Canadian Food Inspection Agency | Agence canadienne d'inspection des aliments

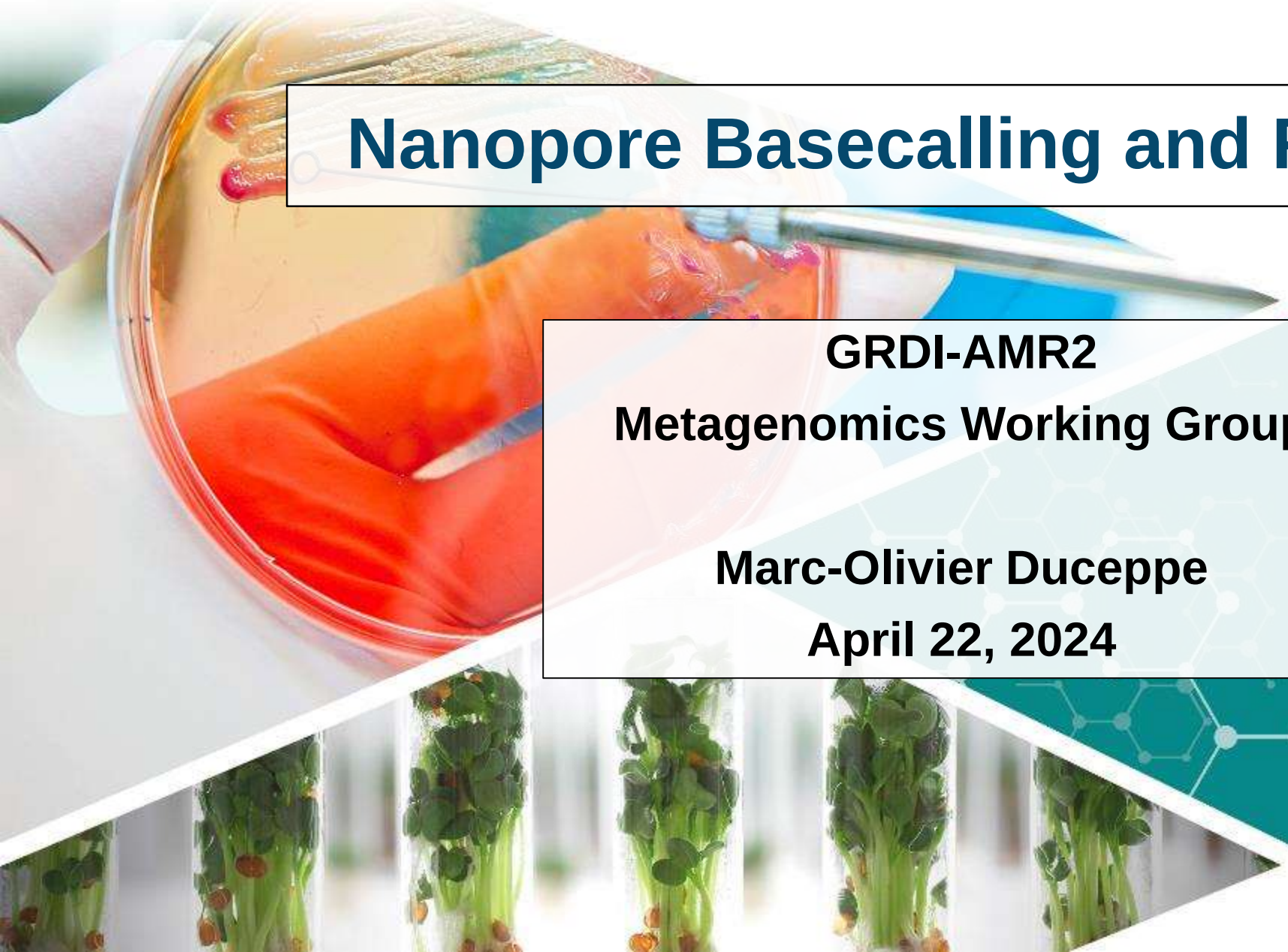
Nanopore Basecalling and Read QC

GRDI-AMR2

Metagenomics Working Group

Marc-Olivier Duceppe

April 22, 2024



Canada

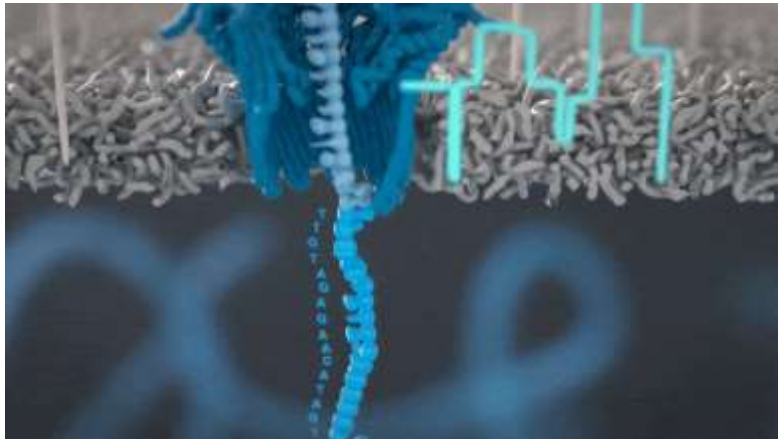
Presentation Plan

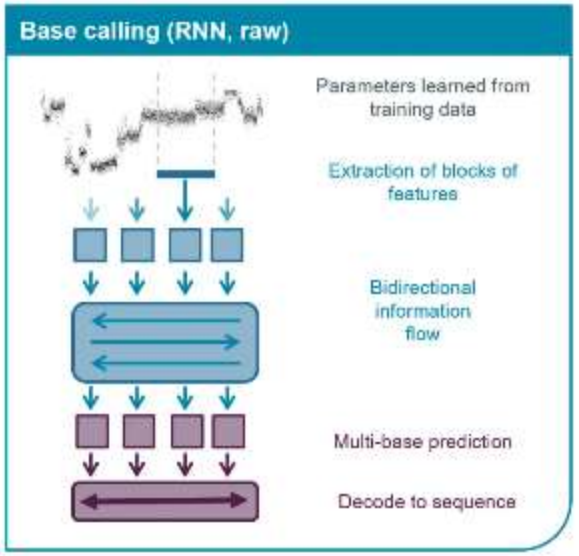
1. Why Nanopore?

2. How to basecall the reads in “super accuracy” mode?

3. How to assess run / read quality?

The Nanopore Technology





- Basecalling is AI based (Deep Neural Network)
- Performed from 5-mers

https://nanoporetech.com/platform/technology/basecalling

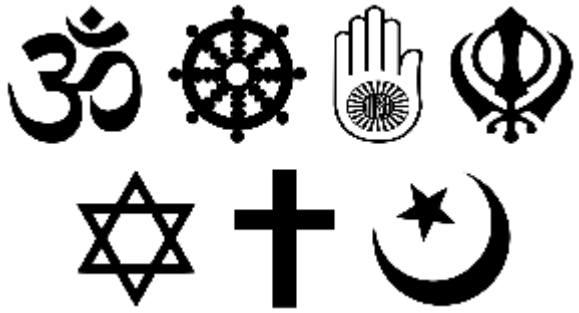
Two Families of Sequencers

- MinION Mk1B
- Flongle
- GridION
- **Cheaper**

- PromethION
- P2 Solo
- P2i
- **Higher yield and Q-score**

https://store.nanoporetech.com/

Do You Believe in Nanopore?



