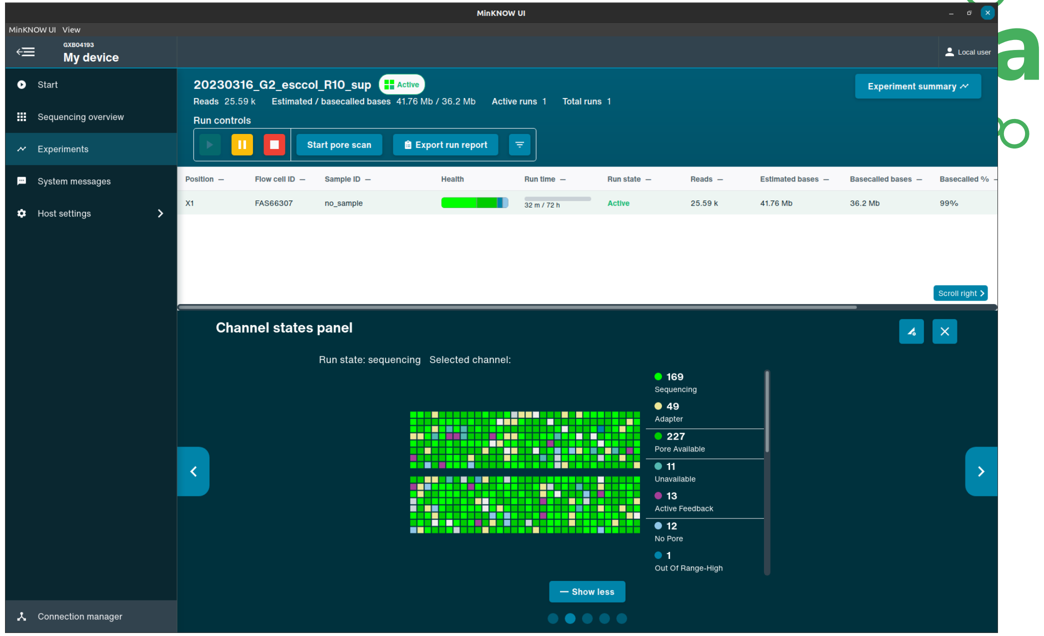
Bio 296: Microbial Bioinformatics Introduction to DNA extraction and Sequencing Technologies

Tim Roloff

Pores sequencing

Green is good!

For a good run Sequencing > Pores available



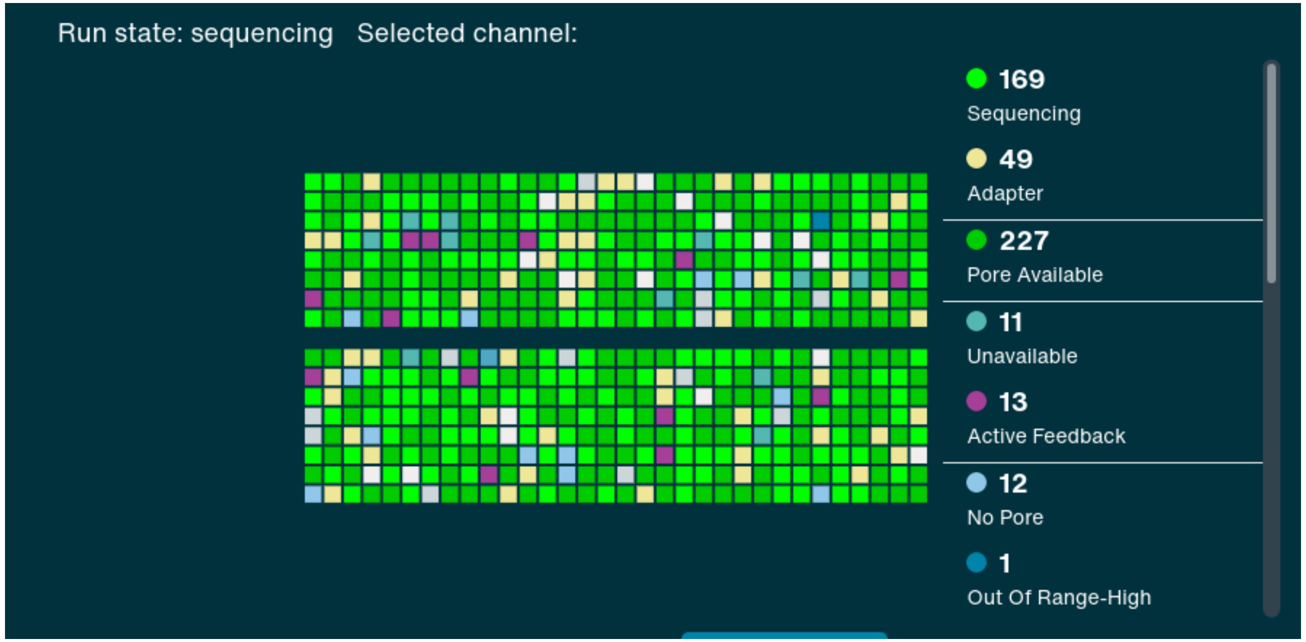
20230316_G2_esccol_R10_sup Active

Reads 25.59 k Estimated / basecalled bases 41.76 Mb / 36.2 Mb Active runs 1 Total runs 1

Position	Flow cell ID	Sample ID	Health	Run time	Run state	Reads	Estimated bases	Basecalled bases	Basecalled %
X1	FAS66307	no_sample		32 m / 72 h	Active	25.59 k	41.76 Mb	36.2 Mb	99%

