**Agilent on Tap**

May 2018

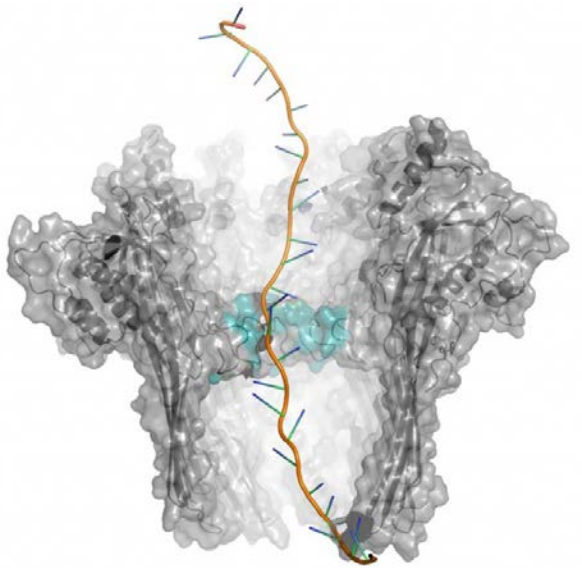
JOHNS HOPKINS
BIOMEDICAL ENGINEERING

# Researching Cancer with the minION: Methylation and Structural Variation

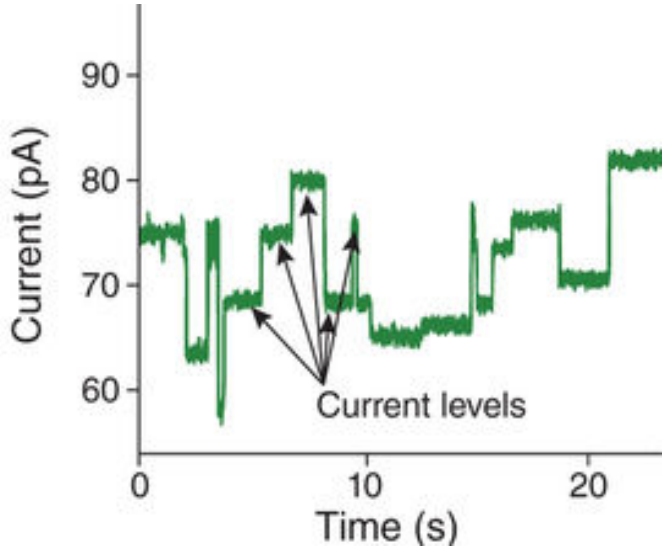
Winston Timp
Department of Biomedical Engineering
Johns Hopkins University

# Nanopore: Single Molecule Sequencing

- 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
- Predicted sequencing output 3-6Gb



Oxford Nanopore Google Hangout March 2016

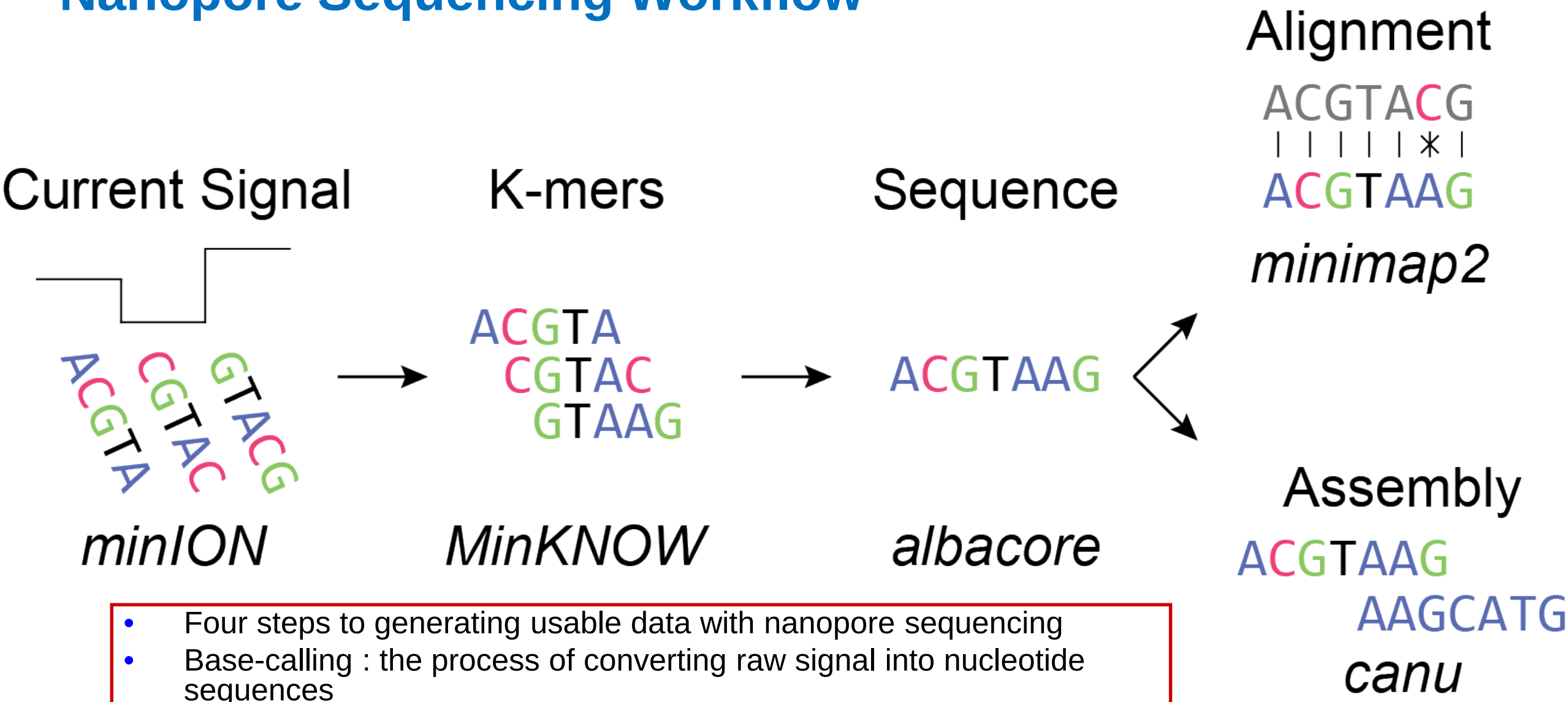



Deamer et al 2016, Nature Biotech

ATCGATCGATAGTAT
TAGATACGACTAGC
GATCAG

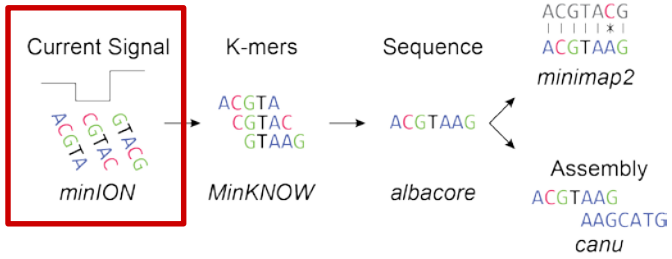
Disclosure: Timp has two patents (US 2011/0226623 A1; US2012/0040343 A1) licensed to ONT

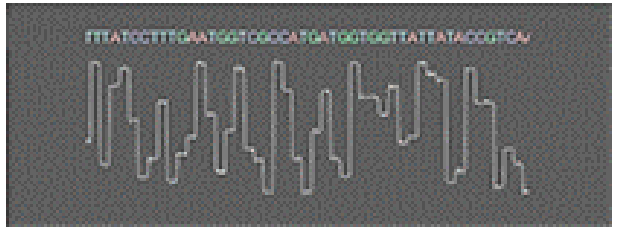
# Nanopore Sequencing Workflow




- Four steps to generating usable data with nanopore sequencing
- Base-calling : the process of converting raw signal into nucleotide sequences
- Nanopolish : uses alignment and current signal to improve base-calls

# Sequencing Operation



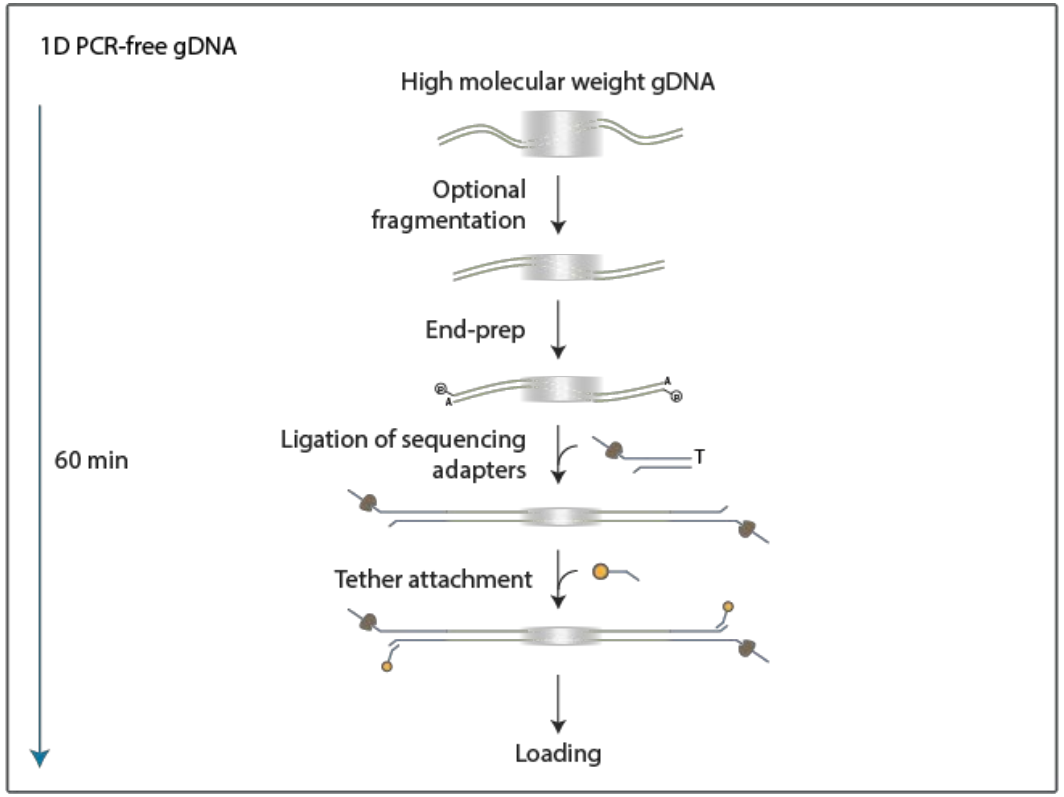





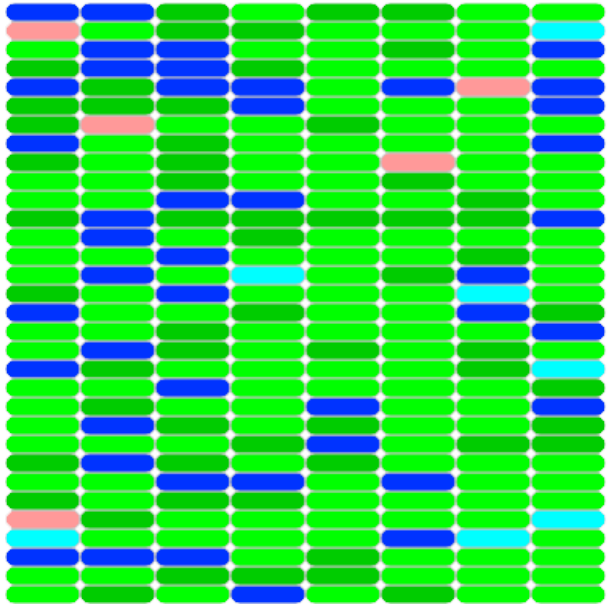
Oxford Nanopore Technologies

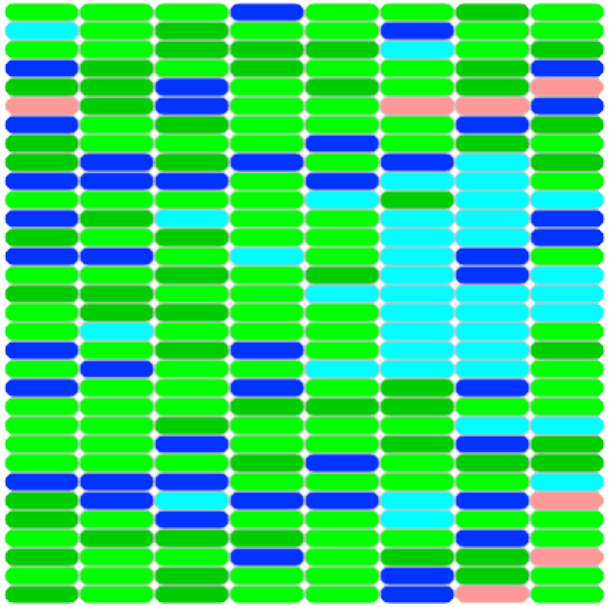
- Protein nanopores on a synthetic polymer
- Multiple base-pairs at a time (“k-mers”)
- Characteristic current signature is converted to nucleotide sequences

