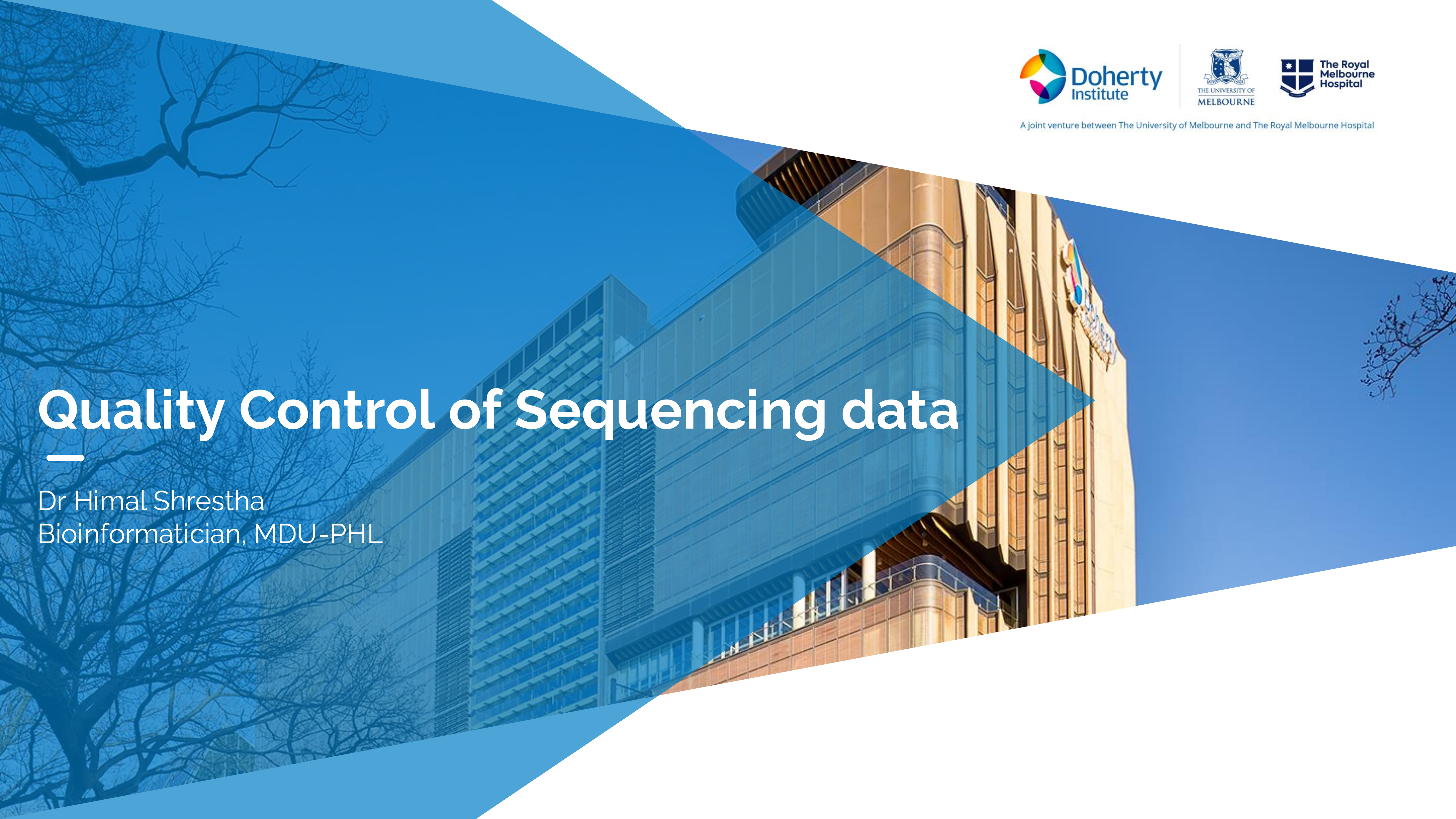
Doherty Institute

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital



# Quality Control of Sequencing data

Dr Himal Shrestha
Bioinformatician, MDU-PHL

# Objectives

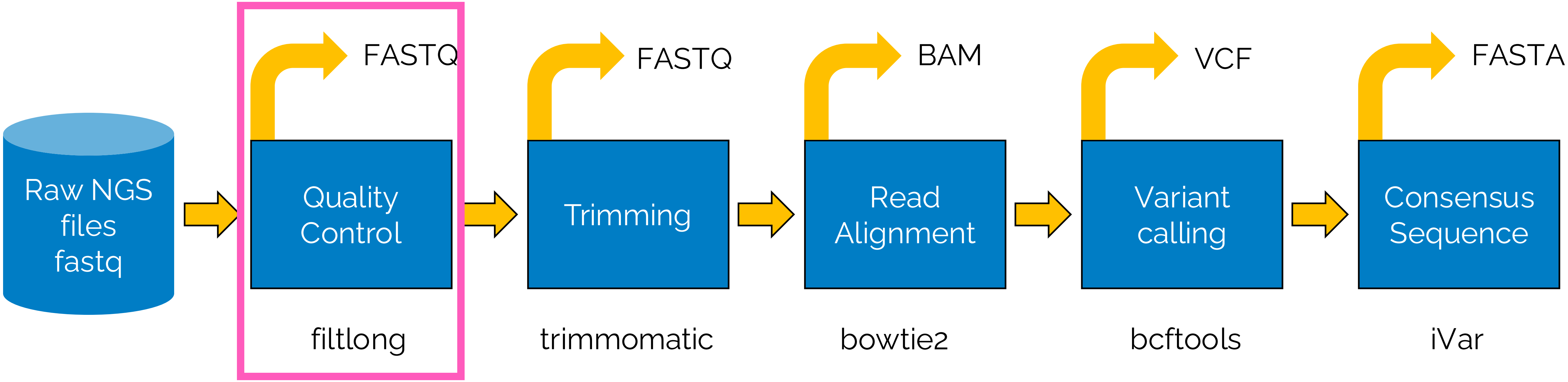
- Understand the need for quality control measures
- Different QC metrics
- Assess long reads FASTQ quality using *NanoPlot*

## Questions

- How to perform quality control of NGS raw data?
- What are the quality parameters to check for a dataset?
- How to improve the quality of a dataset?

# Typical Bioinformatics Pipeline

Today we will be doing QC section!



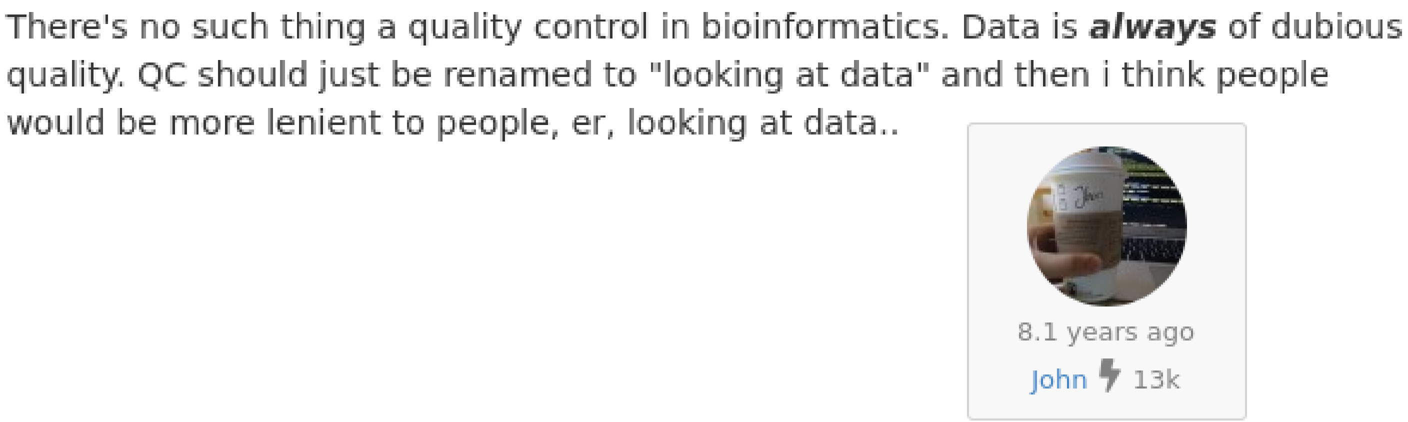
- Many file formats are used throughout a typical bioinformatics pipeline.
- Different programs will input and output different file formats.
- Each file format has different specifications, uses and limitations.