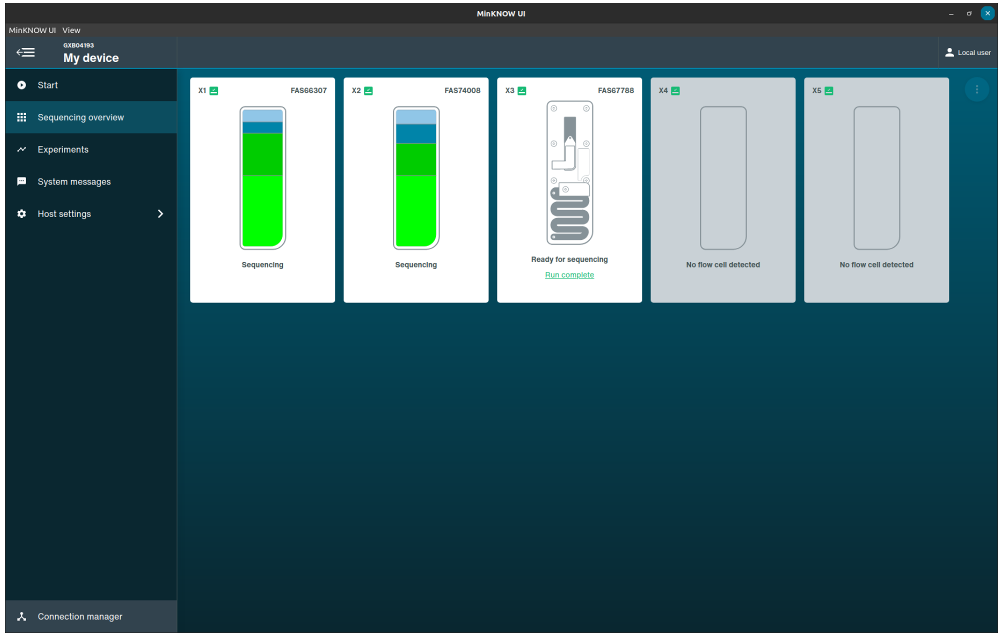
Run state: sequencing Selected channel:

- 169 Sequencing
- 49 Adapter
- 227 Pore Available
- 11 Unavailable
- 13 Active Feedback
- 12 No Pore
- 1 Out Of Range-High



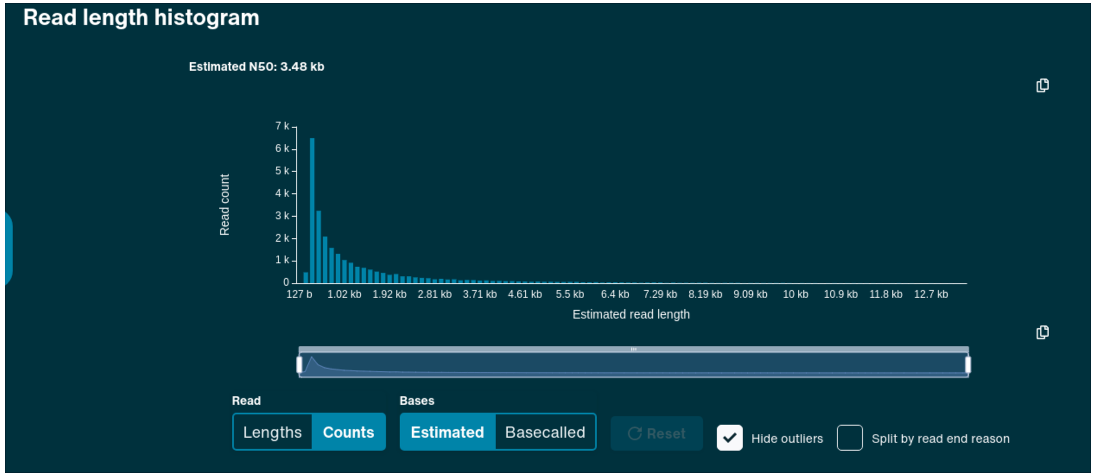
Run state: sequencing Selected channel:

- ● 169 Sequencing
- ● 49 Adapter
- ● 227 Pore Available
- ● 11 Unavailable
- ● 13 Active Feedback
- ● 12 No Pore
- ● 1 Out Of Range-High



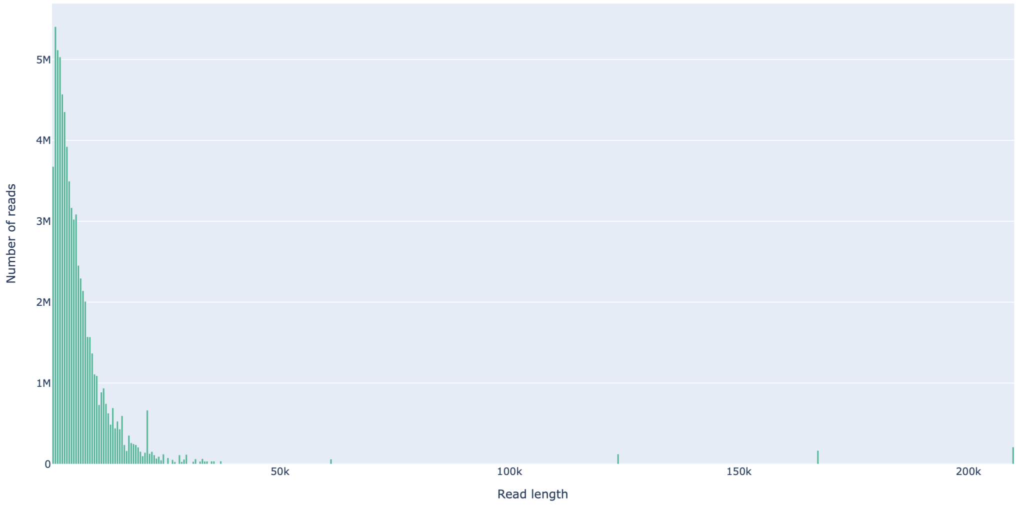
X1	FAS66307	Sequencing
X2	FAS74008	Sequencing
X3	FAS67788	Ready for sequencing Run complete
X4		No flow cell detected
X5		No flow cell detected

