# Bioinformatics workflows for Oxford Nanopore ampliconic data

John A. Juma

Animal and Human Health

International Livestock Research Institute (ILRI)

March, 2023

j.juma@cgiar.org

ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

CGIAR

AFRICA CDC Centres for Disease Control and Prevention Safeguarding Africa's Health

African Union

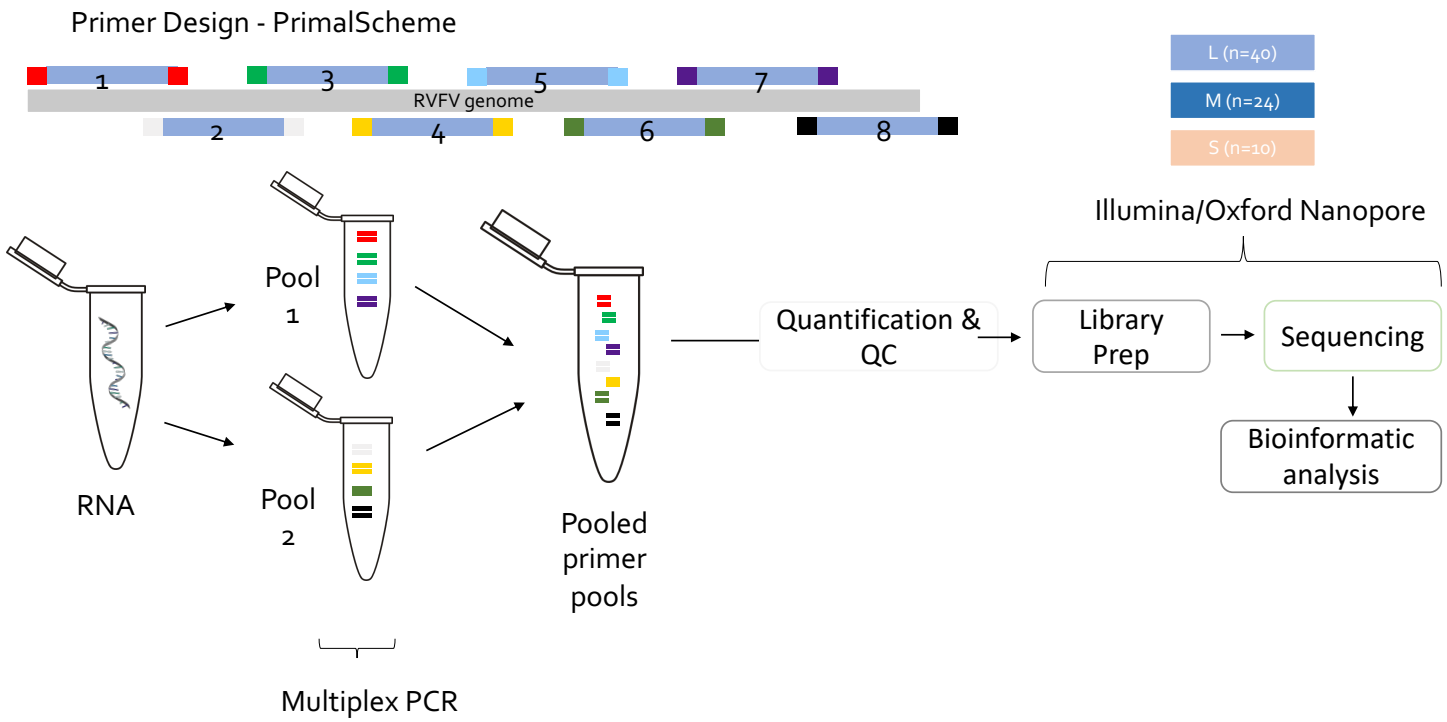
ASLM AFRICAN SOCIETY FOR LABORATORY MEDICINE

mastercard foundation

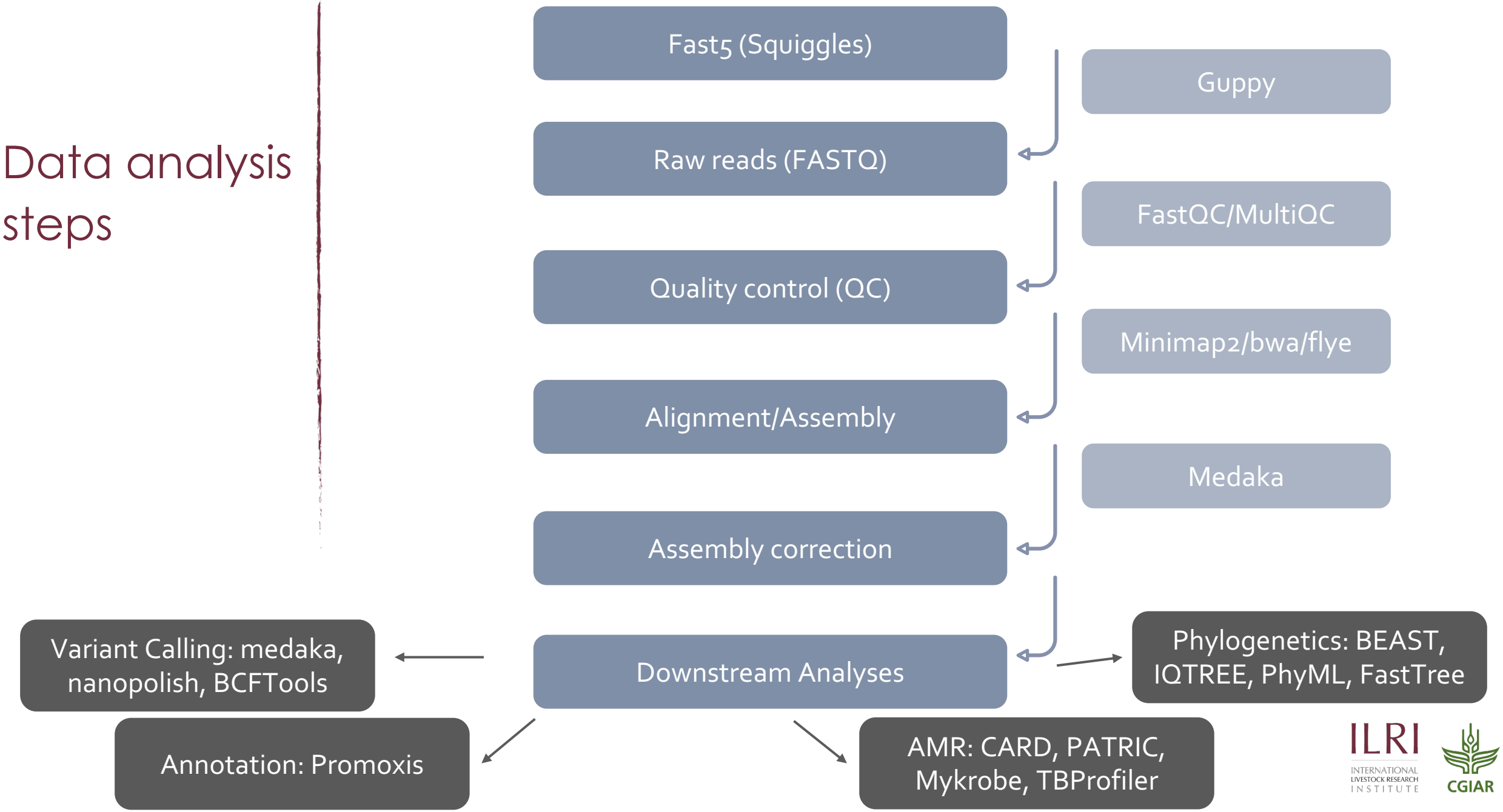
# Objectives

- Quality control procedures for ONT MinION data
- Reference-guided alignment of NGS data
- Variant calling workflows – emphasis on medaka
- Learn common terminology in NGS steps and understand concepts of depth and coverage
- Consensus sequence building