# Nanopore Library Prep



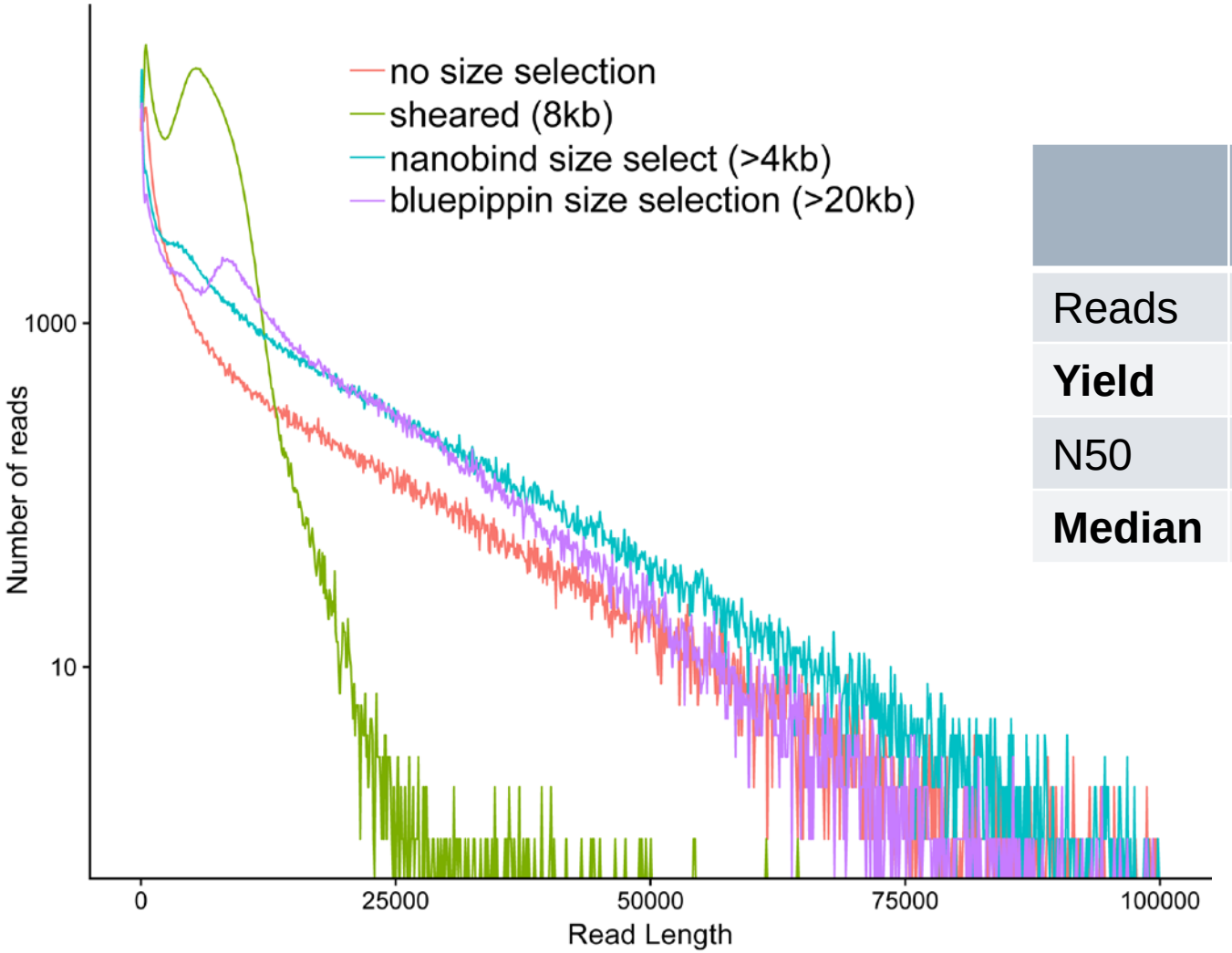






- Library prep is very similar to methods for short-read sequencing
- For DNA shearing we used Covaris gTubes or Diagenode Megaruptor
- After end-repair and A-tailing, leader adapter with motor protein is ligated
- MinION arrays 512 channels (with 4 pores possible per channel) (shown bottom left from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

# Improving Read Lengths: Size selection



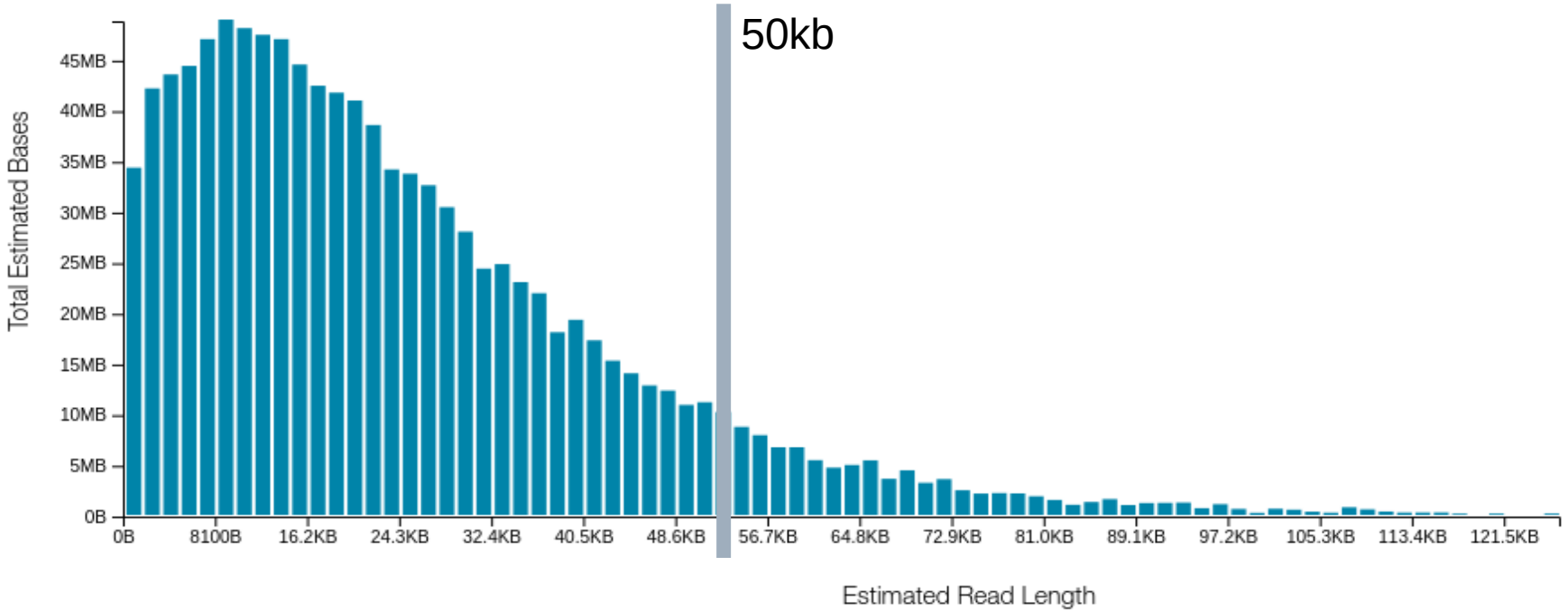

| | None | Sheared | Nanobind SS (4kb) | Blue Pippin SS (20kb) |
|---|---|---|---|---|
| Reads | 353k | 2060k | 400k | 435k |
| **Yield** | **1.71Gb** | **10.1Gb** | **3.57Gb** | **3.65Gb** |
| N50 | 17.3kb | 6.6kb | 15.7kb | 19.0kb |
| **Median** | **1.2kb** | **5.1kb** | **6.8kb** | **4.3kb** |

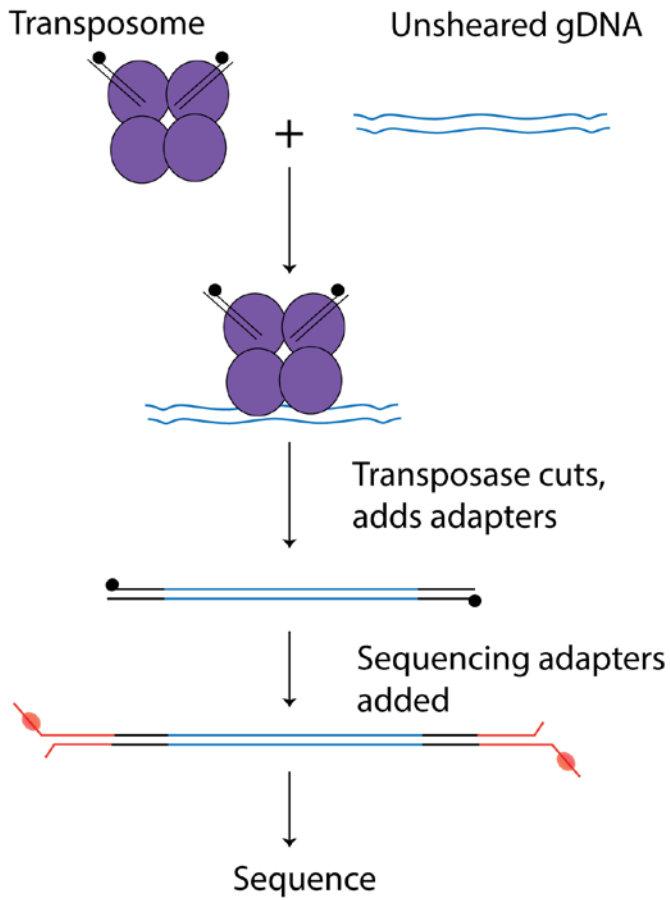
Read length and yield require some optimization and trade-offs

