Association of Biomolecular Resource Facilities

April 25, 2018

JOHNS HOPKINS
BIOMEDICAL ENGINEERING

Bacterial DNA Sequencing with Nanopores: Assembly and Modifications

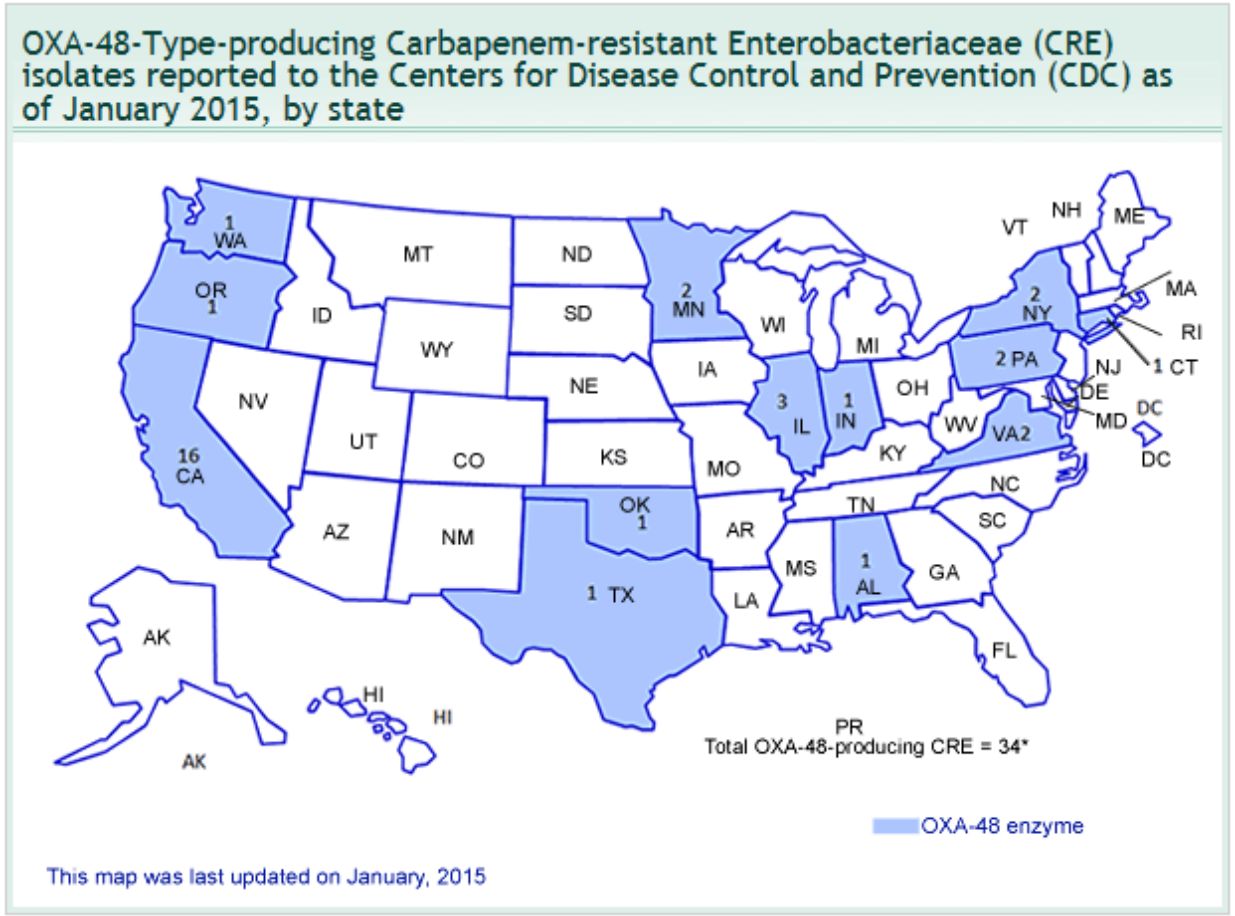
Winston Timp
Department of Biomedical Engineering
Johns Hopkins University

Hypervirulent (hypermucoviscous) *K. pneumoniae*

New variant of *Klebsiella pneumoniae*
First described in the Asian Pacific Rim 1980s
Now increasingly recognized in other countries

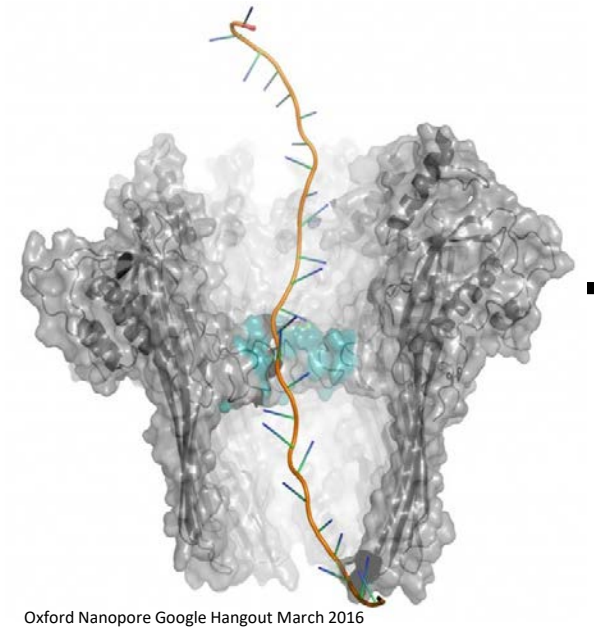
Defining clinical features:

- Serious, life-threatening community-acquired infection in younger healthy hosts
- Liver abscess, pneumonia, meningitis and endophthalmitis
- Metastatic spread

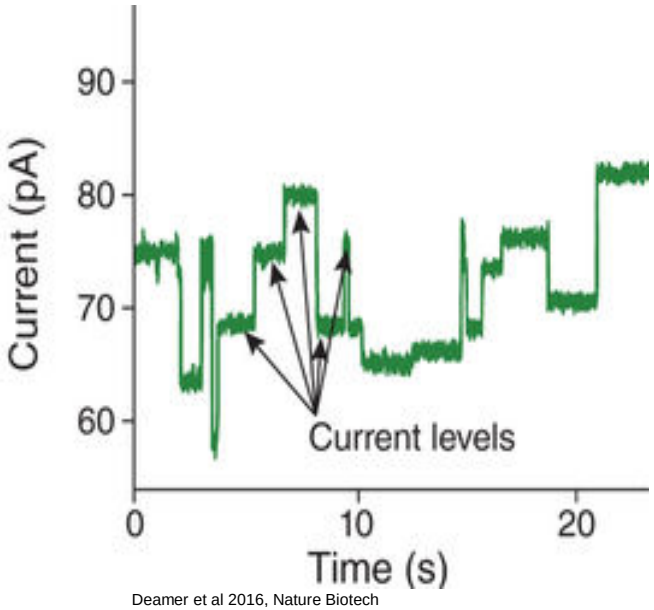


Source: Shon, Rajinda, Russo 2013
CDC
http://www.cdc.gov/hai/organisms/cre/TrackingCRE.html#CREmapOXA

