JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

# Nanopore Data Analysis Update

Ashleigh Thomas

# Topics

- **YACRD:**
  - Yet Another Chimeric Read Detector
- **MetaBAT 3.0**
  - Ideas for future research

# YACRD Background

- **Nanopore sequencing:**
  - Ultra long read sequencer
  - Currently working on adoption process
  - MinION and PromethION
- **Chimeric reads:**
  - Chimeric reads impact downstream analysis process and introduce errors
  - Present in both single-genome sequencing and metagenome sequencing projects, including within-genome chimera and between-genome chimera
- **Currently we don't have established chimeric read removal in our Nanopore analysis pipeline**
  - Will YACRD work correctly?

# YACRD workflow

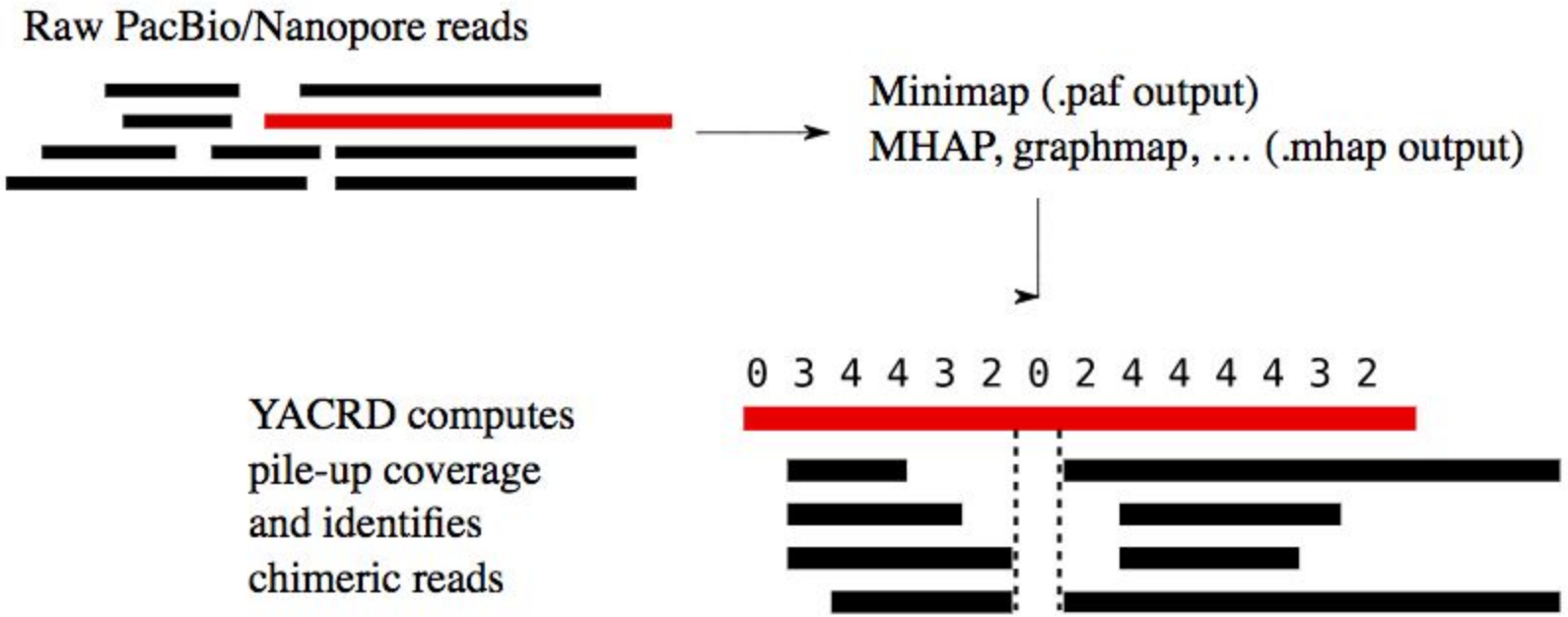



Image from yacrd github

# YACRD Experimental Design Goals

- **First stage:**
  - How good is YACRD? There is no paper on it.
  - Can it successfully identify chimeric reads and low coverage regions of reads?
  - Will this tool improve our analysis of Nanopore runs?
  - Will it work with real data, including metagenomes?
- **Second stage:**
  - Estimate the true positive and false positive rates by using the reference genome
  - Do this with real Nanopore data, both with a single genome and with a metagenome
- **Third stage:**
  - Determine whether or not YACRD should be incorporated into the Nanopore analysis pipeline

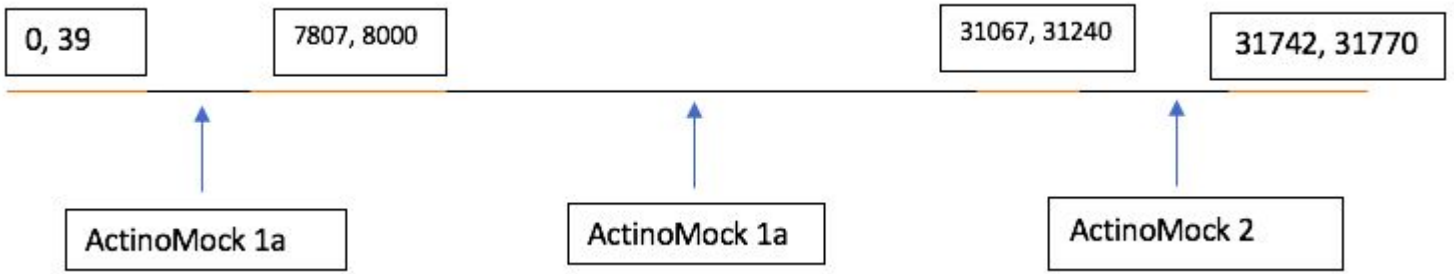
# YACRD Experimental Design Steps

1. **Run minimap2 on dataset (start with E. coli first)**
2. **Run YACRD on the minimap2 output**
3. **Create new fastq file with the following:**
   a. If read is marked by YACRD as not covered, remove it if it the not covered region is over 25% of the total read