**Note**: Example only – not a real pipeline!

# Why perform Quality Control?

First step in the process is always quality control to ensure:

- Raw sequencing contains usable data (reads)
- Data is what you expect – long reads? Paired end data?
- Decide on downstream processing

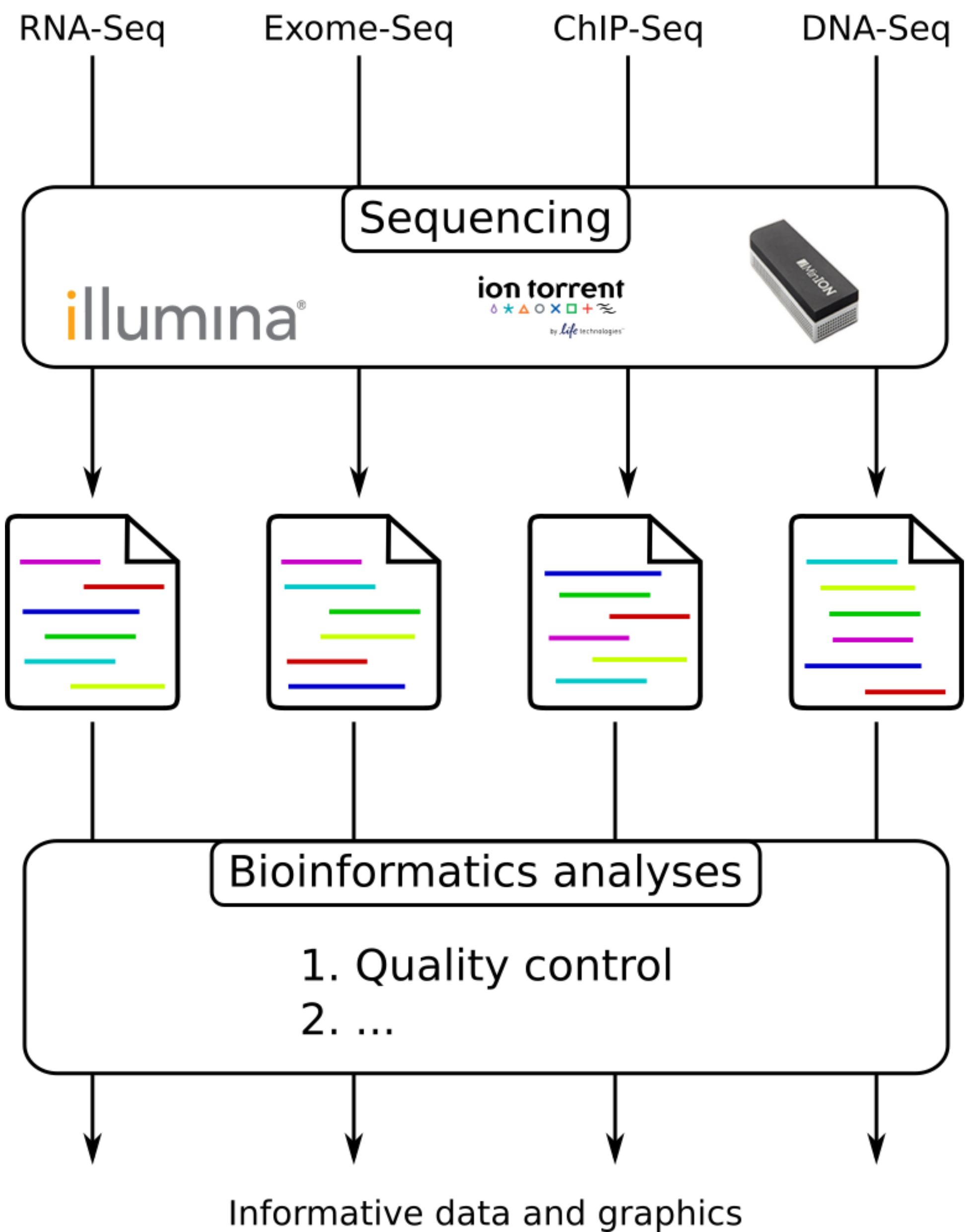


There's no such thing a quality control in bioinformatics. Data is ***always*** of dubious quality. QC should just be renamed to "looking at data" and then i think people would be more lenient to people, er, looking at data..

8.1 years ago

John 13k

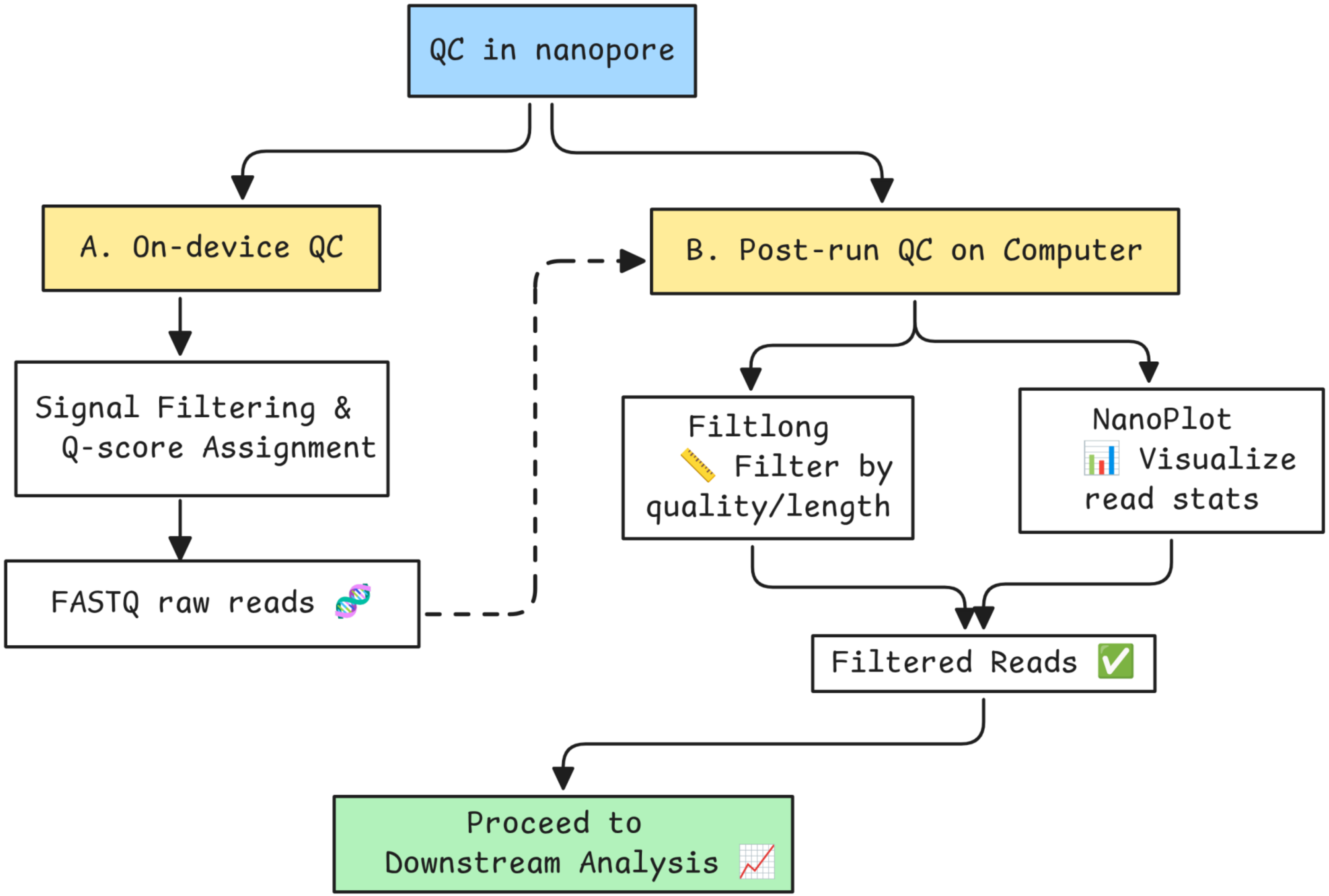
🧠 QC - Not just “Pass/Fail” but “Understand/Decide”