# Primer schemes



# Data analysis steps

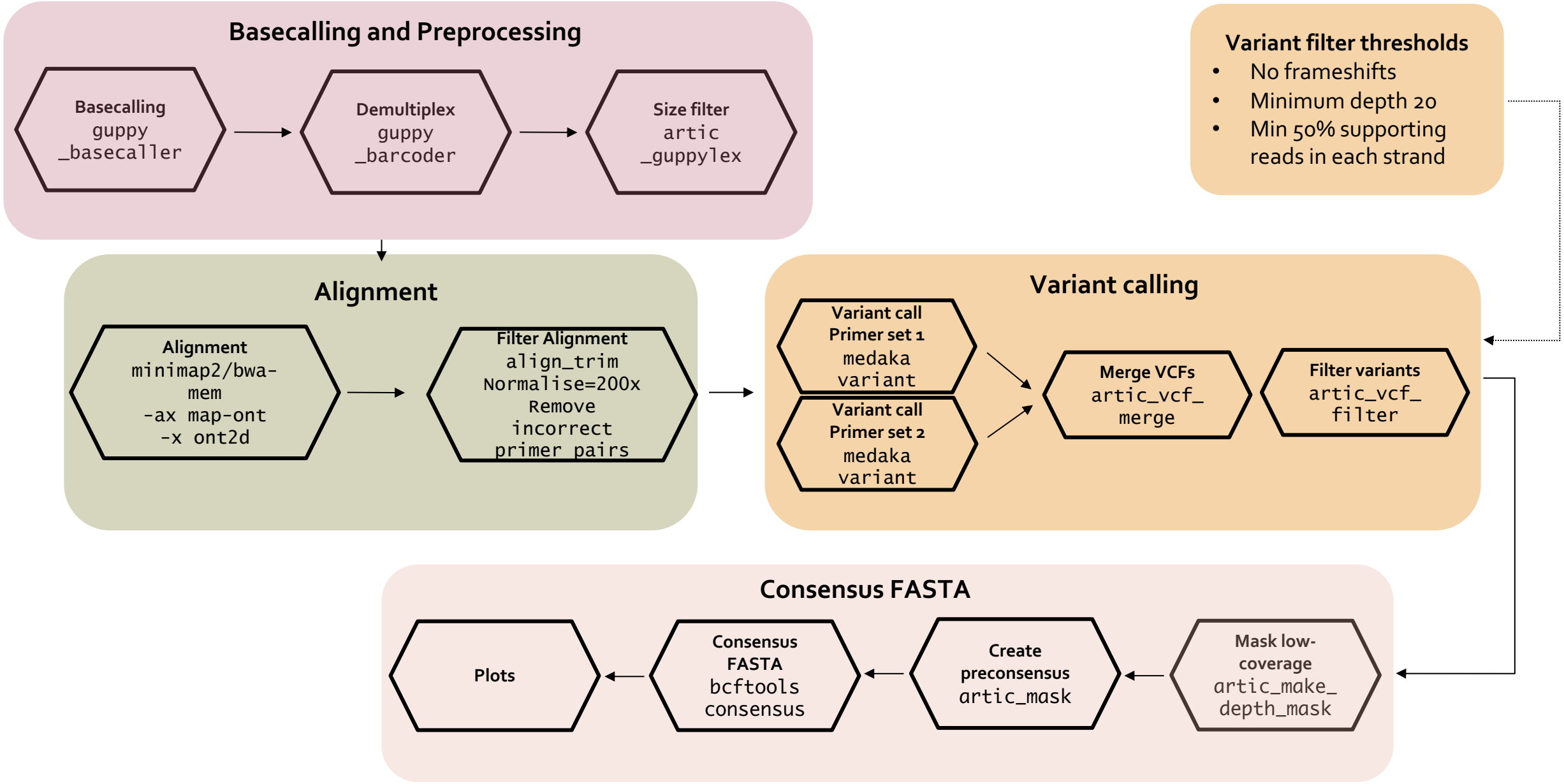



ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

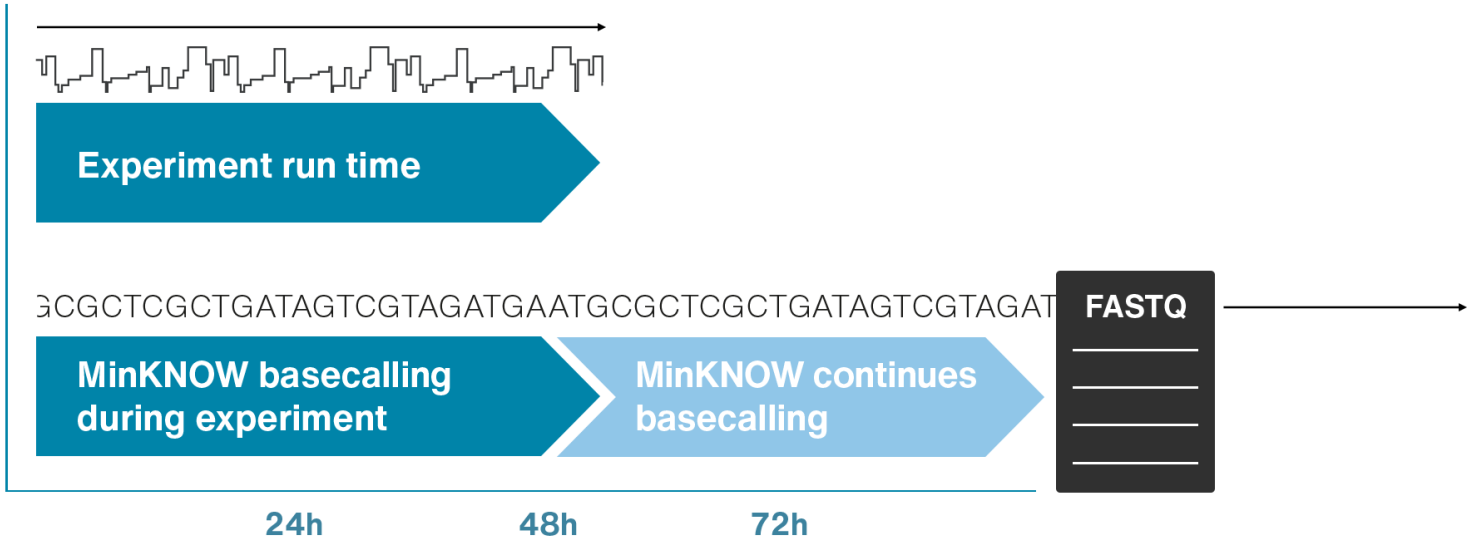
CGIAR

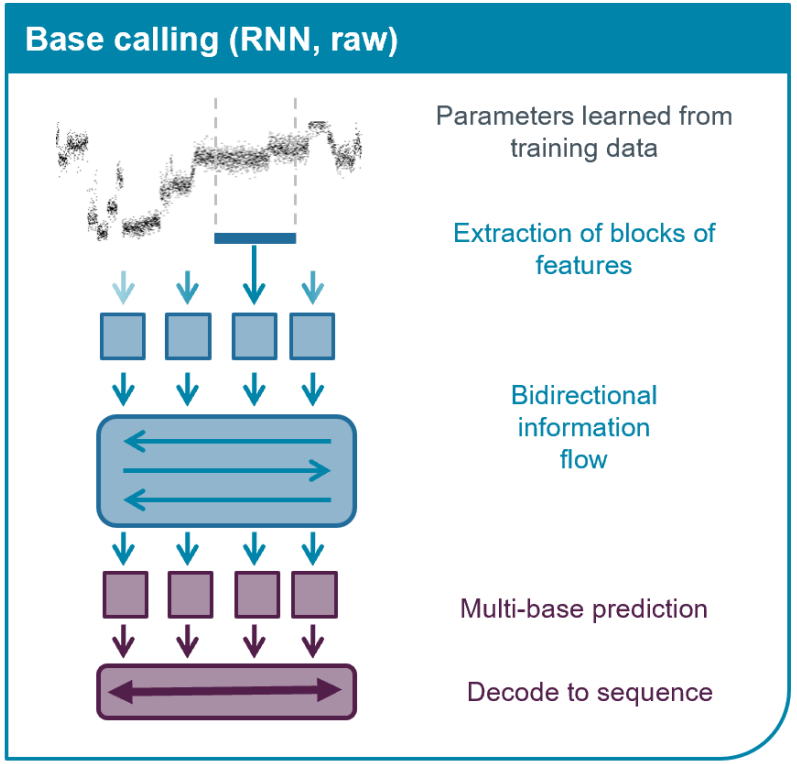
# Artic Bioinformatics pipeline





# Basecalling




- guppy_basecaller: --config dna_r9.4.1_450bps_fast.cfg --barcode_kits "EXP-NBD104 EXP-NBD114" --trim_barcodes).
- guppy_barcoder separate fastq files into barcode folders (demultiplexing)
- guppy_aligner (under development).

