Basecalling – How Good is Good Enough?

David Yarmosh

Lead Bioinformatician, ATCC

Contents

01 Background

02 Design

03 Results

04 Discussion

About ATCC®

- ATCC® was founded in 1925
- 501(c)(3) not-for-profit organization
- World's largest, most diverse biorepository
- Quality Accreditation by multiple industry standards
  - ISO 9001 Certified
  - ISO 13485 Certified
  - ISO/IEC 17025 Accredited
  - ISO 17034 Accredited
- Standards development partner with multiple industry working groups
  - ANSI Standards Working Groups
  - AOAC International Working Group
  - IMMSA/NIST Microbiome Standards

- Global supplier of authenticated cell lines, microorganisms, and molecular standards
- Sales and distribution to 150+ countries
- Talented team of 600+ employees

Thousands of authenticated biomaterials

- 5000+ cell lines & primary tissue
- 2900+ viruses
- 17000+ bacteria & archaea
- 53000+ fungi
- 2000+ protists

Visit **atcc.org**

ATCC®

Oxford NANOPORE Technologies

Background

NCM Boston

Basecalling Need to Know

Nanopore basecalling is, by default, performed via the high accuracy (HAC) model

- Fast and super accurate (SUP) models are also available
- Dorado is designed to produce identical results at higher speeds compared to the Guppy basecaller
  - Also offering additional models for methylation analysis

ATCC®'s Role

High-throughput (>1000/year) ONT® sequencing

- Using standard recommended software and settings (HAC)
- Sequences wide array of microbes across all taxa
- GridION® platform
- Hybrid assembly workflow with Illumina® data

Oxford NANOPORE Technologies

Benchmarking Goals

NCM Boston

Set expectations and recommendations for sequencing labs of variable:

- Throughput demand
- Organism diversity
- Upgrade capacity

Evaluate the analysis of dorado and guppy FAST/HAC/SUP basecalling models in terms of:

- Output read quality scores
- Basecalling speed
- Read length and count
- Illumina®-hybrid and ONT®-only *de novo* assembly

Oxford NANOPORE Technologies

Study

NCM Boston

More benchmarking than done before

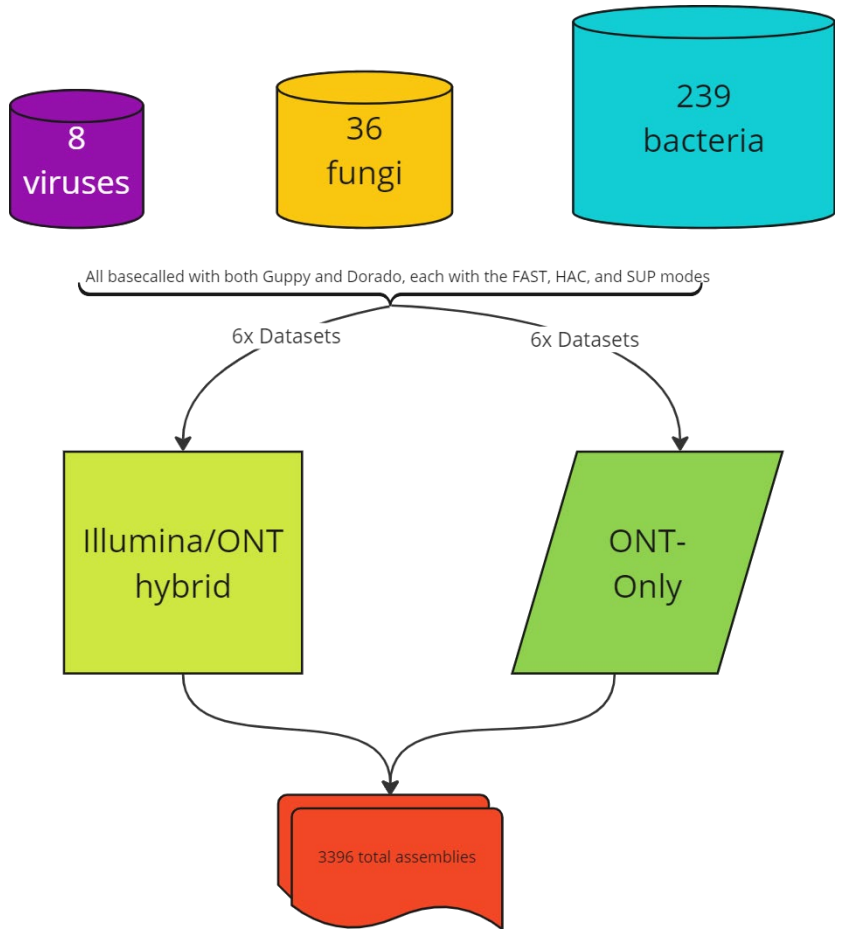
Taking a mix of samples we had available:

- Compare the performance of both guppy and dorado in terms of speed and quality at the read level
  - Across basecalling models fast, HAC, and SUP
- Compare the impact of these different basecallers on the downstream assembly quality
  - ONT®-only assembly with flye
  - ONT®-first assembly with flye, polished with Illumina® data
  - Illumina®-first assembly with unicycler, scaffolded with ONT® data
  - Measured by N50, assembly completion score, gene annotation content

- Establish sense of value vs sacrifice of each mode

Minimal experimental variability across 283 different organisms

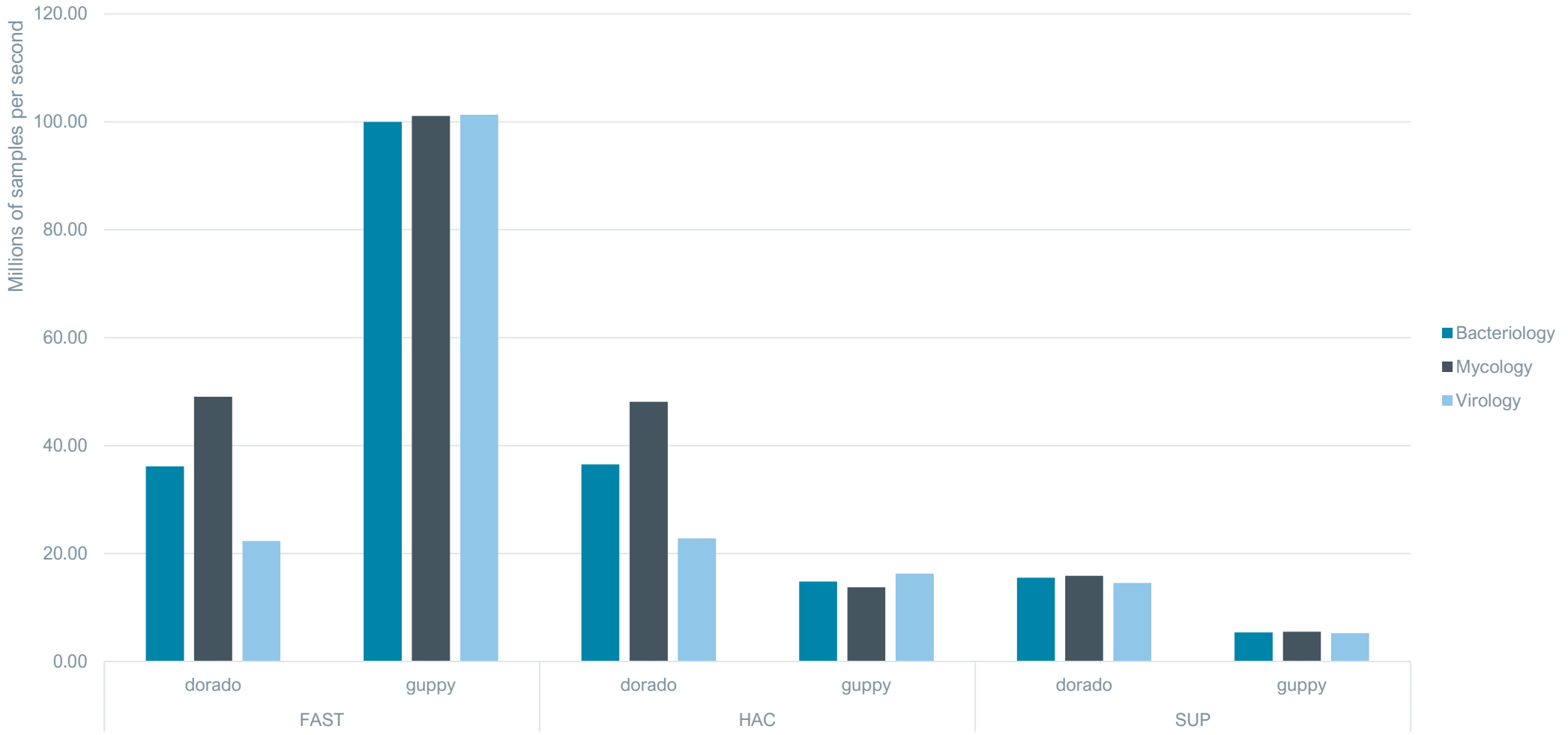
- Sequenced by the same user
- Same handling and preparation (as the organisms allow)
- Same flow cell
- Same physical instrument
- Same basecalling hardware



