AACR-ONT Workshop

Washington, DC; April 2017

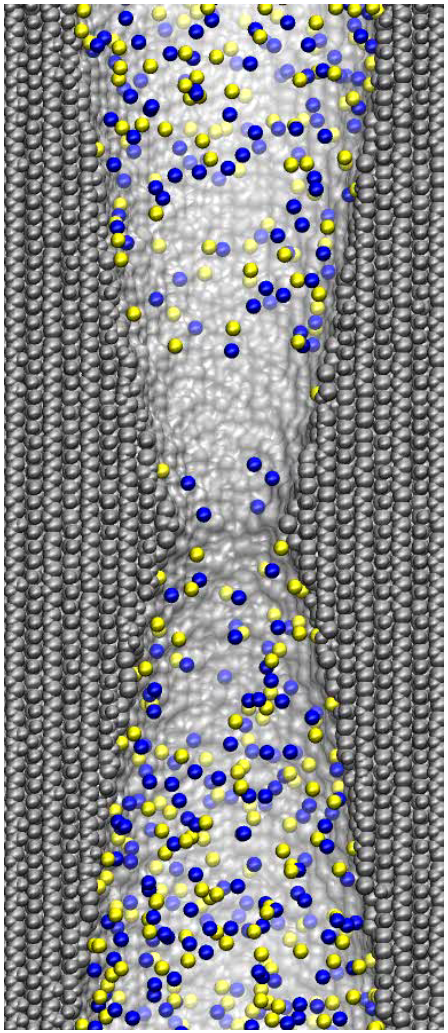
JOHNS HOPKINS
BIOMEDICAL ENGINEERING

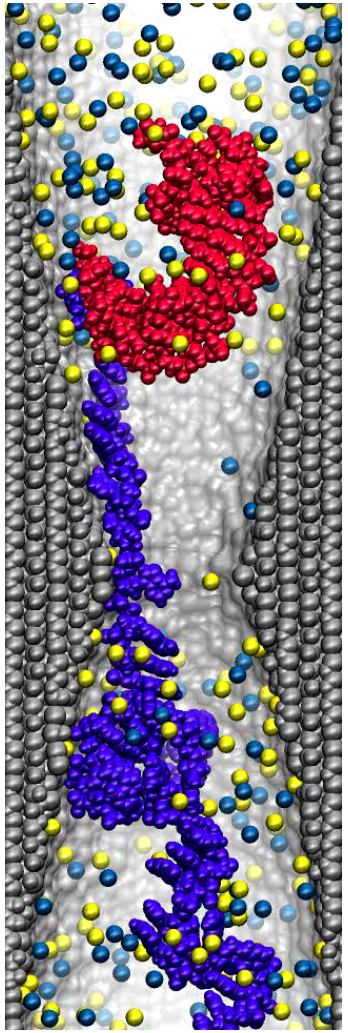
Nanopore Sequencing: Methylation and Structural Variant Calling