# Improving Read Lengths: Rapid kit RAD004

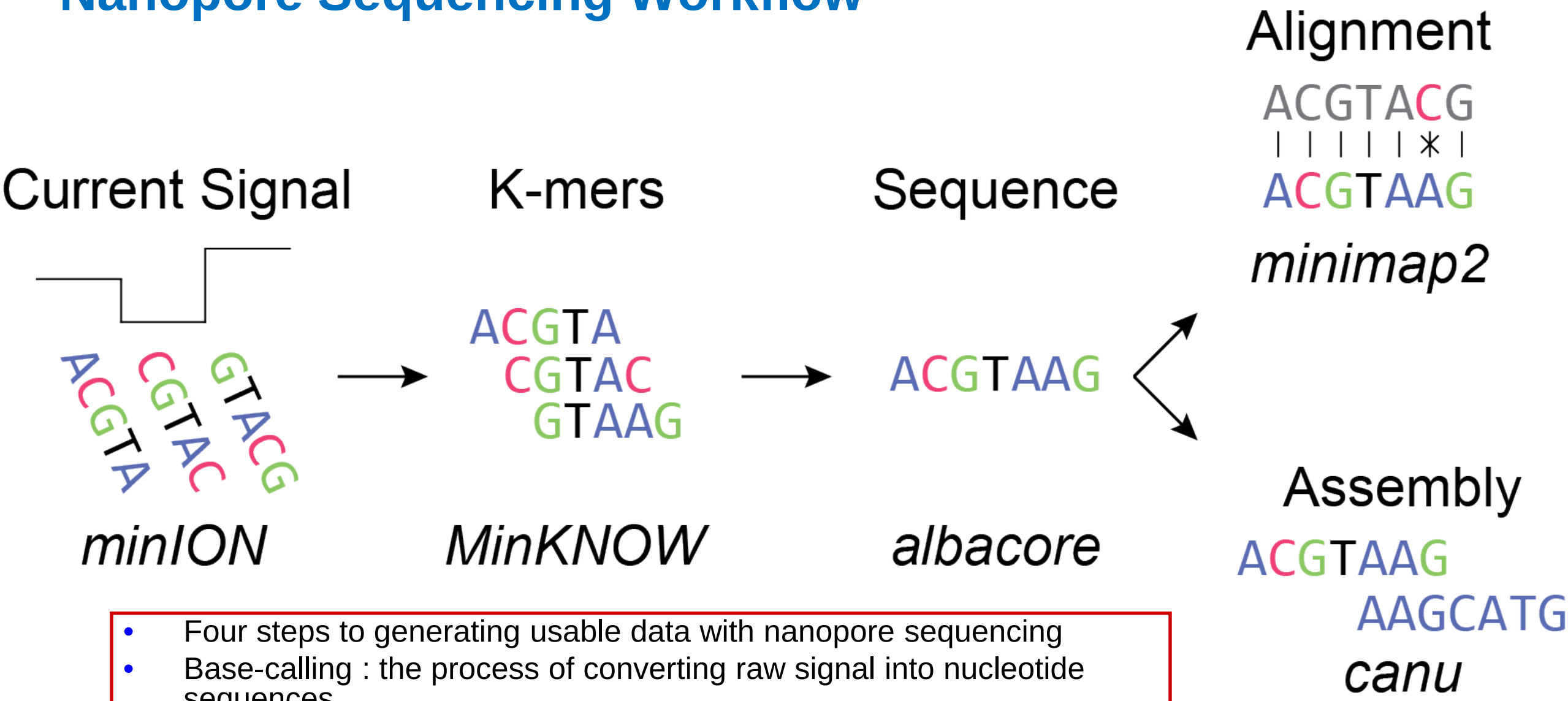
**15 minute protocol**

**Yield:**
3Gb from 150K reads
0.89Gb from >50kb reads



# Nanopore Sequencing Workflow

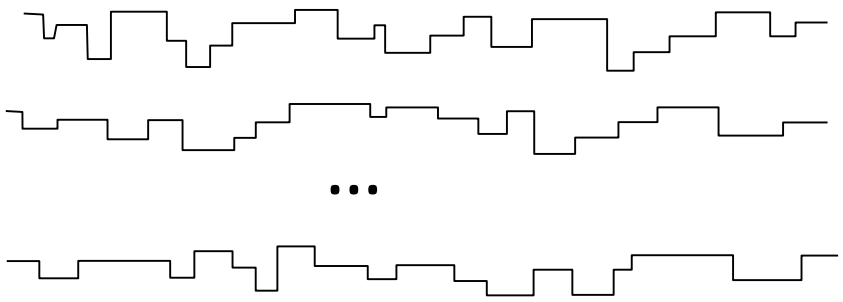



- Four steps to generating usable data with nanopore sequencing
- Base-calling : the process of converting raw signal into nucleotide sequences
- Nanopolish : uses alignment and current signal to improve base-calls

# Nanopolish tools

github.com/jts/nanopolish

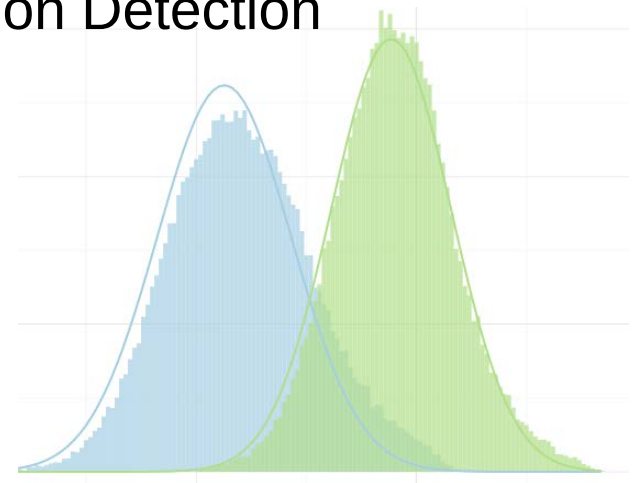
## Consensus Calling



## Reference-based SNP Calling

```
chr20   44921212        T       C       165.9   1/1
chr20   44921404        A       T       381.3   1/1
chr20   44922637        A       C       354.0   1/1
chr20   44934236        G       A       24.3    0/1
chr20   44960481        C       T       39.1    0/1
chr20   44963260        G       A       99.1    0/1
chr20   44963607        T       C       207.3   0/1
```

## Methylation Detection

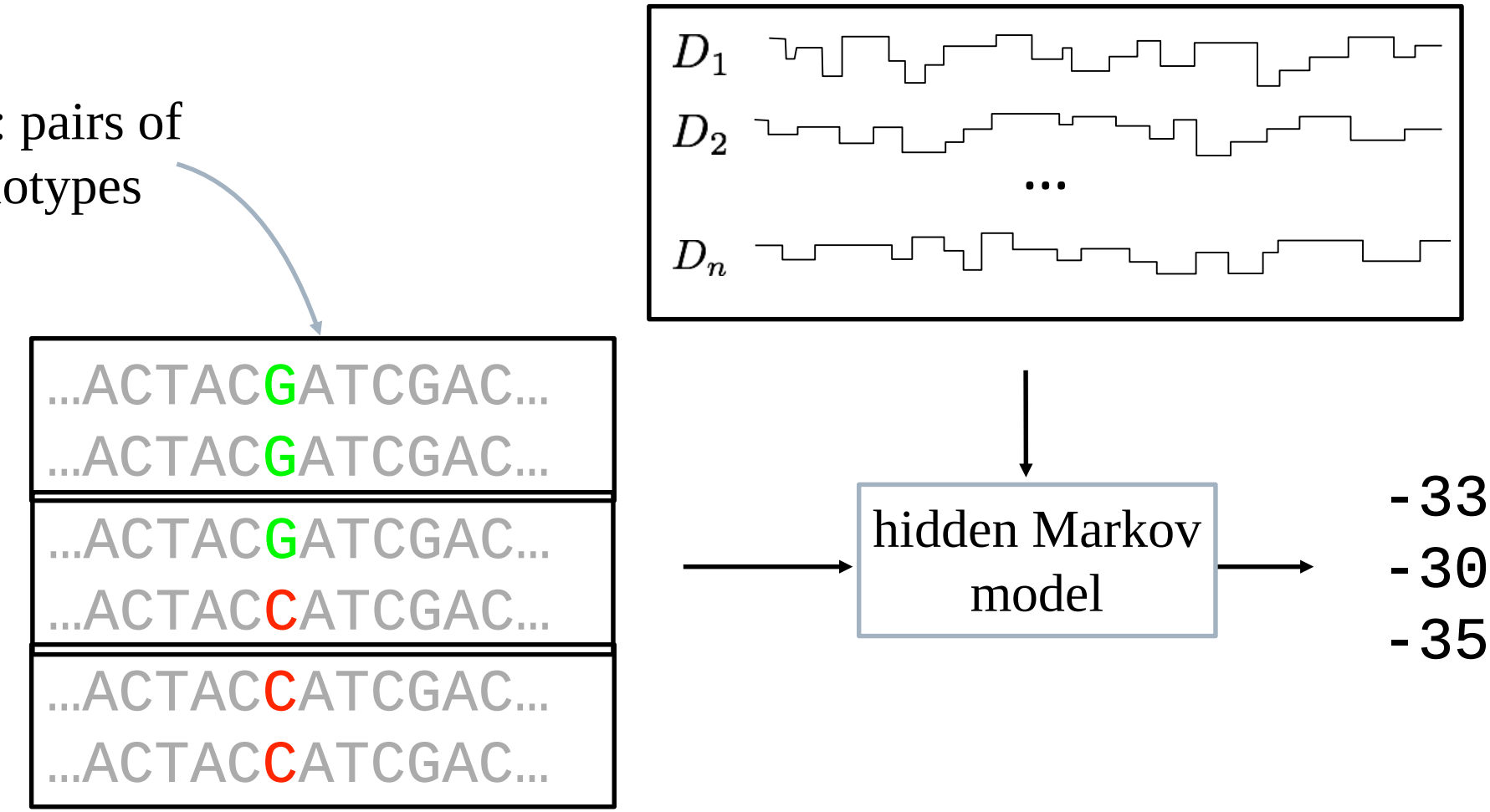


## Read Phasing

```
TAGAAGATATCATGTATAGTACGAT
TAGAAGATATCATG
TAGCAGATATCATGTATATTACGAT
          CATGTATATTACGAT
```

# Nanopolish SNP Calling and Genotyping

**input**: pairs of haplotypes




**output**: genotype call

# Human Genotyping Results

| | | Platinum (Illumina) Genotype | | |
|---|---|---|---|---|
| | | 0/0 | 0/1 | 1/1 |
| Nanopolish Genotype | 0/0 | 727598 | 1730 | 75 |
| | 0/1 | 3217 | 29096 | 914 |
| | 1/1 | 601 | 49 | 21718 |