Oxford NANOPORE Technologies

Basecalling Speeds

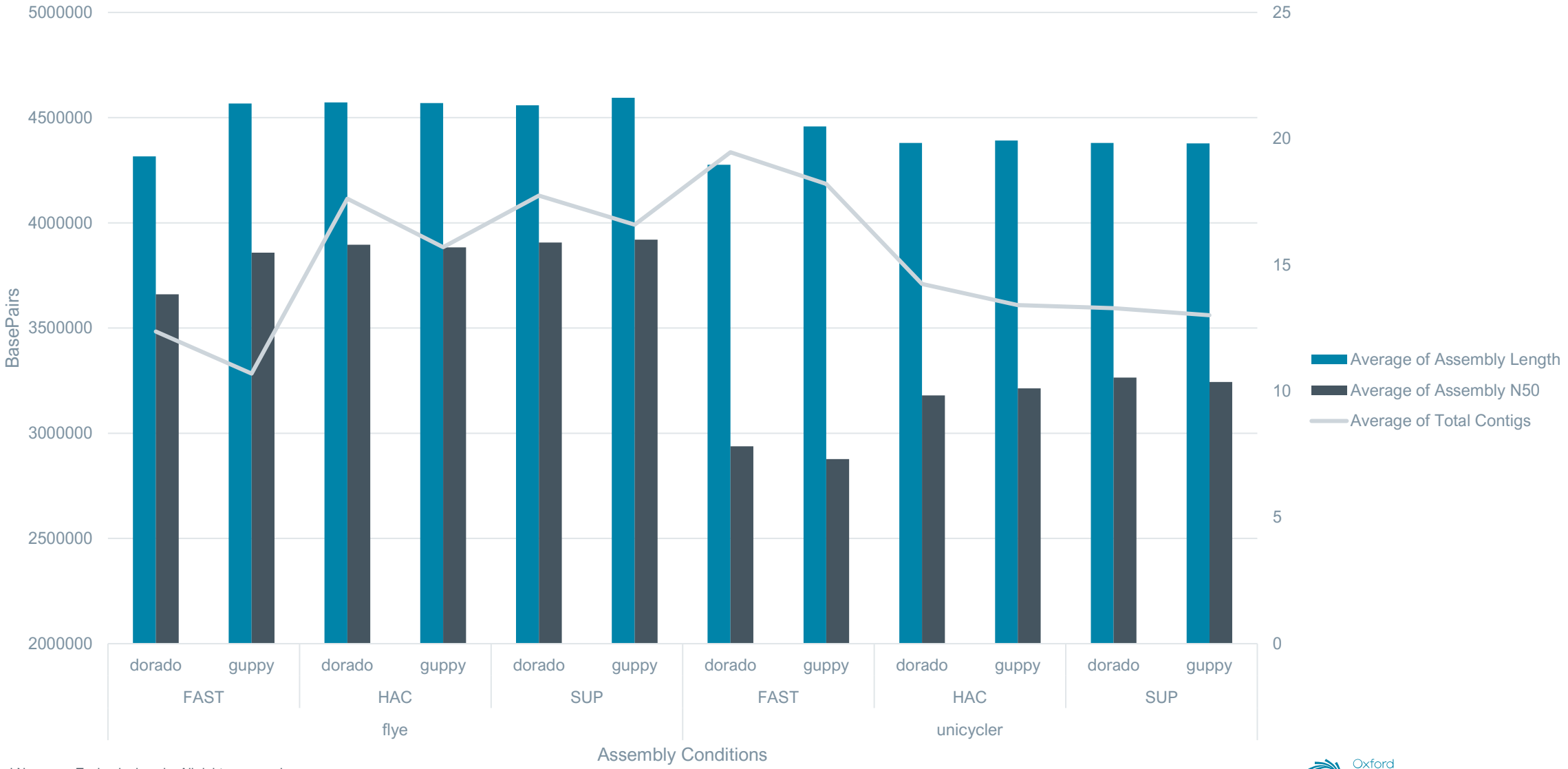
NCM Boston



Oxford NANOPORE Technologies