Nanopore: Single Molecule Sequencing



Oxford Nanopore Google Hangout March 2016



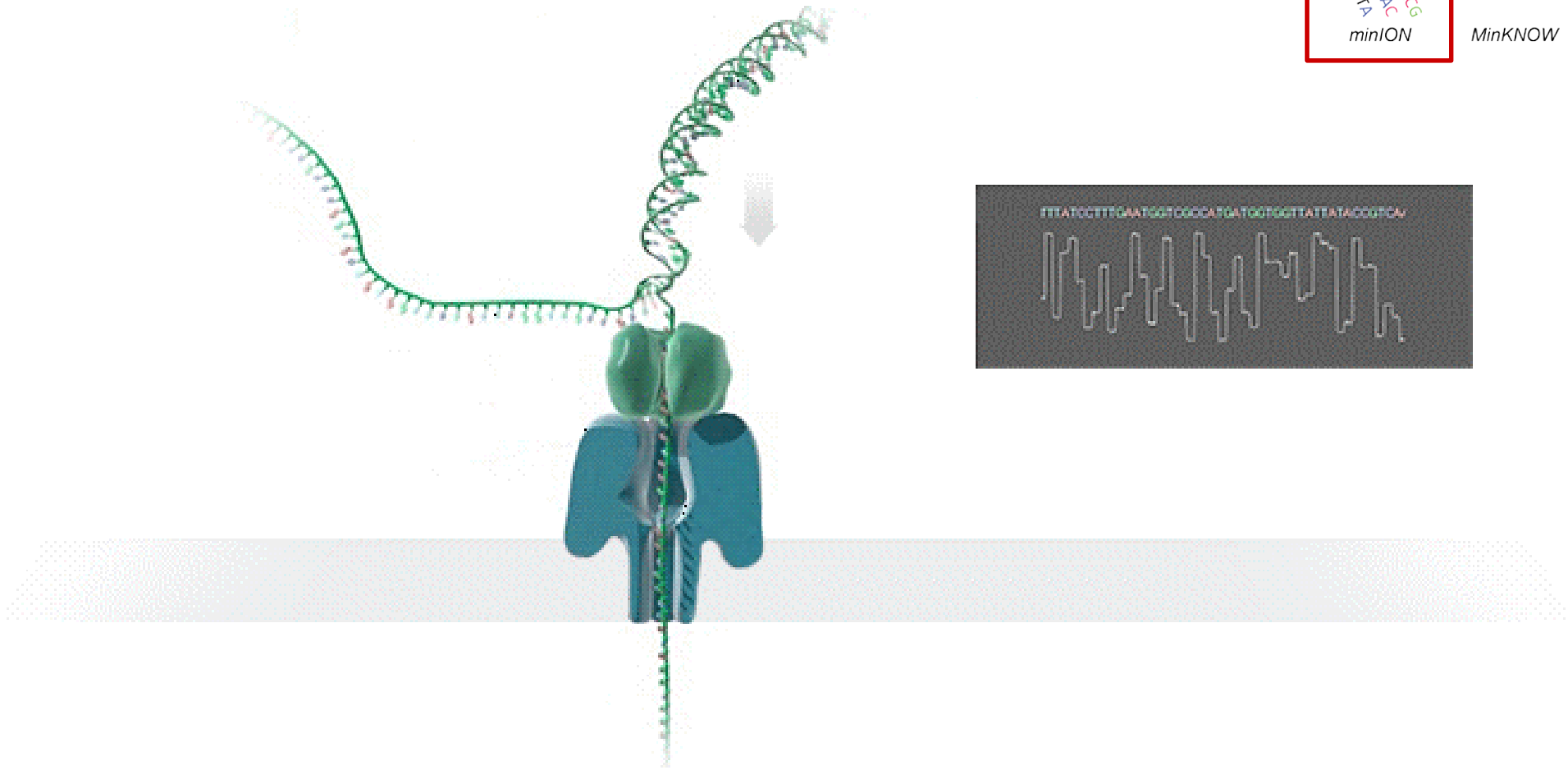
Deamer et al 2016, Nature Biotech

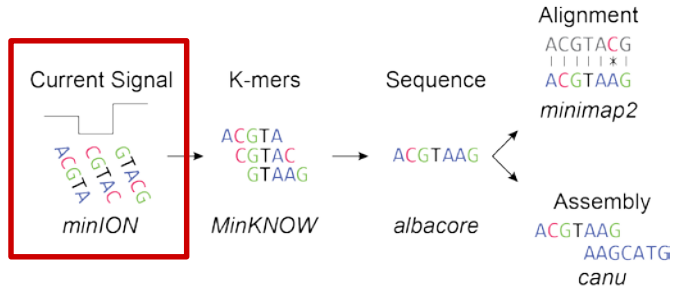
ATCGATCGATAGTA
TTAGATACGACTAG
CGATCAG

- Oxford Nanopore Technologies, CsgG biological pore
- No theoretical upper limit to sequencing read length, practical limit only in delivering DNA to the pore intact
- Palm sized sequencer
- Sequencing output 5-10Gb

Disclosure: Timp has two patents (US 2011/0226623 A1; US2012/0040343 A1) licensed to ONT