Source: https://nanoporetech.com/how-it-works/basecalling

ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE CGIAR

# Basecalling

Basecalling can be performed "**live**" or in **real time** while sequencing.

However,  it is often useful to separate the sequencing from basecalling.

One advantage of "**offline**" basecalling is that the basecaller can use significant amounts of compute and read/write resources which may **slow the sequencing process** and, in some cases, even lead to **loss of sequencing data**.

**Guppy** uses significant amounts of compute resources/time if run on a processor (CPU), especially if using the *High-Accuracy* (hac) models. Graphical processor units (GPUs) allow for parallelization of the basecalling process.

Predecessor basecallers: Albacore, Metrichor

Source: https://nanoporetech.com/how-it-works/basecalling

# Quality control

Quality control on next-generation sequencing (NGS)

I. On the starting nucleic acid samples (spectrophotometric, fluorometric, Gel electrophoresis, RIN)
II. After library preparation (Bioanalyzer)
III. Post-sequencing (FastQC, MultiQC)

The **Ultimate best QC** will likely come from the **sequence data**.

# Quality control

Quality control on next-generation sequencing (NGS)

I. On the starting nucleic acid samples (spectrophotometric, fluorometric, Gel electrophoresis, RIN)
II. After library preparation (Bioanalyzer)
III. Post-sequencing (FastQC, MultiQC)

The **Ultimate best QC** will likely come from the **sequence data**.

# Quality control with FastQC

## FastQC Report

### Summary

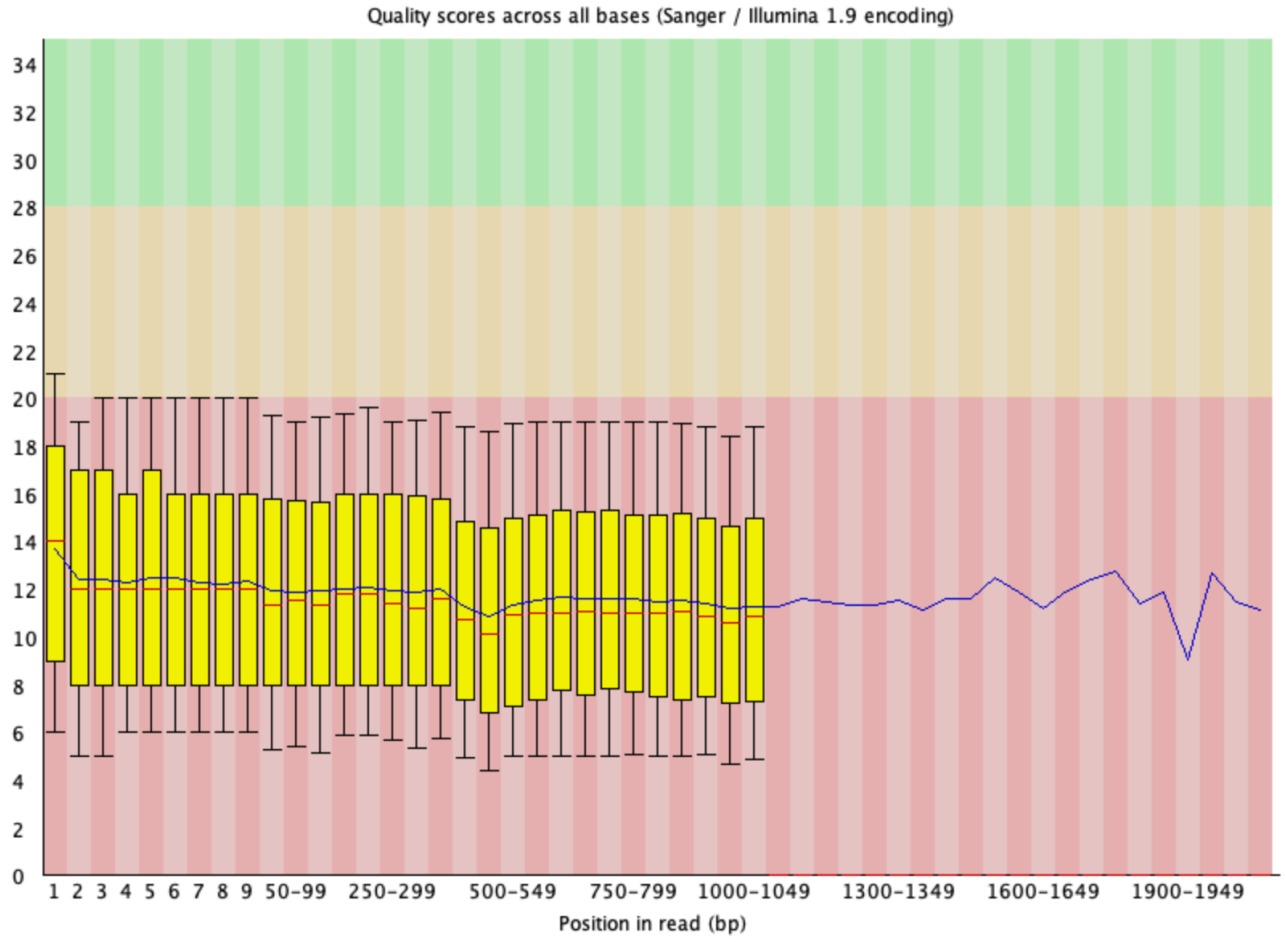
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

## Basic Statistics

| Measure | Value |
|---|---|
| Filename | ERR3790220.fastq |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 51554 |
| Sequences flagged as poor quality | 0 |
| Sequence length | 100-2078 |
| %GC | 46 |