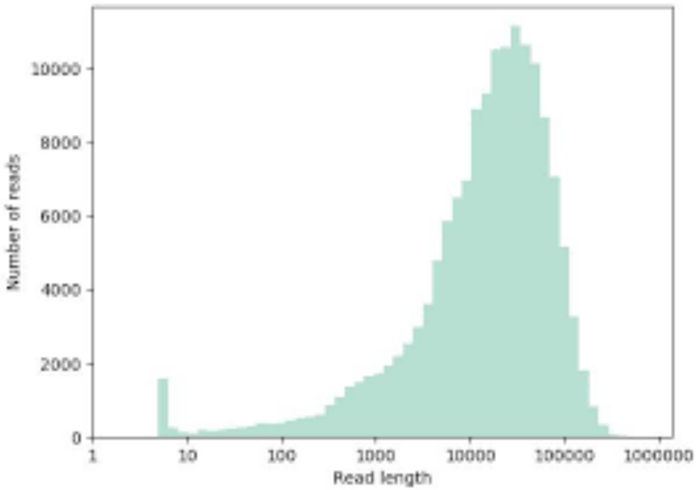
Read length distribution



☑ Hide outliers ☐ Split by read end reason

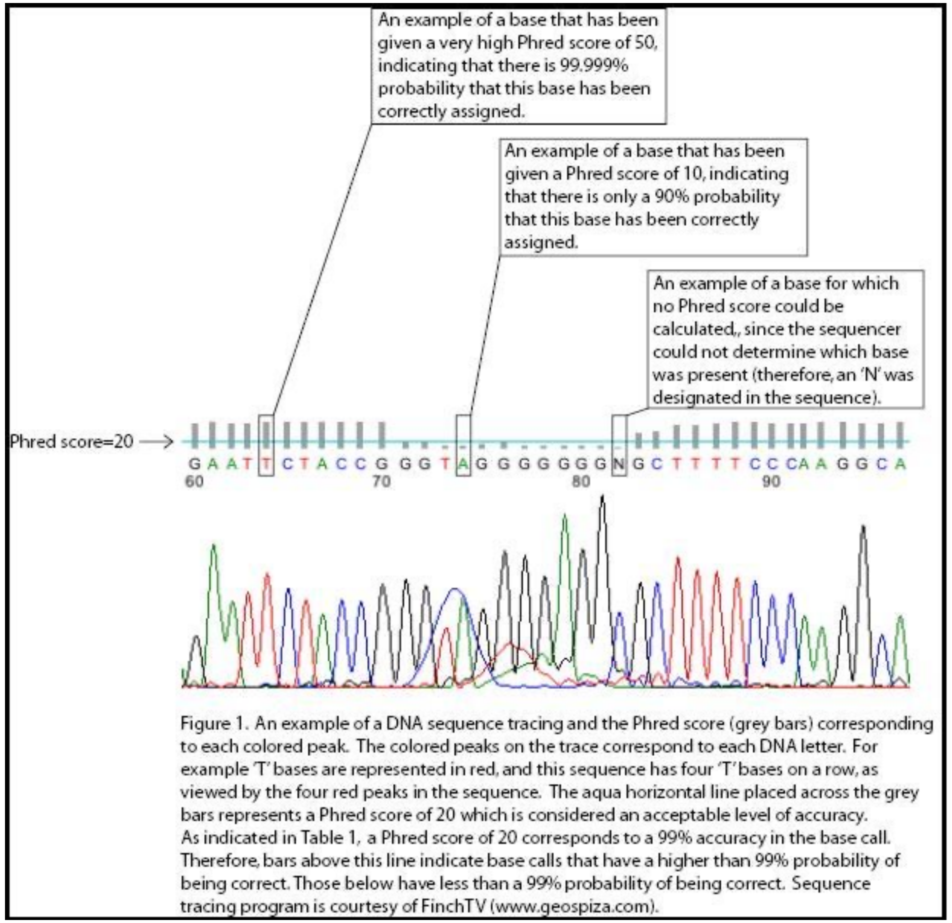
- Shows the distribution of read lengths
- For Illumina data typically 1 bar as all reads are of the same length
- For Nanopore data heavy tailed





Phred score

Origin of Phred score in Sanger sequencing data



An example of a base that has been given a very high Phred score of 50, indicating that there is 99.999% probability that this base has been correctly assigned.

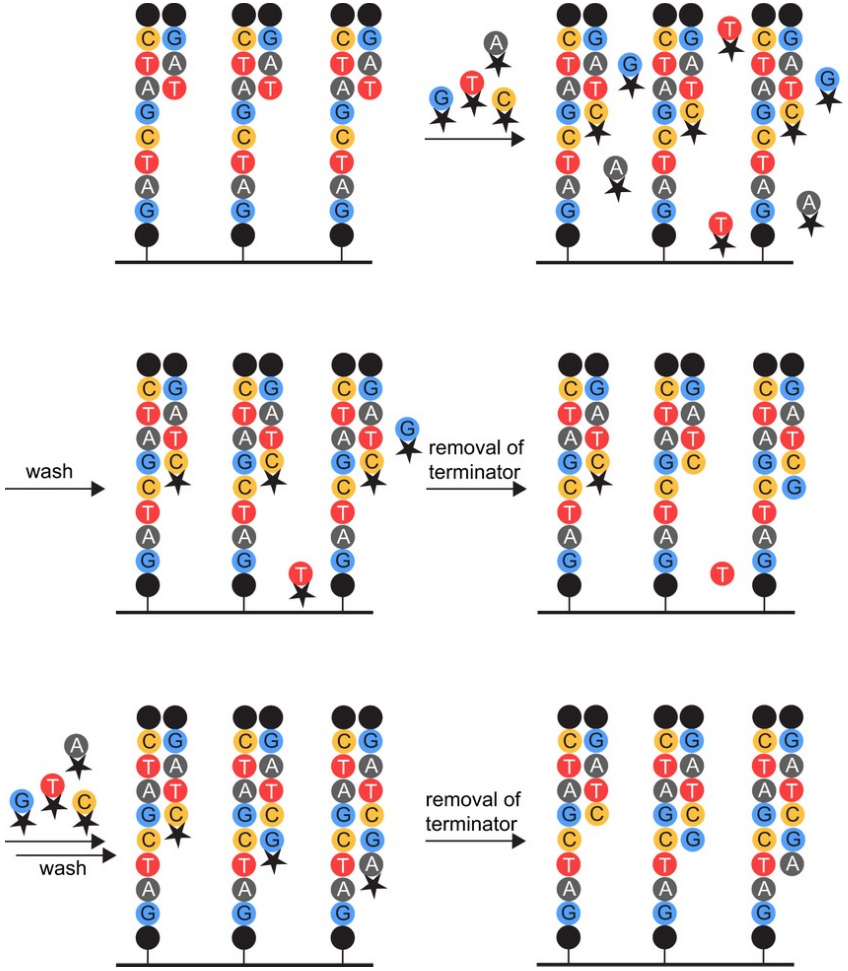
An example of a base that has been given a Phred score of 10, indicating that there is only a 90% probability that this base has been correctly assigned.