# QC in Nanopore

For nanopore data, QC is performed **during sequence acquisition (basecalling)** on the sequencer and **after using dedicated tools**:

- Filtlong
- NanoPlot

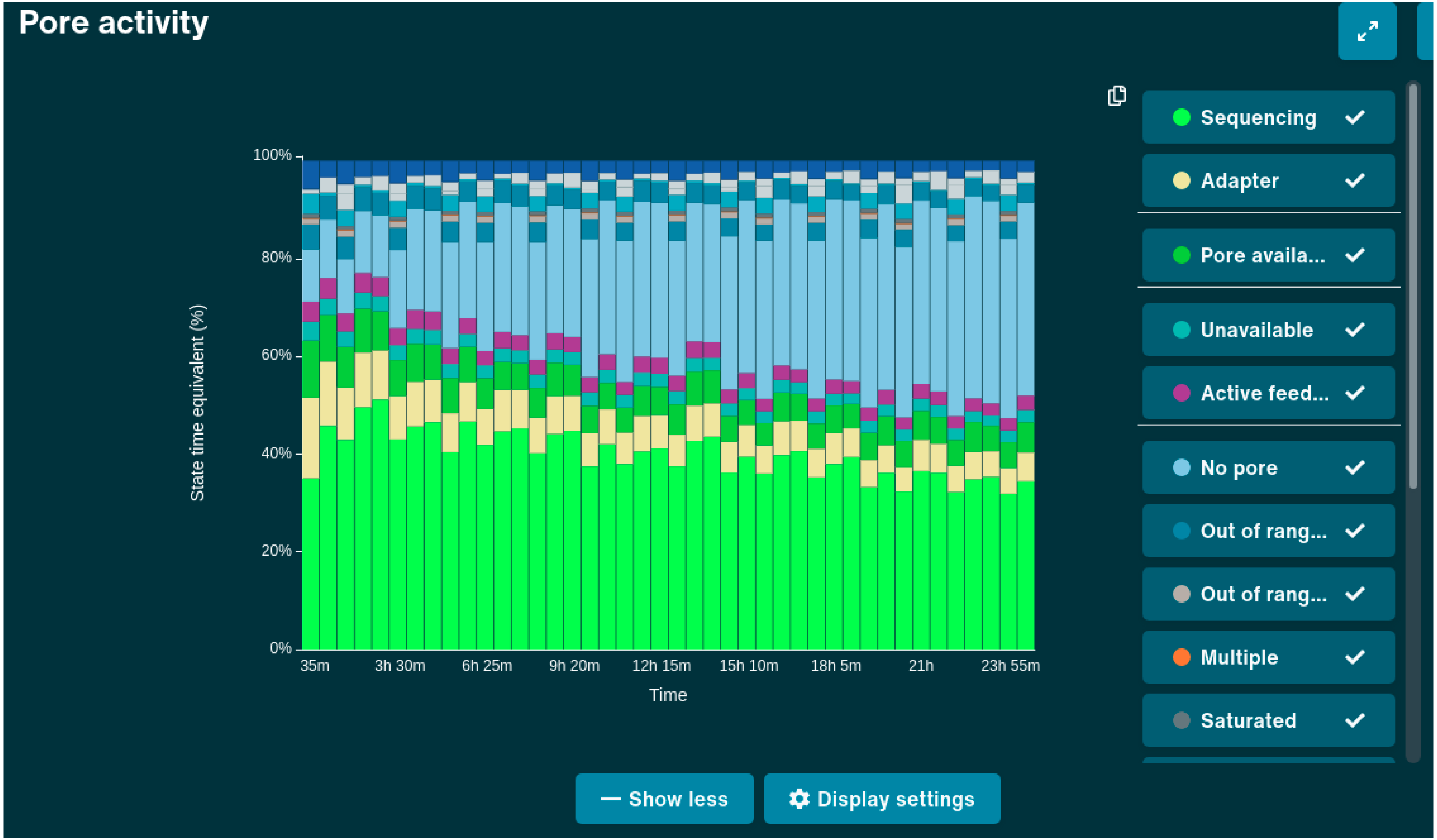

QC in nanopore
A. On-device QC
Signal Filtering & Q-score Assignment
FASTQ raw reads 🧬
B. Post-run QC on Computer
Filtlong 📏 Filter by quality/length
NanoPlot 📊 Visualize read stats
Filtered Reads ✅
Proceed to Downstream Analysis 📈

# On Machine Quality Metrics - MinKNOW

After run completion, on the computer with MinKNOW software, look at:

- Pore Activity
- Total number of reads generated
- N50 and distribution of read lengths
- Median Phred Qualities generated
- Read count per Barcode

Note: HTML file is generated in the output folder named 'report_{Minknow-Run-ID}.html' which contains these data.



Note: Old flowcell was used – possibility non-typical outputs demonstrated here

# 📊 Making sense of the pore activity plot

## ✅ What You Want to See

💚 More Green = More Sequencing = Great!

- •Green: Pores actively sequencing DNA
- •First 30–60 min: Mostly green → healthy run

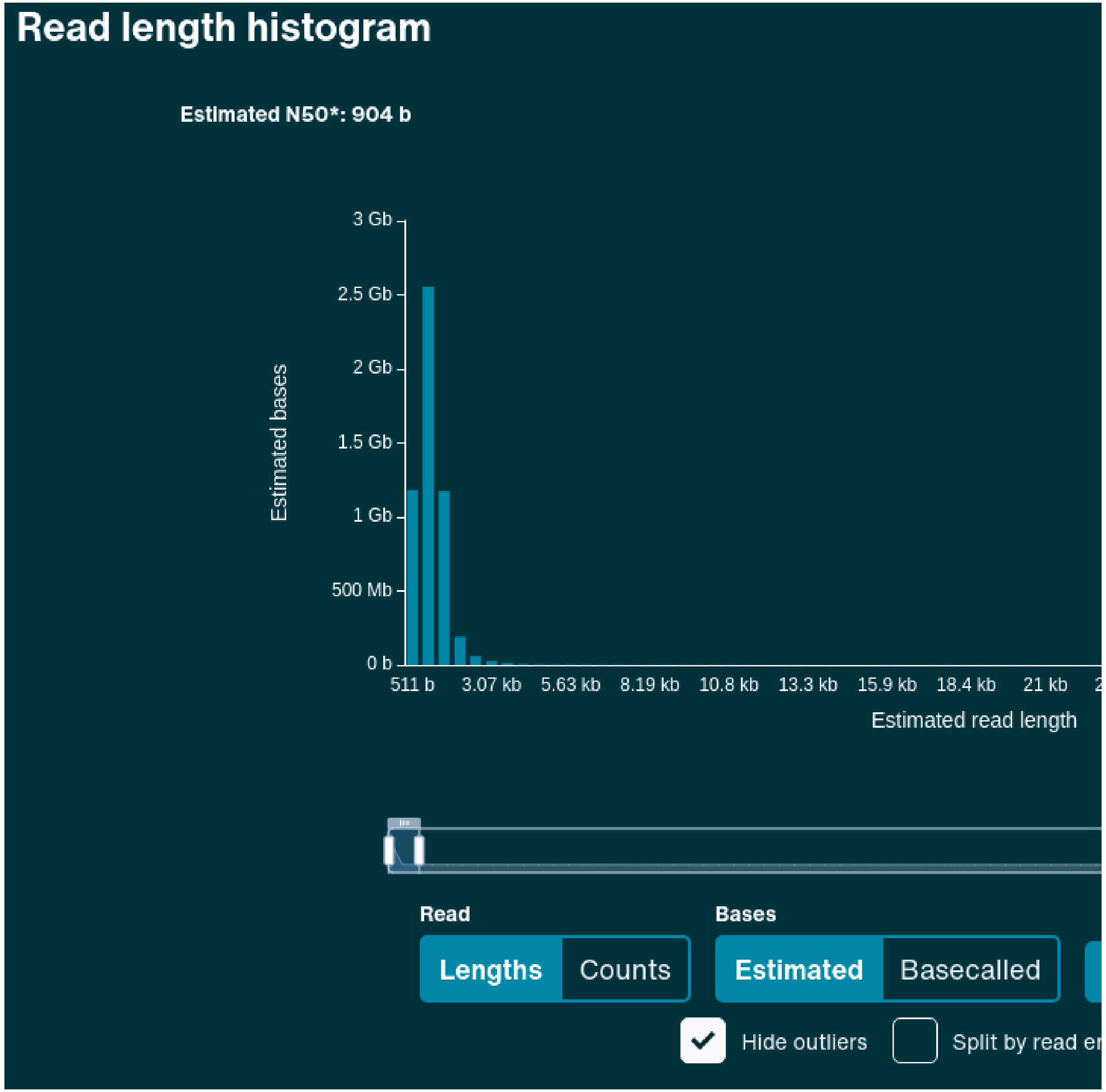
Pore available: Ready but idle → Normal if underloaded