Sequencing Operation

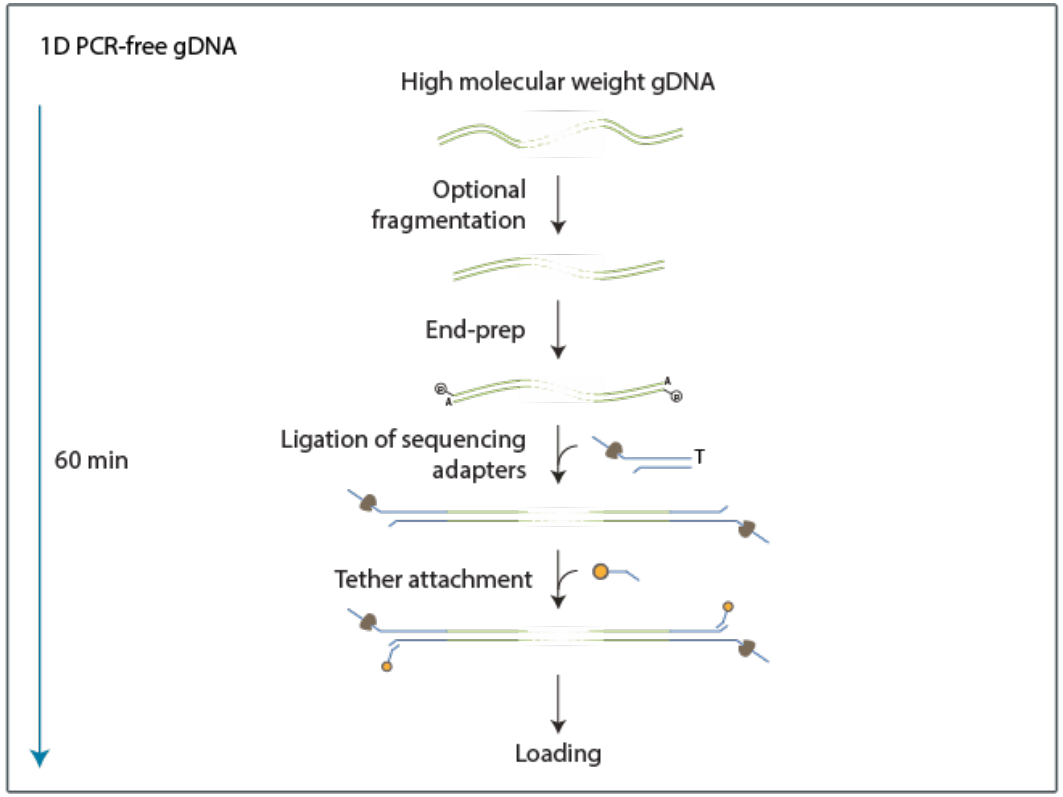


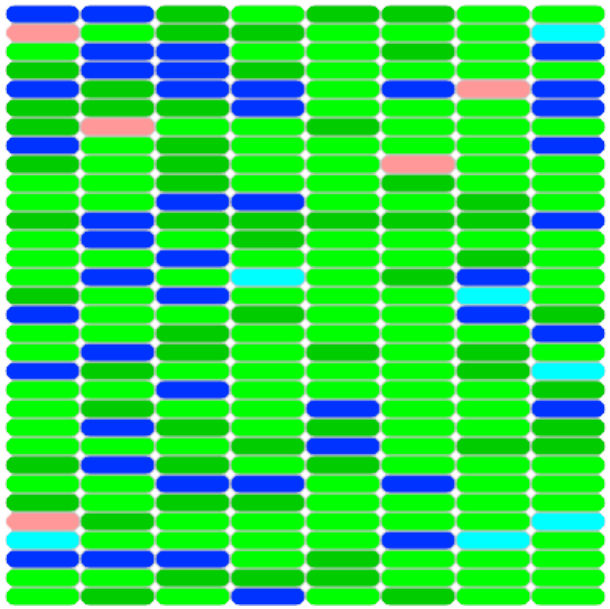


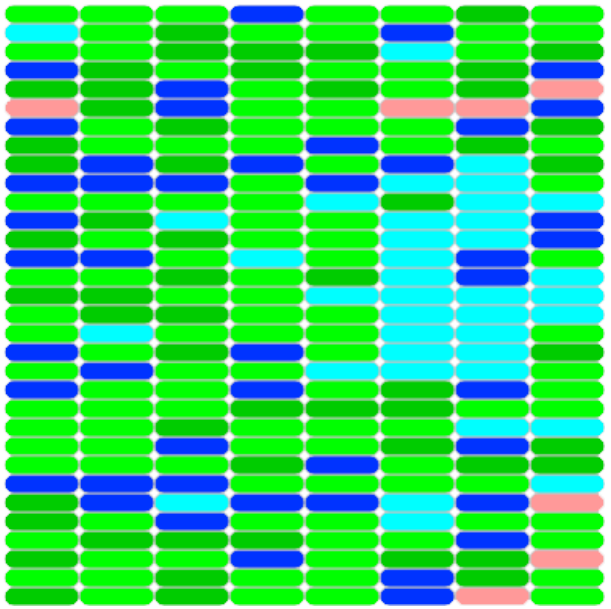
Oxford Nanopore Technologies

- Protein nanopores on a synthetic polymer
- Multiple base-pairs at a time (“k-mers”)
- Characteristic current signature is converted to nucleotide sequences

Nanopore Library Prep







- Library prep is very similar to methods for short-read sequencing
- For DNA shearing we use Covaris gTubes or Diagenode Megaruptor
- After end-repair and A-tailing, leader adapter with motor protein is ligated
- MinION arrays 512 channels (with 4 pores possible per channel) (shown bottom left from running software); dark green pores are sequencing, light green available, other colors inactive.