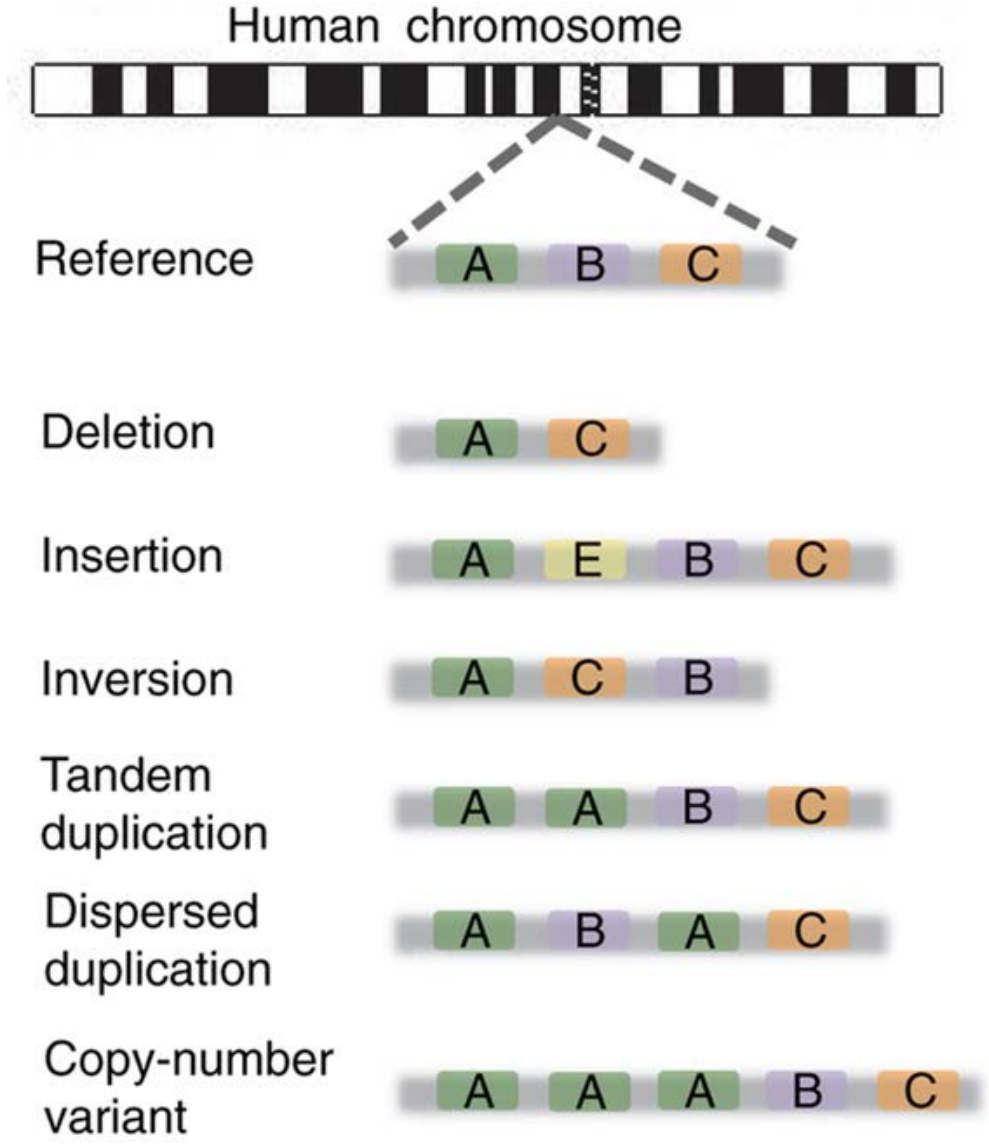
Genotype accuracy at all sites: 99.2%
Genotype accuracy at variable sites: 94.8%



Jain et al *Nature Biotech* 2018

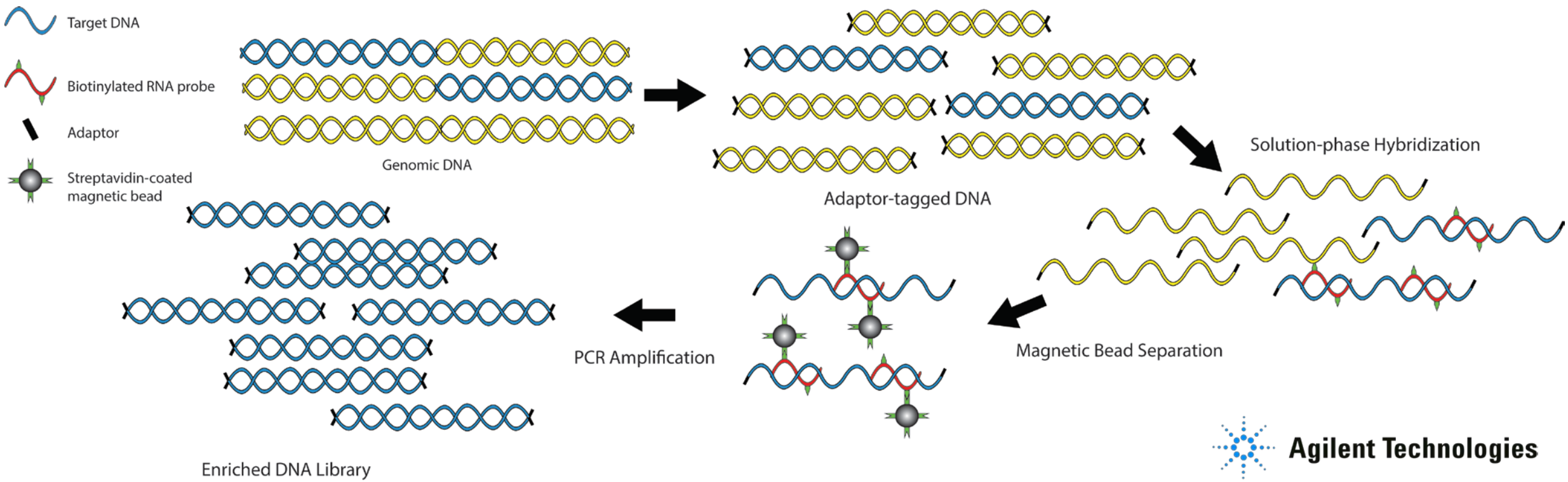
# Structural Variation

- Defined as an abnormality in large region (50b-3mb) of a chromosome
- Pervasive in cancer – 50% of pancreatic ductal adenocarcinoma (PDAC) have SVs
- Common in tumor suppressor genes such as *CDKN2A* and *SMAD4*
- Short-read sequencing has difficulty resolving SVs, but nanopore sequencing long reads can stretch across them.
- High coverage needed to detect, but yield from nanopore sequencing still relatively low per flowcell (~3-5Gb)




Baker, Monya. 2012. Nature Methods

# Solution-phase Hybridization Capture




**Use of Agilent SureSelectXT Targeted Sequencing System in cancer research**

- ~90 bps biotinylated RNA probes complementary to target sequence
- Biotin-streptavidin interaction to enrich for the targeted region
- Optimization for long-reads : > 2 kb

Agilent Technologies

# Targeted Capture Optimization

- Trial 1
  - Probe tiling, No empty spaces between probes
  - Target region
    - CDKN2A : 1.5 Mbp
  - Low stringency to allow mismatches
  - Result: 2.28 % on-target
- Trial 2
  - No tiling, average 400 bp space between probes
  - Target regions
    - CDKN2A : 1.5 Mbps
    - SMAD4 : 850 Kbps
  - High stringency to limit off-target capture
  - Consideration of known SV breakpoints
  - PDAC SVs from James Eshleman lab
  - Result: 30 % on-target

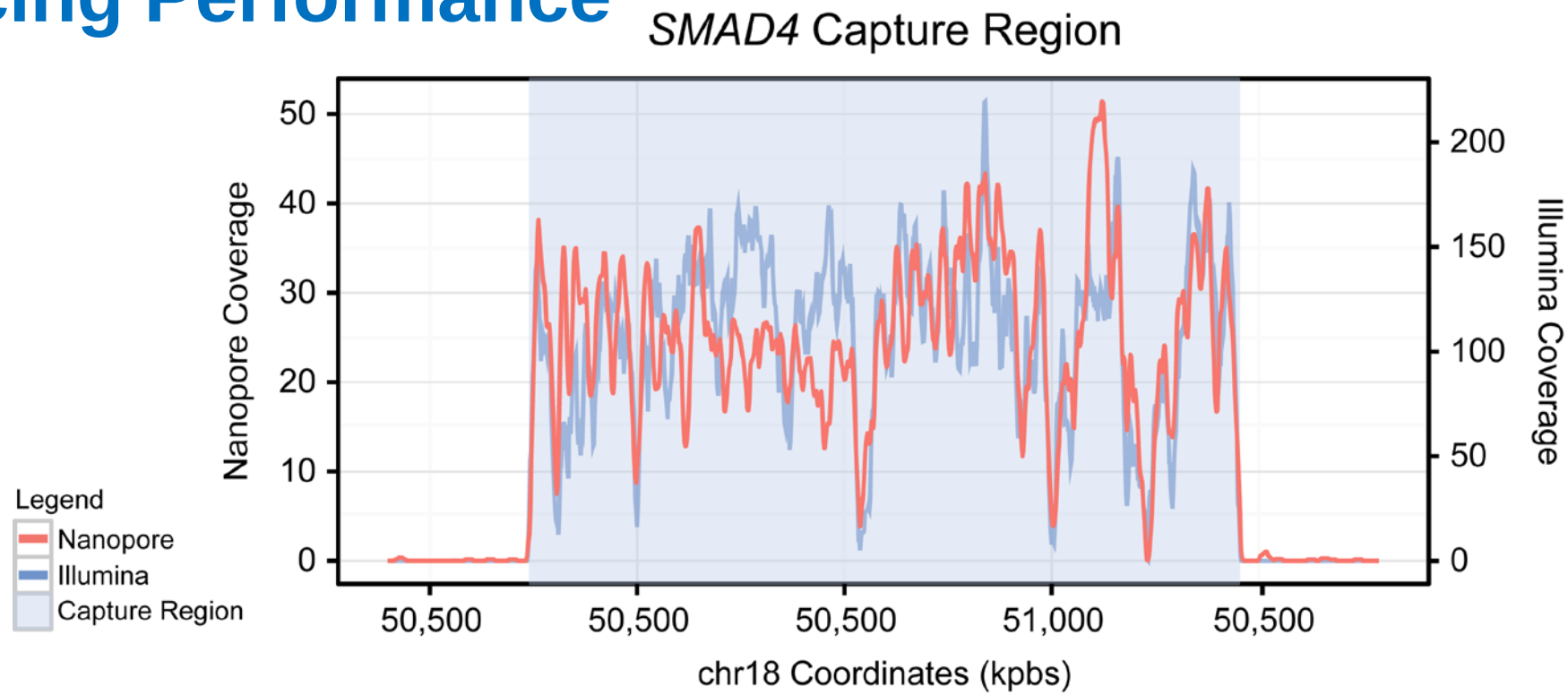
Agilent Technologies



Collaboration with Josh Wang from Agilent

# Targeted Sequencing Performance

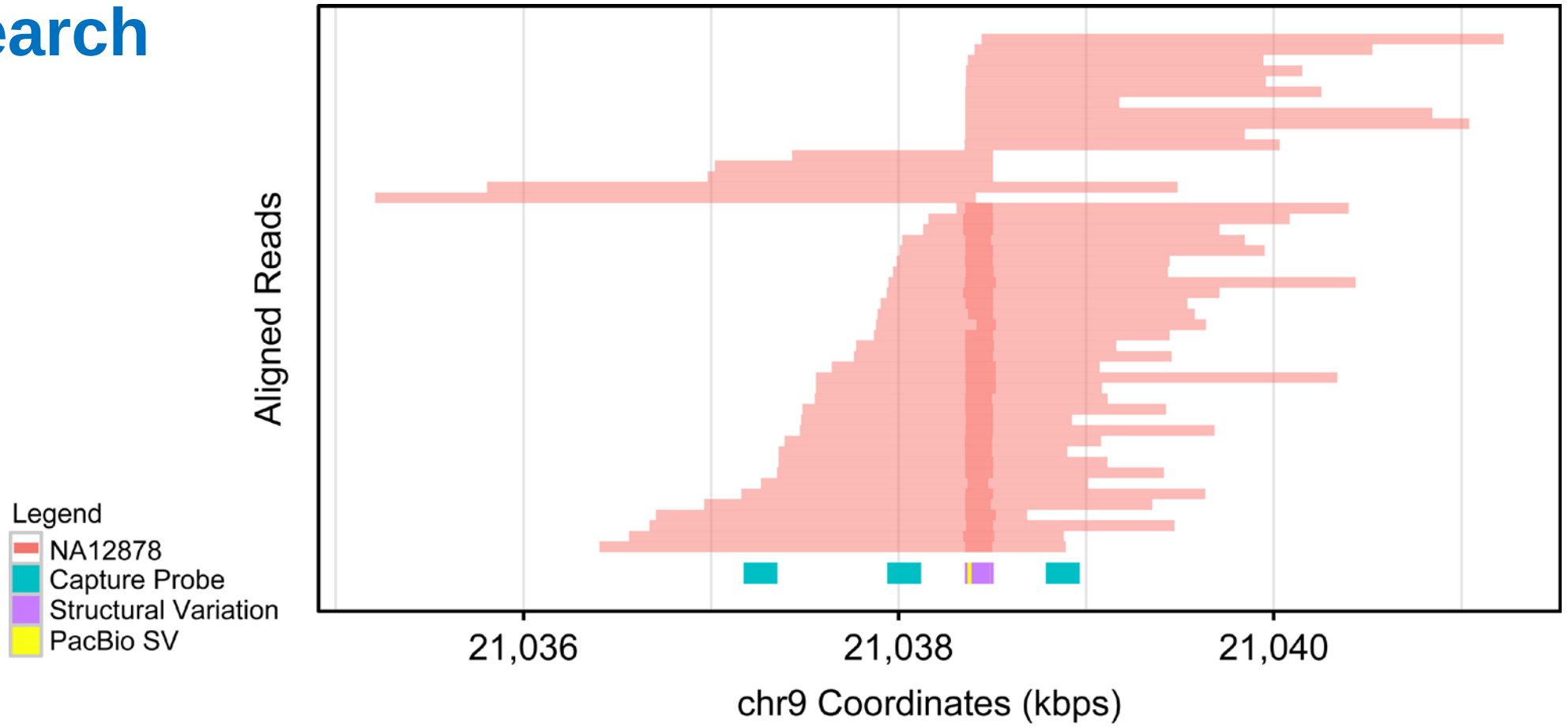
- Control : NA12878 lymphoblast
- Sample : PDAC from Eshleman lab
- Illumina short-read targeted sequencing for comparison
- > 300-fold enrichment
- > 20X average coverage
- Agilent App Note: https://goo.gl/8V2Fei