   b. If the read is marked as chimeric, split up the reads that correspond to different regions of a single genome
4. **Randomly sample 20 reads from the final fastq file**
   a. Determine if each is a true or false positive by mapping them to the reference genome, and comparing their distances apart

# Chimeric read illustration

Typical example of chimeric read where adapter is making it chimeric


Example of chimeric read with adapter and cross-genome chimeric read



# YACRD Experiment Results

## Chimeric read statistics

| Sample name | Sample type | New # reads | Original # chimeric reads |
|---|---|---|---|
| X0161 | E. coli | 380,000 | 4296 |
| ActinoMock X0163 | Metagenome | 500,000 | 10,337 |

# YACRD Experiment Results

## Random sampling results

| Sample name | Sample type | Number samples | # True positives | # False positives |
|---|---|---|---|---|
| X0161 | E. coli | 20 | 20 | 0 |
| ActinoMock X0163 | Metagenome | 20 | 20 | 0 |

# YACRD Experiment Conclusion

- Overall, YACRD is good at detecting chimeric reads and no coverage regions
- 0 out of 20 random samples are false positives, so the false positive rate should be less than or equal to 5%
- If JGI decides to adopt Nanopore sequencing, YACRD should be added to the Nanopore data analysis pipeline

# Future Research

- **MetaBAT 3.0 exploration:**
  - Want to improve the accuracy of MetaBAT
  - Explicitly want to improve the sensitivity, as the specificity is already quite good
- **Will explore using new metrics for improving metagenome binning results**
  - May utilize machine learning

# Questions

## Thank you

Zhong Wang, Rob Egan, Volkan Sevim

Link to YACRD analysis repo and summary:
https://github.com/JGI-Bioinformatics/yacrd_analysis