An example of a base for which no Phred score could be calculated, since the sequencer could not determine which base was present (therefore, an 'N' was designated in the sequence).

Phred score=20
G A A T T C T A C C G G G T A G G G G G G N G C T T T T C C C A A G G C A
60
70
80
90

Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy. As indicated in Table 1, a Phred score of 20 corresponds to a 99% accuracy in the base call. Therefore, bars above this line indicate base calls that have a higher than 99% probability of being correct. Those below have less than a 99% probability of being correct. Sequence tracing program is courtesy of FinchTV (www.geospiza.com).

– Pre-/Post-phasing in Illumina data



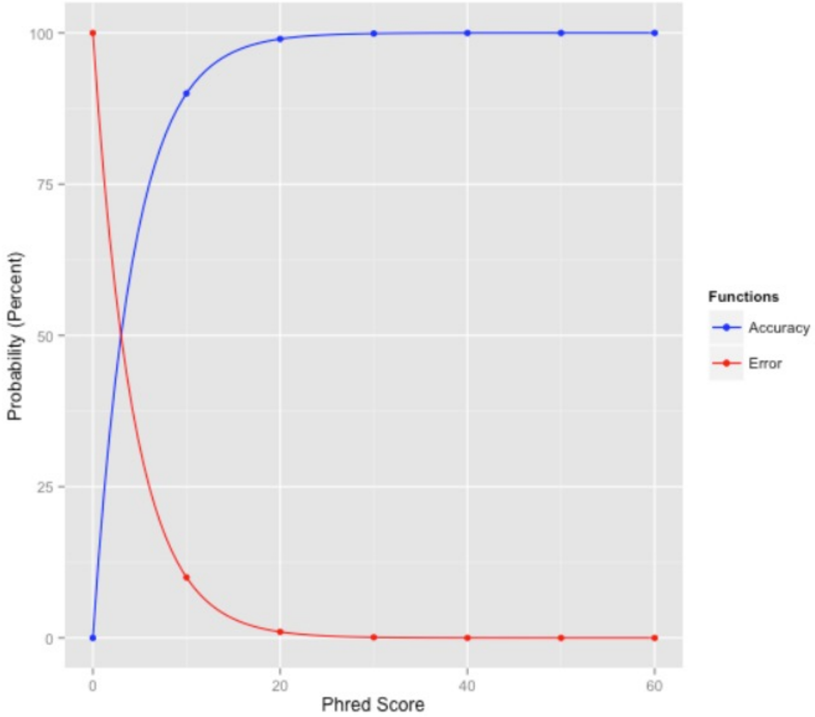
Phred score

- Q-score or Phred Quality Score
- Q30 considered gold standard for Illumina sequencing
  - 1 error in 1000 sequenced bases

$$E = 10^{-\left(\frac{Q}{10}\right)}$$

$$Q = -10 \log E$$

Phred score is inversly correlated to accuracy



Phred Quality Score	Error	Accuracy (1 - Error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%
50	1/100000 = 0.001%	99.999%
60	1/1000000 = 0.0001%	99.9999%

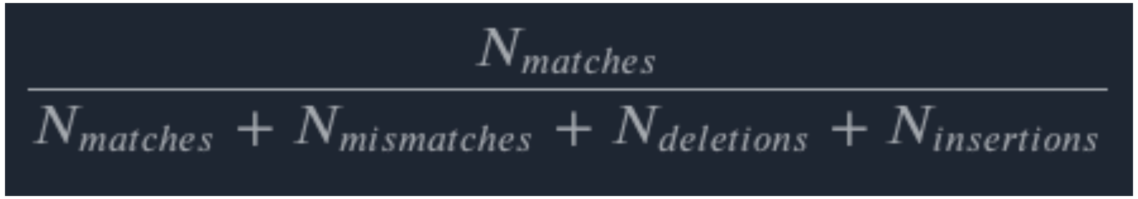
Q score for Nanopore data

- For Nanopore data there are no intensities to calculate Q-scores from

- Nanopore per base Q-scores are calculated based on then output of the neural network used by the base caller (e.g. guppy)

- Nanopore single read accuracies can be calculated by aligning reads to a reference sequence

- For more details and formulas see https://labs.epi2me.io/quality-scores/



$$\frac{N_{matches}}{N_{matches} + N_{mismatches} + N_{deletions} + N_{insertions}}$$

Chastity filter

Illumina only

% PF = clusters passing filter

Illumina run statistics show % PF

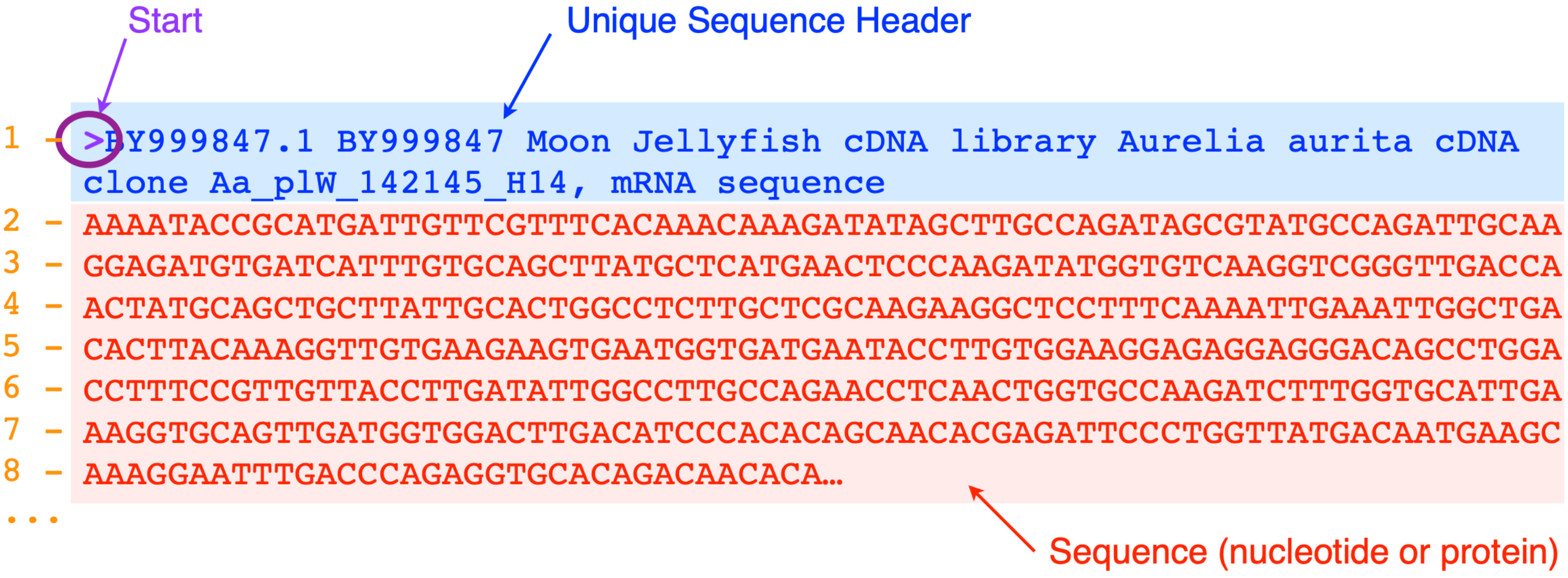
Different machines / sequencing kits have different % PF cutoffs