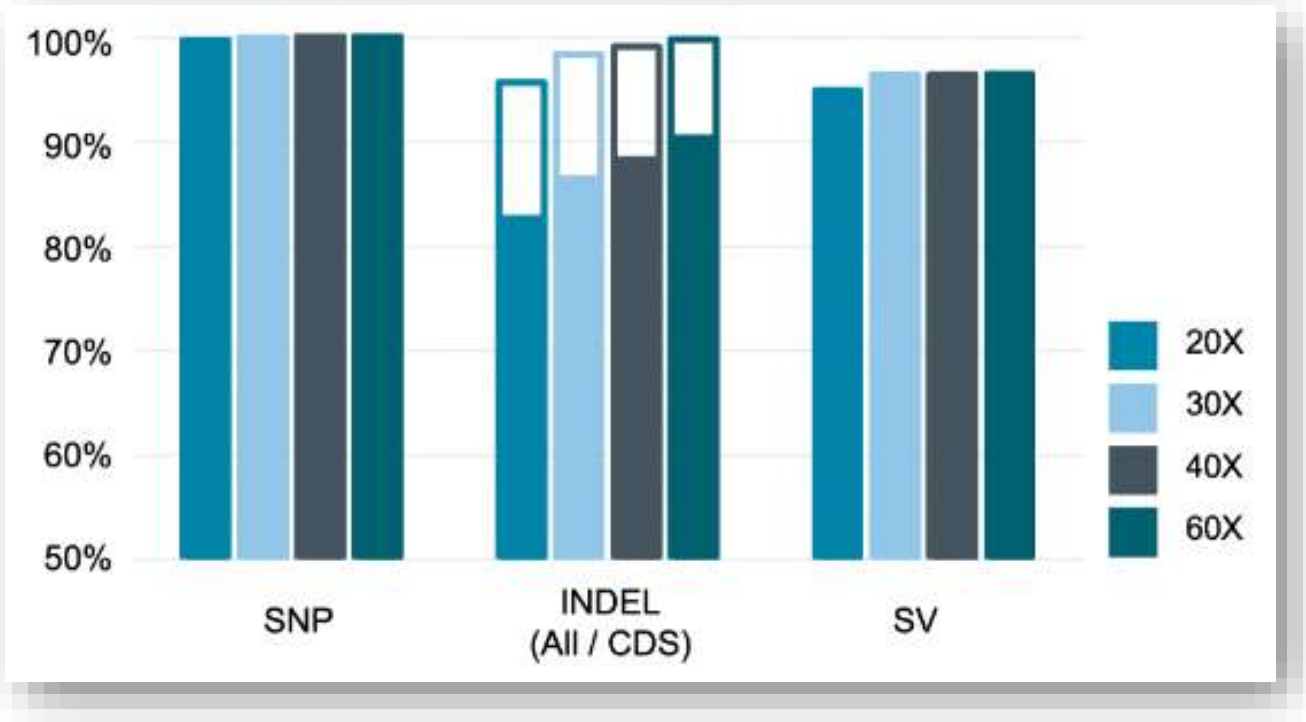
Pros

- $1k / 87g
- **20bp - 100+kb**
- Native DNA / RNA
- Real-time sequencing
- Reusable flowcells
- Adaptive sequencing
- Growing community

Cons

- Keeps changing (QA)
- **Error-prone**
- Hard to analyze
  - Computationally intensive (GPU)
  - Somewhat specialized tools