## ⚠️ What to Watch Out For

🟥 Red / Yellow: Trouble!

- •'Unavailable', 'No Pore' → Clogging, contamination, bad sample prep
- •'Adapter', 'Active Feed…' → DNA stuck, poor ligation

'Multiple', 'Saturated' → Overloaded or signal issues



Read length histogram

Estimated N50*: 904 b

Read: **Lengths** Counts

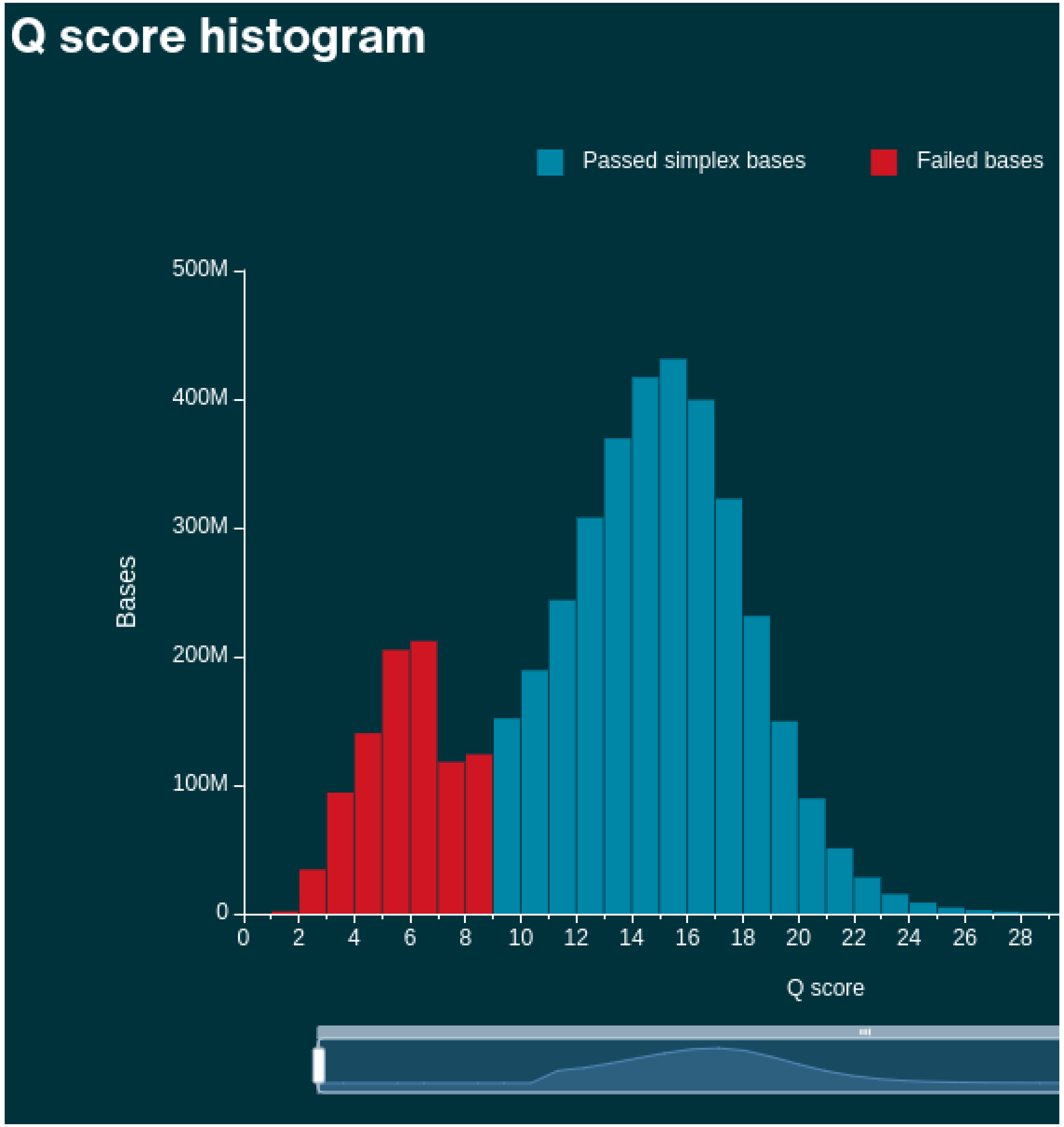
Bases: **Estimated** Basecalled

☑ Hide outliers □ Split by read er

Questions to ask:

- How long of a product did I sequence?
  - Tiled Amplicon is limited in size
  - MPXV schemes are generally ~2500
- What kit (ligation vs rapid) did I use?
  - Rapid produces shorter and more fragmented reads.

Note: Old flowcell was used – possibility non-typical outputs demonstrated here



R10 flow cells should produce reads a histogram like this one

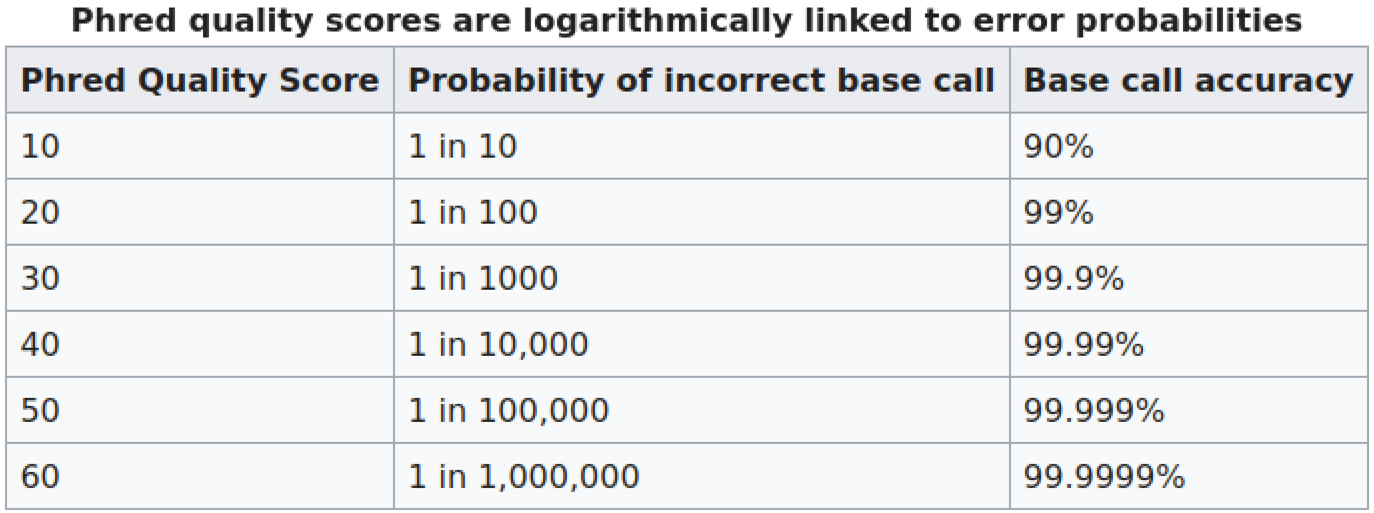
- Q Score around 15 (~97% accurate)
- Red is below threshold (usually 9)