$$\frac{Ia}{Ia + Ib} > 0.6$$

Ia = Intensity of brightest base

Ib = Intensity of second brightest base

FASTA file format



```
>BY999847.1 BY999847 Moon Jellyfish cDNA library Aurelia aurita cDNA
clone Aa_plW_142145_H14, mRNA sequence
AAAATACCGCATGATTGTTCGTTTCACAAACAAAGATATAGCTTGCCAGATAGCGTATGCCAGATTGCAA
GGAGATGTGATCATTTGTGCAGCTTATGCTCATGAACTCCCAAGATATGGTGTCAAGGTCGGGTTGACCA
ACTATGCAGCTGCTTATTGCACTGGCCTCTTGCTCGCAAGAAGGCTCCTTTCAAAATTGAAATTGGCTGA
CACTTACAAAGGTTGTGAAGAAGTGAATGGTGATGAATACCTTGTGGAAGGAGAGGAGGGACAGCCTGGA
CCTTTCCGTTGTTACCTTGATATTGGCCTTGCCAGAACCTCAACTGGTGCCAAGATCTTTGGTGCATTGA
AAGGTGCAGTTGATGGTGGACTTGACATCCCACACAGCAACACGAGATTCCCTGGTTATGACAATGAAGC
AAAGGAATTTGACCCAGAGGTGCACAGACAACACA…
```

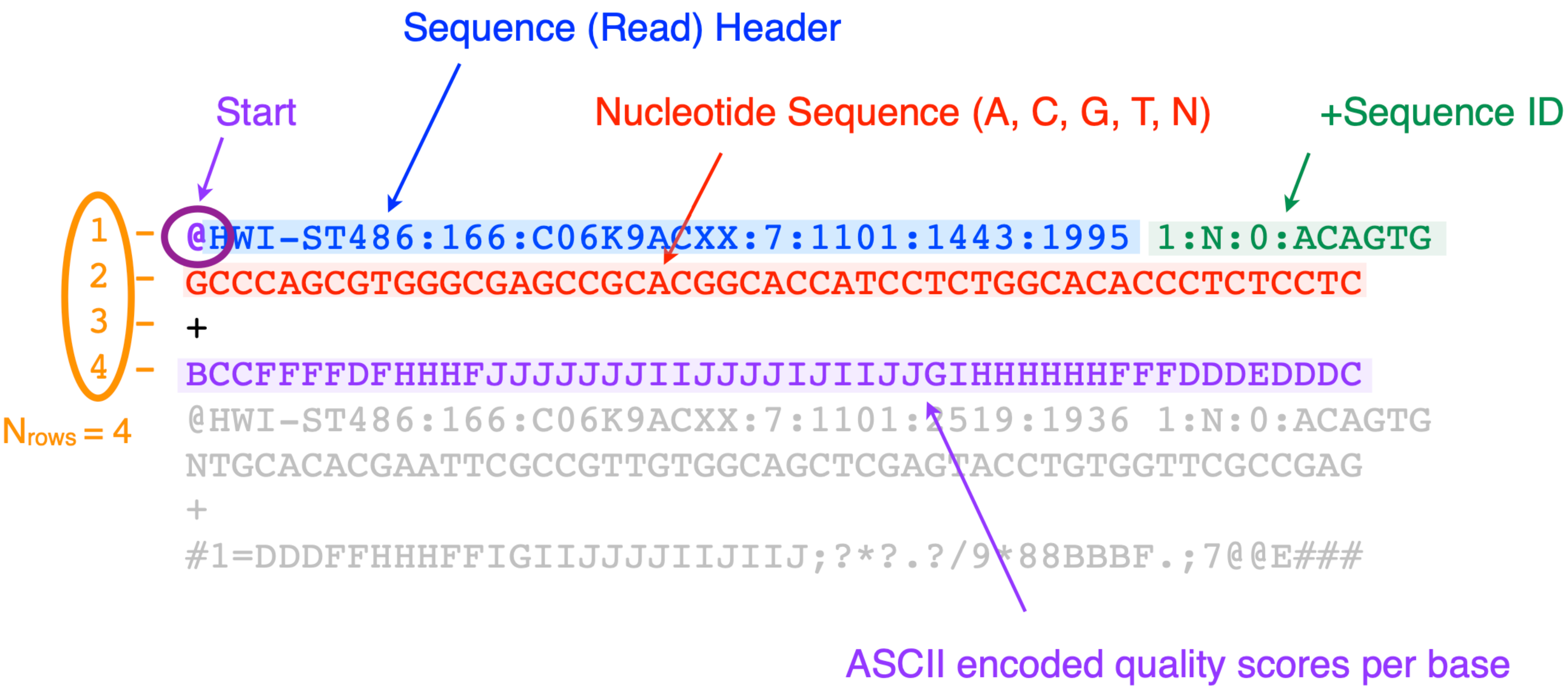
Sequence (nucleotide or protein)

File Suffix: sequence(s).fa, sequence(s).fasta

Special cases: sequences.mfa (multiple - aligned - sequences)
sequences.afa (aligned sequences)

https://www.gdc-docs.ethz.ch/GeneticDiversityAnalysis/GDA20/handouts/03_GDA20_NGS_QC.pdf