| | Total yield (reads) | On-target | On-target percentage | Fold enrichment | Coverage |
|---|---|---|---|---|---|
| Illumina NA12878 | 4.4m | 3.7m | 85% | 641X | 113X |
| Nanopore NA12878 | 107k | 32k | 30% | 353X | 27X |
| Nanopore PDAC | 56k | 20k | 26% | 332X | 20X |

Agilent Technologies

# Nanopore Structural Variation Detection in Cancer Research

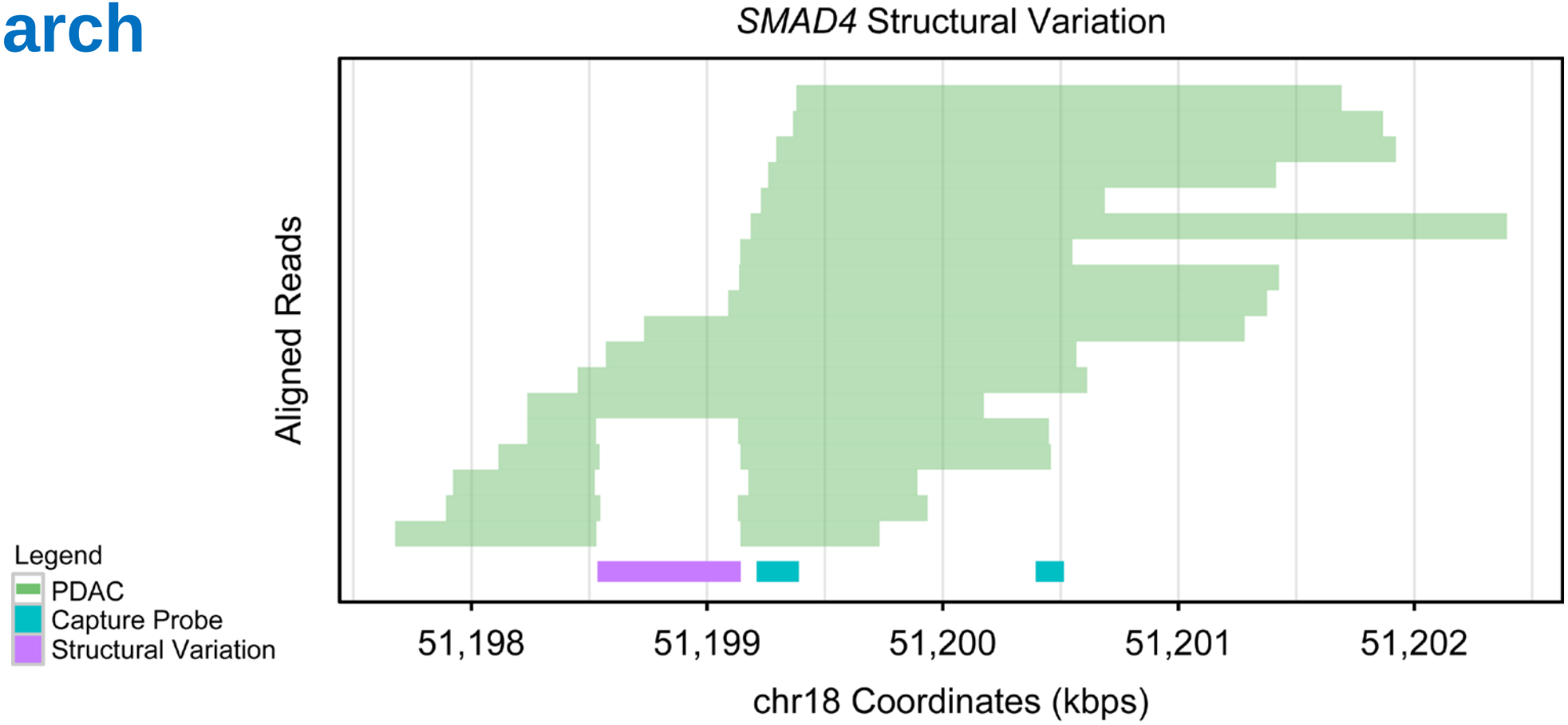



- NA12878 (ENCODE Human lymphoblast cell line)
- SVs detected with Sniffles (Schatz lab)
- chr9:21,038,354 - 21,038,506; 152 bps duplication
- Validated with PacBio data from Genome in a Bottle (Mt. Sinai School of Medicine)



w/Josh Wang (Agilent) & Jim Eshleman (JHMI)

# Nanopore Structural Variation Detection in Cancer Research

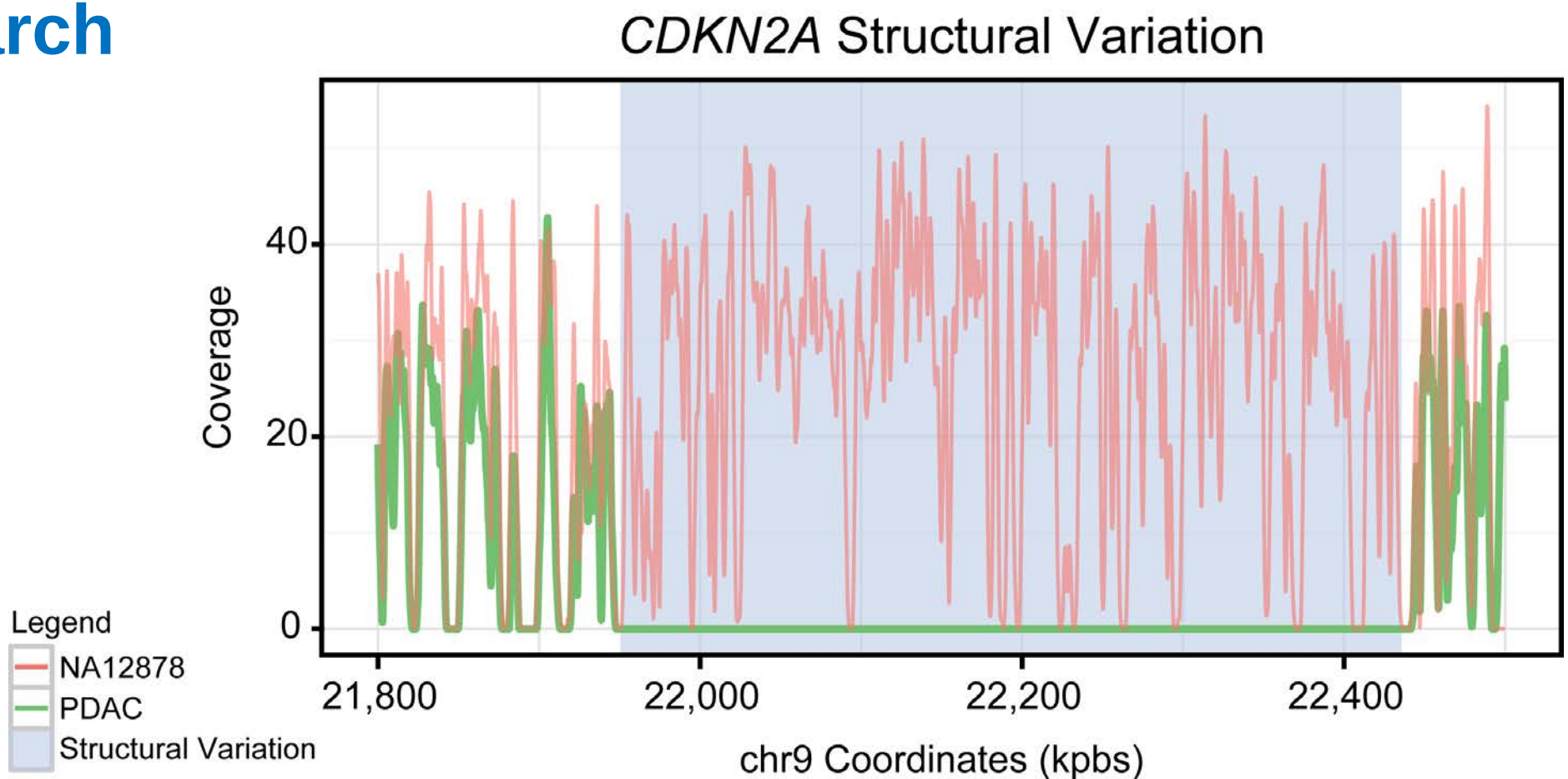



- PDAC cell line (Eshleman): Novel, putative SVs detected from PDAC
- chr18: 51,198,535 – 51,199,143; 600 bps deletion
- Possibly allele-specific SV



w/Josh Wang (Agilent) & Jim Eshleman (JHMI)

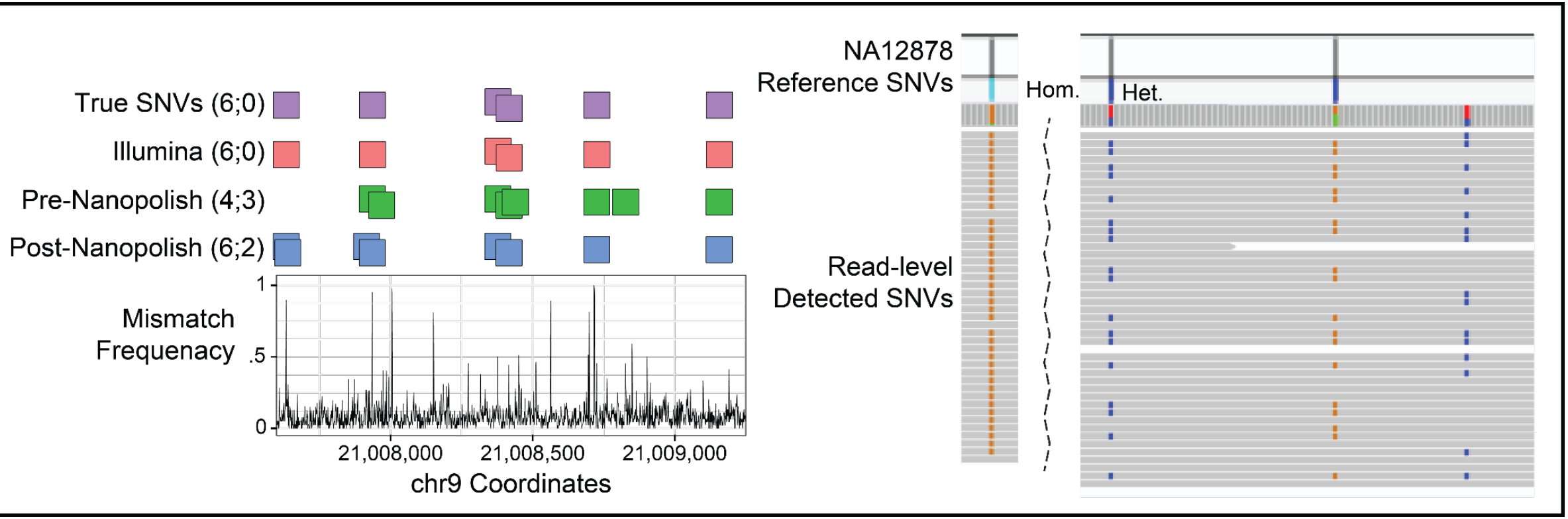
# Nanopore Structural Variation Detection in Cancer Research