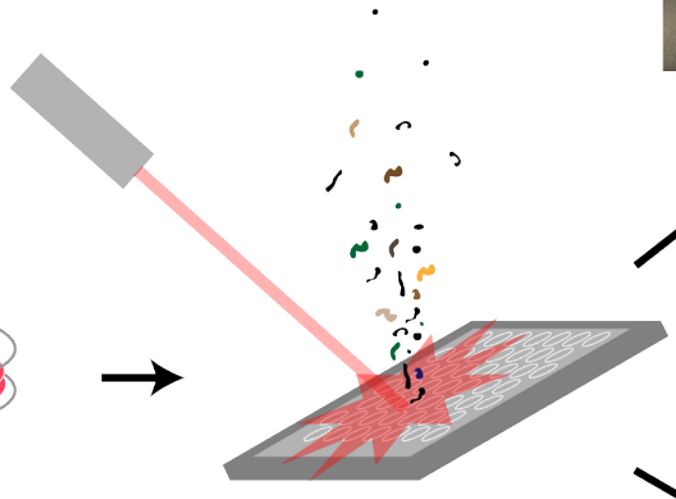
Project Overview

➔ Correlate resistance phenotypes with genomic features

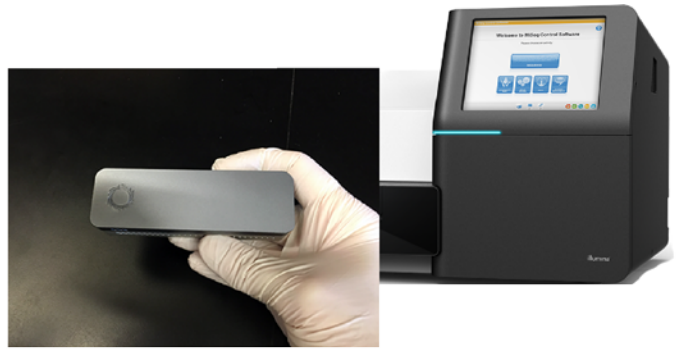
Patients

Rectal/
Perirectal
Swab

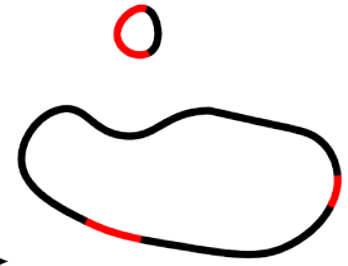
Culture



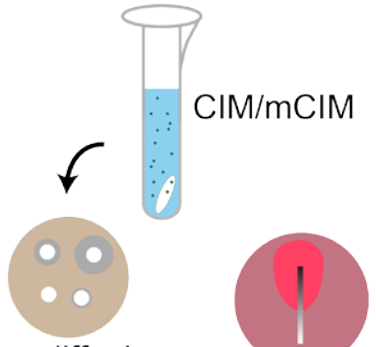
MALDI-TOF to
ID organisms



Nanopore (+ Illumina)
DNA Sequencing



Assemble genome,
find resistance genes
and mutations

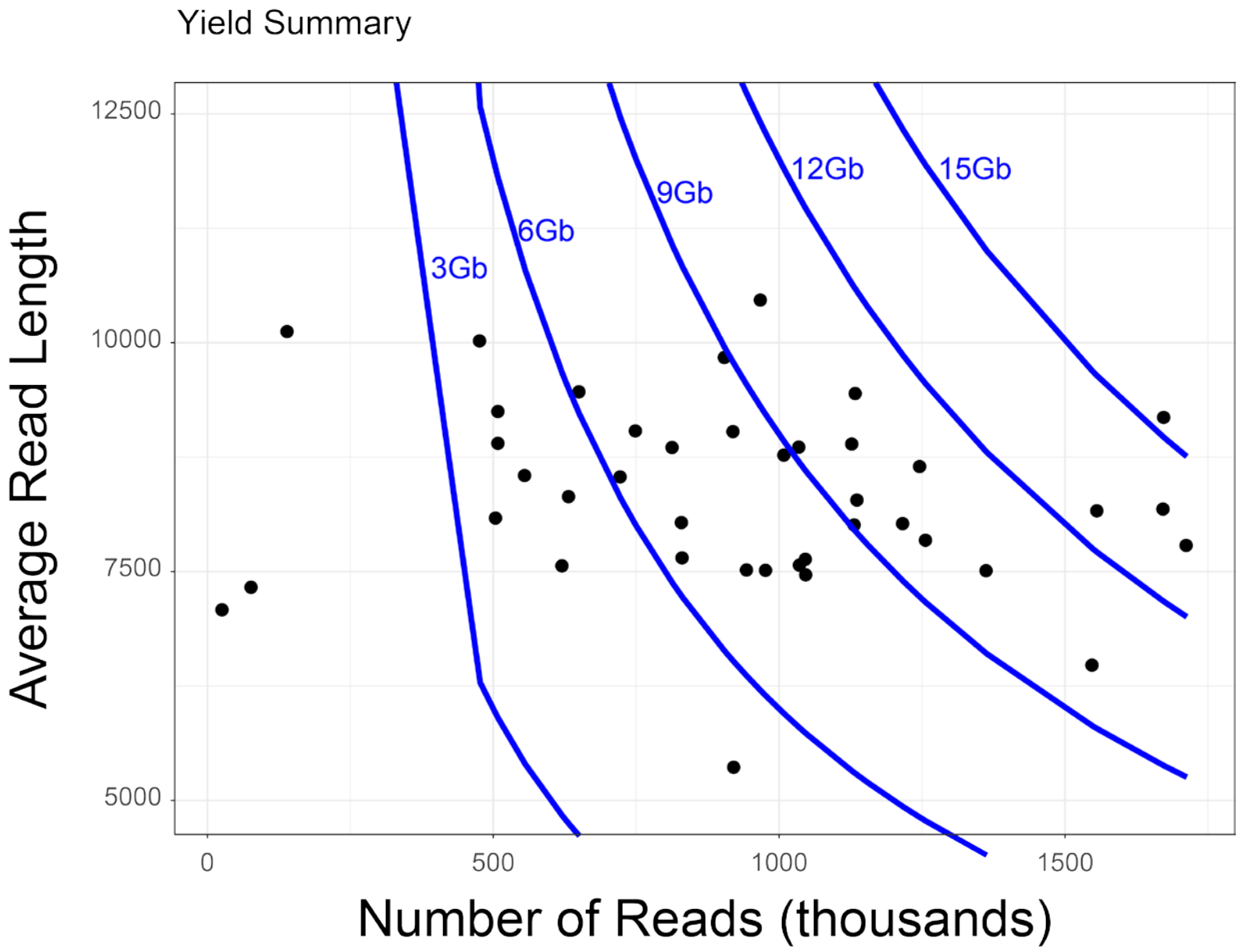


Standard of Care Tests

Antibiotic
Susceptibility
Profiles

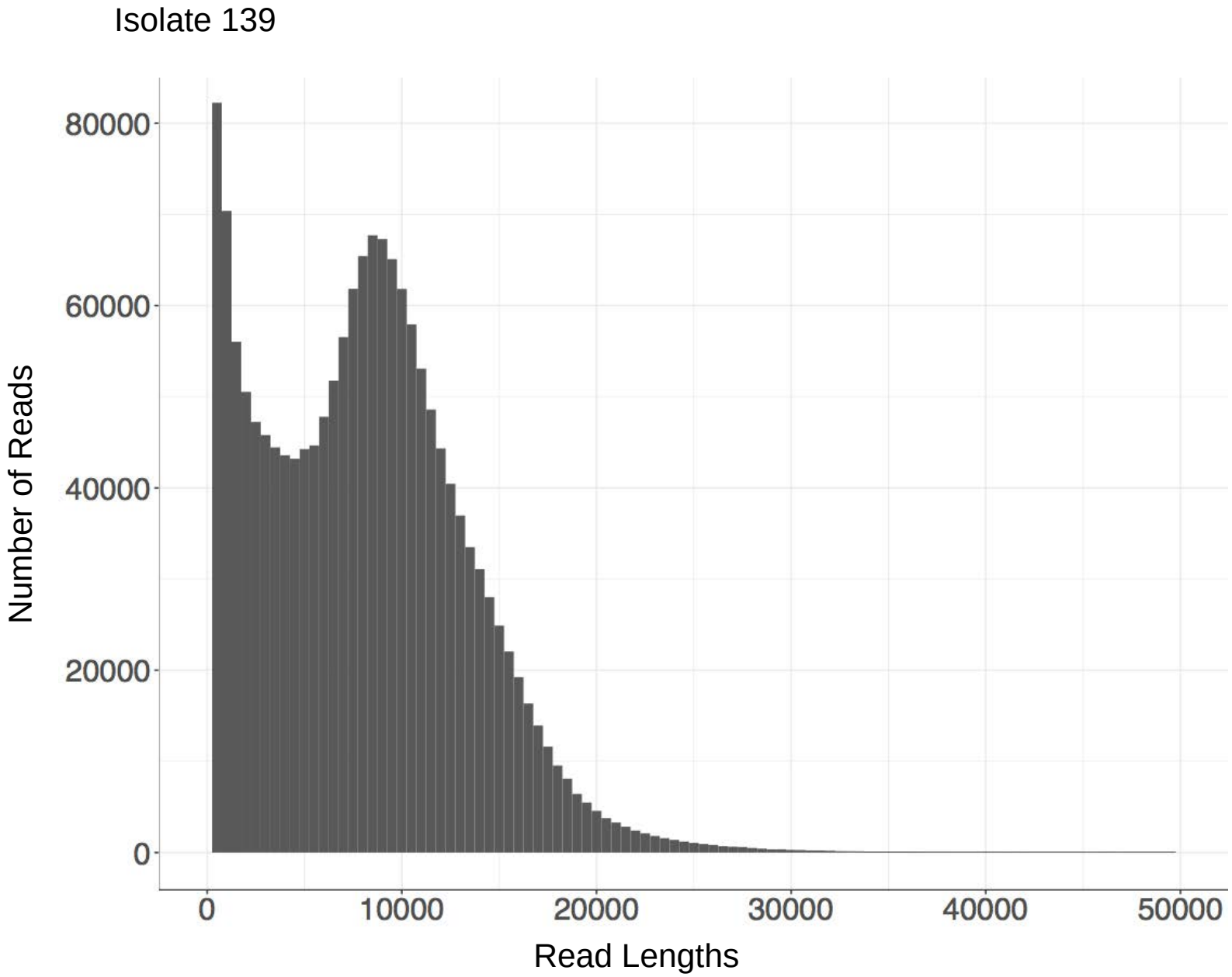
Run data

- Nanopore sequencing of bacterial isolates from the Simner Lab achieves base pair yields often >10Gb per flowcell.
- For reference (in bulk) costs per flowcell are ~$600 per
- For a nanopore flowcell with average yield of recent bacterial runs (8.9Gb) the cost is $500 or $56.18 per Gb.