# Quality control with FastQC

**Per base sequence quality**



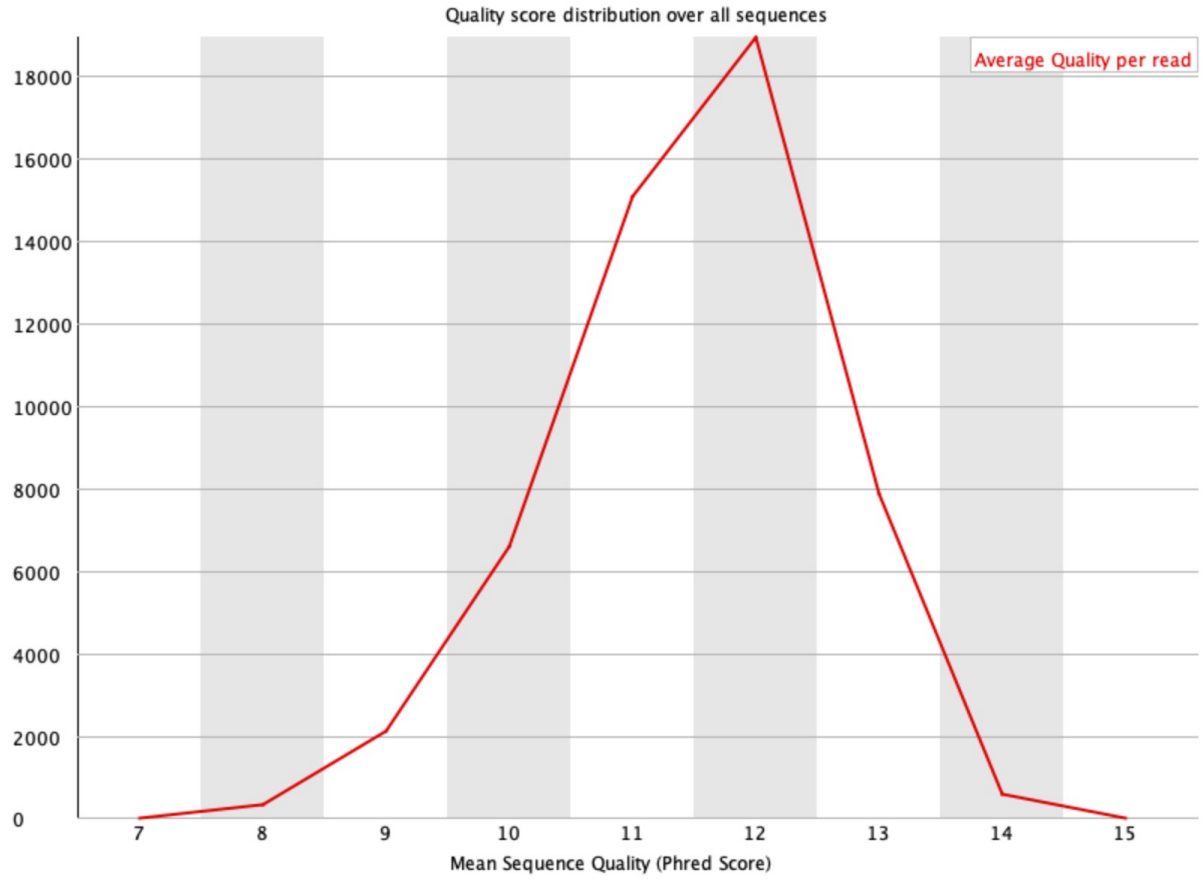

Per Base sequence quality module

- Shows mean and standard deviation of sequencing quality for each position in all reads of the data set.
- Shows low quality regions that may have to be removed, low quality reads filtered.

ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE CGIAR

# Quality control with FastQC

**Per sequence quality scores**

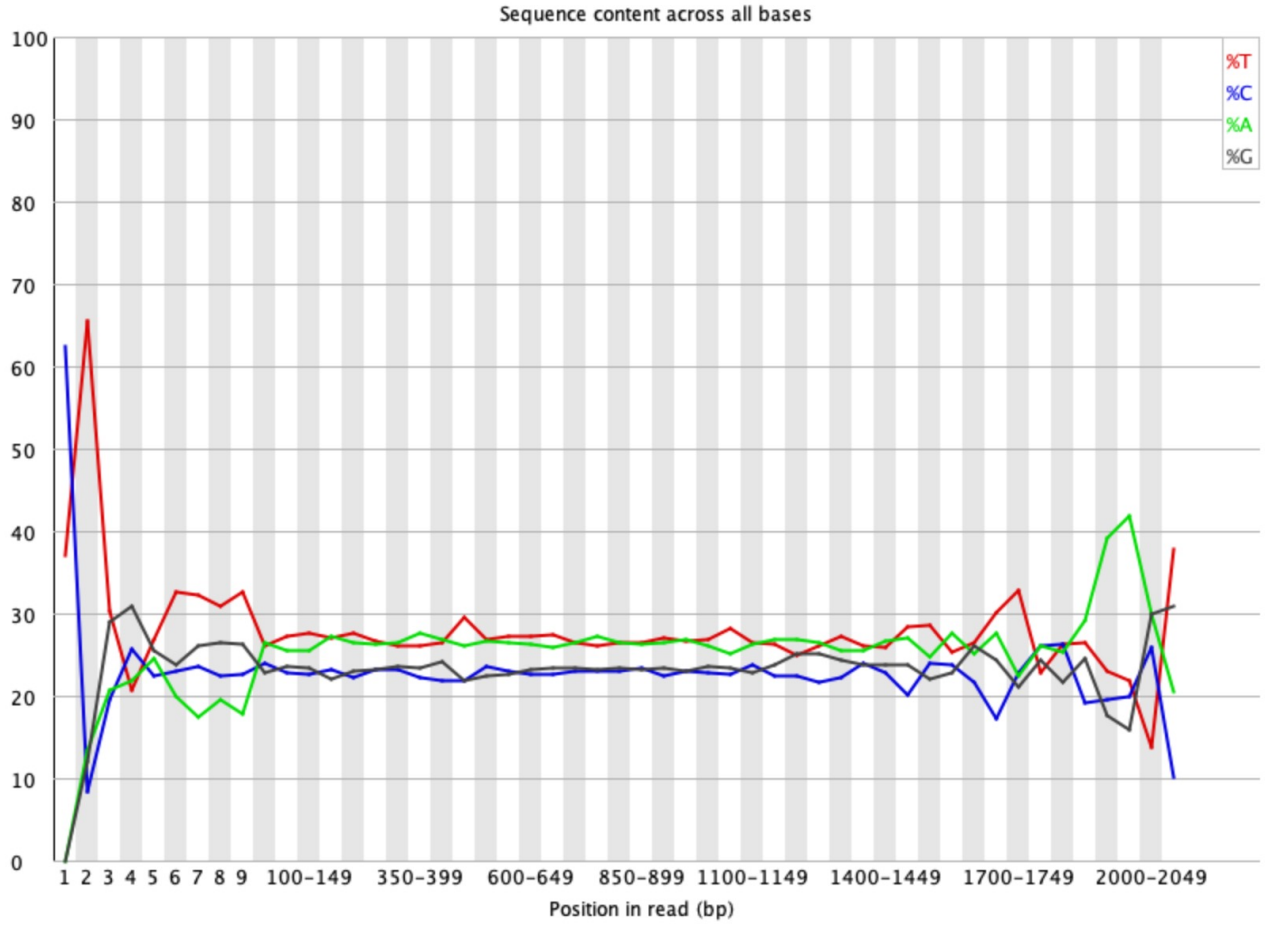



Per sequence quality scores module

- Shows the average quality score distribution of the reads.

# Quality control with FastQC

**Per base sequence content**



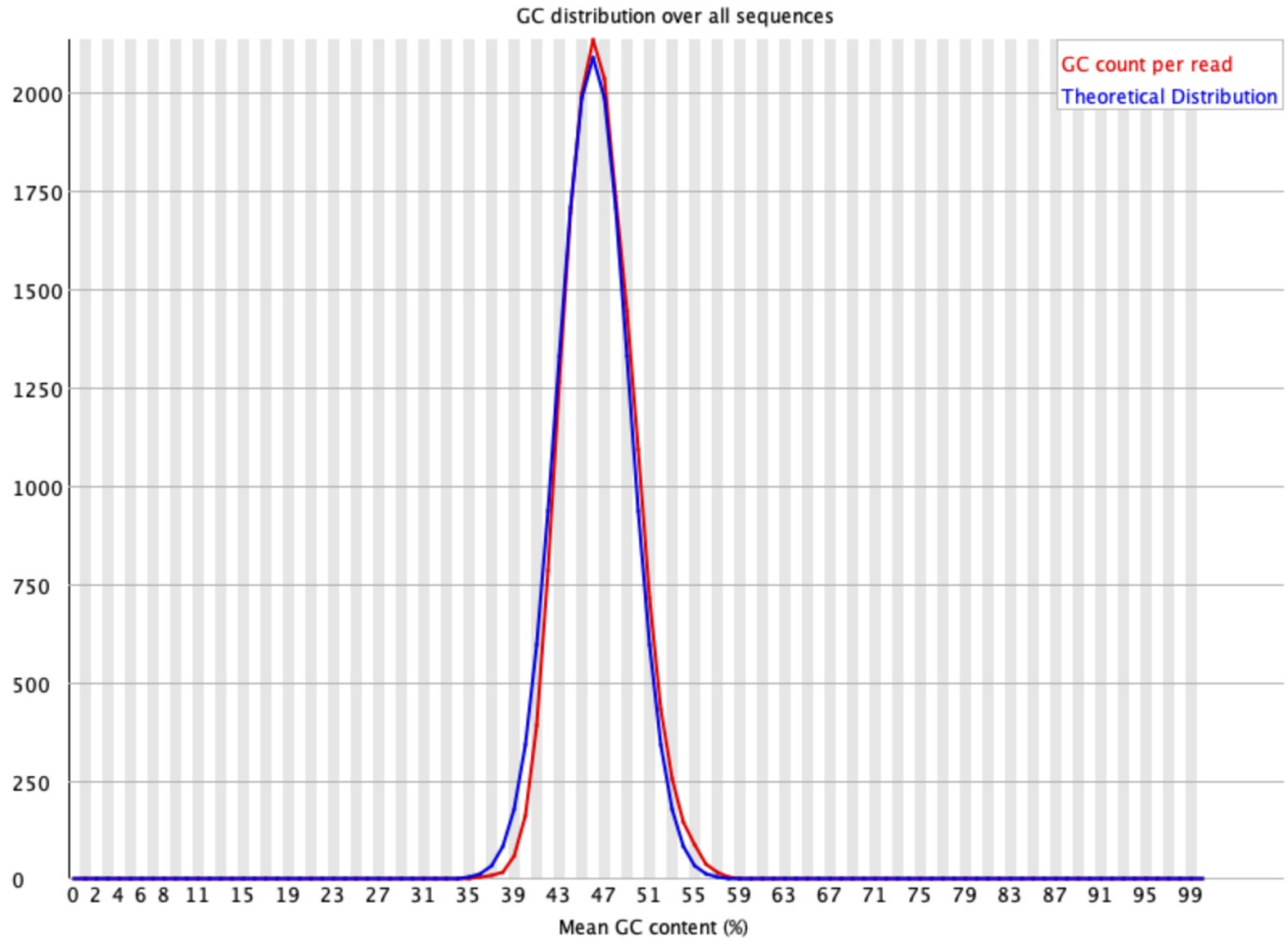

Per base sequence content module

- Show the average ratio of As, Ts, Cs and Gs in your data set. "Clean" data without sequencing errors should show almost parallel/flat lines for all four nucleotides.
- Trailing or leading peaks indicate sequencing problems and may have to be trimmed.

ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE CGIAR

# Quality control with FastQC

## Per sequence GC content




ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE CGIAR

# Aggregating quality reports with MultiQC

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

file:///Users/jjuma/trainings/africacdc-ilri-aslm-2023/output-dir/denv-1/multiqc/multiqc_report.html

# Quality control with PycoQC

- PycoQC is a data visualisation and quality control tool for nanopore data.
- It requires a **specific file**, the sequencing_summary.txt file generated by Oxford nanopore basecallers such as Guppy or the older albacore basecaller.
- Provides read statistics, sequencing and flow cell information (sequencing run, yield over time, number of active pores)

## Summary

All Reads

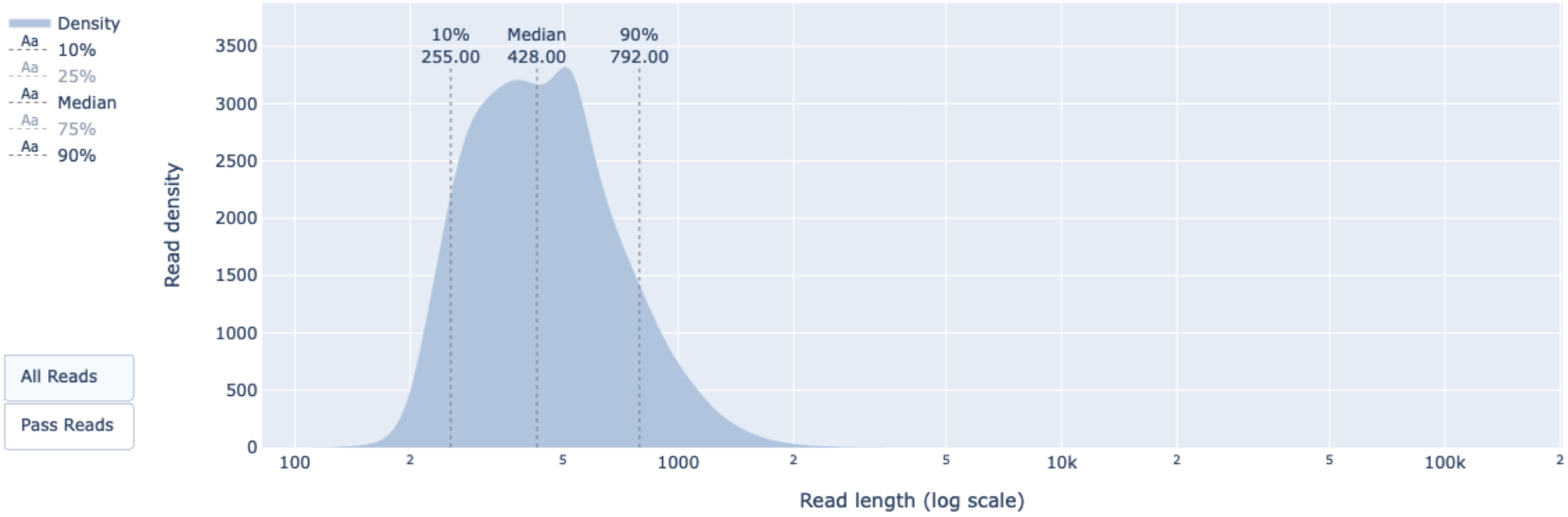
Pass Reads

| Run_ID | Reads | Bases | Med Read Length | N50 Length | Med Read Quality | Active Channels | Run Duration (h) | Unique Barcodes |
|---|---|---|---|---|---|---|---|---|
| All Run_IDs | 1,905,822 | 931,643,122 | 428.00 | 528.00 | 10.12 | 450 | 33.29 | 53 |
| 97e5b0bb021af3e8855b0058e5b9a1b6ccb1 | 1,905,822 | 931,643,122 | 428.00 | 528.00 | 10.12 | 450 | 33.29 | 53 |

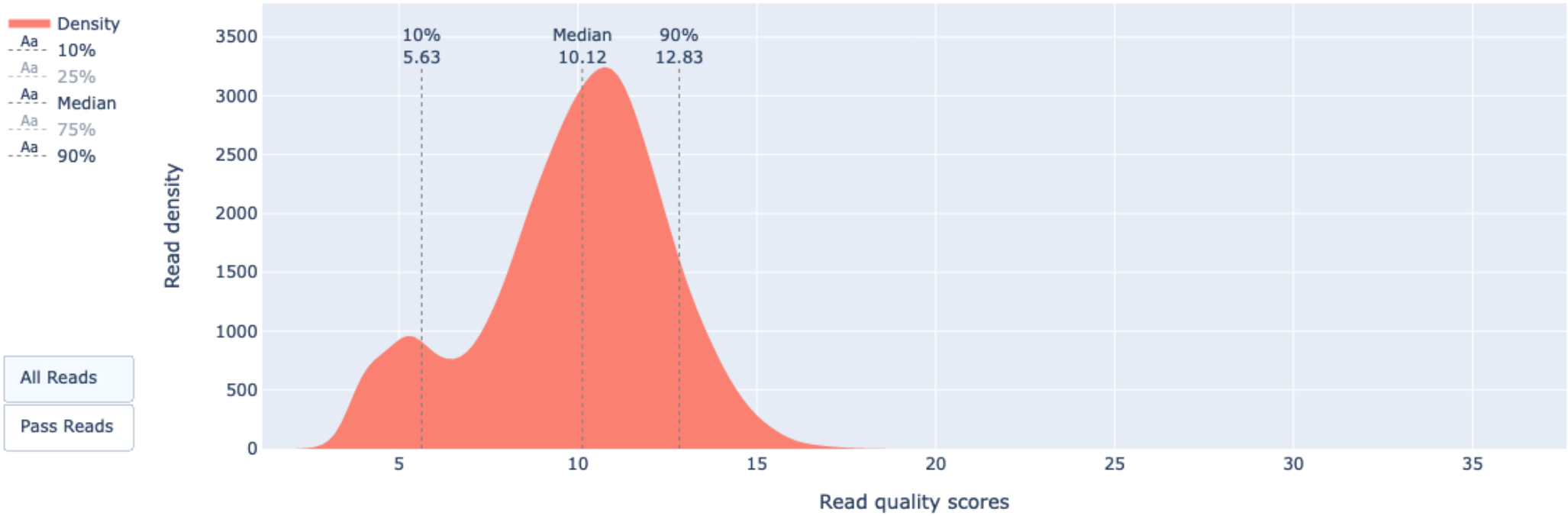
ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