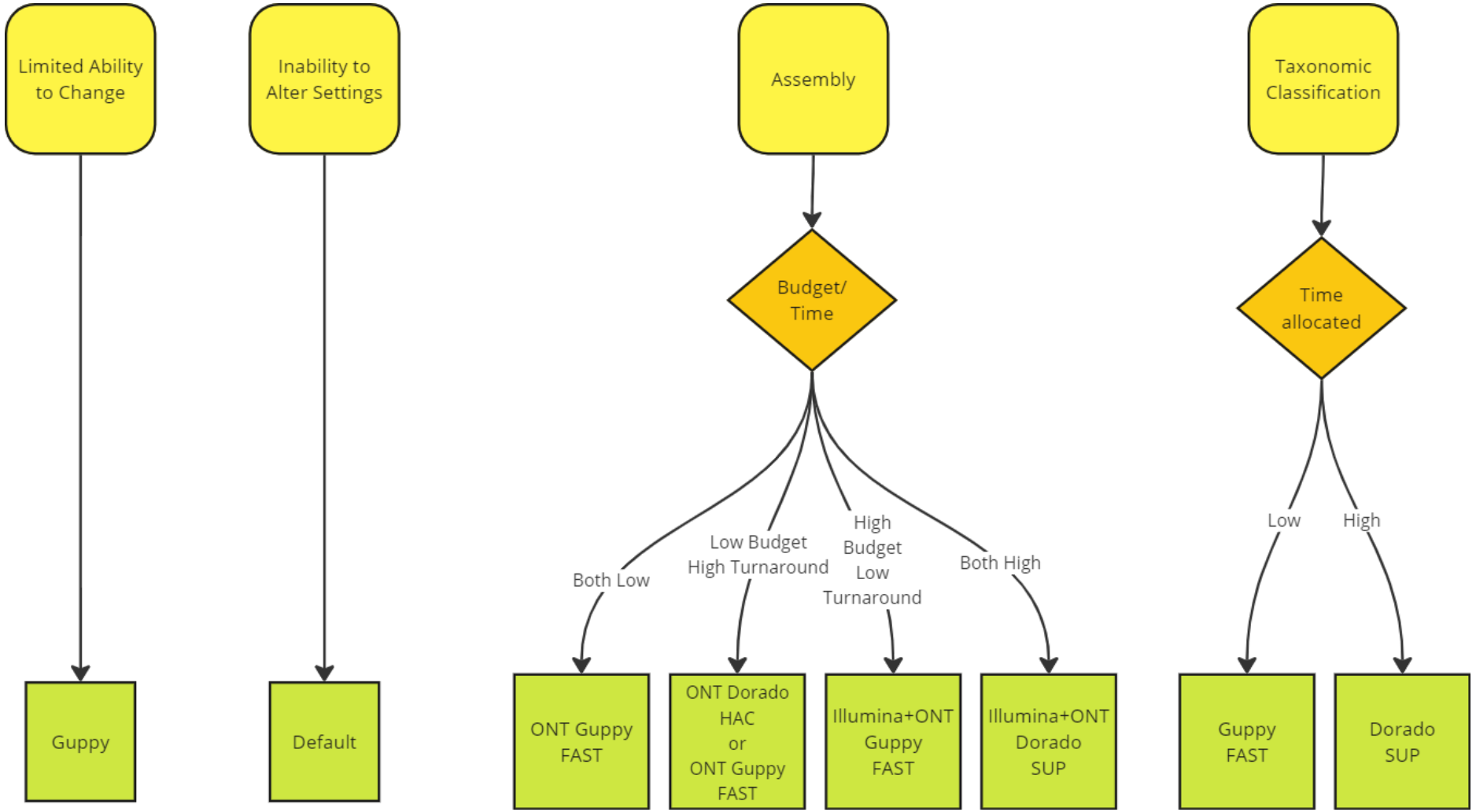
Assembly Statistics

NCM Boston



Oxford NANOPORE Technologies

Application (preliminary)