Winston Timp
Department of Biomedical Engineering
Johns Hopkins University

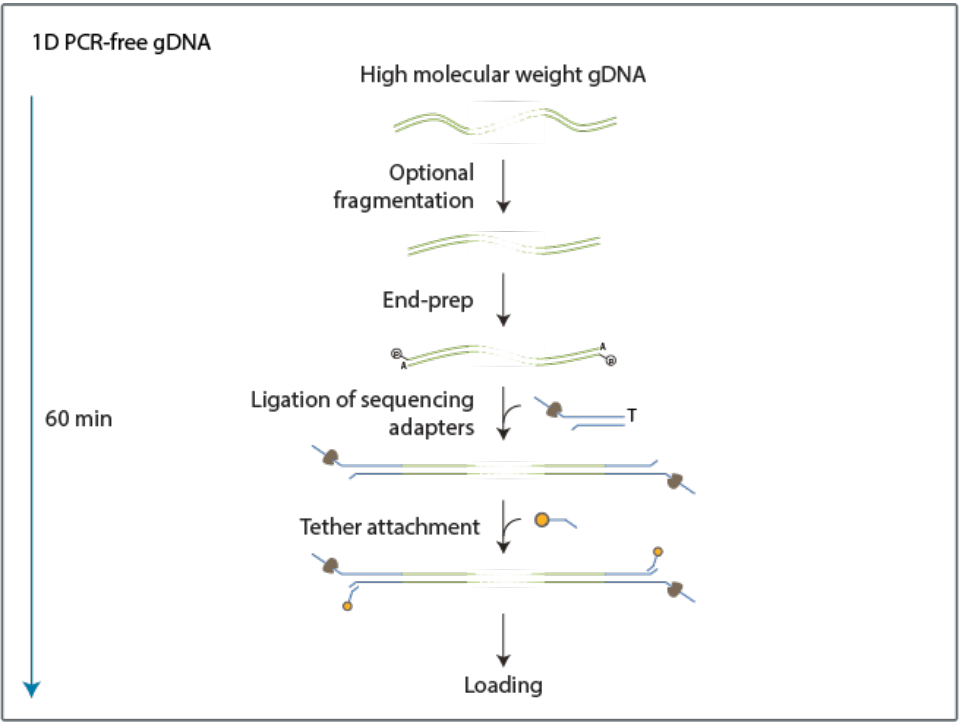
Nanopore: Current Blockage

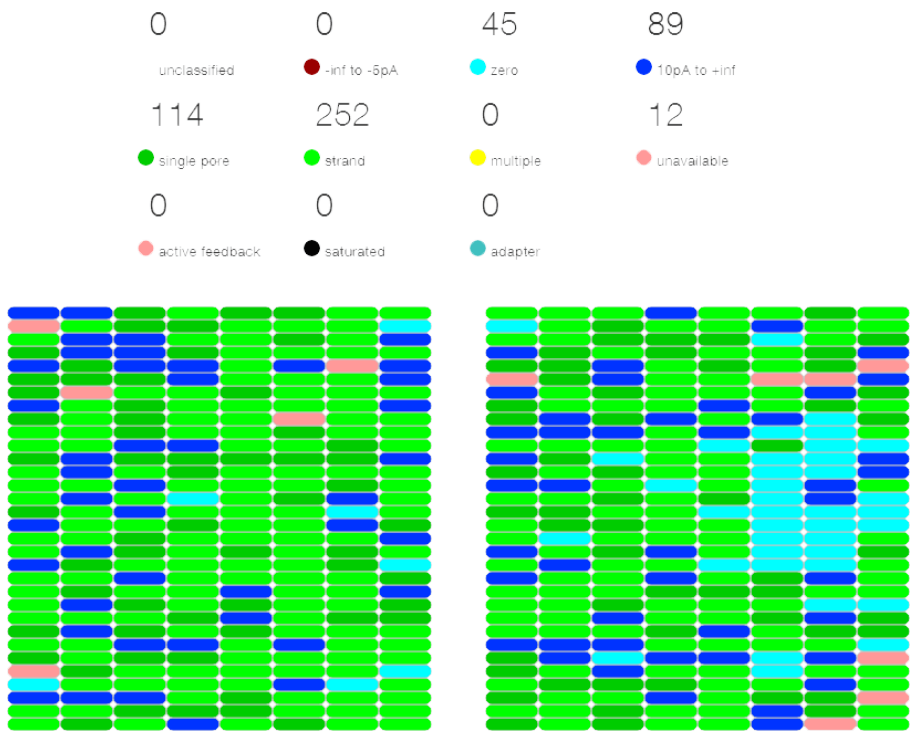
- Basic concept is similar to Coulter counter – detect and characterize objects based on their blockage of a pore
- Current nanopore work is performed either with ion channel proteins (α-hemolysin, MspA) or fabricated in silicon nitride or graphene membranes.
- Most work focuses on distinguishing what is in the pore based on the duration and level of blockade events





Nanopore Library Prep

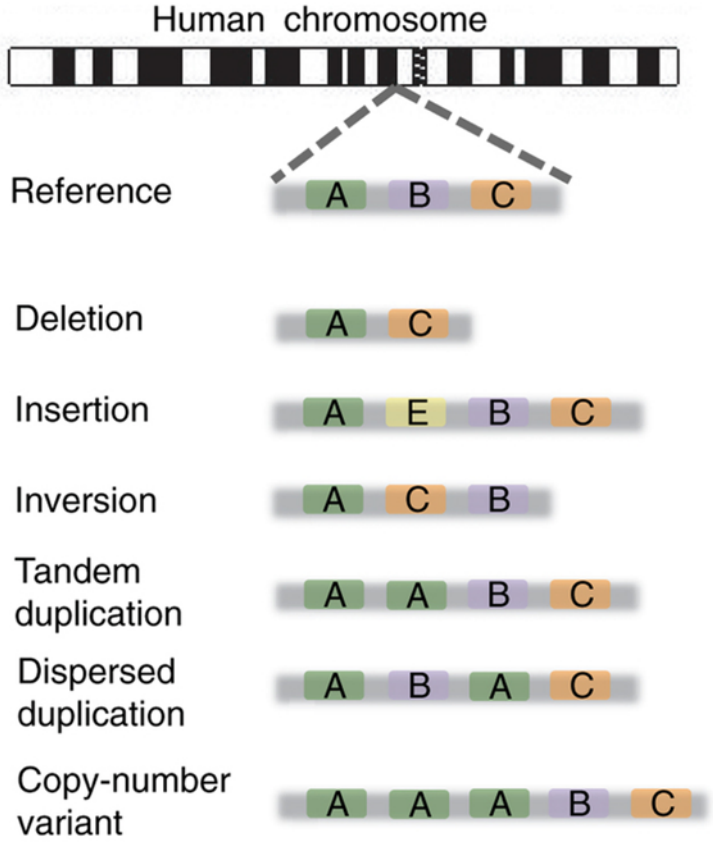




- Library prep is very similar to methods for short-read sequencing
- For DNA shearing we used Covaris gTubes
- After end-repair and A-tailing, leader adapter with motor protein is ligated
- MinION arrays 512 channels (with 4 pores possible per channel) (shown bottom left from running software); dark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