Note: Old flowcell was used – possibility non-typical outputs demonstrated here

# Phred Scores Explained

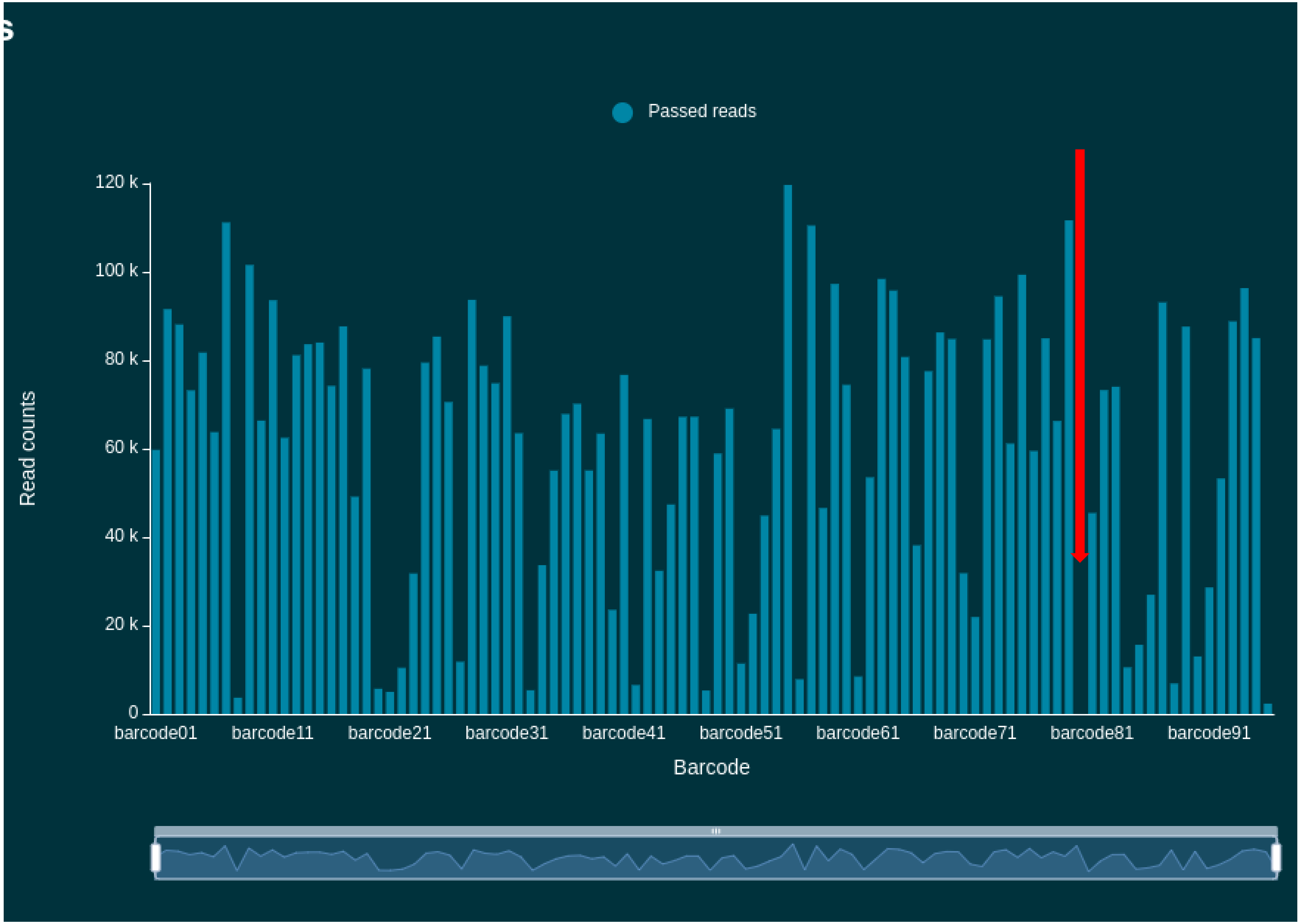
A **Phred quality score** is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing.

$$Q_{\text{sanger}} = -10 \log_{10} p$$



Phred quality scores are logarithmically linked to error probabilities

| Phred Quality Score | Probability of incorrect base call | Base call accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |
| 60 | 1 in 1,000,000 | 99.9999% |



Counts of reads per barcode is related to how well you've balanced your barcodes, and the quality and quantity of input material.

**Question to ask yourself:**

- Which are the positive and negative controls? Does the read count for controls make sense?
- Can you guess what the barcode number 80 is with no reads? (red arrow)

Note: Old flowcell was used – possibility non-typical outputs demonstrated here

# Questions? + Resources

ONT Video on performing QC:

- https://community.nanoporetech.com/nanopore_learning/lessons/introduction-to-read-quality-assessment-and-filtering

Galaxy Training Network – hands on QC (long, short reads)

- https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#histogram-of-read-lengths

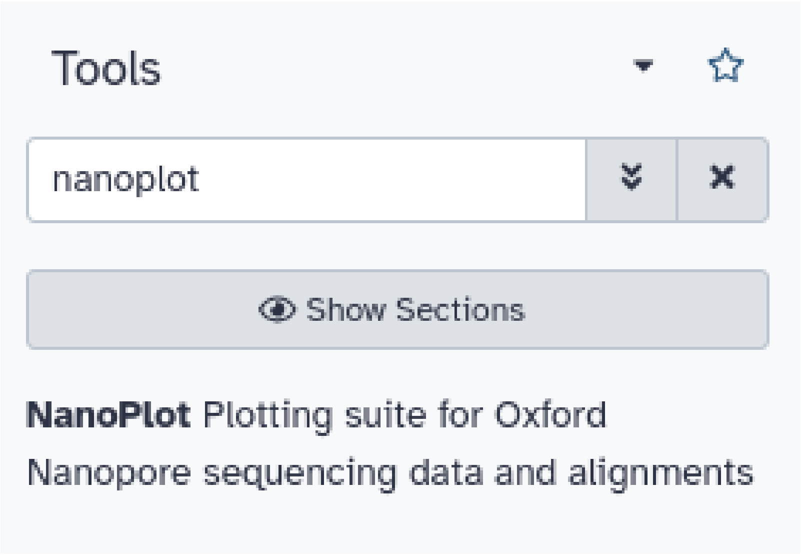
SIB Course – QC Lecture (Video) for Illumina:

- https://sib-swiss.github.io/NGS-introduction-training/latest/day1/quality_control/

SIB – Course QC (CLI) for Nanopore

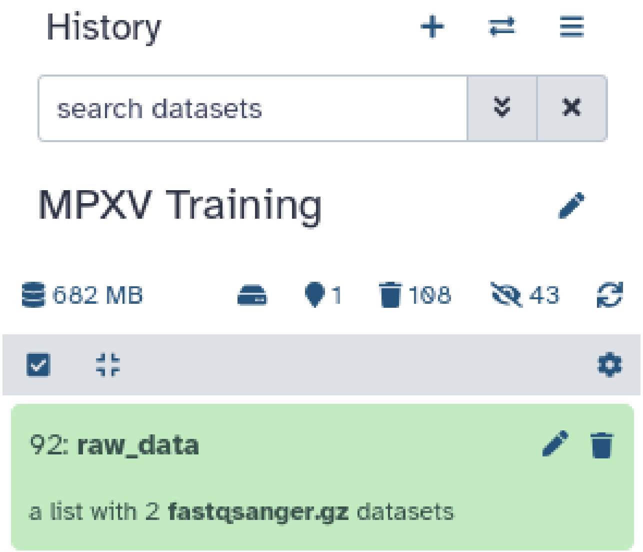
- https://sib-swiss.github.io/NGS-longreads-training/latest/course_material/qc_alignment/