CGIAR

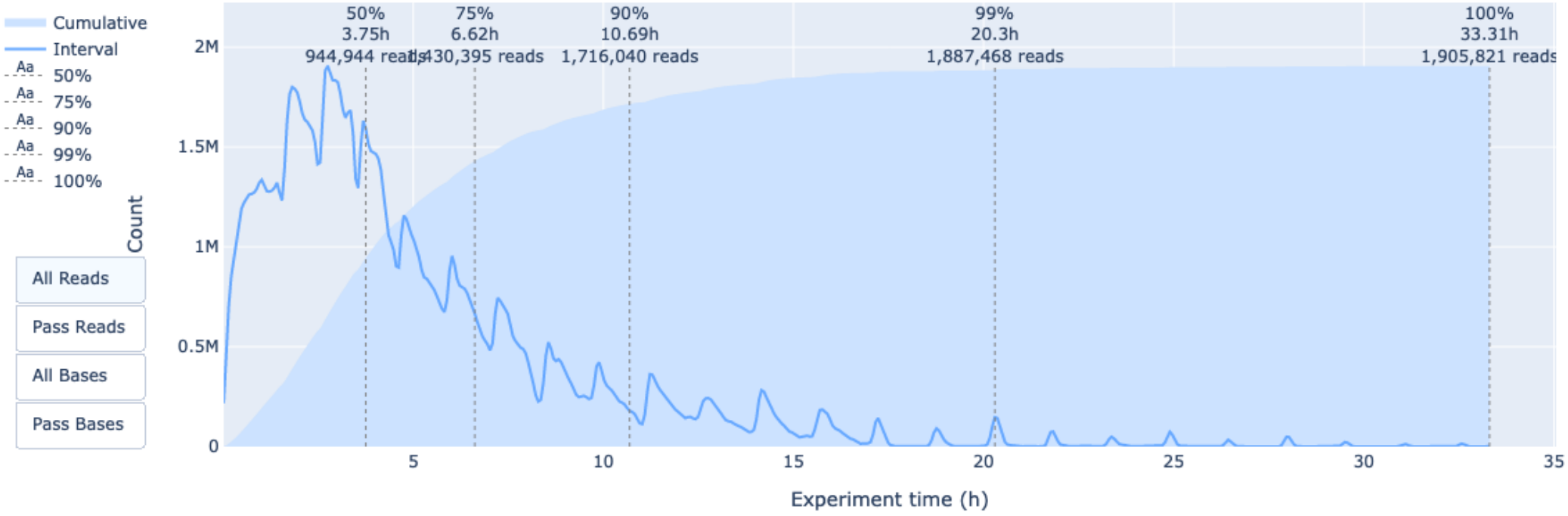
# PycoQC – Distribution of read length



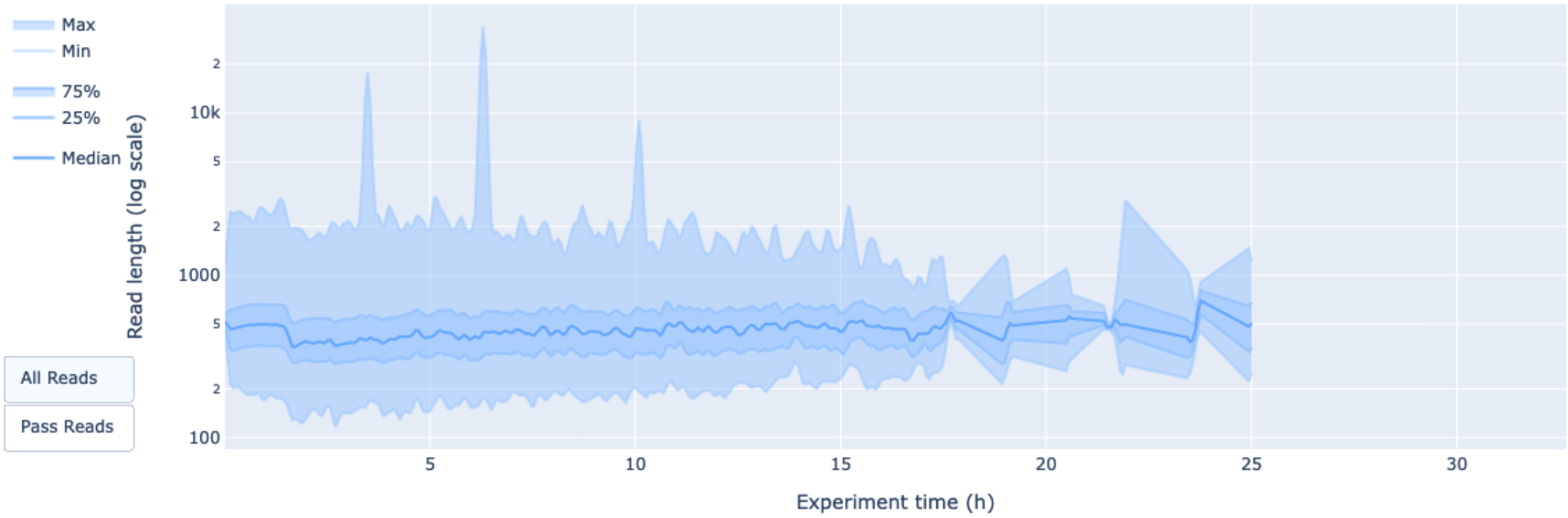
# PycoQC – Distribution of read quality



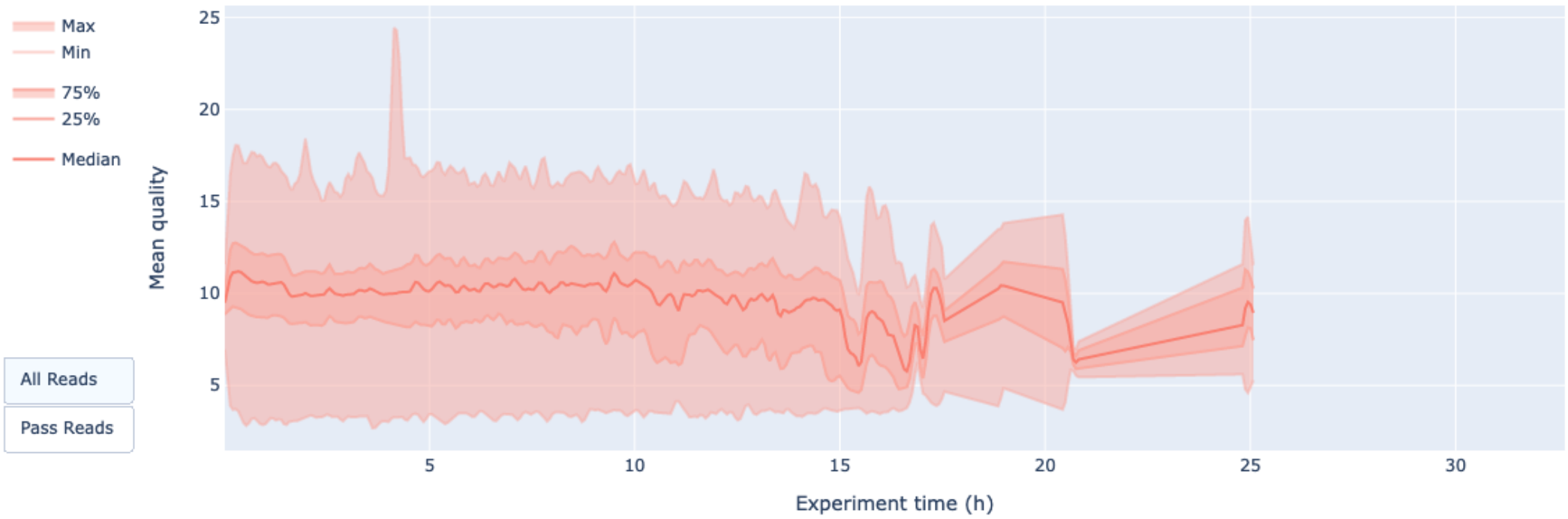
# PycoQC – Output over experiment time



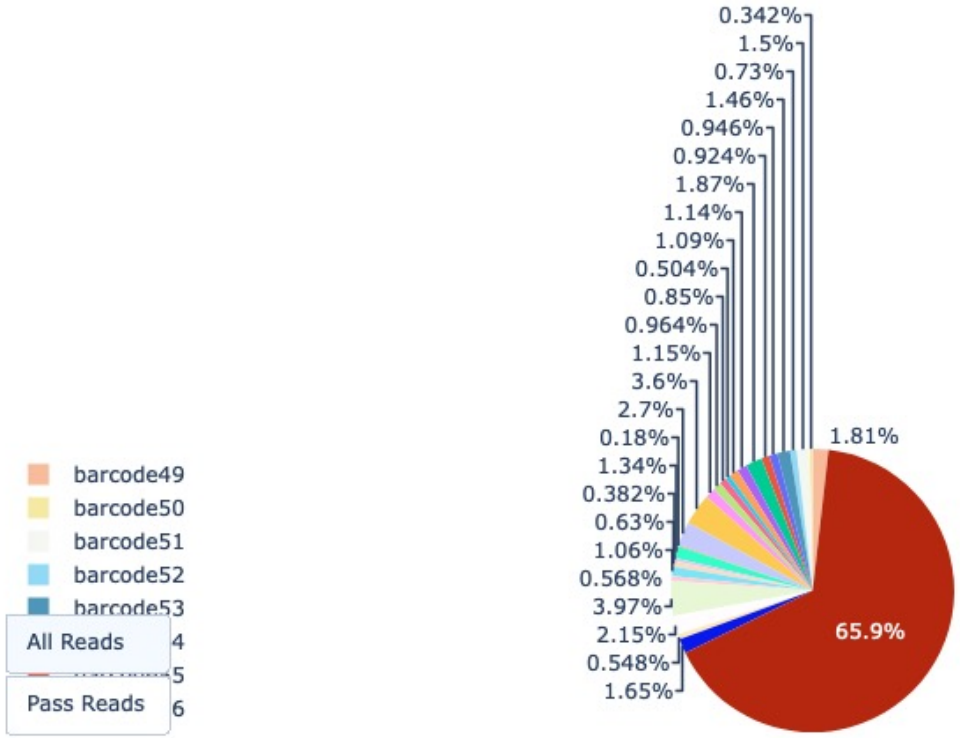
# PycoQC – read length over time



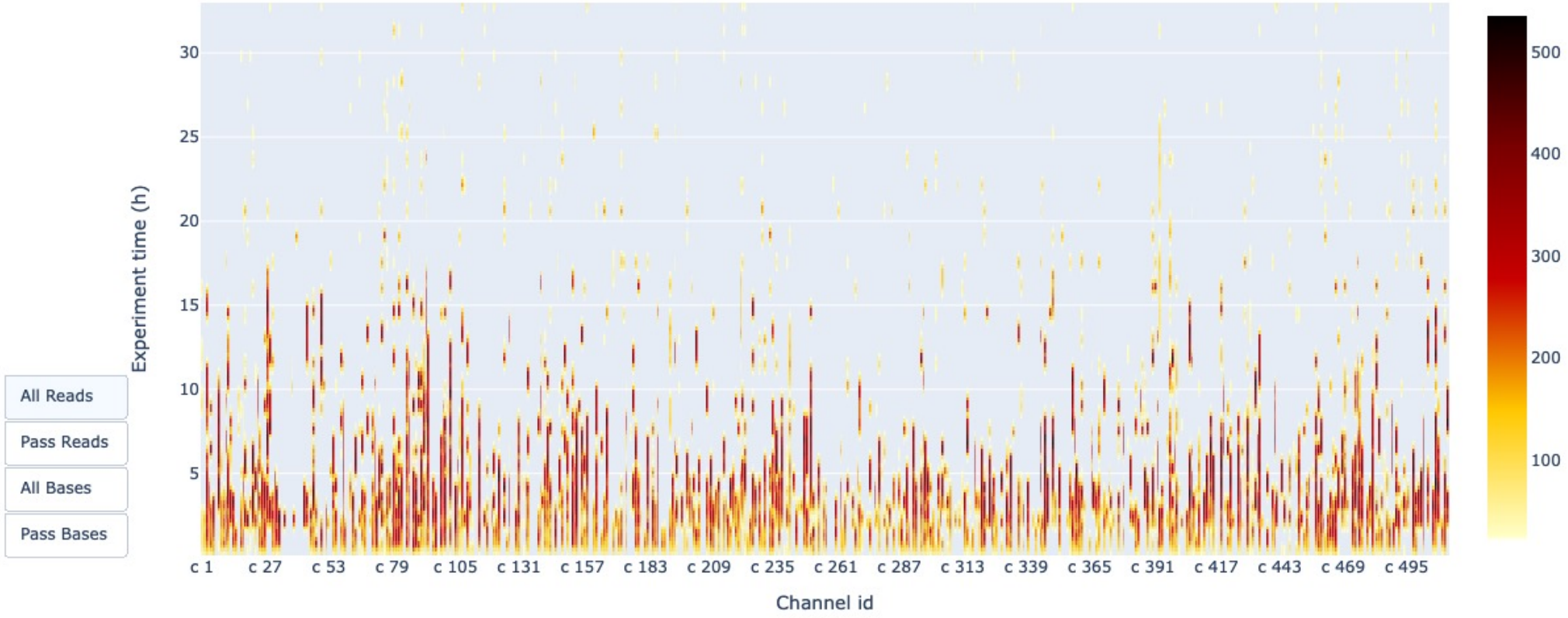
# PycoQC – mean read quality over time



# PycoQC – number of reads per barcode



# PycoQC – channel activity over time




ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

CGIAR

# Quality control using MinION_QC

- Ability to compare multiple sequencing runs, e.g., to compare different library preparation or DNA extraction protocols used.
- Takes in multiple sequencing summary files.

# Read filtering: trimming and adapter removal

- 3rd generation sequencing platforms contain linker, barcode or adapter sequences at the read beginning or end.
- Additionally, 2D library preparation protocols for ONT contain adapters in the middle.
- Porechop – trim adapters if 2D libraries use the option `-discard_middle`
- adapter removal is not necessary as they have no impact on current assemblers. (https://f1000research.com/articles/8-2138)
- NanoFilt – Read trimming and filtering.

```
NanoFilt -l 500 --headcrop 10 < porechopped.fastq > nanofilt_trimmed.fastq
```

# Alignment

A reference alignment of the basecalled reads against the reference sequence is the performed by **minimap2** or **bwa**. Both aligners use their respective ONT presets. The alignments are filtered to keep only mapped reads, and then sorted and indexed.

Nearly all alignment methods rely on pre-processing the reference genome into a data-structure (like a suffix tree) that provides an ***index*** which **makes it fast to find the place in a genome where a query sequence matches.**

Such indexes can take up a large amount of computer memory. **The Burrows-Wheeler Transform** (BWT) provides a way of decreasing the size of such indexes. The BWT is an operation that takes a sequence of characters (in this case DNA bases) and re-orders them so that similar characters tend to appear together in long runs of the same character.