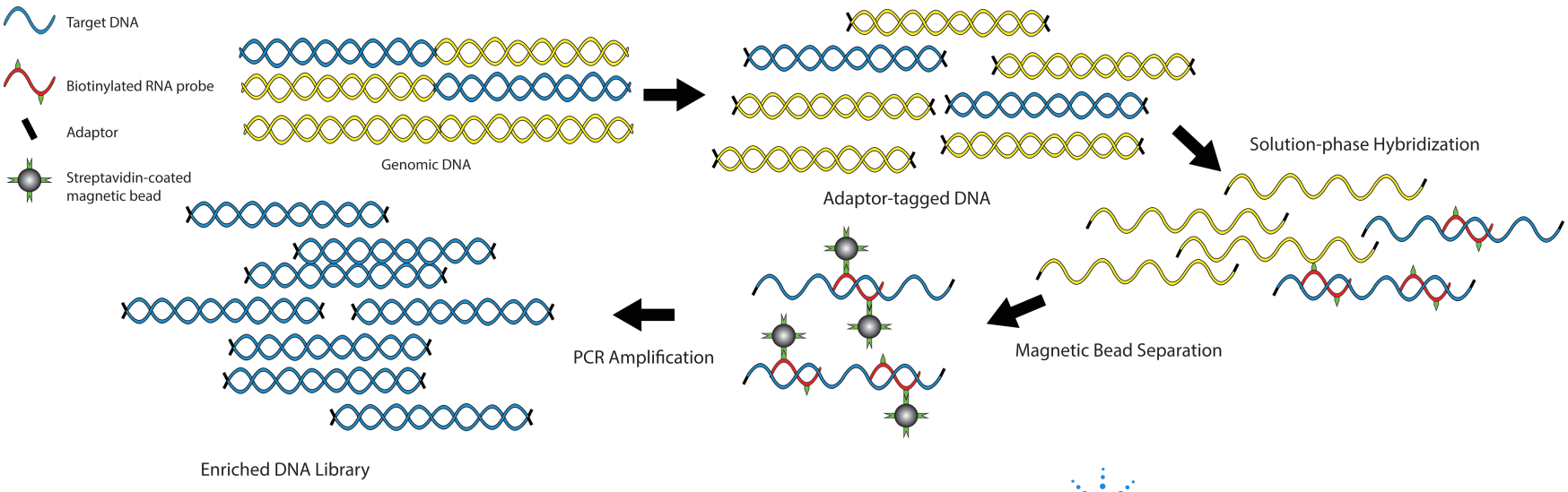
Structural Variation

- Abnormality in large region (50b-3mb) of a chromosome
- Pervasive in cancer – 50% of pancreatic ductal adenocarcinoma (PDAC)
- Common in tumor suppressor genes such as *CDKN2A* and *SMAD4*
- Nanopore sequencing can resolve SVs
  - *But* High coverage desired for heterogeneous samples
  - and size of human genome : 3 Gbps
- We know where SVs tend to occur