FASTQ file format



```
@HWI-ST486:166:C06K9ACXX:7:1101:1443:1995 1:N:0:ACAGTG
GCCCAGCGTGGGCGAGCCGCACGGCACCATCCTCTGGCACACCCTCTCCTC
+
BCCFFFFDFHHHFJJJJJJIIJJJJIJIIJJGIHHHHHHFFFDDDEDDDC
@HWI-ST486:166:C06K9ACXX:7:1101:2519:1936 1:N:0:ACAGTG
NTGCACACGAATTCGCCGTTGTGGCAGCTCGAGTACCTGTGGTTCGCCGAG
+
#1=DDDFFHHHFFIGIIJJJJIIJIIJ;?*?.?/9*88BBBF.;7@@E###
```

Sequence (Read) Header
Start
Nucleotide Sequence (A, C, G, T, N)
+Sequence ID
1
2
3
4
$N_{rows} = 4$
ASCII encoded quality scores per base

File Suffix: reads.fq, reads.fastq

Special cases: read_R[12].fq (> paired reads)

read_I[12].fq (> index)

https://www.gdc-docs.ethz.ch/GeneticDiversityAnalysis/GDA20/handouts/03_GDA20_NGS_QC.pdf

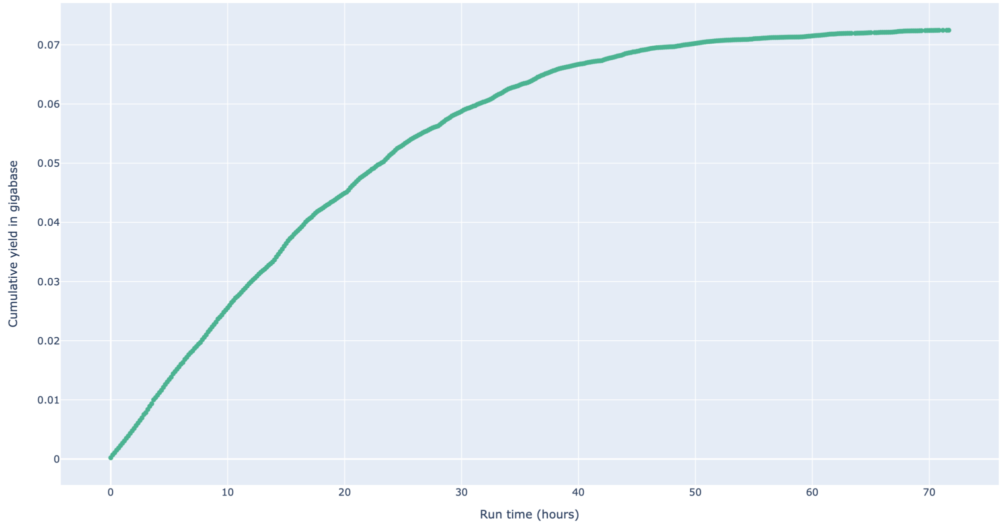
Amount of data generated

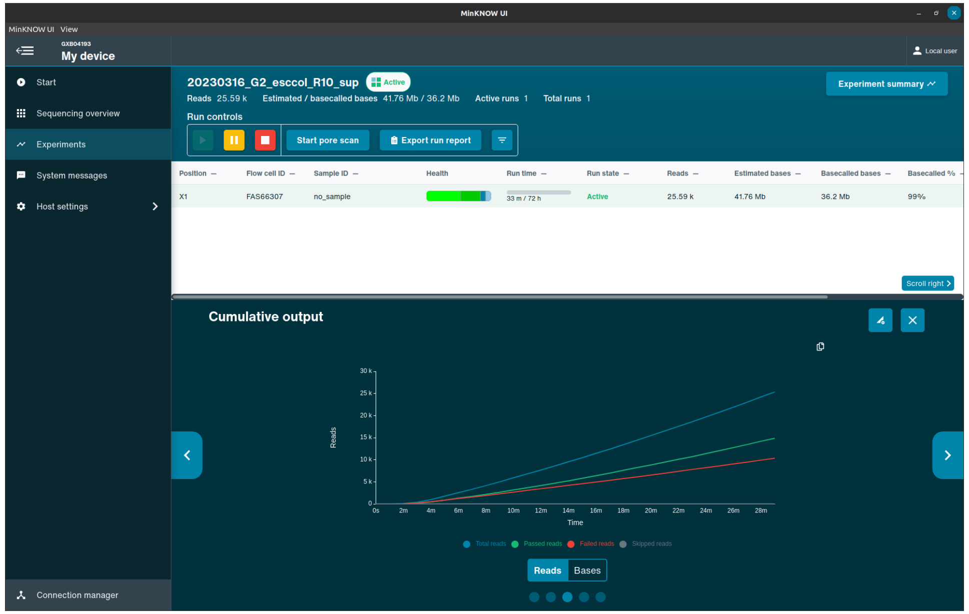
Data accumulates over time

Files of 4000 sequences saved

Max run time 72 hours

Typical output > 25 GB / flow cell





GXB04193

My device

20230316_G2_esccol_R10_sup Active

Reads 25.59 k Estimated / basecalled bases 41.76 Mb / 36.2 Mb Active runs 1 Total runs 1

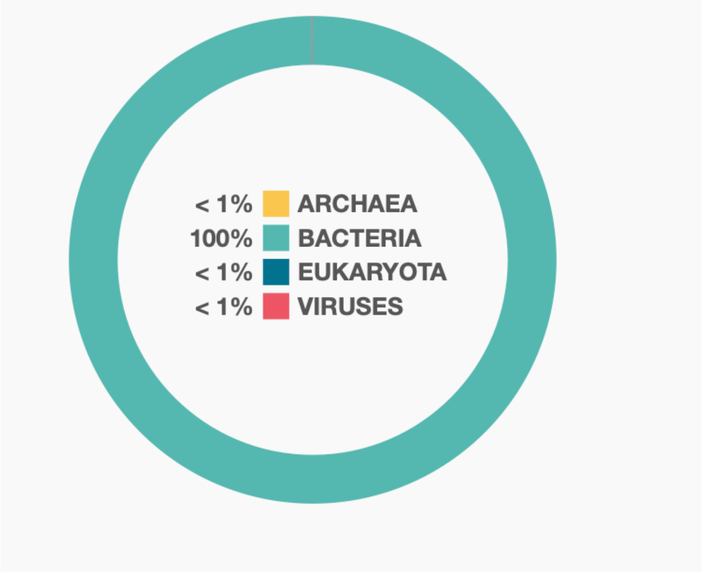
Run controls

Position	Flow cell ID	Sample ID	Health	Run time	Run state	Reads	Estimated bases	Basecalled bases	Basecalled %
X1	FAS66307	no_sample		33 m / 72 h	Active	25.59 k	41.76 Mb	36.2 Mb	99%