Recap

NCM Boston

ATCC's Collection offers unparalleled insight into Nanopore sequencing efficacy

ATCC®

- With the goal of optimizing our usage of the GridION® platform, ATCC® has found that for our usage, Dorado run with its SUP mode best serves our needs:
  - It is substantially faster than the Guppy equivalent
  - It allows for one GridION run per week, compared to two if in HAC or FAST modes
  - Assembly accuracy is slightly superior to HAC and greatly superior to FAST

- Future work includes:
  - Describing virology and mycology collections in more detail
  - Tuning parameters beyond default settings
  - Methylation analysis

Other Considerations

- Other use cases may find other requirements:
  - Guppy's FAST mode remains substantially faster (nearly 3x faster than Dorado) at the cost of higher bacterial contamination
    - Dorado is otherwise nearly 3X faster than Guppy

- Hybrid assembly is relatively unaffected by basecaller or basecalling speed

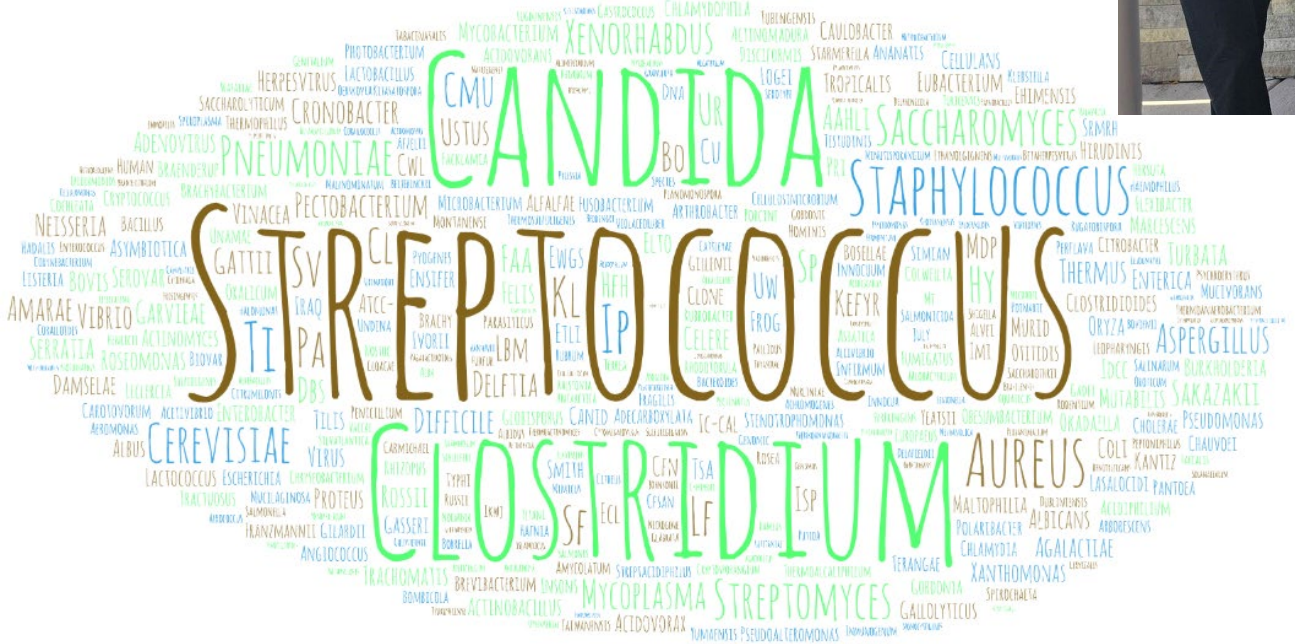
- ONT®-only *de novo* assembly is still outperformed by hybrid assembly methods
  - Using simple default parameters
    - Tuning parameters for trimming and assembly to be performed in future work

- Dorado offers superior methylation analysis capability

Oxford NANOPORE Technologies

Acknowledgements

- ATCC's Sequencing and Bioinformatics Center
- Corina Tabron



Thank you