Baker, Monya. 2012. Nature Methods

Solution-phase Hybridization Capture



Agilent Technologies

Agilent SureSelectXT Targeted Sequencing System

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

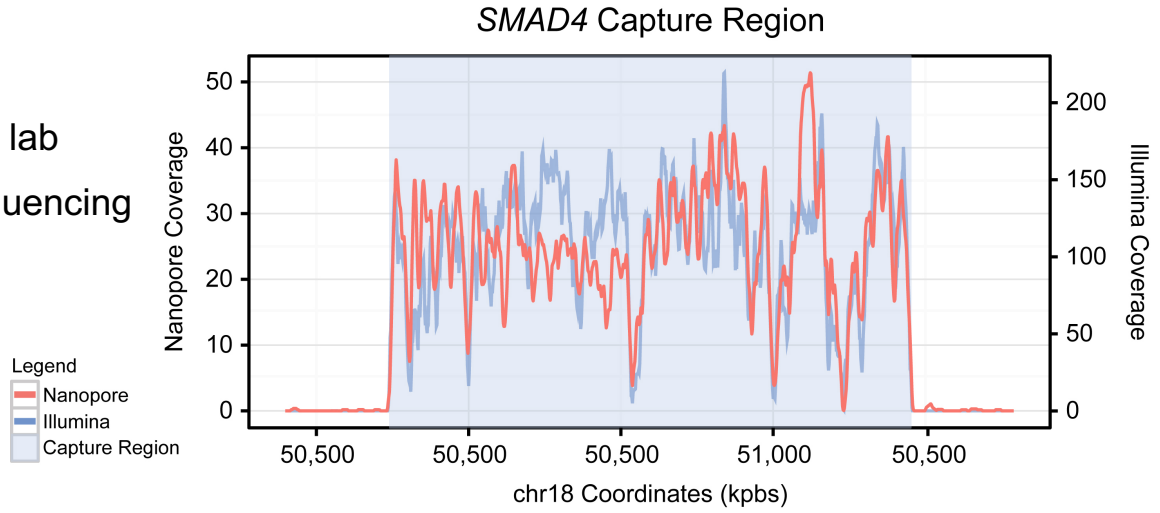
Targeted Capture Optimization

- 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

Collaboration with Josh Wang from Agilent

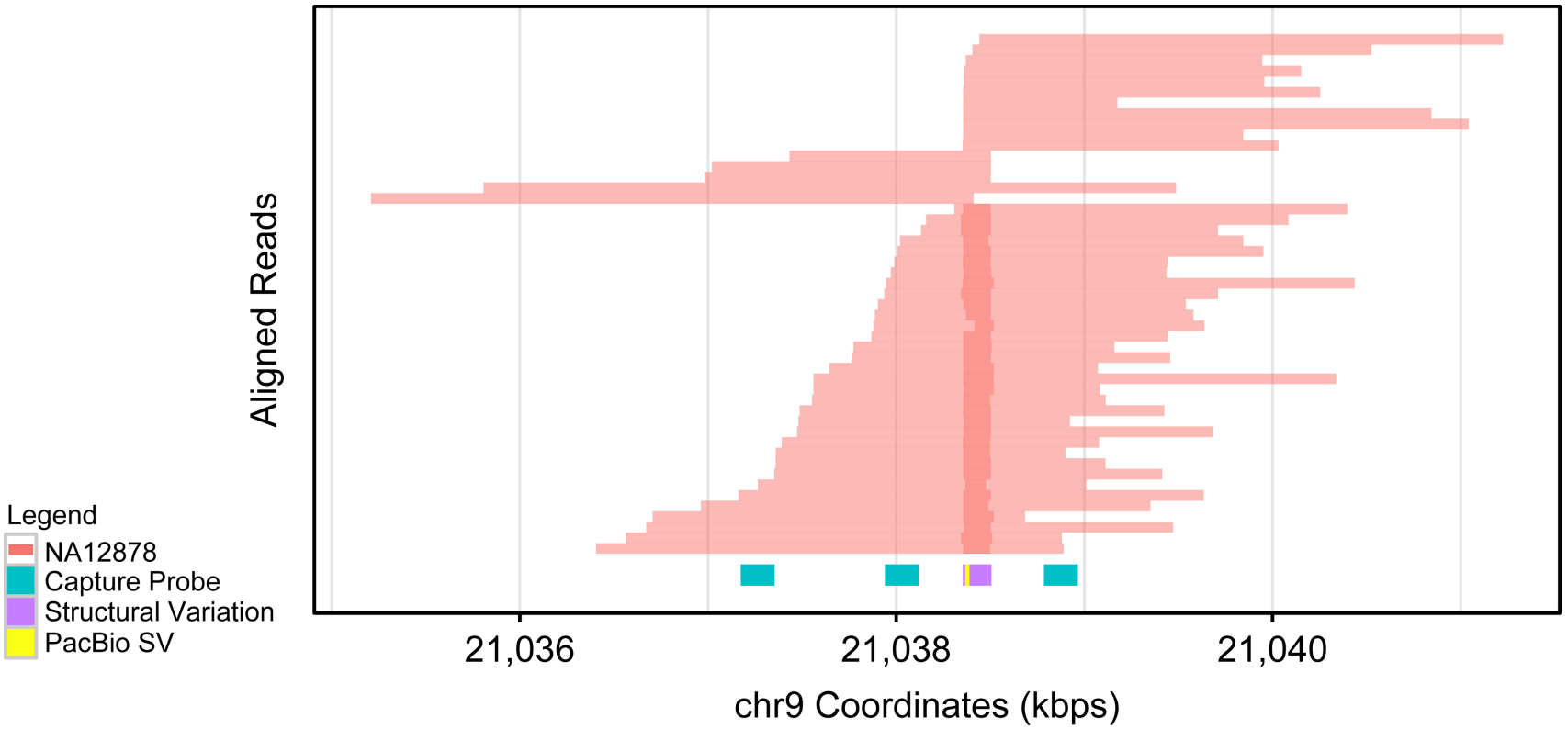
Targeted Sequencing Performance

- Control : NA12878 lymphoblast
- Sample : PDAC from Eshleman lab
- Illumina short-read targeted sequencing for comparison
- > 300-fold enrichment
- > 20X average coverage



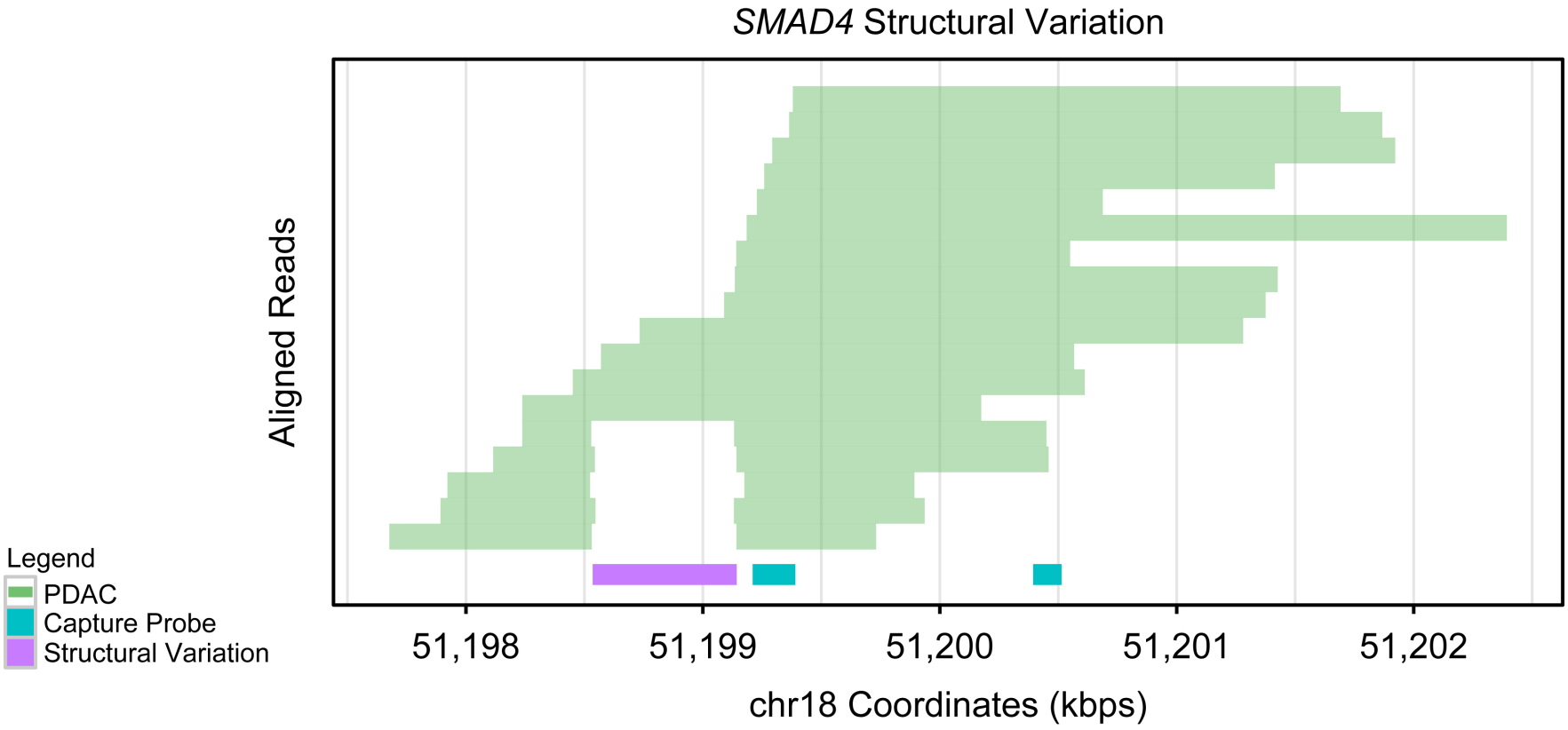
	Total yield (reads)	On-target	On-target percentage	Fold enrichment	Coverage
Illumina NA12878	4.4m	3.7m	85%	641X	113X
Nanopore NA12878	107k	32k	30%	353X	27X
Nanopore PDAC	56k	20k	26%	332X	20X