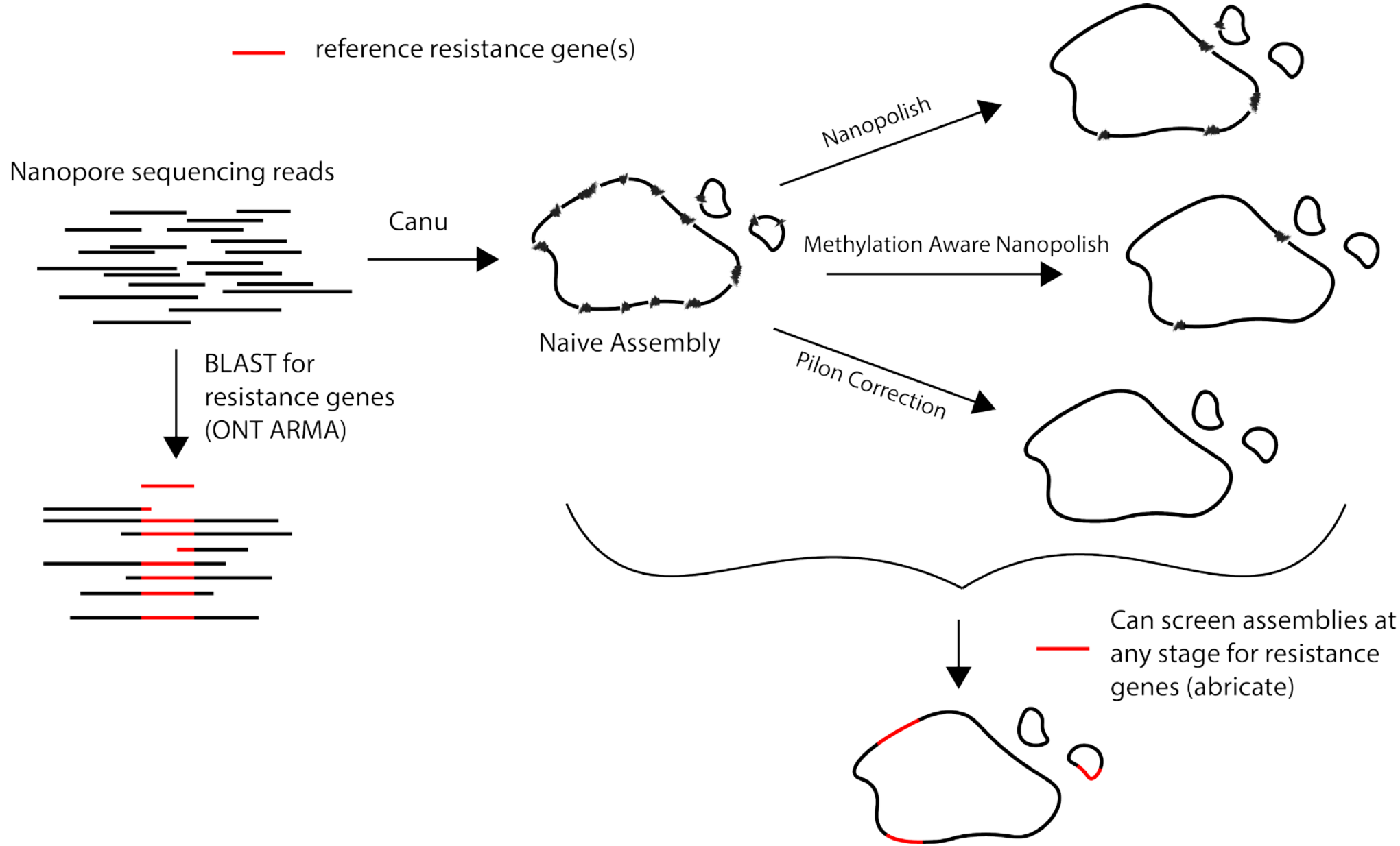


Read length histogram

- Read length for nanopore sequencing was relatively short
- Due to prep methods:
  - HMV was hard to extract effectively, bead beating likely broke DNA
  - We also sheared (Covaris gTubes)
  - To get plasmid sequences, either tagmentation or shearing is likely required to get smaller circular sequences

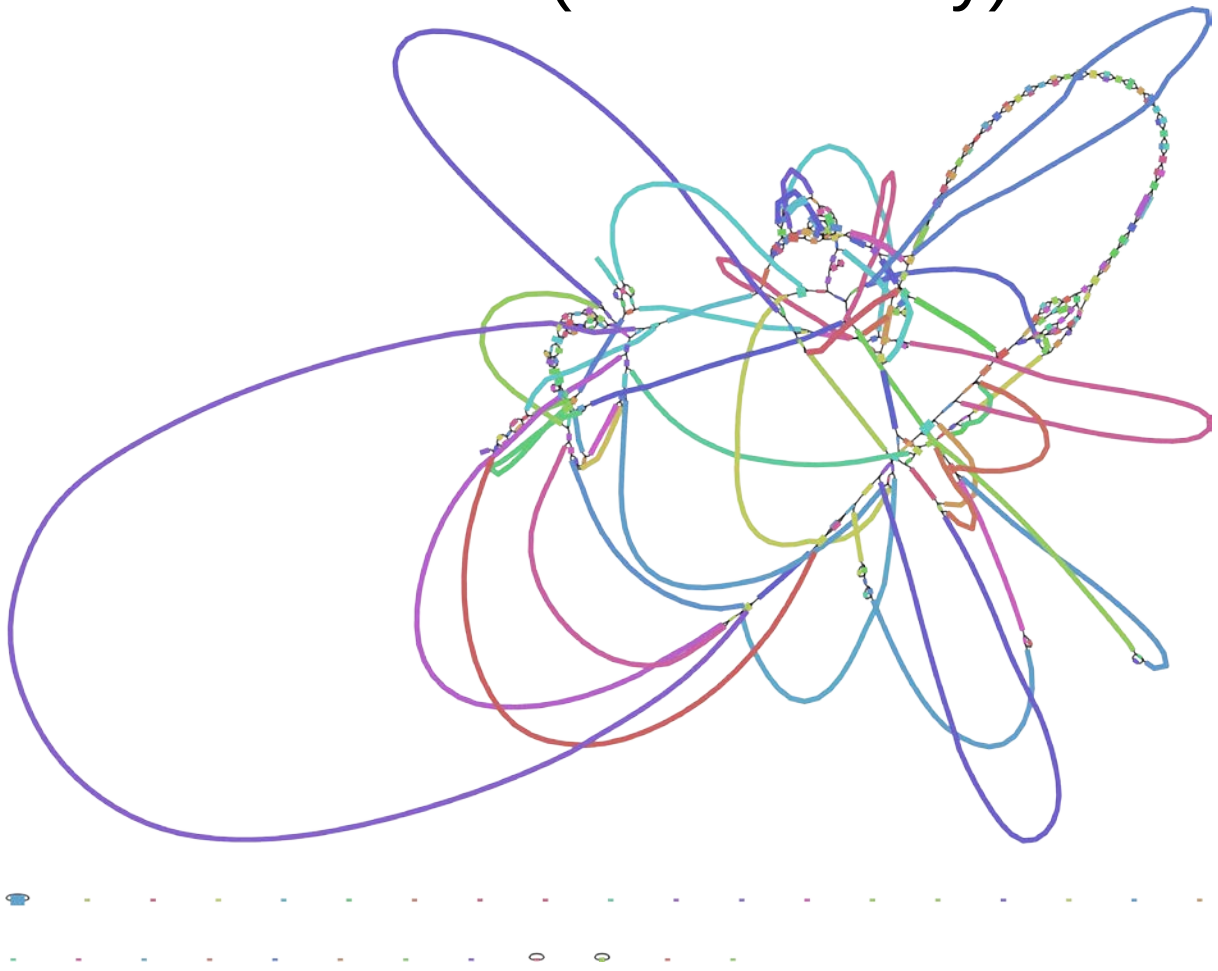


The Pipeline so far

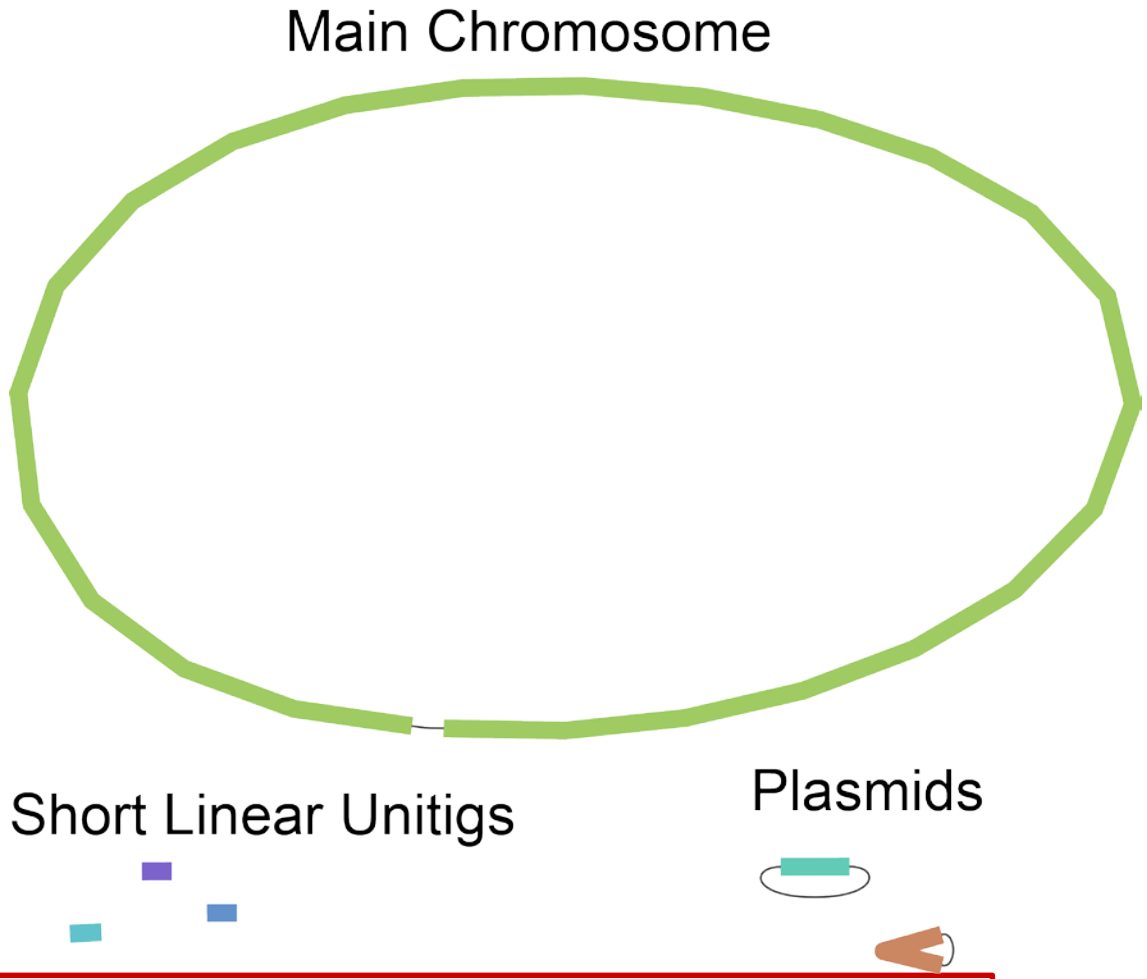


Assemblies

SPAdes (Illumina only)