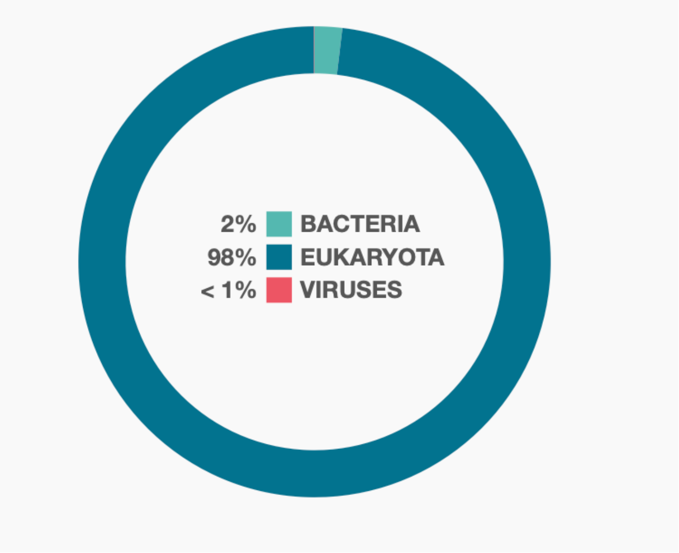
Cumulative output

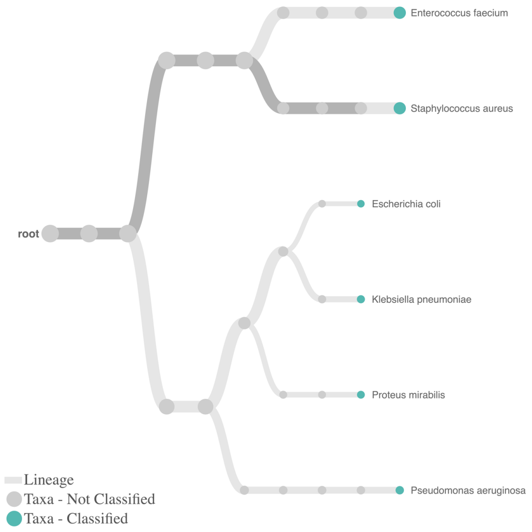
Species detected

Pure Culture



Tissue sample





Various approaches to identify species from sequencing data

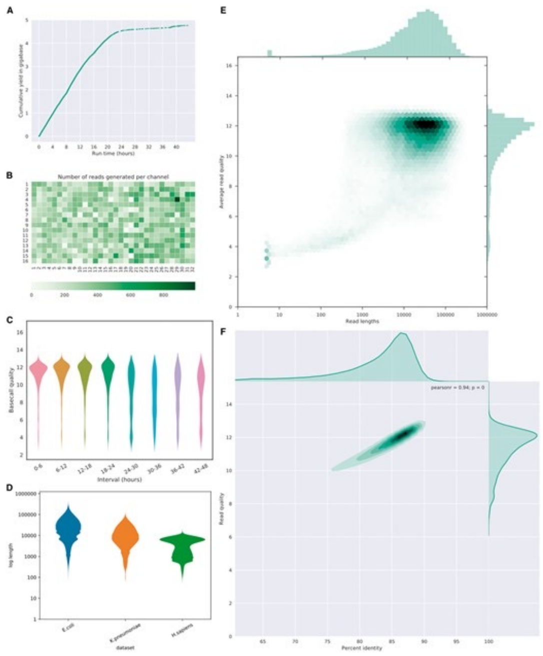
rMLST

WIMP - what is in my pot

FASTQ screen – screen against a fiven set of databases

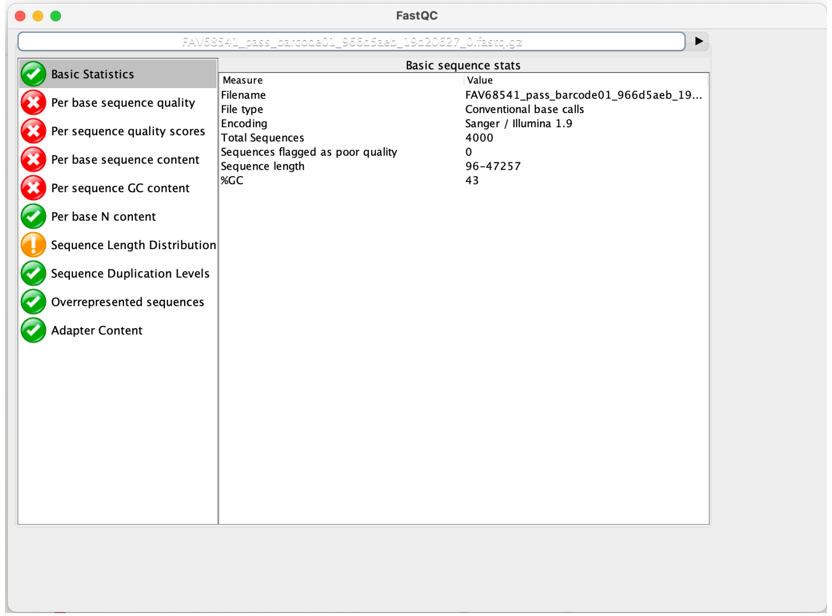
KRAKEN2 – use Kmer approach to

Software for QC



Nanoplot - tool for quality control analysis for nanopore reads.