Nanopore Structural Variation Detection



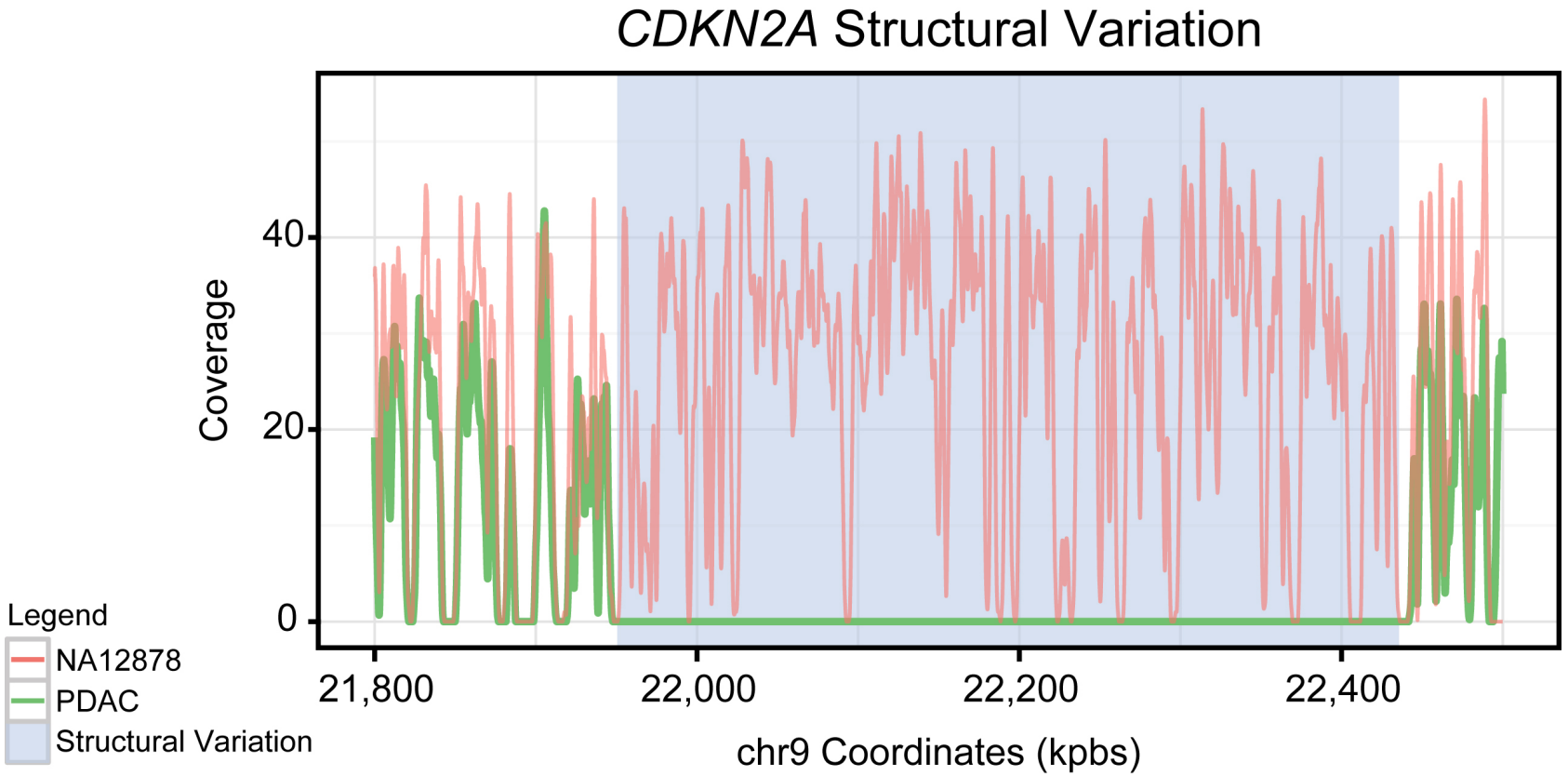
- NA12878 (ENCODE Human lymphoblast cell line)
- SVs detected with Sniffles (Schatz lab)
- chr9:21,038,354 - 21,038,506; 152 bps duplication
- Validated with PacBio data from Genome in a Bottle (Mt. Sinai School of Medicine)