Accuracy



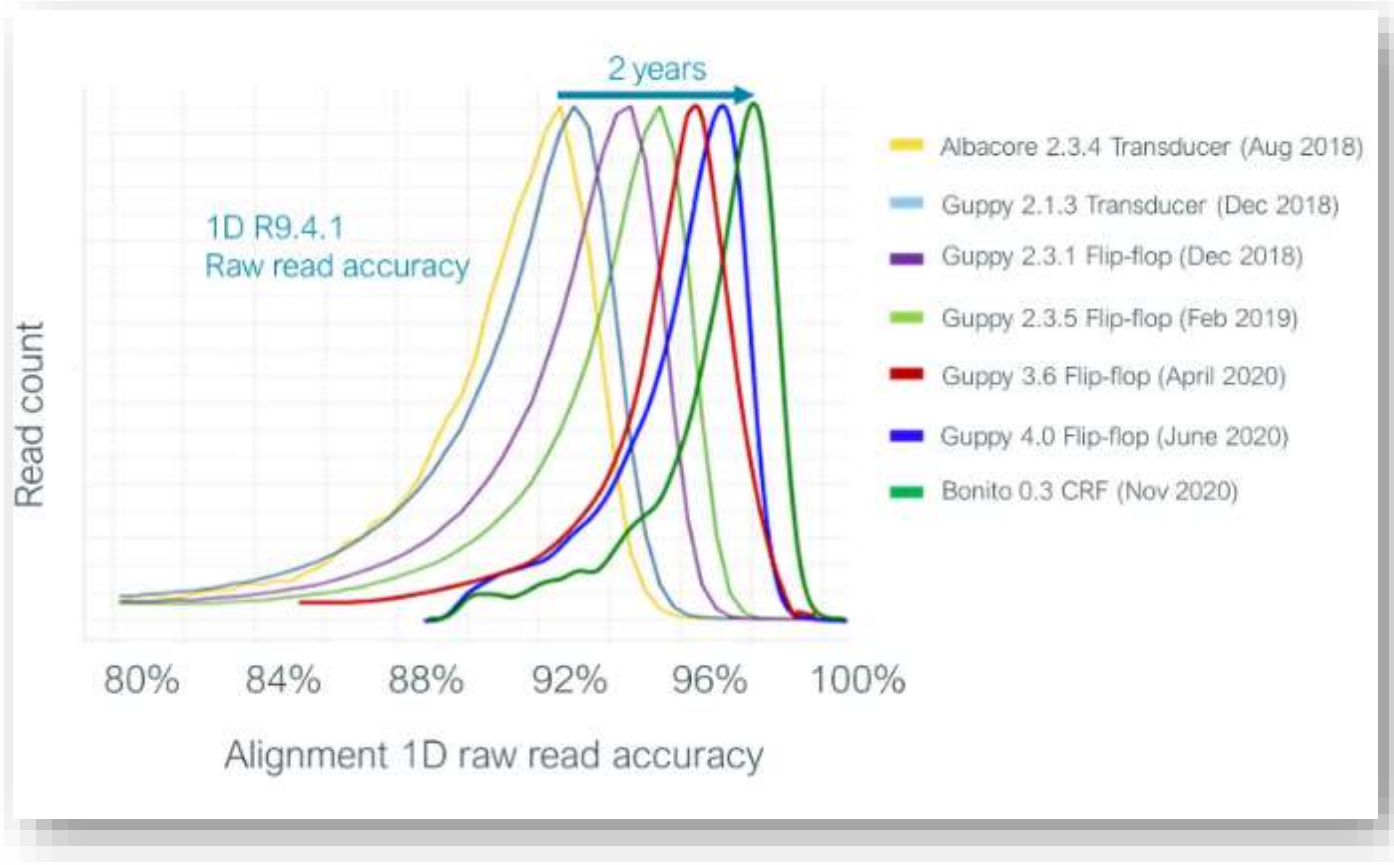
- Can’t quite resolve with accuracy long homopolymers stretches
- Leads to lots of INDELs
- Often long with more errors is better than short with less errors
- AMR SNP-based:
- AMR functional gene:

Evolution of Basecallers

- Albacore: 2017 – 2018
- Guppy: 2018 – 2023
- Dorado: 2023 – present

Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 2019 Jun 24;20(1):129. doi: 10.1186/s13059-019-1727-y. PMID: 31234903; PMCID: PMC6591954.