- Input - fastq or sequencing summary file.
- Output - HTML report, summary table and plot images
- https://github.com/wdecoster/NanoPlot


FastQC
FAV68541_pass_barcode01_966d5aeb_19d20627_0.fastq.gz
✓ Basic Statistics
☒ Per base sequence quality
☒ Per sequence quality scores
☒ Per base sequence content
☒ Per sequence GC content
✓ Per base N content
! Sequence Length Distribution
✓ Sequence Duplication Levels
✓ Overrepresented sequences
✓ Adapter Content

Basic sequence stats

Measure	Value
Filename	FAV68541_pass_barcode01_966d5aeb_19...
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4000
Sequences flagged as poor quality	0
Sequence length	96-47257
%GC	43

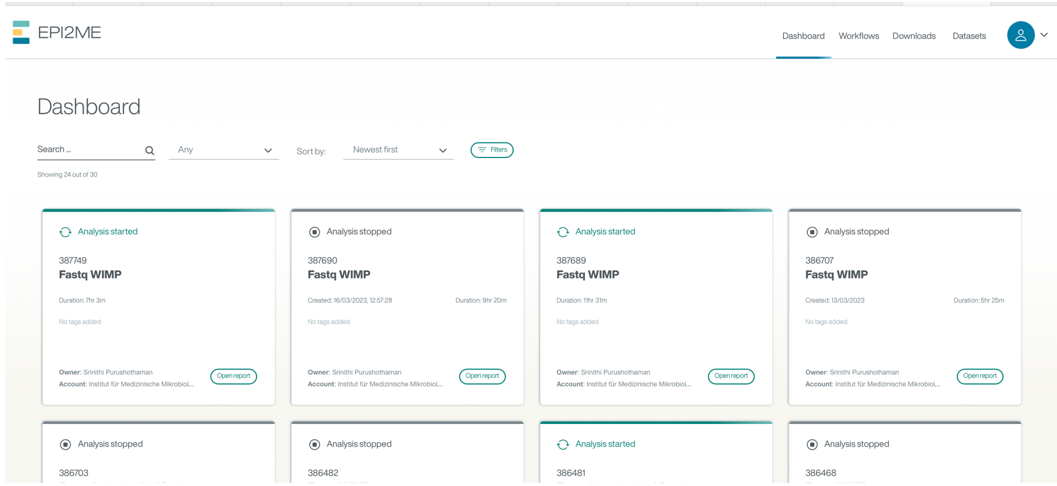
FastQC – tool for quality control analysis for fastq files

- Input – fastq files
- Output – HTML report with different graphs
- https://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc

Software from ONT for data analysis

Epi2me – online platform for on the fly data analysis with various modules

EPI2MELabs downloadable version for local analysis



Dashboard

Showing 24 out of 30

Analysis started
387749
Fastq WIMP
Duration: 7hr 3m
No tags added
Owner: Srinithi Purushothaman
Account: Institut für Medizinische Mikrobiol...

Analysis stopped
387690
Fastq WIMP
Created: 16/03/2023, 12:57:28 Duration: 9hr 20m
No tags added
Owner: Srinithi Purushothaman
Account: Institut für Medizinische Mikrobiol...

Analysis started
387689
Fastq WIMP
Duration: 11hr 31m
No tags added
Owner: Srinithi Purushothaman
Account: Institut für Medizinische Mikrobiol...

Analysis stopped
386707
Fastq WIMP
Created: 13/03/2023 Duration: 5hr 25m
No tags added
Owner: Srinithi Purushothaman
Account: Institut für Medizinische Mikrobiol...

Analysis stopped
386703

Analysis stopped
386482

Analysis started
386481

Analysis stopped
386468

https://epi2me.nanoporetech.com/report-387749

https://labs.epi2me.io/downloads/

Useful worklfows:

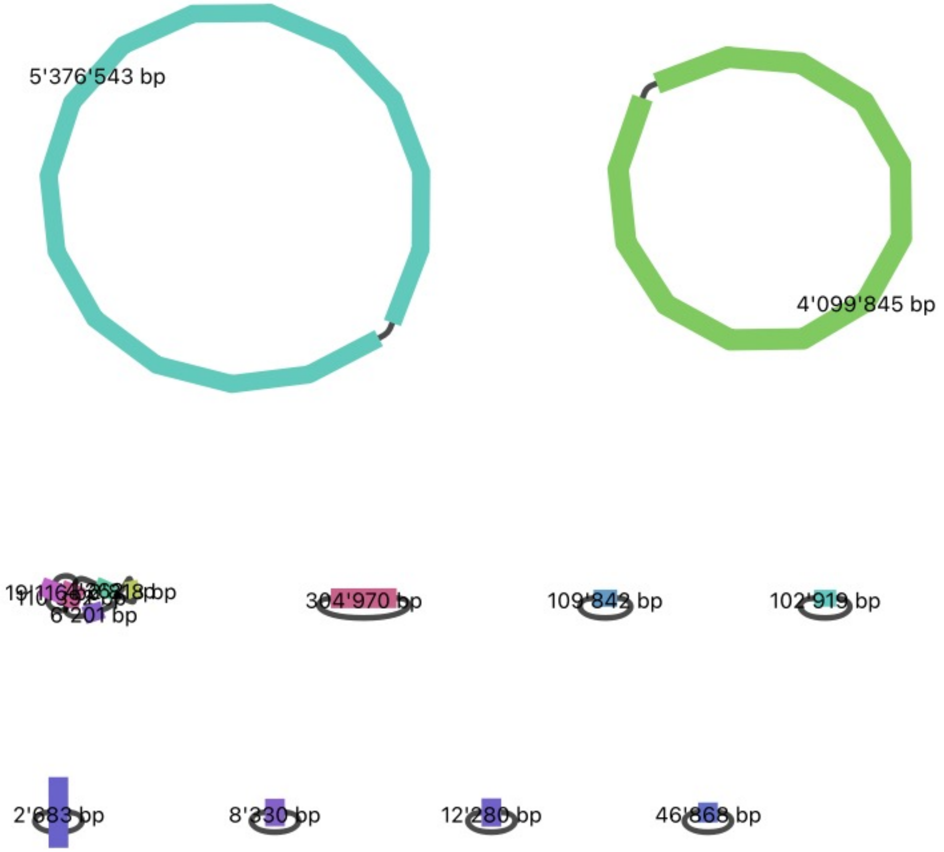
FASTQ – WIMP

- QC of reads
- What is in my pot (WIMP): bacterial classification using Centrifuge classification engine by Johns Hopkins University

FASTQ – Antimicrobial Resistance

- QC of reads
- What is in my pot (WIMP): bacterial classification using Centrifuge classification engine by Johns Hopkins University
- Antimicrobial resistance: search against CARD database

Some examples



– What has happened here?