Nanopore Structural Variation Detection



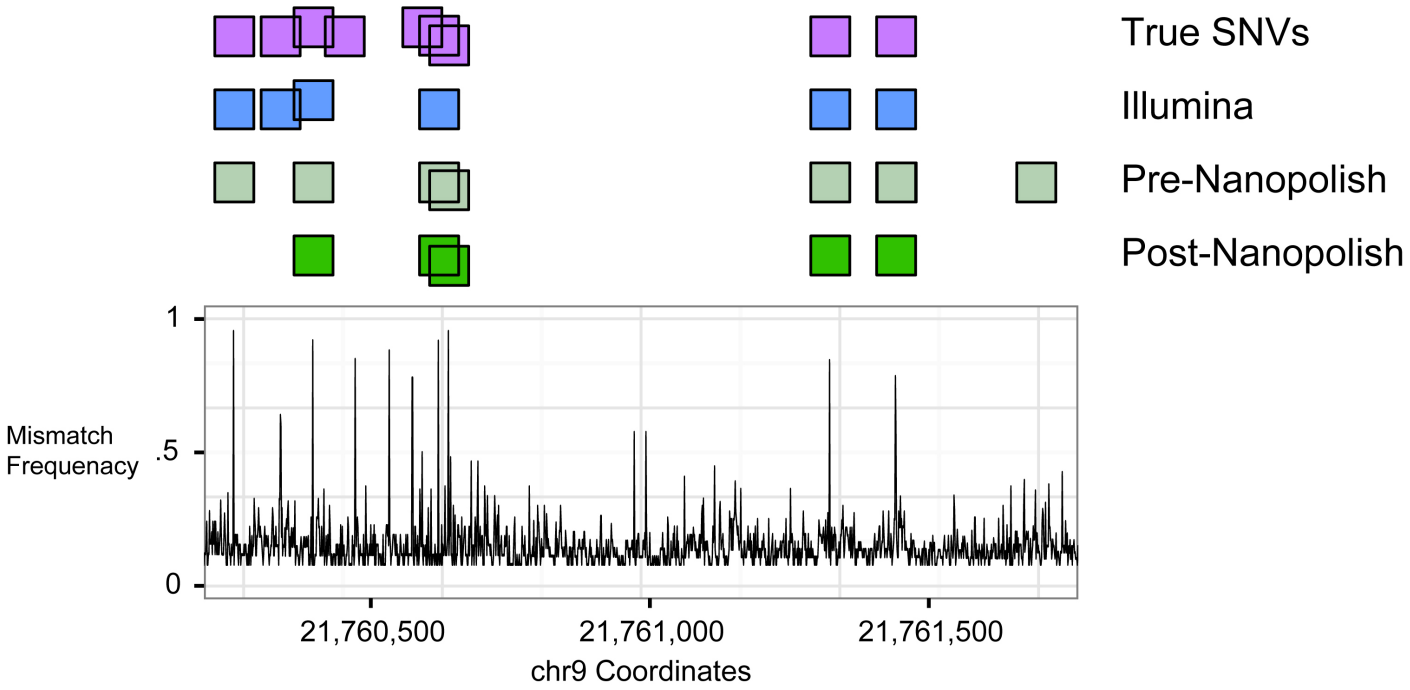
- PDAC cell line (Eshleman): Novel, putative SVs detected from PDAC
- chr18: 51,198,535 – 51,199,143; 600 bps deletion
- Possibly allele-specific SV