Read Accuracy Keeps Improving



- In 2024
  - R10.4.1 v14
  - Dorado v0.6
  - Super accuracy mode
- Simplex: 99.5% (Q23)
- Duplex: >99.9% (Q30)
- **More Q15 – Q20**

https://nanoporetech.com/platform/accuracy

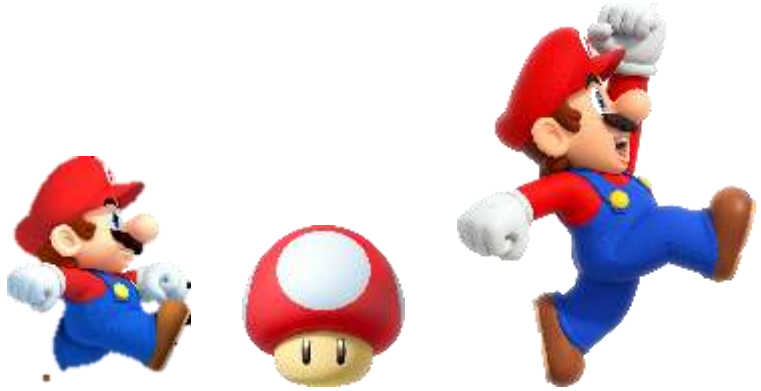
To Guppy or not to Guppy

- Why you'd still want to use Guppy
  - Don't know how to run Dorado
  - You have an easier way to run Guppy
  - Part of your current SOP / certified workflow
  - Downstream analysis requirements
    - "Old" tools with "old" data (e.g. base modification)
  - Lower end GPU (< 8GB VRAM)

- Why you don't want to use Guppy anymore
  - Legacy (no more updates)
  - Slower
  - Less accurate
  - New chemistries won't be supported

Basecalling Recommendations

- While sequencing: “Fast” mode
  - Enables more real-time run metrics
  - Do not overload host computer



- Post sequencing: “Super accuracy” mode
  - Higher Q-scores
  - Less “fail” reads
  - Better barcode assignment
  - Need computer with high-end GPU
  - Lose real-time benefit

Guppy - Legacy

- Input: Fast5 and kits (“final_summary”)
- Output: Fastq and “sequencing_summary”

- https://cdn.oxfordnanoportal.com/software/analysis/ont-guppy_6.5.7_linux64.tar.gz
- https://github.com/duceppemo/basecall_nanopore

Dorado – Two Flavours

Standalone

- Input: Pod5 | Fast5
- Output: a single uBAM | SAM | Fastq

- Needs to be demultiplexed separately
- Fastq headers only contain “ReadID“

- https://github.com/nanoporetech/dorado (v0.6.0)

Dorado – Two Flavours

Basecall Server

- Input: Pod5 | Fast5
- Output: Fastq and “sequencing_summary.txt”

- Demultiplexing built in
- Fastq headers with “all” info (like Guppy)

- https://cdn.oxfordnanoportal.com/software/analysis/ont-dorado-server_7.3.9_linux64.tar.gz

Nanopore read QC

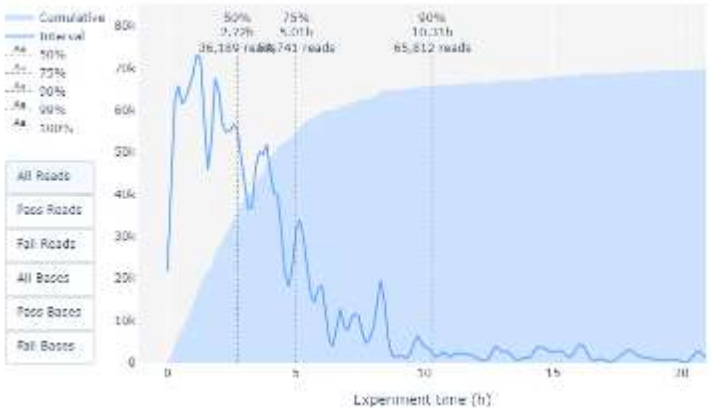
- pycoQC v2 (https://github.com/a-slide/pycoQC)
  - Designed for Guppy
  - Can work with Dorado
    - Standalone: need to run additional commands to generate a compatible "seq_summaray.txt" file
    - Server: as is