# Hands-on with NanoPlot for QC



nanoplot

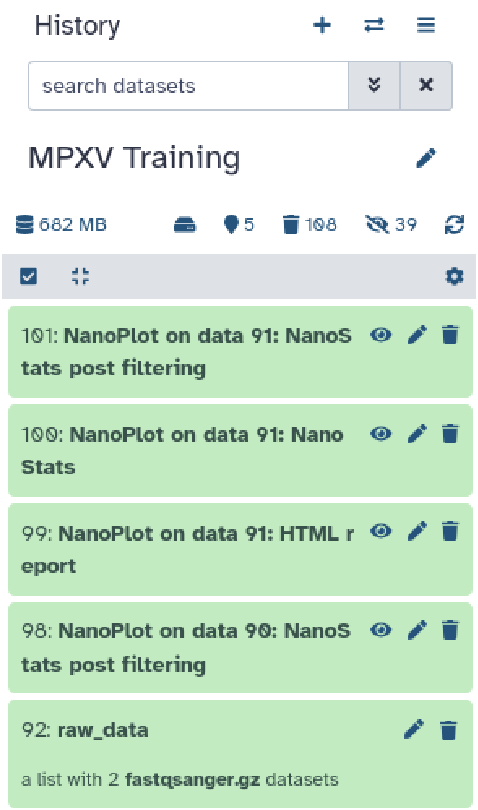
**NanoPlot** Plotting suite for Oxford Nanopore sequencing data and alignments



MPXV Training

92: **raw_data**

a list with 2 **fastqsanger.gz** datasets



MPXV Training

101: **NanoPlot on data 91: NanoStats post filtering**

100: **NanoPlot on data 91: NanoStats**

99: **NanoPlot on data 91: HTML report**

98: **NanoPlot on data 90: NanoStats post filtering**

92: **raw_data**

a list with 2 **fastqsanger.gz** datasets

- On the course webpage see the TUTORIAL link for how to run Nanoplot.
- Run NanoPlot on your raw data – ~10 minutes
- Once submitted, we can discuss NanoPlot output

# Instruction to run NanoPlot for QC

**1. Prepare Your Input**

- If you have multiple FASTQ files, combine them into a Collection

**2. Find NanoPlot Tool:**

- In the Tools panel, search for “NanoPlot” and click to open it.

**3. Specify Input:**

- For single file: Select your FASTQ file.
- For multiple files: Enable Batch Mode, and select the FASTQ Collection.

**4. Customize Your Plot:**

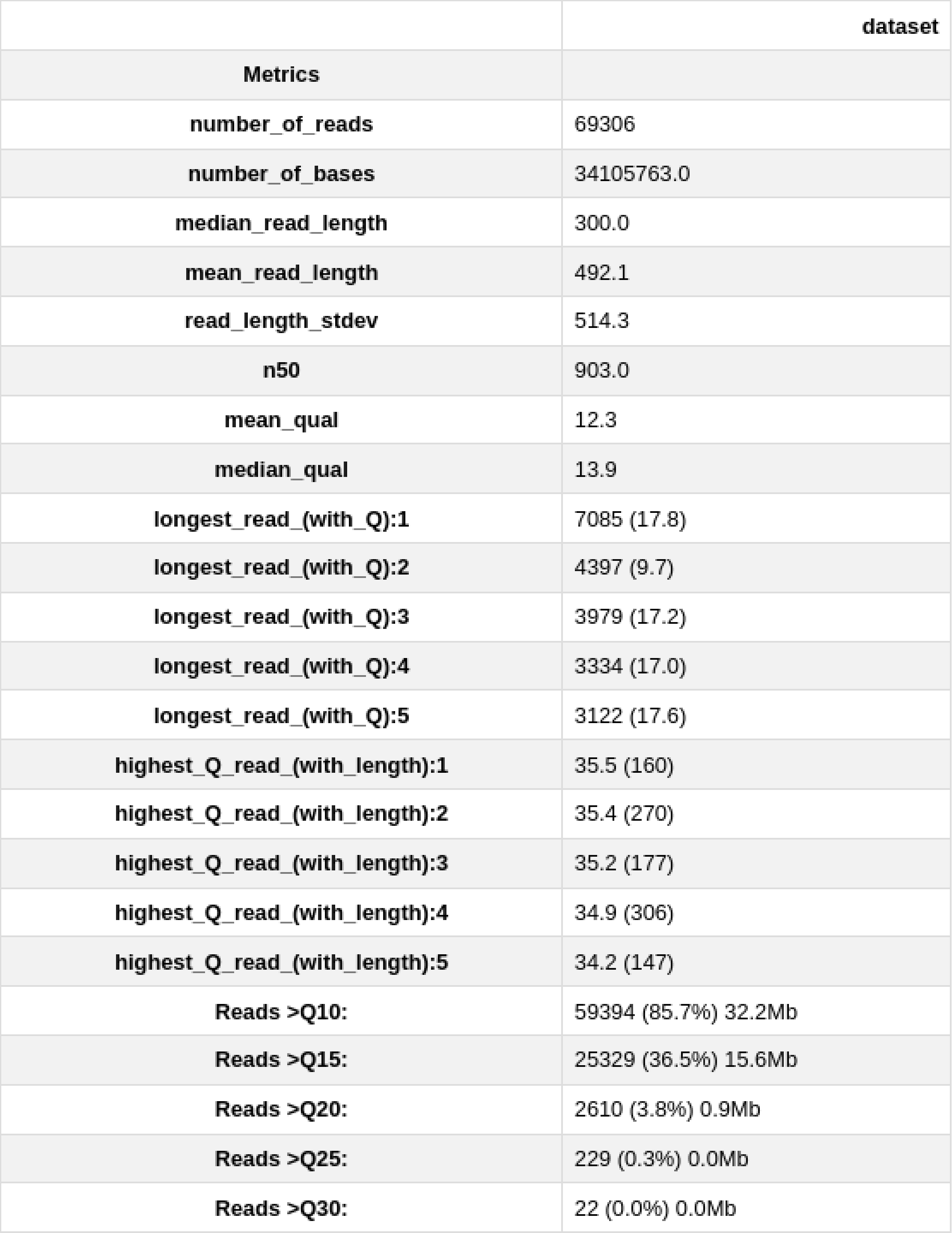
- Choose bivariate plot format: dot (for individual reads) or kde (for smooth density view).
- Enable N50 marker on read length histogram: Set to Yes.

**5. Run the Tool and explore the output plots & stats**

# Nanoplot Output - Table

## Questions:

1. What is the total number of reads?
2. What is the mean quality? Is this what you expect from your data?
3. How many reads are above Q10? Q20?



|  | dataset |
|---|---|
| Metrics |  |
| number_of_reads | 69306 |
| number_of_bases | 34105763.0 |
| median_read_length | 300.0 |
| mean_read_length | 492.1 |
| read_length_stdev | 514.3 |
| n50 | 903.0 |
| mean_qual | 12.3 |
| median_qual | 13.9 |
| longest_read_(with_Q):1 | 7085 (17.8) |
| longest_read_(with_Q):2 | 4397 (9.7) |
| longest_read_(with_Q):3 | 3979 (17.2) |
| longest_read_(with_Q):4 | 3334 (17.0) |
| longest_read_(with_Q):5 | 3122 (17.6) |
| highest_Q_read_(with_length):1 | 35.5 (160) |
| highest_Q_read_(with_length):2 | 35.4 (270) |
| highest_Q_read_(with_length):3 | 35.2 (177) |
| highest_Q_read_(with_length):4 | 34.9 (306) |
| highest_Q_read_(with_length):5 | 34.2 (147) |
| Reads >Q10: | 59394 (85.7%) 32.2Mb |
| Reads >Q15: | 25329 (36.5%) 15.6Mb |
| Reads >Q20: | 2610 (3.8%) 0.9Mb |
| Reads >Q25: | 229 (0.3%) 0.0Mb |
| Reads >Q30: | 22 (0.0%) 0.0Mb |

# 📏 Mean vs Median vs N50 – what's the difference?

- Mean: Average read length (add all, divide by count)
- Median: Middle read length when sorted
- N50: The read length where 50% of the total bases are in reads of this length or longer (sort from longest to shortest)

### 💡 Why does it matter?

- Mean can be skewed by very long reads (outliers)
- Median gives the middle of the read length distribution (ignores base count)
- N50 gives information on data usability
  - If high N50, it means you have long reads that cover a lot of the genome, which is good for assembly and variant detection.
  - If low N50, it means there are many short reads.

