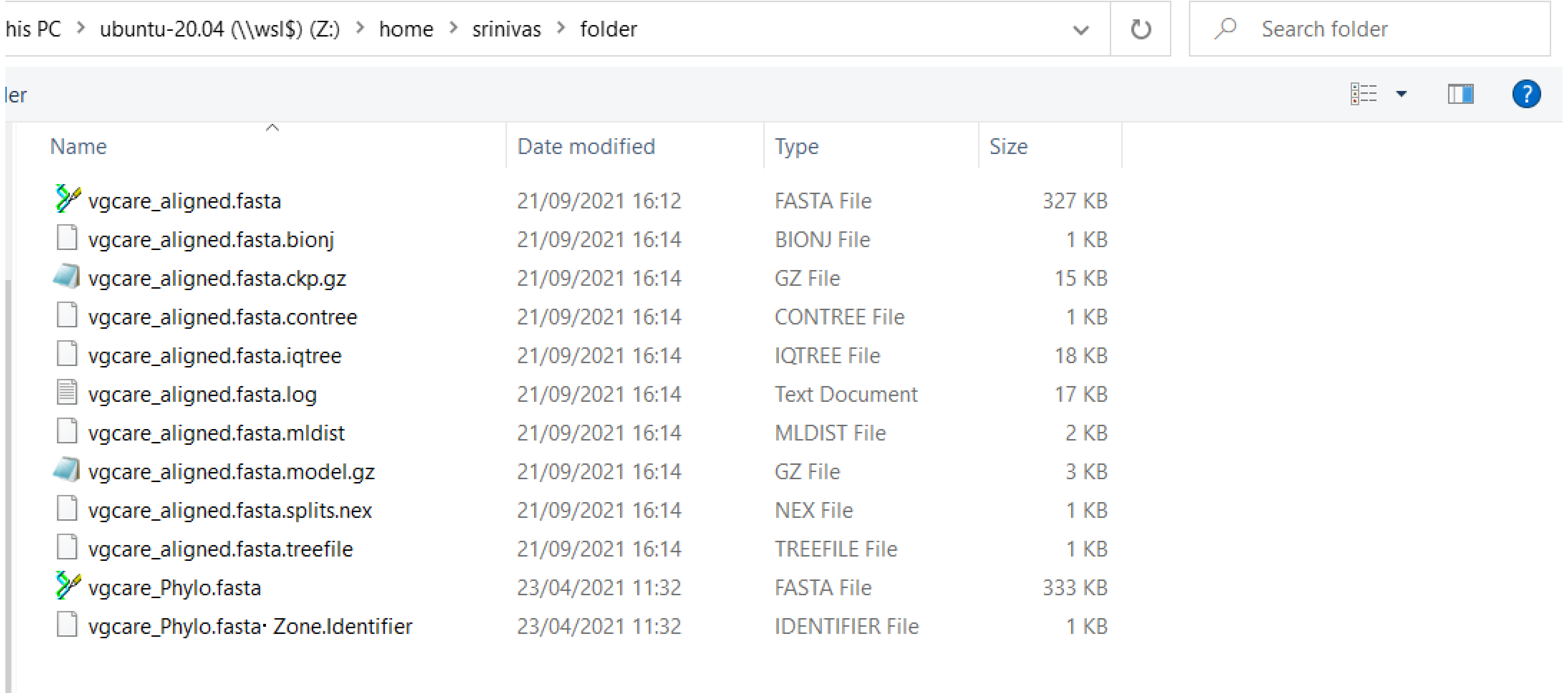
srinivas@DESKTOP-H9B3T2K: ~/folder

```
Progressive alignment 2/2...
STEP    10 / 10  f
done.

disttbfast (nuc) Version 7.453
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
0 thread(s)


Strategy:
 FFT-NS-2 (Fast but rough)
 Progressive method (guide trees were built 2 times.)

If unsure which option to use, try 'mafft --auto input > output'.
For more information, see 'mafft --help', 'mafft --man' and the mafft page.

The default gap scoring scheme has been changed in version 7.110 (2013 Oct).
It tends to insert more gaps into gap-rich regions than previous versions.
To disable this change, add the --leavegappyregion option.

(base) srinivas@DESKTOP-H9B3T2K:~/folder$ ls
vgcare_Phylo.fasta  vgcare_Phylo.fasta:Zone.Identifier  vgcare_aligned.fasta
(base) srinivas@DESKTOP-H9B3T2K:~/folder$ iqtree -s vgcare_aligned.fasta -bb 1000
```