# Burrows-Wheeler Transform

```
                ttccggat$a    $attccggat
                tccggat$at    at$attccgg
                ccggat$att    attccggat$
                cggat$attc    ccggat$att
attccggat$      ggat$attcc    cggat$attc    tg$tcgcata
                gat$attccg    gat$attccg
                at$attccgg    ggat$attcc
                t$attccgga    t$attccgga
                $attccggat    tccggat$at
                attccggat$    ttccggat$a
```

M. Burrows and D. J. Wheeler, "A Block-Sorting Lossless Data Compression Algorithm," Technical Report, 124, Digital Equipment Corporation, 1994.

# Sequence Alignment Map

SAM – human readable format of the alignment
BAM – Binary format of the alignment, machine-readable format

| Read ID | FLAG | Reference | Position | MAPQ | CIGAR |
|---|---|---|---|---|---|
| ERR3790222.10477 | 16 | NC_001474.2 | 11 | 60 | 6M1D13M2D14M1I24M1D46M2I14M1D30M2D11M1D22M2I14M1D28M2I24M1D19M2D24M2I9M1D12M |
| ERR3790222.18544 | 16 | NC_001474.2 | 11 | 60 | 85M1I9M1D12M2I14M1D4M3I3M1D15M1D47M1D8M1D4M2I11M5I68M1D61M1D26M1D54M1S * |
| ERR3790222.19035 | 16 | NC_001474.2 | 11 | 60 | 1S26M1D11M2D40M3D18M3D95M2I25M1D8M1I11M1D4M1I4M1D18M4D10M1D13M1D31M1D16M74S |

# Sequence Alignment Map

SAM – human readable format of the alignment
BAM – Binary format of the alignment, machine-readable format

| Read ID | FLAG | Reference | Position | MAPQ | CIGAR |
|---|---|---|---|---|---|
| ERR3790222.10477 | 16 | NC_001474.2 | 11 | 60 | 6M1D13M2D14M1I24M1D46M2I14M1D30M2D11M1D22M2I14M1D28M2I24M1D19M2D24M2I9M1D12M |
| ERR3790222.18544 | 16 | NC_001474.2 | 11 | 60 | 85M1I9M1D12M2I14M1D4M3I3M1D15M1D47M1D8M1D4M2I11M5I68M1D61M1D26M1D54M1S * |
| ERR3790222.19035 | 16 | NC_001474.2 | 11 | 60 | 1S26M1D11M2D40M3D18M3D95M2I25M1D8M1I11M1D4M1I4M1D18M4D10M1D13M1D31M1D16M74S |

**Note:**
Different alignment tools will output differently sorted SAM/BAM, and different downstream tools require differently sorted alignment files as input.