# 📏 Mean vs Median vs N50 – what's the difference? (contd.)

## 🧮 Dummy example

- Read lengths (bp): [900, 1000, 1100, 2100, 2200]

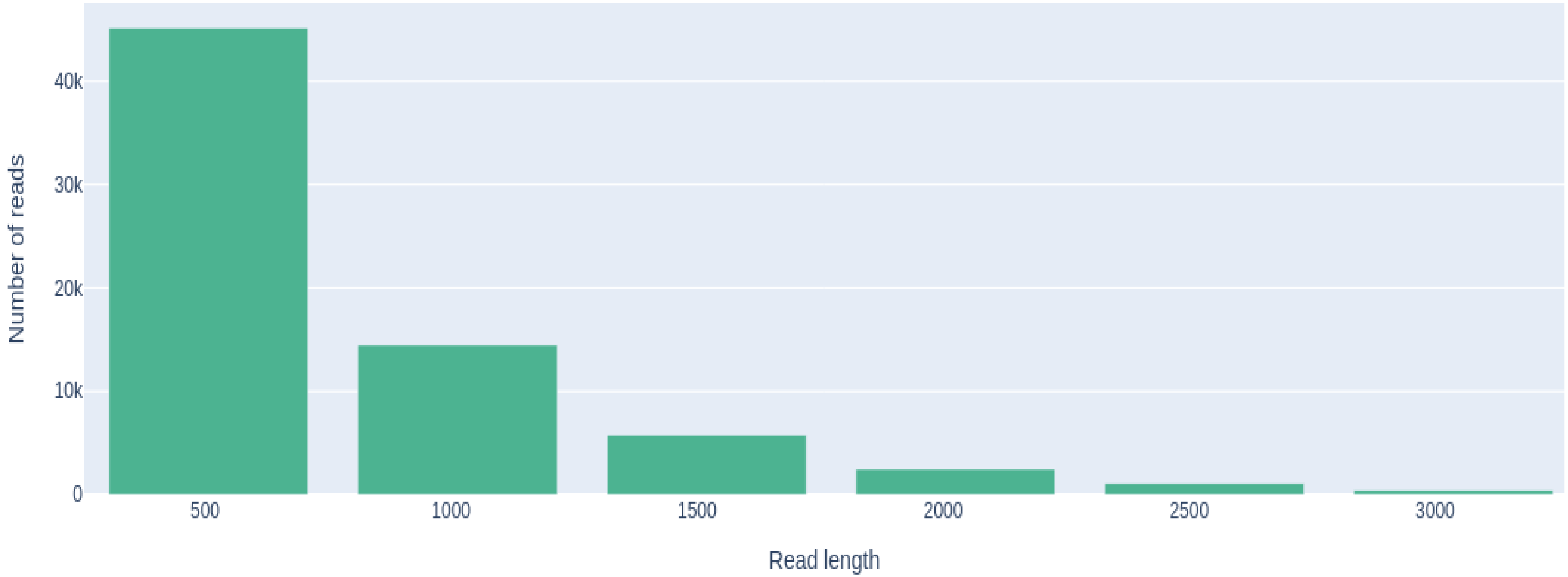
## 🔢 Calculations

- **Mean** = (900 + 1000 + 1100 + 2100 + 2200) / 5 = 7300 / 5 = **1460 bp**
- **Median** = **1100 bp** (middle value when sorted)
- **N50**:
  - Descending: [2200, 2100, 1100, 1000, 900]
  - Total bases = 7,300 → half = 3,650
  - **Cumulative**:
    - 2200 → 2200 (below half)
    - +2100 → 4300 (above half) ✅ → **N50 = 2100 bp**

## 🧠 Takeaway

- Mean and median shows a middle value but does not represent the base content
- N50 reveals that 50% of total bases come from just the two longest reads — informative for downstream analysis.

# Nanoplot Output – Histogram of Read Lengths

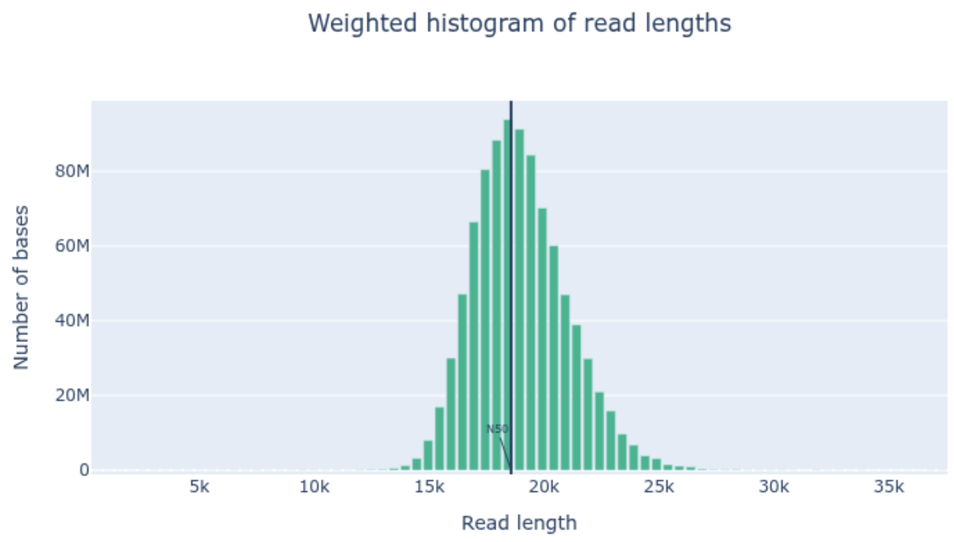


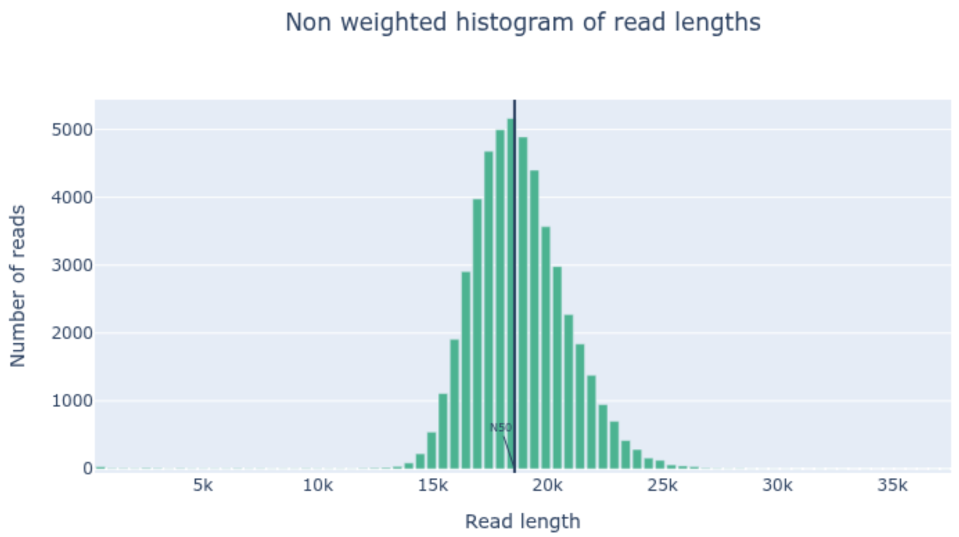
This plot shows the distribution of fragment sizes in the file that was analyzed. Long reads have a variable length and this will show the relative amounts of each different size of sequence fragment.