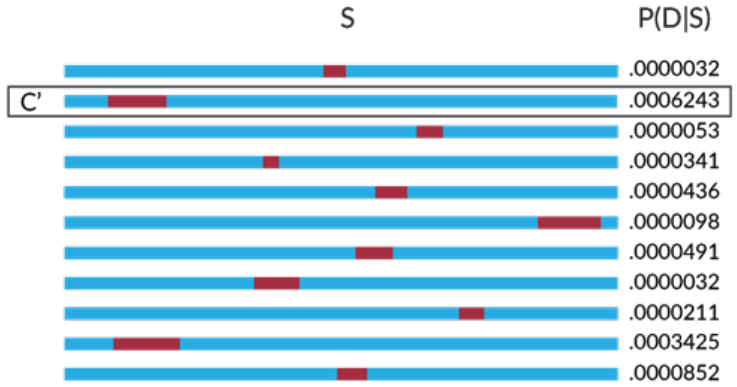
Canu (Nanopore only)



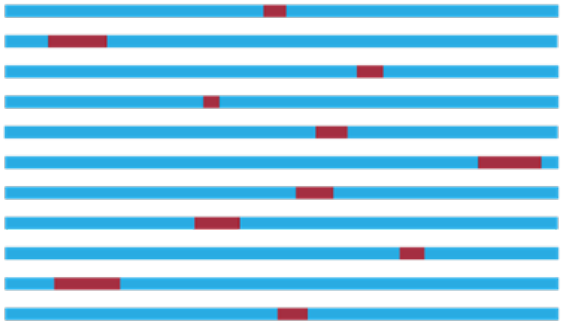
Long reads **really** help in getting complete assemblies – full single contig chromosomes and plasmids identified cleanly.

Nanopolish

Set the best S as the new C, and repeat

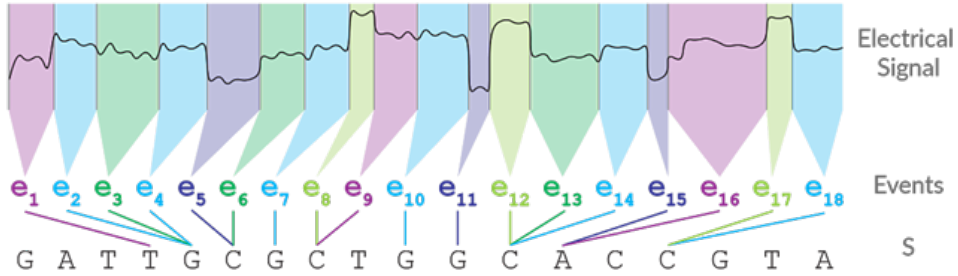


Based on non-matching stretches of sequence between the assembly and the reads, generate a list of candidate improvements to C, called S

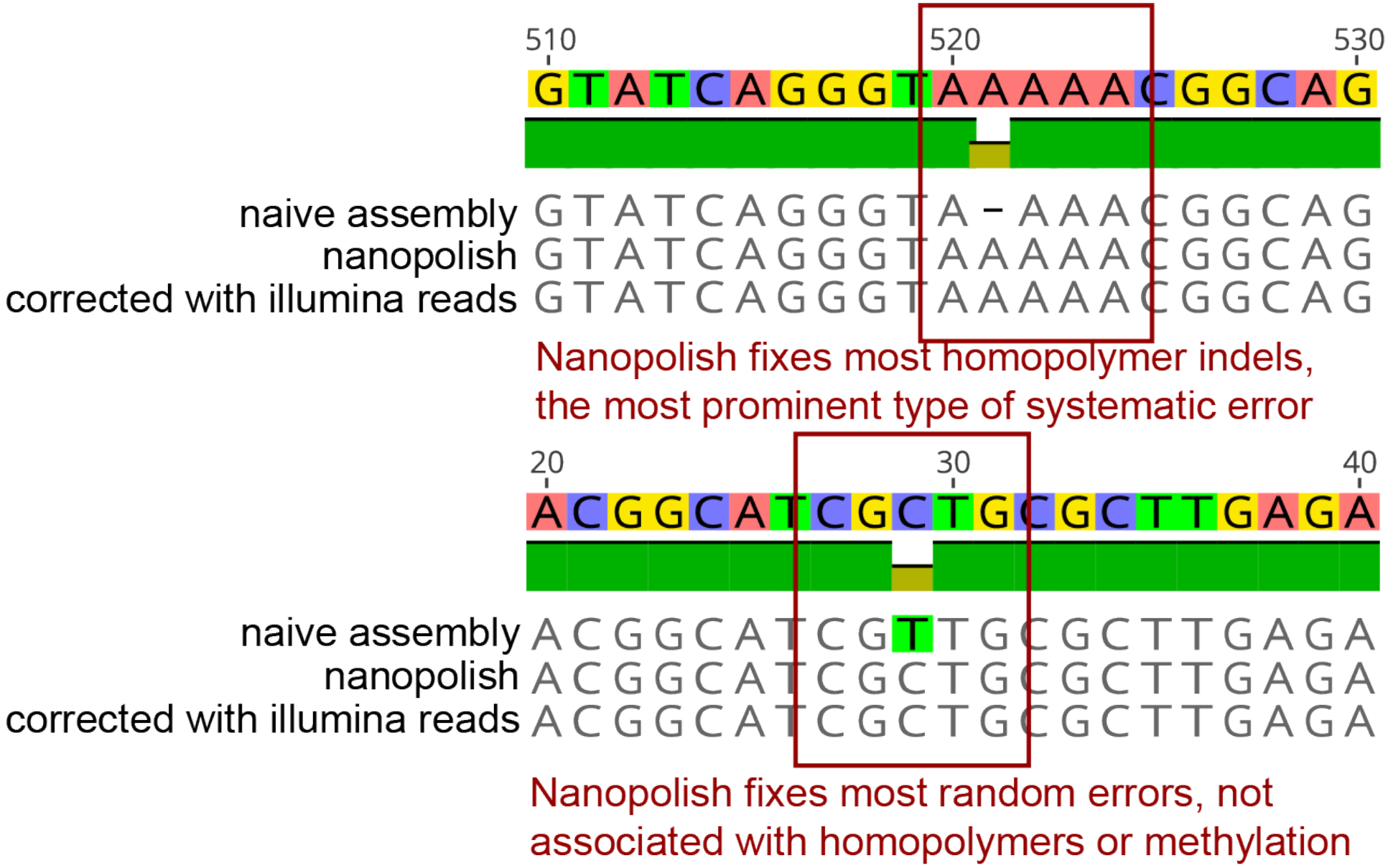


Use a Hidden Markov Model to align the electrical data from the reads to each S, and compute the probability of observing the event data given S, P(D|S)