Nanopore Structural Variation Detection



- 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

Single Nucleotide Variation Detection

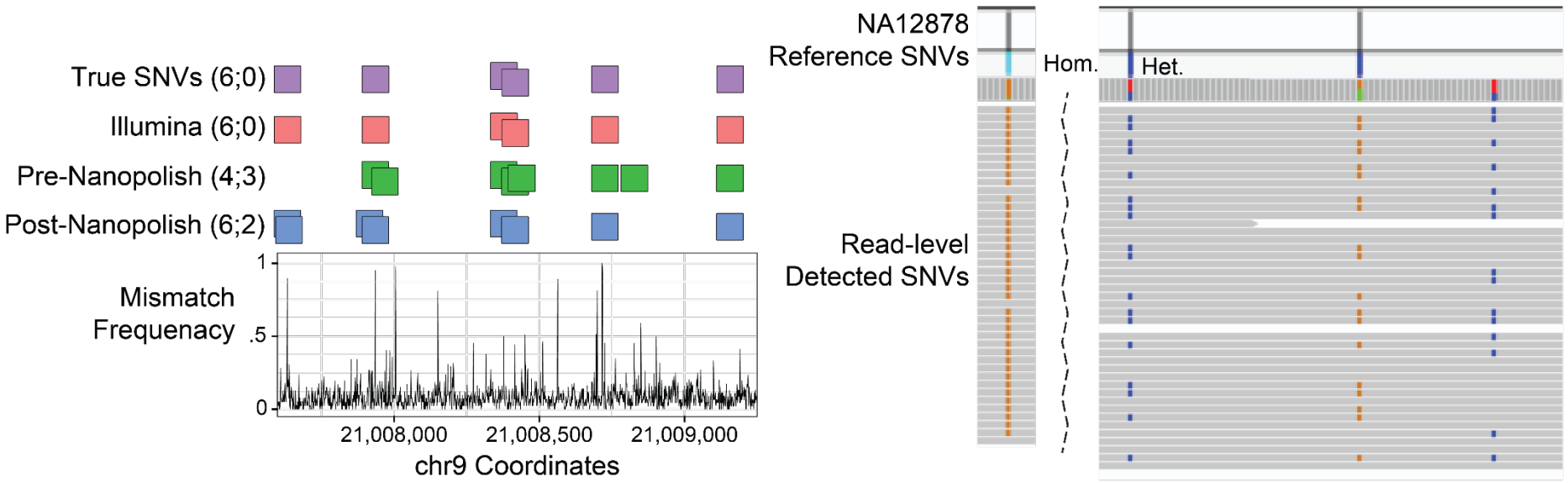


- Nanopolish (Simpson Lab) to improve SNV calling
- Nanopore SNV detection after nanopolish is comparable to Illumina
- Phased SNV analysis is possible with coverage from targeted sequencing

	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)

Single Nucleotide Variation Detection

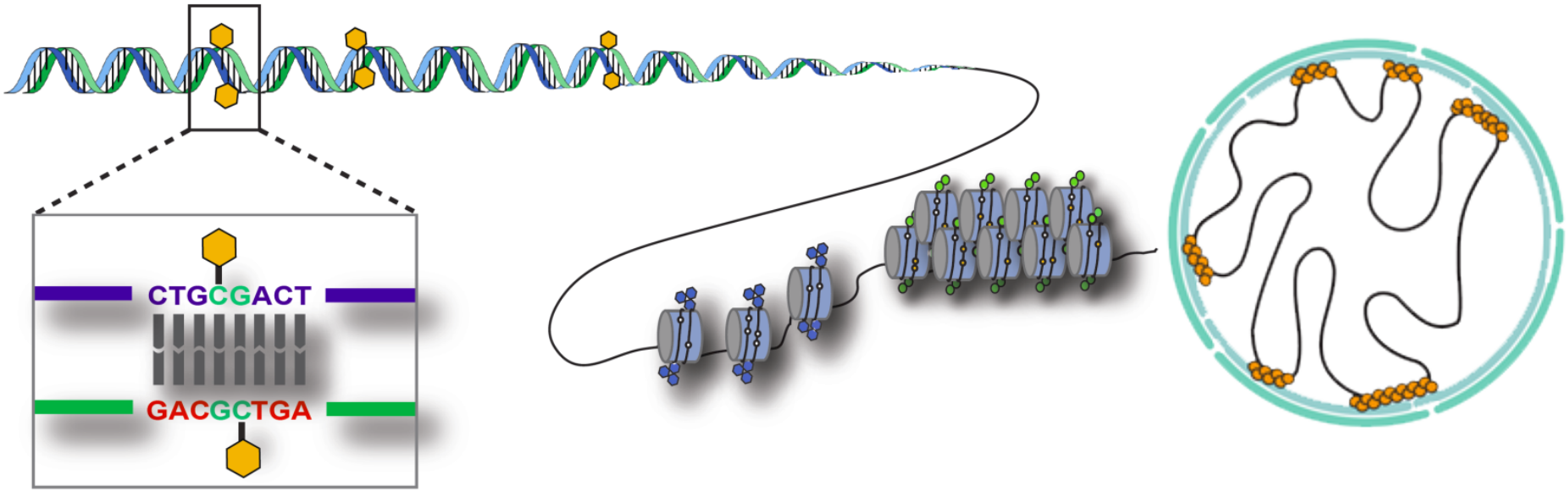


- Nanopolish (Simpson Lab) to improve SNV calling
- Nanopore SNV detection after nanopolish is comparable to Illumina
- Phased SNV analysis is possible with coverage from targeted sequencing

	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)