- Large window of coverage; Absence in CDKN2A region
- chr9:21,950,000 -22,436,000; 486 kbps SV
- Homozygous Deletion previously identified with Illumina data
- Sniffles did not detect this SV; No reads covering either breakpoint, unlucky coincidence of probes

w/Josh Wang (Agilent) & Jim Eshleman (JHMI)

# Single Nucleotide Variation Detection in Cancer Research



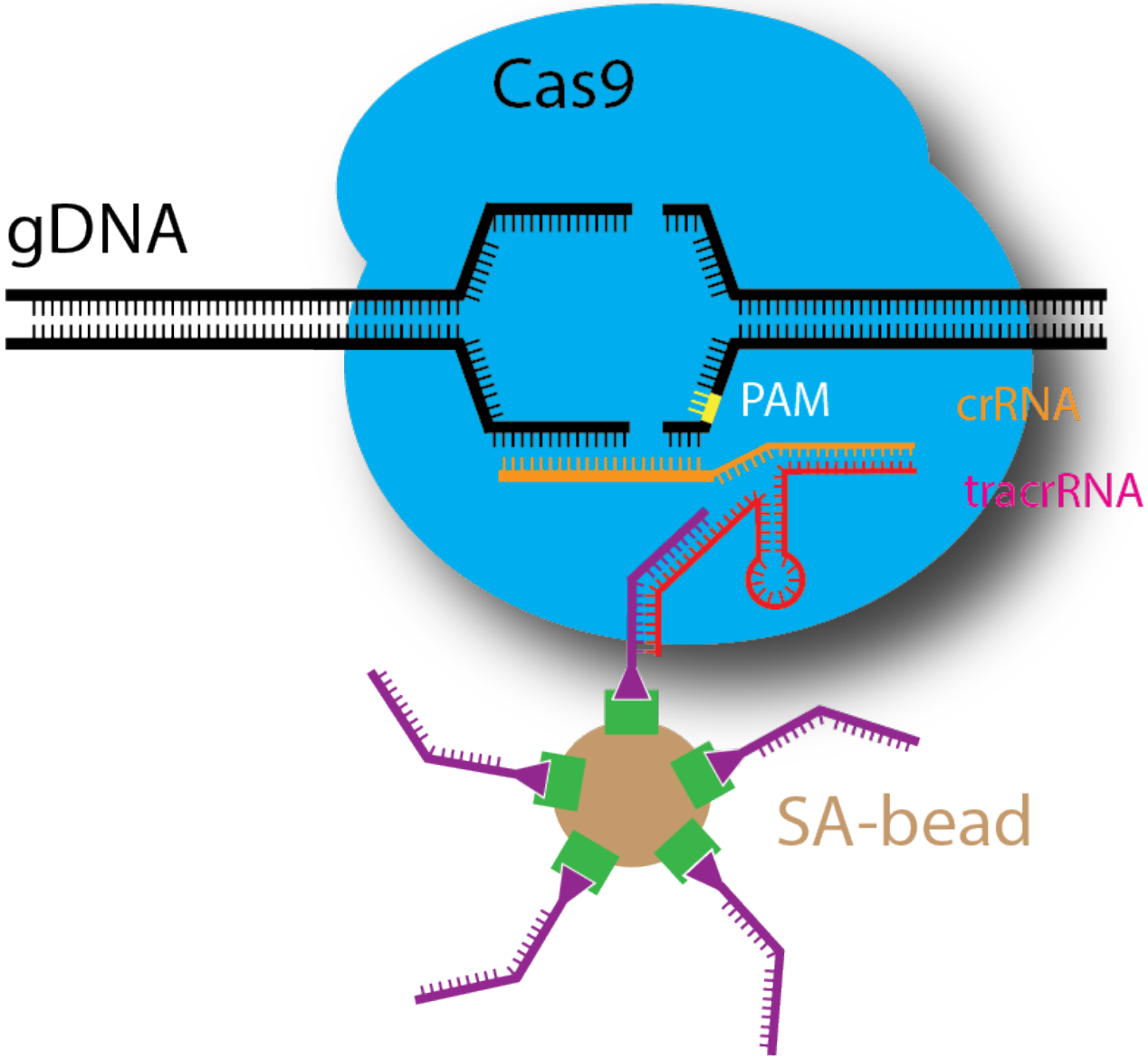

Phased SNV analysis is possible with coverage from targeted sequencing

| | Illumina | Pre-polish | Post-polish |
|---|---|---|---|
| Avg. Coverage | 113 | 27 | 27 |
| Correct | 1133 | 2485 | 947 |
| Total | 1211 | 4138 | 1017 |
| Precision | **94%** | 60% | **93%** |
| Sensitivity | **32%** | 69% | **26%** |

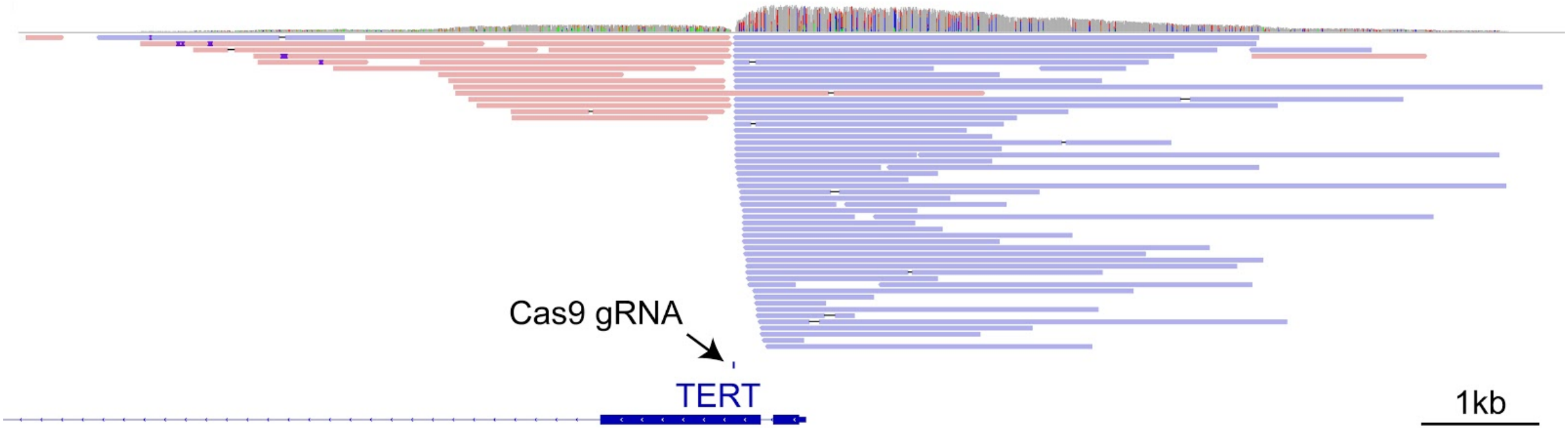
Number of True SNVs: 3587(Eberle,et al. bioRxiv, 2016)

# Cas9 Capture

- Instead of probe hybridization, we can use Cas9
- Cas9 holds on to DNA quite well even after cutting
- We are trying several different strategies
- These results are from the IDT Alt-R strategy we are working on with ONT