Choose the S that maximizes the probability of observing the event data

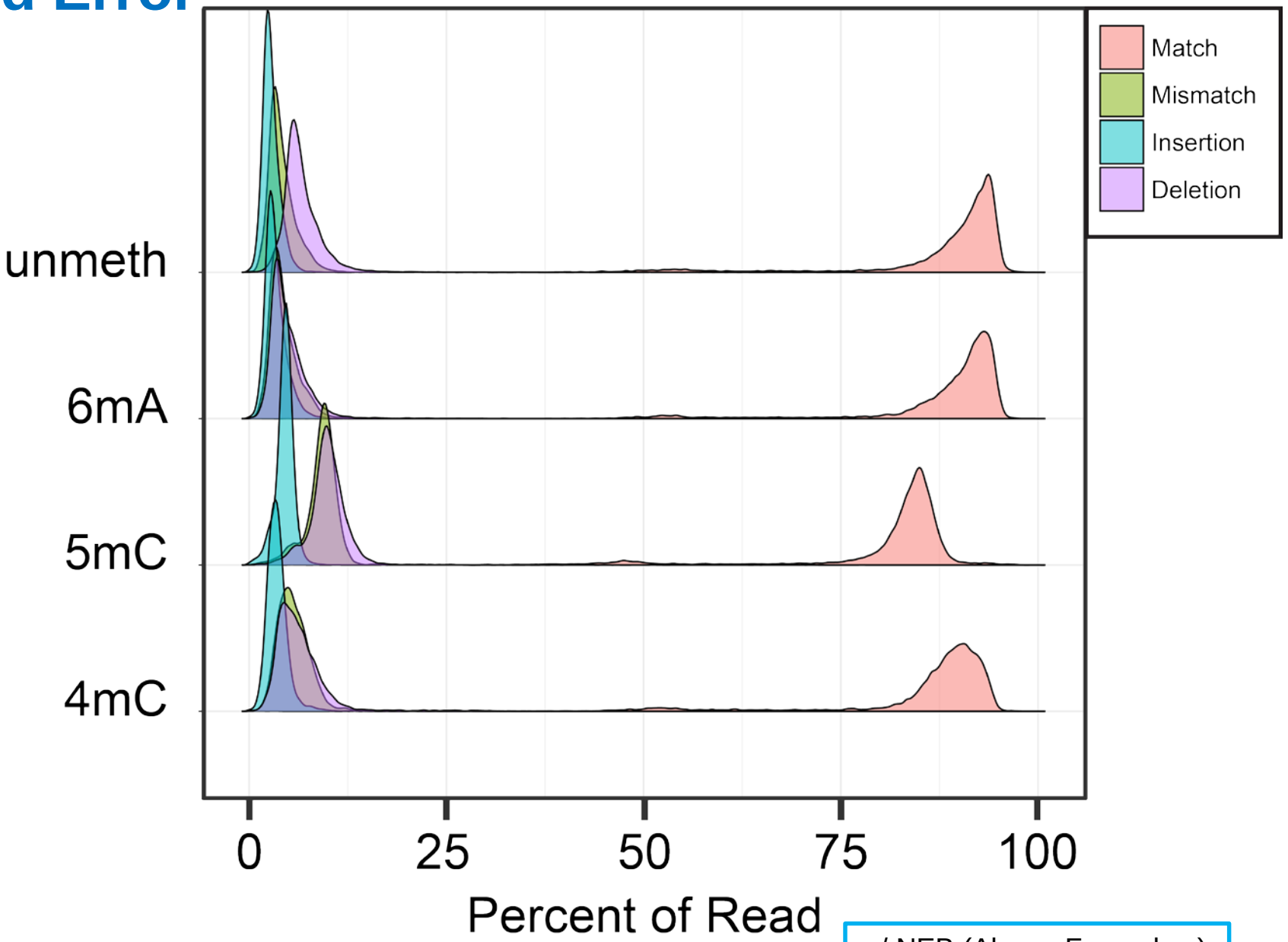


Assembly Using Signal to polish



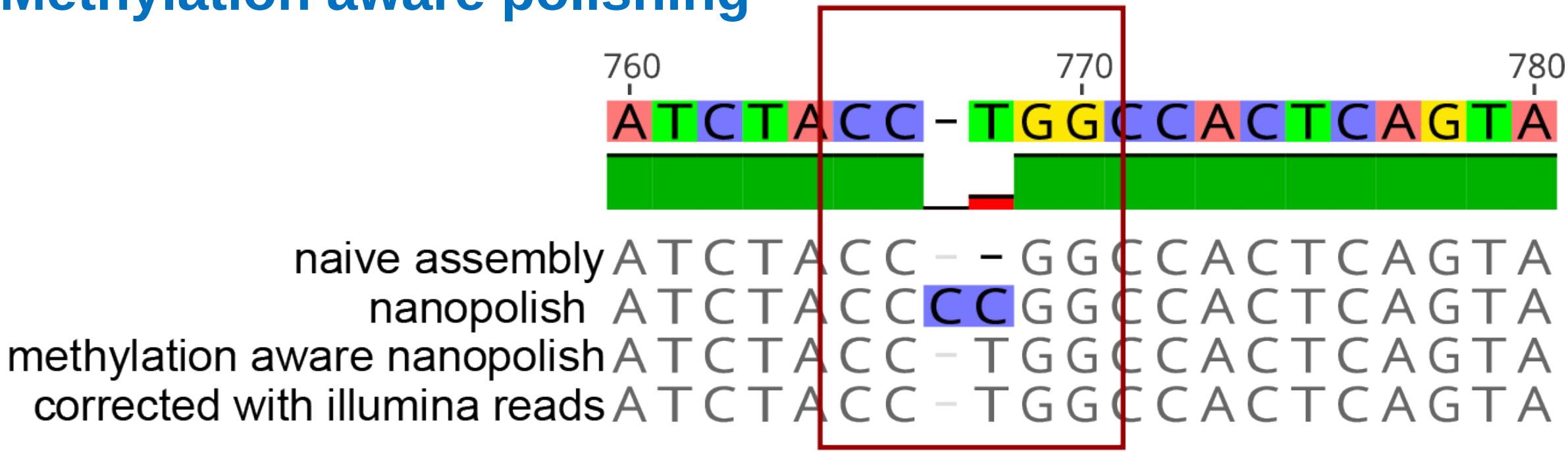
Nanopore: Methylated Error

- We sequenced samples from NEB ER2796 (E. Coli with KO of dam/dcm)
- Different methyltransferases are transformed in.
- Notably, mismatch error rate and deletions seem higher on methylated samples than unmethylated.
- The lower shift in 4mC and 6mA may be do to relative infrequency of those motifs.



w/ NEB (Alexey Fomenkov)

Methylation aware polishing



models need to be trained on methylated k-mers in order to correct MTase motifs (dcm MTase= **CCWGG**)

Raw Assembly	Nanopolish Corrected	Methylation Aware Nanopolish Corrected
98.89%	99.57%	99.76%

Differences report

Mutation or sequencing artifact?

```
>KLPN_133_tig00000001:3594157-3595237- coverage:100.00 score=2120 edit_distance=6
D  115   39(M) A:   115 GAAAATGGTC   AAAAATGGT  homopolymer  premature stop at AA63
I  123   41(G) :G   123 GTCGGgCGAGC  TCGGCGAGC  [unknown]    premature stop at AA89
S  294   98(L) G:A  295 CGTCTaGCGTT  GTCTGGCGT  [unknown]    synonymous
S  688  230(W) T:C  689 AAGCCcGGGCG  AGCCTGGGC  motif CCTGG  AA230|W:R
D  946  316(W) T:   946 GTACCGGTAC   TACCTGGTA  motif CCTGG  premature stop at AA323
I 1049  350(Q) :G  1049 CGACCgAGGCG  GACCAGGCG  motif CCAGG  frame shift QAAV+:RGGR+
```

1. event type
2. WT base loc
3. Amino acid
4. base changes
5. ctg base loc
6. ctg context (5+m+5)
7. WT context (4+1+4)
8. context status
9. mutation impact

Conclusions/Future Directions

- Nanopore sequencing could be useful as an aid in providing proper treatment for infectious diseases
  - Get full coverage of pathogenic organisms with one flowcell
  - Rapid time of detection of acquired resistence genes of interest
- Chromosomal Mutations – a bit harder
  - Application of nanopolish to improve nanopore-only assemblies
  - Application of nanopolish to call methylation in assemblies