4. $ several files generated and use “ls” command to check and copy all the files to the desktop

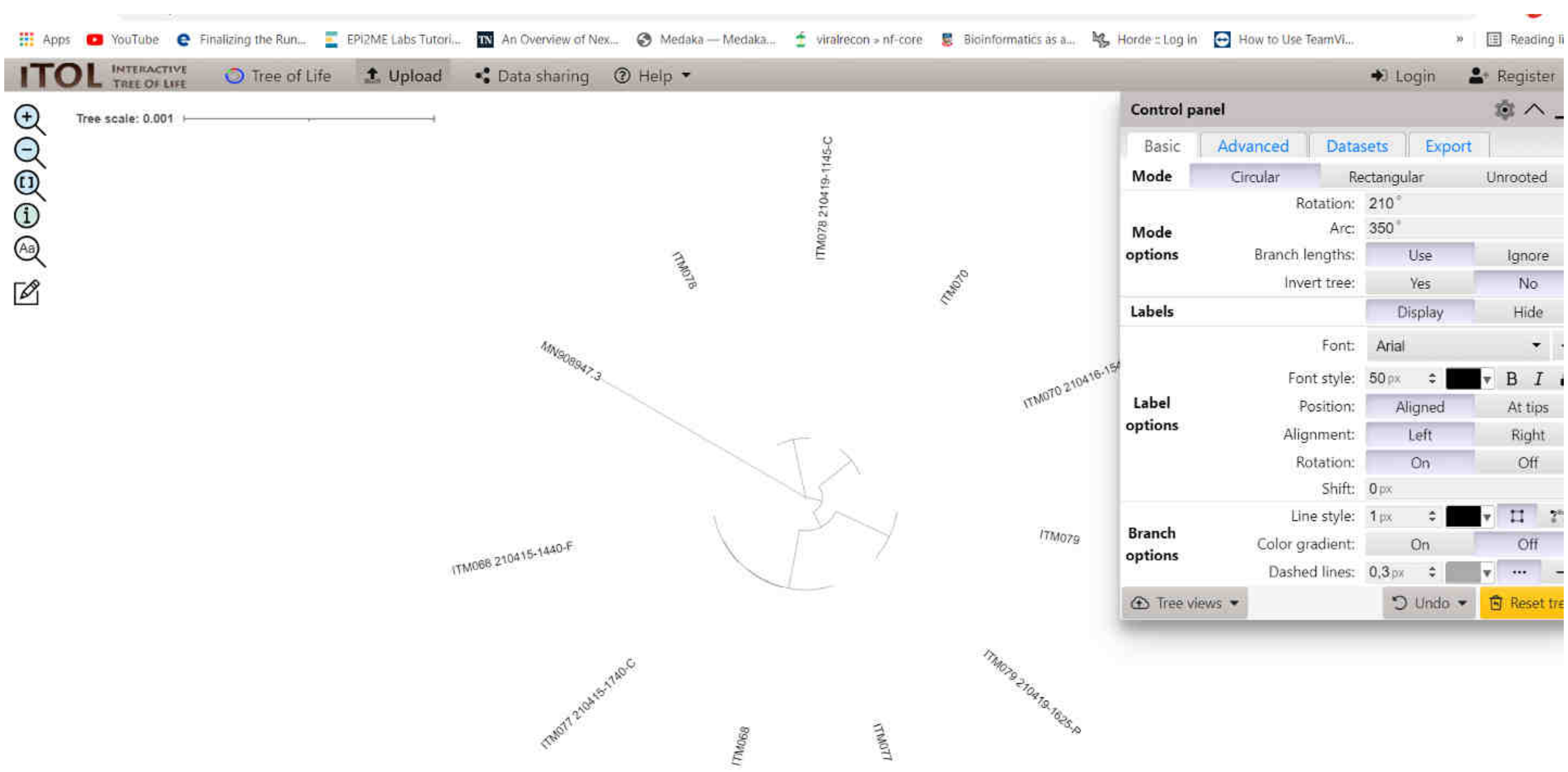


4. Copied files on the desktop



5. Go to https://itol.embl.de/upload.cgi and upload file “vgcare_aligned.fasta.treefile”

It can be edited and exported



Questions and discussion