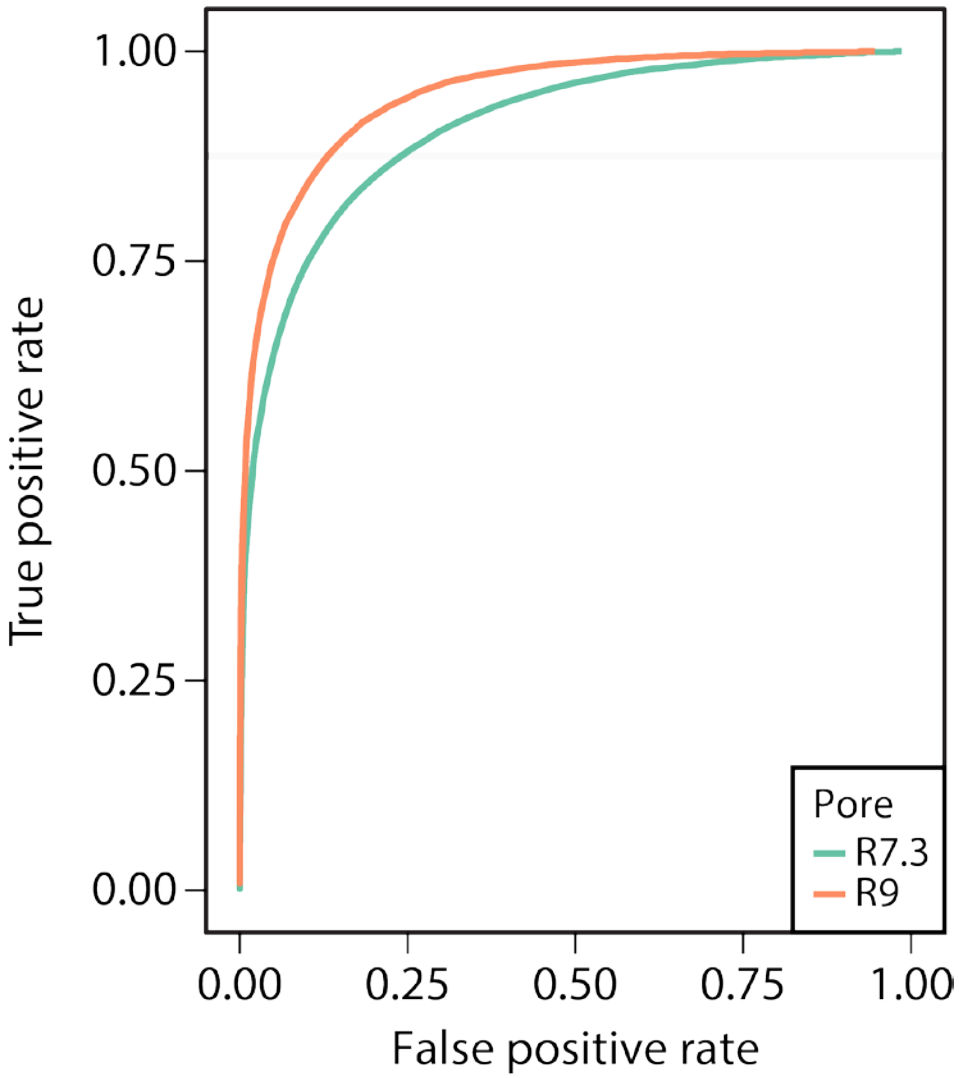


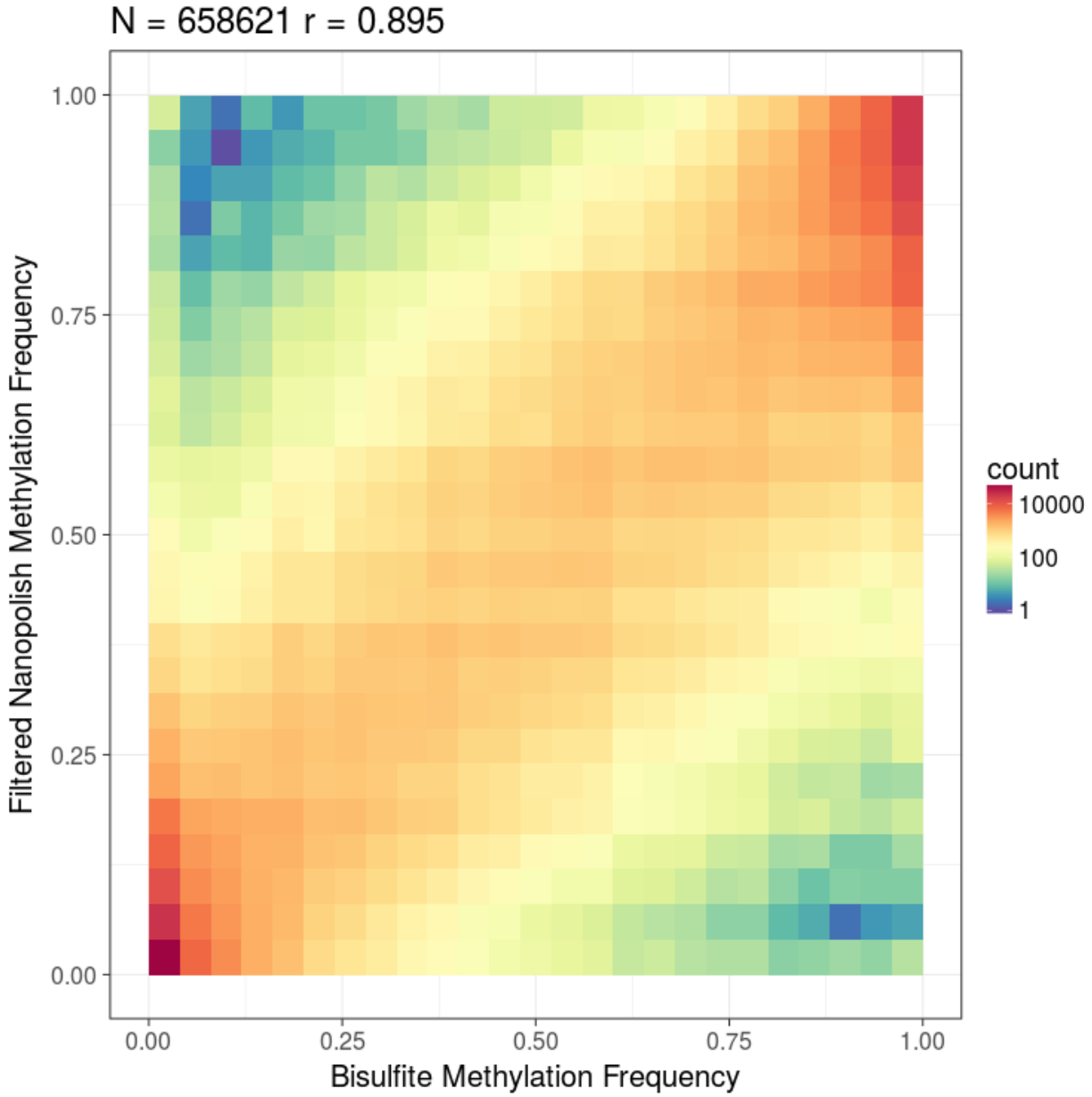
# Cas9 Enrichment around target




- Capture around the hTERT promoter, region with aberrant methylation in many cancers
- gDNA source from a BCPAP thyroid cancer cell line (poorly differentiated papillary thyroid carcinoma
- Hard to amplify with bisulfite PCR because of high CG-density, required many iterations of primer design

# Nanopolish Methylation



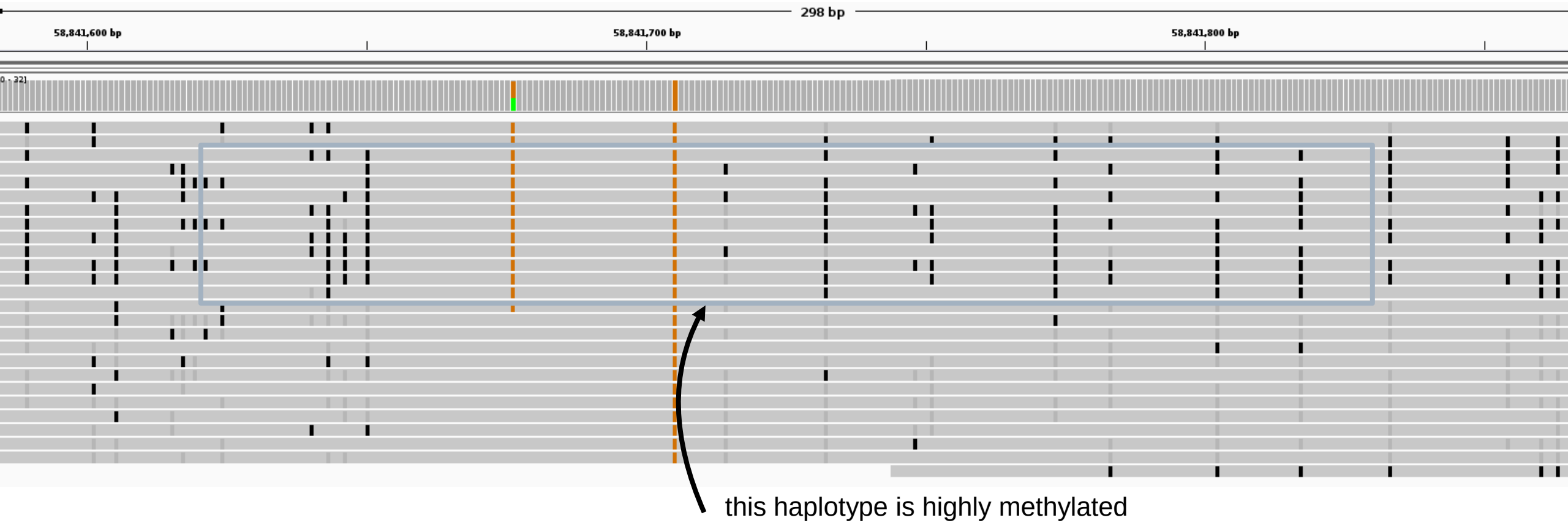




R9 calculates methylation 94% accurate at 77% of sites
NA12878 data shows .895 correlation with bisulfite

Jain et al *Nat Biotech* (2018)
Simpson et al *Nat Methods* (2017)