Acknowledgments

- **Timp Lab – Johns Hopkins University**
- Winston Timp
- Yunfan Fan

- **Simner Lab – Johns Hopkins School of Medicine**
- Patricia (Trish) Simner, PhD
- Yehudit Bergman

- **Ontario Institute for Cancer Research**
- Jared Simpson, PhD
- P.C. Zuzarte, PhD
- Matei David, PhD
- L. J. Dursi, PhD

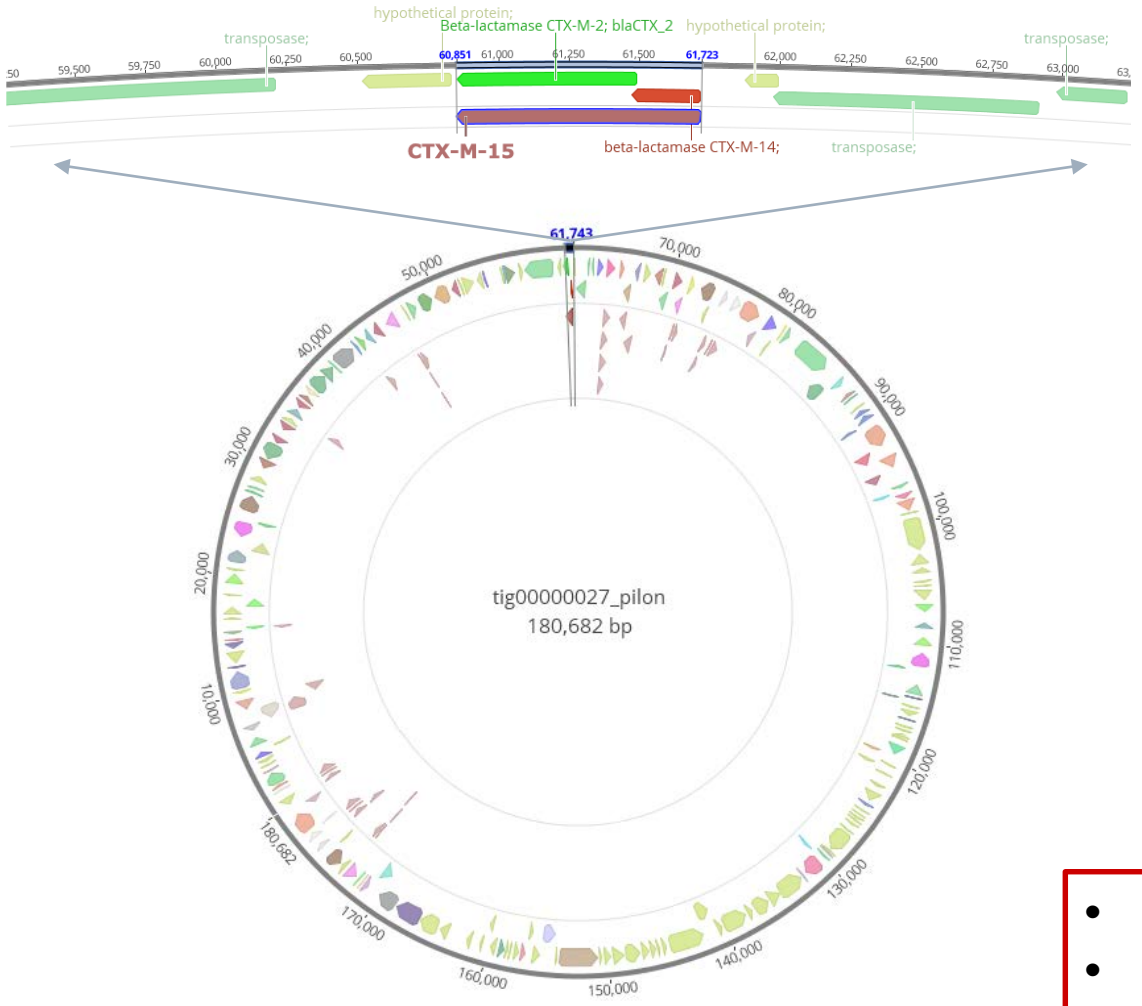
National Institute of Allergy and Infectious Diseases

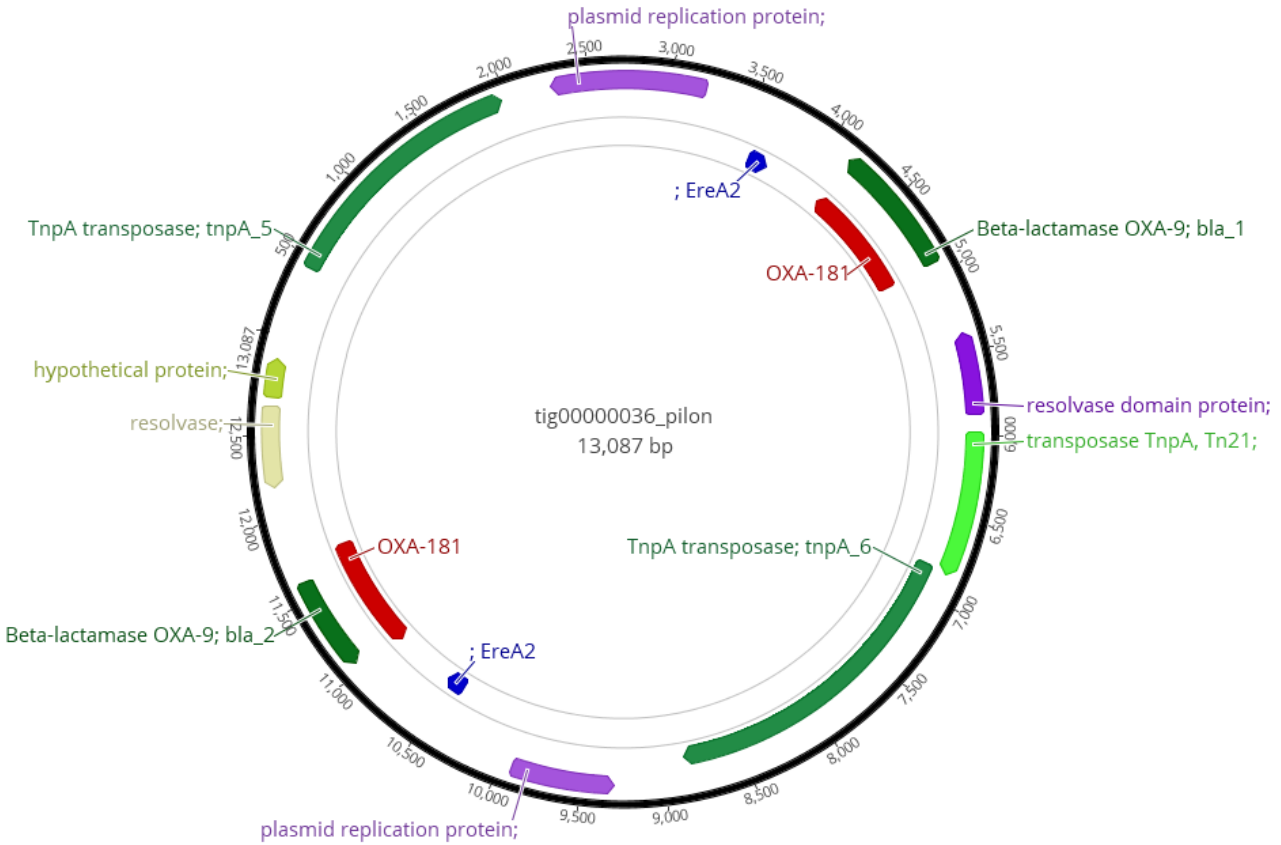
1R21AI130608-01 (Simner)

National Human Genome Research Institute

1R01HG009190-01A1

AMR genes





- Annotated assemblies with prokka and CARD databse
- Two specific AMR genes (CTX-M-15 and OXA-181) were discovered and placed within their genetic context (plasmids)