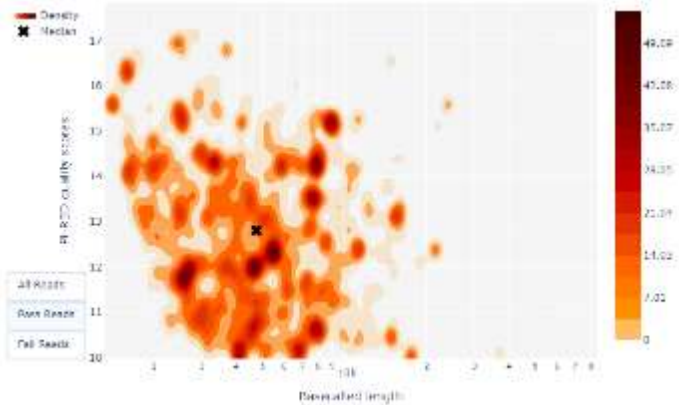
Added Features

- pycoQC v3 (https://github.com/duceppemo/pycoQC)
  - Work in progress
  - Designed to accept all dorado output file types
  - Additional scripts:
    - Pod5_to_seq_summary (no basecall information)
    - Bam_to_seq_summary (slow)
    - **Fastq_to_seq_summary (recommended)**
  - Starting from Fastq triggers additional GC plots
  - Display “All”, “Pass” or “Fail” reads

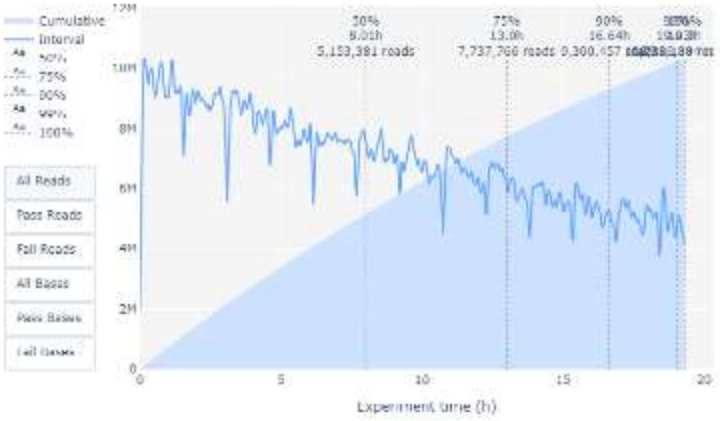
Suboptimal Loading

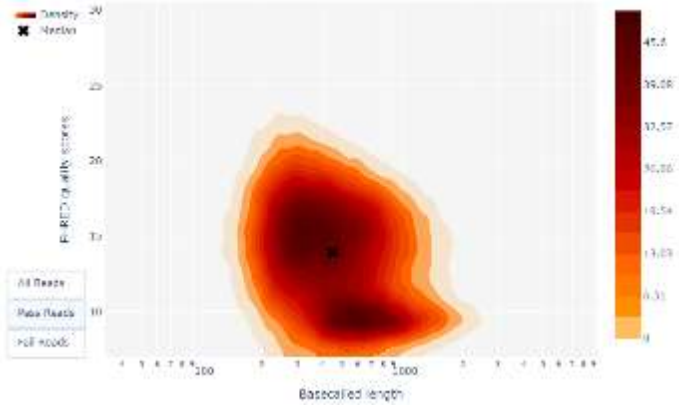
- High quality library
- Not enough loaded
  - Ran out after 10h





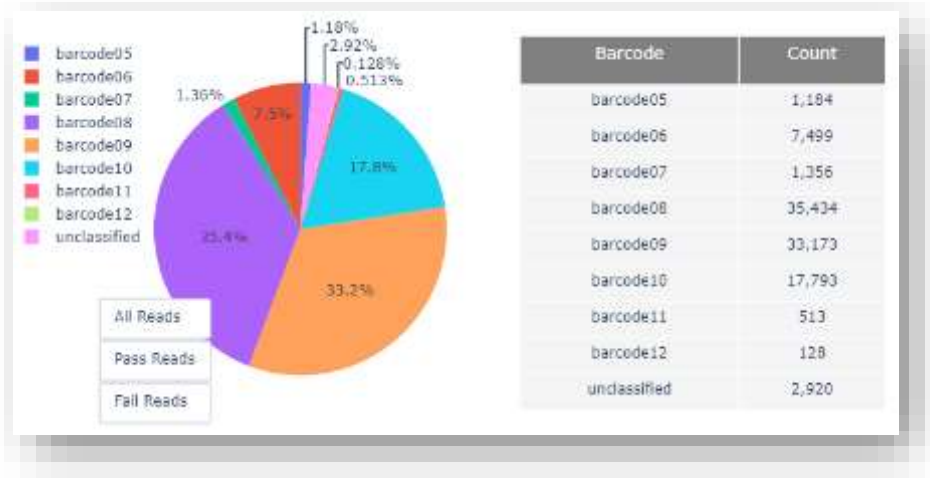
- High quality library
- Run stopped too early