# Post-alignment processing

The purpose of alignment post-processing is:

- assign each read alignment to a derived amplicon
- using the derived amplicon, assign each read a **read group** based on the primer pool
- softmask read alignments within their derived amplicon

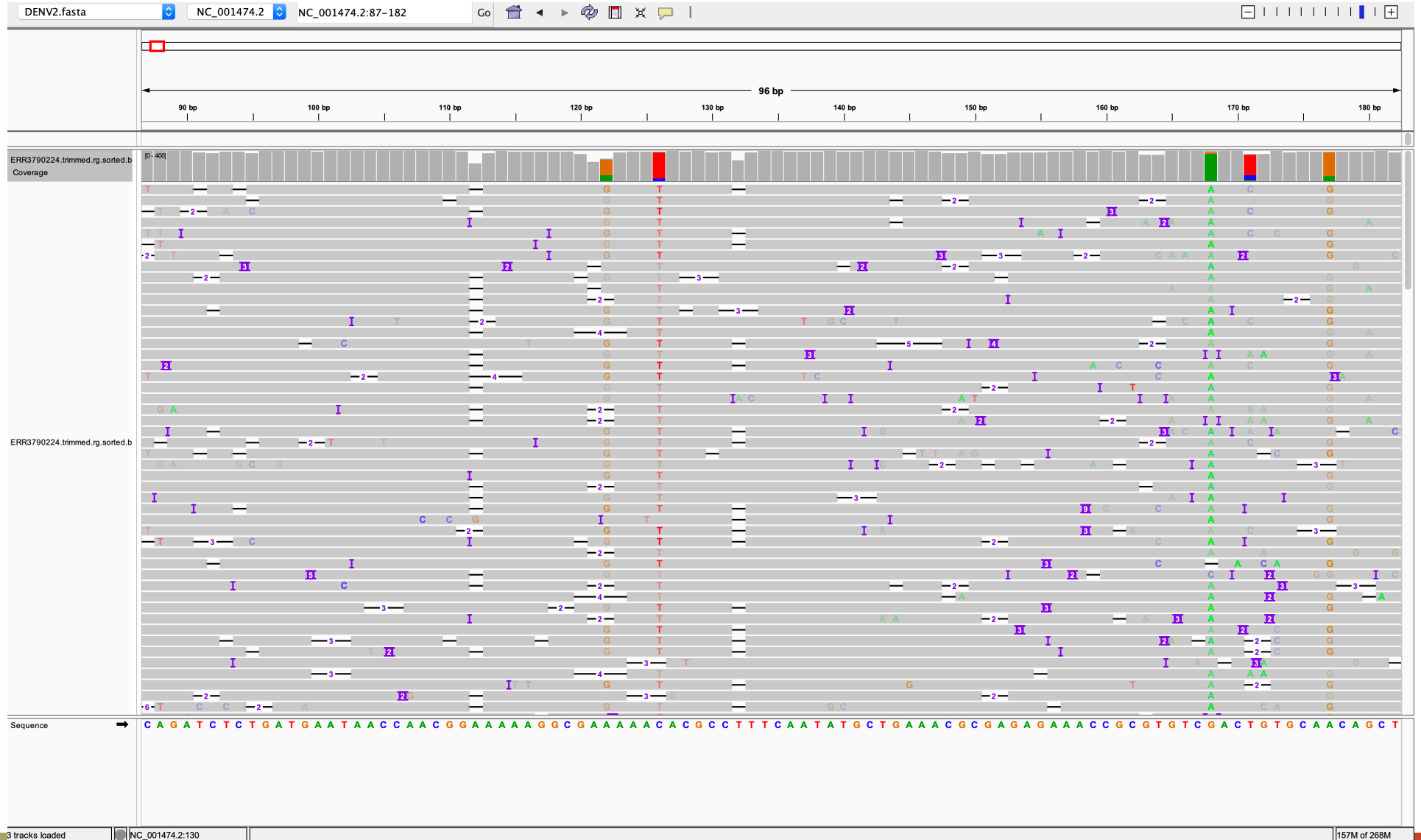
Also, there is the option to:

- remove primer sequence by further softmasking the read alignments
- normalise/reduce the number of read alignments to each amplicon
- remove reads with imperfect primer pairing, e.g., from amplicon read through

# Post-alignment processing steps

- For **each read** we **find the amplicon from which it originated** by **selecting the amplicon with the largest overlap with the read**, we also find the **next closest match**.
- We **discard the read** if the **next closest match** is a **large proportion of the mutual overlap of the two amplicons**. This is a guard against **chimeric reads**, either from library preparation or faults in the sequencing platform control software.
- There is an option to **only allow those reads that extend across the whole amplicon**, if set to true then we check whether the alignment **extends to the primer at each end**, this is a "correctly paired" read.
- To normalise we take the passing reads, **sort by the amount of coverage** they provide and take the 1st *n* reads.

# Visualizing alignment




ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

CGIAR

# Variant calling

Variant calling entails identifying **single nucleotide polymorphisms (SNPs)** and **small insertions and deletion (indels)** from next generation sequencing data.

The artic bioinformatics pipeline can use two variant callers:

1. Nanopolish
2. Medaka

If calling variants with Nanopolish, the raw data file having squiggles is REQUIRED.

# Variant calling

For each **read group** (based on primer pools), variants are called then merged using `artic_vcf_merge` module.

The `artic_vcf_filter` module categorizes the variants as either **PASS** or **FAIL**. Variants that have passed are used in downstream steps.

# Consensus building

- Post-processed alignment is used to check each position of the reference sequence for sample coverage. Any position that is not covered by at least **20 reads** from either read group are marked as low coverage. This is computed using `artic_make_depth_mask` module. This step produces coverage information for each read group and produces a coverage mask to tell us which coordinates in the reference sequence failed the coverage threshold.

- A consensus sequence is built, first as a pre-consensus sequence based on the input reference sequence. The preconsensus has low quality sites masked out with **N**'s using the coverage mask and the sample `fail.vcf` file. `bcftools consensus` is used to combine the preconsensus with the `pass.vcf` variants to produce a consensus sequence for the sample.

# Depth

```
A T G C T A G C C T A G T A C G A T G G G T A C C A G T T T A C C G A T A C G A T A C G G
A T G C T A G C C T A G T A C G A T G G
                        T G G G T A C C A G T T T A
                  T A C G A T G G G T A C C A G T T T A C C G A T A C G
A T G C T A G C C T A G                                   T A C G A T A C G G
                            G G G T A C C A G T T T A C C G A T A C G A T A C G
                                  T A C C A G T T T A C C G A T A C G A T A C G
  T G C T A G C C T A G
```

Reference genome

2X 3X 4X 1X

# Coverage

```
ATGCTAGCCTAGTACGATGGGTACCAGTTTACCGATACGATACGG      Reference genome

ATGCTAGCCTAG
ATGCTAGCCTAGTACGAT

ATGCTAGCCTAGT                  CGATACGATACGG
                             ACCGATACGATACGG      39/45*100 = 86.66% ~87%

ATGCTAGCCTAG                         TACGATACGG

                       TACCAG      ACCGATACGATA
                             ACCGATACGATACGG
ATGCTAGC                             TACGATACGG

ATGCTAGCCTAGTACGATNNNTACCAGNNNACCGATACGATACGG      Consensus genome
```

# Acknowledgments

African Union

AFRICA CDC
Centres for Disease Control and Prevention
Safeguarding Africa's Health

# References and further reading

ILRI
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE
CGIAR