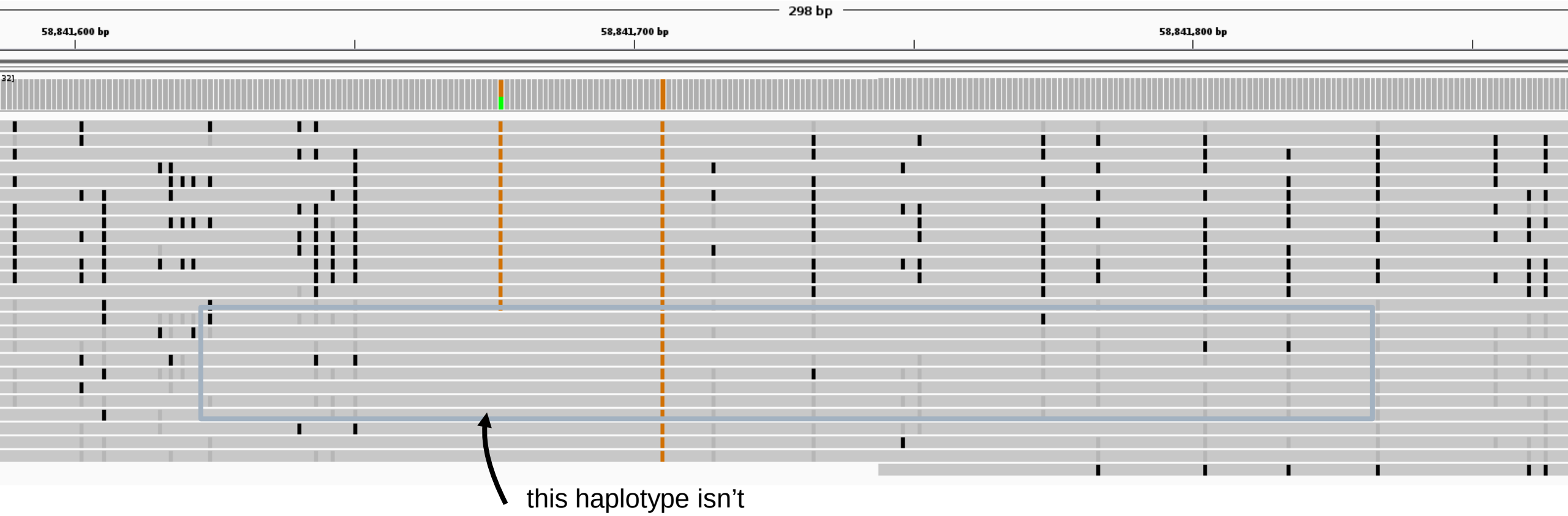
# Haplotype-Phased Methylation

nanopolish has experimental support for phasing methylation patterns



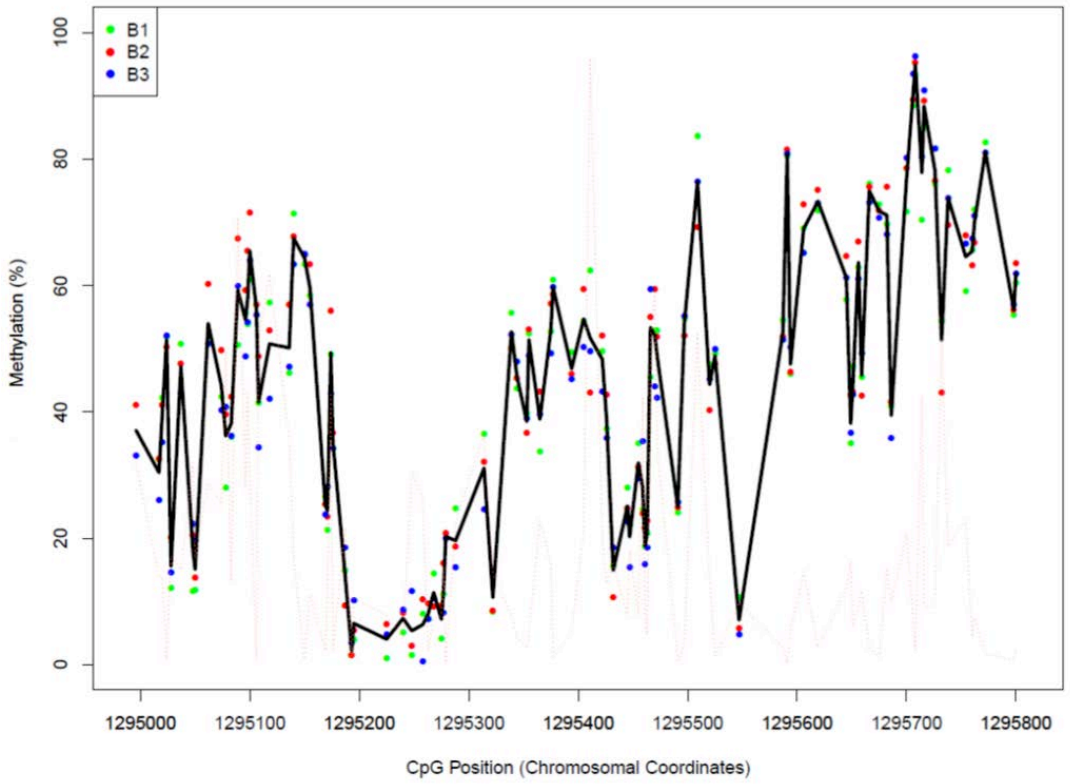
# Haplotype-Phased Methylation

nanopolish has experimental support for phasing methylation patterns

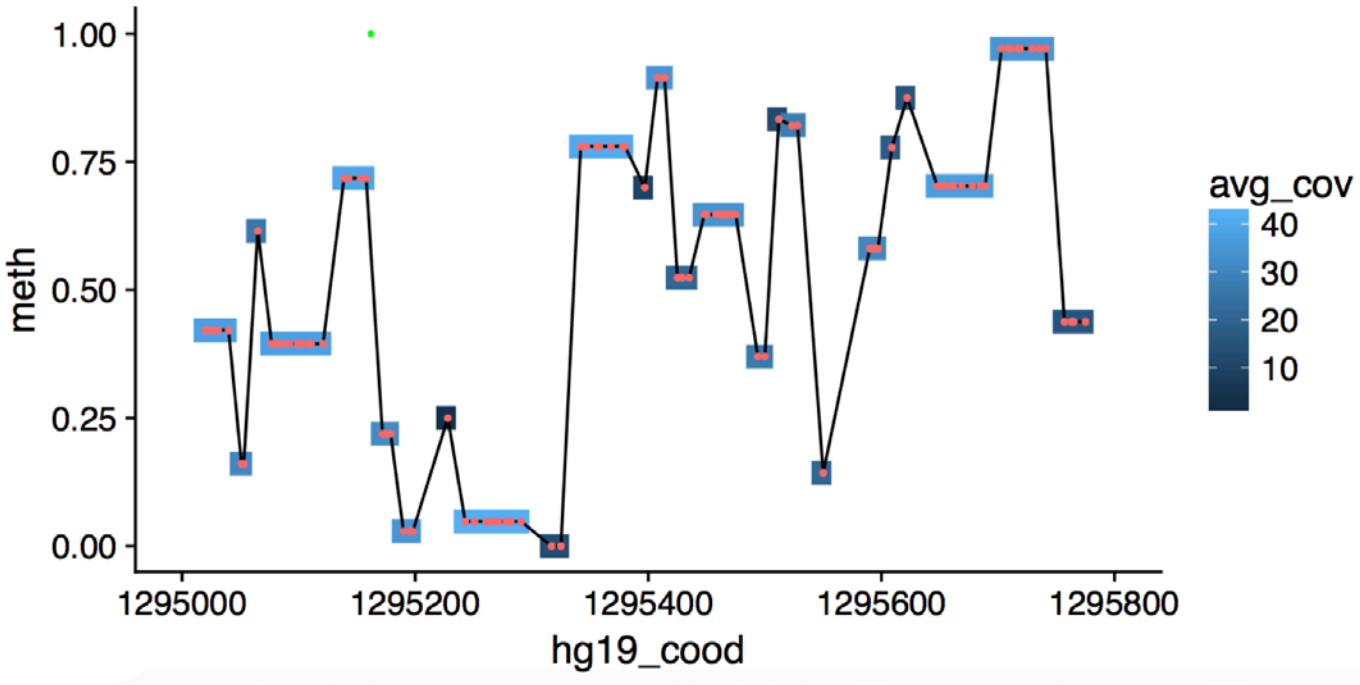


# Methylation compare of capture/bisulfite

illumina

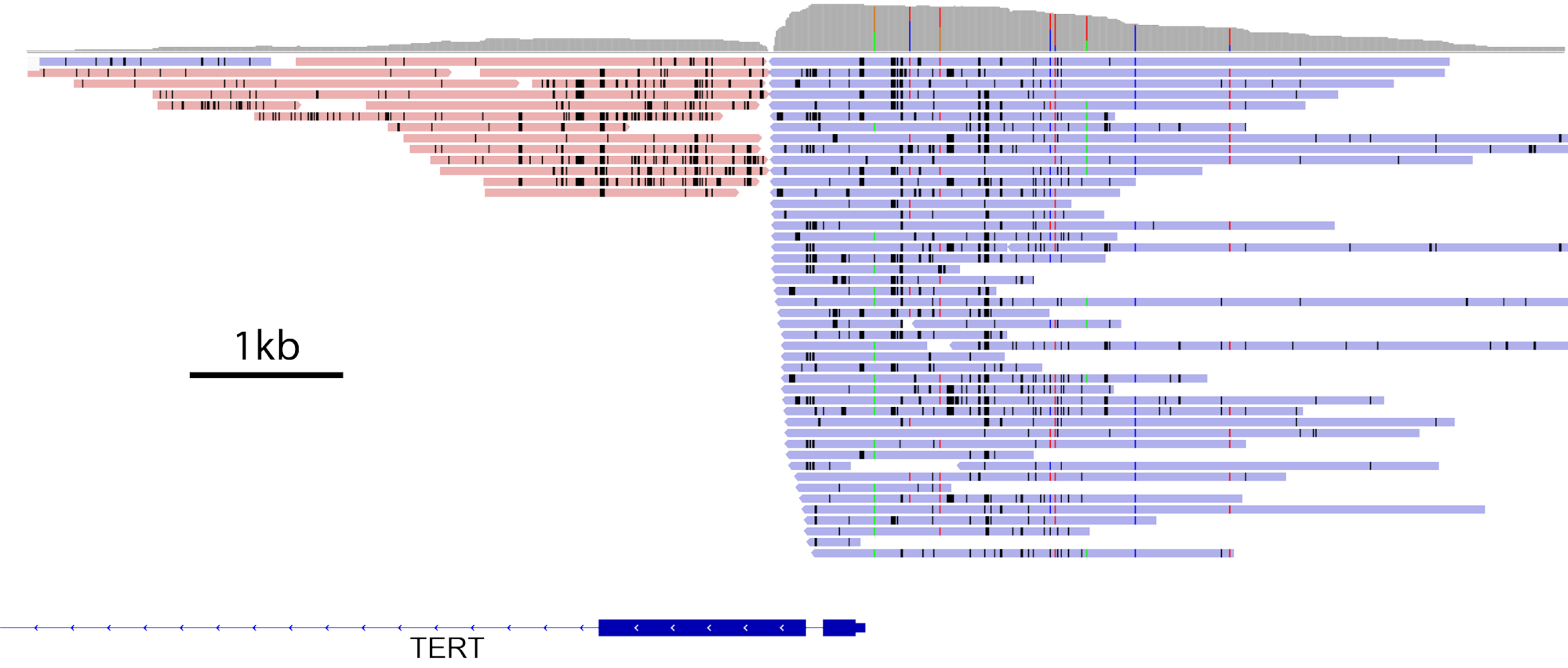



Nanopore



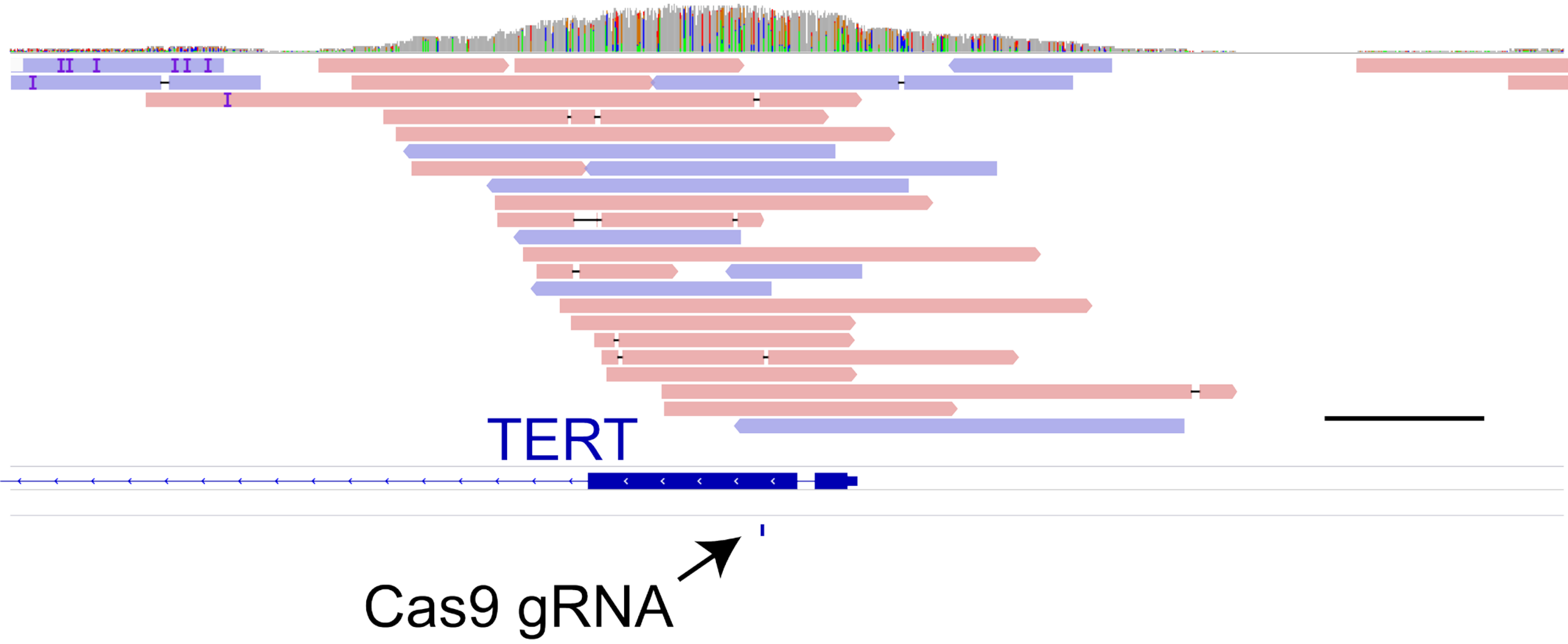

Preliminary data indicates methylation patterns largely concordant between bisulfite and nanopore

# Phased Methylation




- Some of these are C->T SNPs which are impossible to resolve with bisulfite sequencing

# Dead Cas9 Capture




- Capture with a dead Cas9 (same guide RNA target)
- Works, but needs more development

# Summary

- Nanopore technology is full of potential for sequencing, but always choose the right tool for the right job. Often multiple approaches with complementary data yield the best results.
- Multiple bases affect the electrical signal from nanopores; rather than a problem, this can be an advantage, as each base is interrogated multiple times. Using a hidden Markov model and Viterbi algorithm, we can decode the electrical signal to a DNA base sequence.
- Long-read sequencing is great at detection of structural variants in cancer samples used in clinical research – when coupled with hybridization capture this can be a powerful approach.
- Modifications to the primary DNA sequence (e.g. cytosine methylation) can be detected directly using nanopores, as compared to the gold standard of chemical modification (bisulfite treatment).
- Using PCR-less enrichment with Cas9 – we can also detect methylation on long reads with nanopore.

# Acknowledgments


- **Timp Lab**
- Winston Timp, PhD
- Rachael Workman, MS
- Norah Sadowski
- Timothy Gilpatrick
- Yunfan Fan
- Isac Lee

- **Eshleman Lab**
- James Eshleman, MD, PhD
- Alexis Norris, PhD

- **Umbricht Lab**
- Brittany Avin

**Agilent Technologies**

**Agilent Technologies**
Josh Wang, PhD
Jonathan Levine, PhD

National Human Genome Research Institute
1R01HG009190-01A1

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


Looking for Postdocs!!