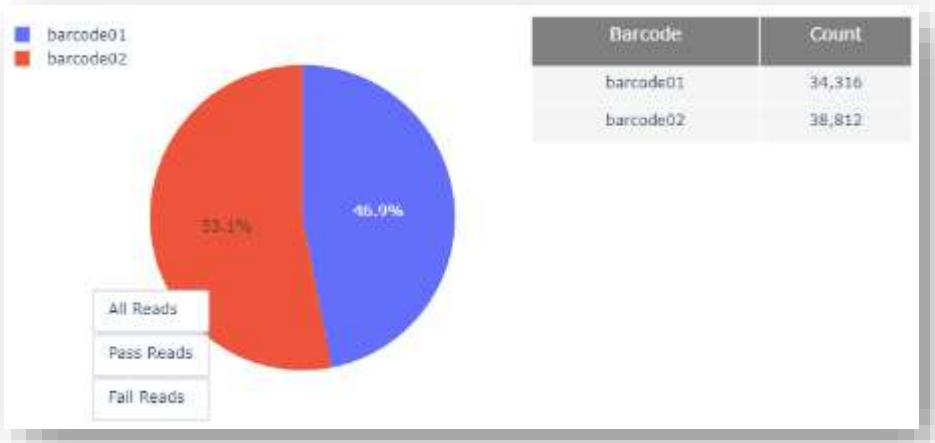
Suboptimal Sample Normalization

- Uneven barcode distribution
  - Poor normalization
  - Adapter ligation problems
    - Wrong molarity



Barcode	Count
barcode05	1,184
barcode06	7,499
barcode07	1,356
barcode08	35,434
barcode09	33,173
barcode10	17,793
barcode11	513
barcode12	128
unclassified	2,920

- Equal barcode distribution
  - MinKNOW now uses adaptive sequencing to help balance barcodes



Barcode	Count
barcode01	34,316
barcode02	38,812

Run Diagnosis

- Very useful to find what went wrong
  - DNA extraction (contaminants, short DNA fragments, etc.)
  - Library prep (adapter-to-DNA ratio, sample normalization, etc.)
  - Sequencing (flowcell quality, amount loaded, overheating, script error, etc.)

You Can Do It!

- Github repo with basic code to install and run Dorado basecall server

- https://github.com/OLF-Bioinformatics/2024-04-22_GRDI_basecalling_presentation

Nanopore Metagenomics Resources

- https://github.com/duceppemo/mashID
  - Quick species ID from metagenomics samples
  - Raw reads or assemblies
  - Pre-compiled DB available for bacteria

- https://github.com/duceppemo/seqcounter
  - Ultra-rapid preliminary AMR detection from raw reads

Take-Home Message

- Switch to Dorado!
- Why basecall you runs using Dorado SUP?
  - Access higher accuracy
  - It keeps getting better (chemistry and software)
  - Better downstream analysis results
- DNA quality (and its size) matters
  - Garbage in = garbage out
  - The longer the better
  - Amplified DNA sequences better (number of reads and q-score)
- Bioinformaticians are not magicians
  - Bad runs need to be scraped (like for diagnostics)
- The truth is in the reads
  - Read QC will help you find what went wrong and areas of improvement

Thank You!

Q & A