In this example, the distribution of read length is skewed towards 500 bp but the results can be very different depending of your experiment.

**Question**: Our amplicon scheme is ~2500-3000bp, but the read length distribution is smaller. Why?

# Weighted vs. Non-Weighted Histograms





## Weighted Histogram

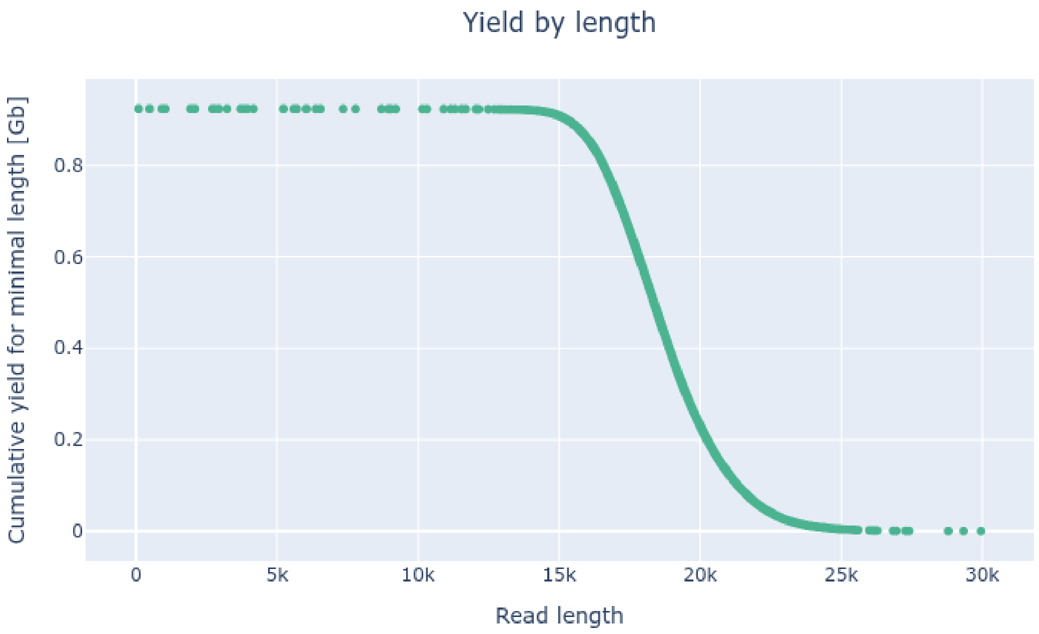
- Focuses on total base output
- Y-axis = Total number of bases per read length bin
- Highlights which read lengths contribute most to total yield

## Non-Weighted Histogram

- Focuses on read count
- Y-axis = Number of reads per read length bin
- Shows most common read sizes in your dataset

💡 Weighted plot helps reveal - longer reads even though fewer can still dominate the total base output

# Nanoplot Output – Cumulative yield by read length

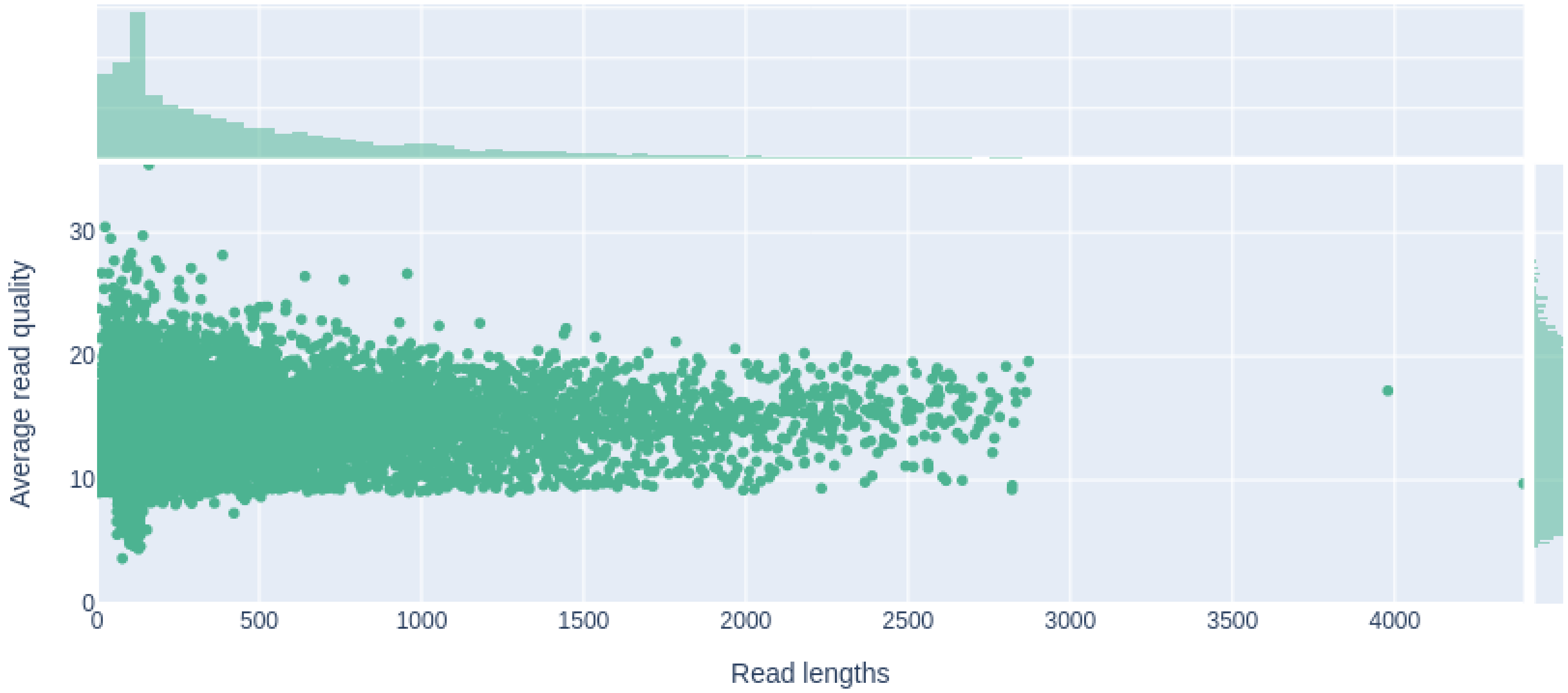


**What Does This Plot Show?**
Y-axis: Total number of bases (in Gb) → Includes only reads **equal to or longer** than each corresponding read length on the X-axis.

Helps assess how much of your yield comes from long reads.

🧠 Insight:
If most of your sequencing output is from long reads, you'll see a sharper decline at the right side of the plot (which is good ✅ )

# Nanoplot Output – Read Lenghts vs Average read quality



This plot shows the distribution of fragment sizes according to the Qscore in the file which was analysed.

In general, there is **no link between read length and read quality** but this representation allows to visualise both information into a single plot and detect possible aberrations.

In runs with a lot of short reads the shorter reads are sometimes of lower quality than the rest.

# Exercise - NanoPlot

99: NanoPlot on data 91: HTML report

- Choose one (1) sample
- Inspect the HTML file
- Answer the below questions

**Questions:**

1. What is the total number of reads?
2. What is the mean quality? Is this what you expect from your data?
3. How many reads are above Q10? Q20?

**Question**: Our amplicon scheme is ~2500-3000bp, but the read length distribution is smaller. Why?

**Question:** Looking at “Read lengths vs Average read quality plot using dots plot”. Did you notice something unusual with the Qscore?

# Thank you!

Any questions?