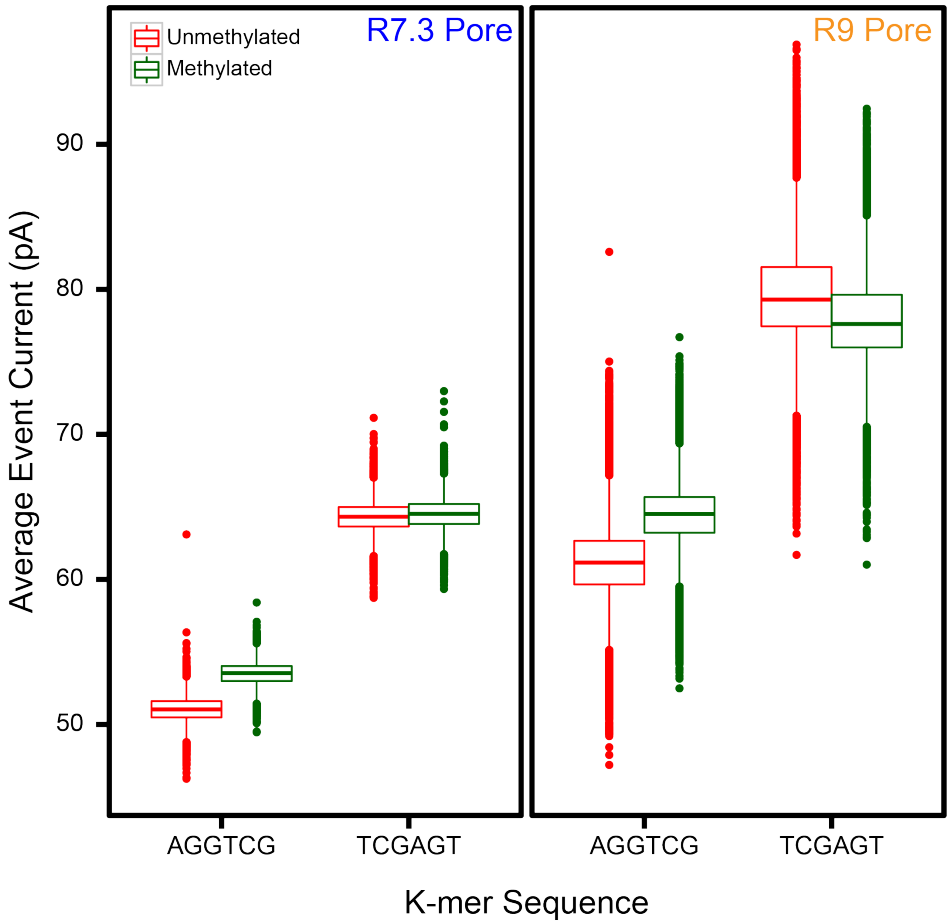
Epigenetics: Modern



- Modern Definition of epigenetics involves heritable changes other than genetic sequence, e.g., positive feedback, high order structure, chromatin organization, histone modifications, DNA methylation.
- An analogy to a computer system:
  - DNA Sequence = Hardware
  - User input = Environment
  - Systems Biology = Running programs
  - **Epigenetics = RAM**

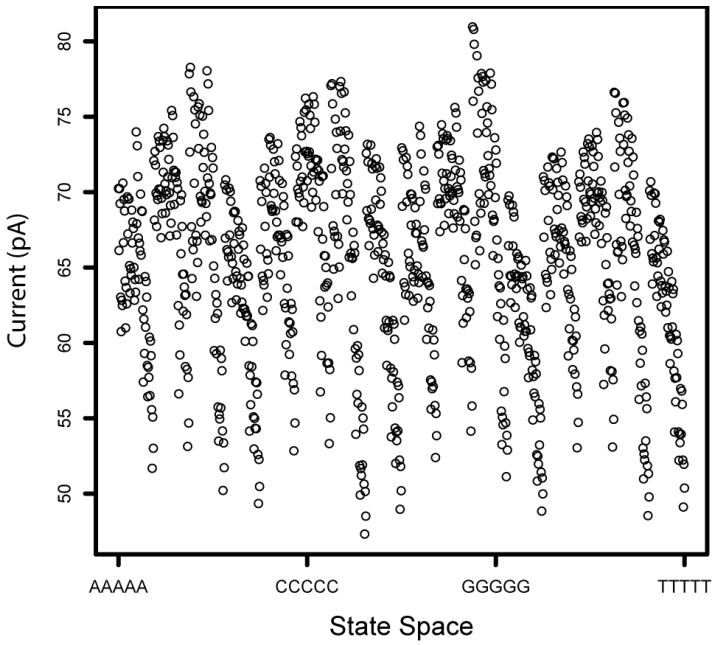
Emission Probabilities

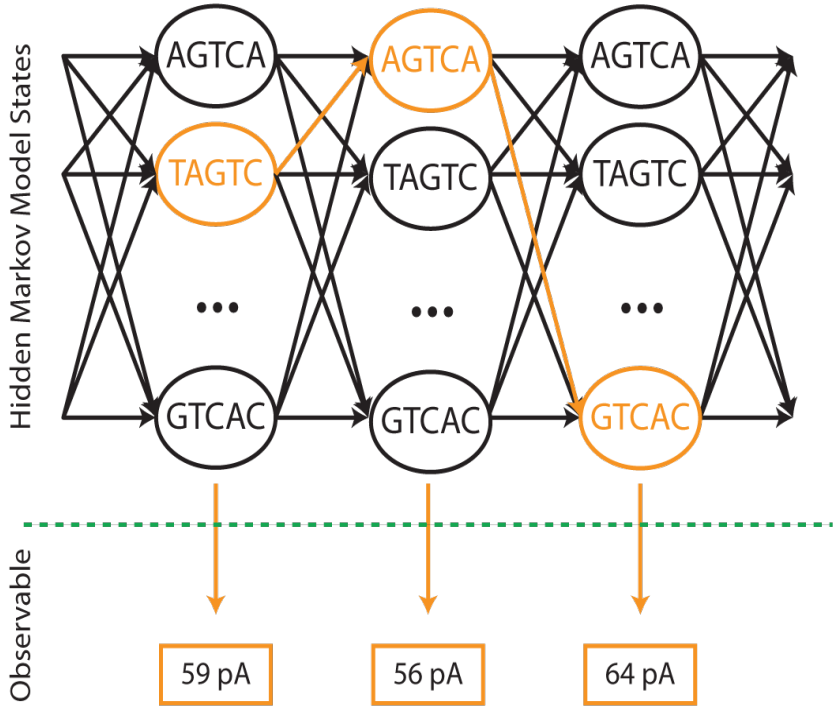
- We measured distributions of current for k-mers from *E. Coli* M.SssI treated (methylated; green) and untreated (unmethylated; red) samples on two different sets of pores - R7.3 and R9 flowcells.
- Boxplots of AGGTCG and TCGAGT k-mers which both contain CGs show significant differences in current in some cases (AGGTCG R7.3) and little to none in others (TCGAGT R7.3)
- R9 current distribution seem wider in both cases, but gives better discrimination in TCGAGT.



Simpson, Workman, et al. Nature Methods (2017)

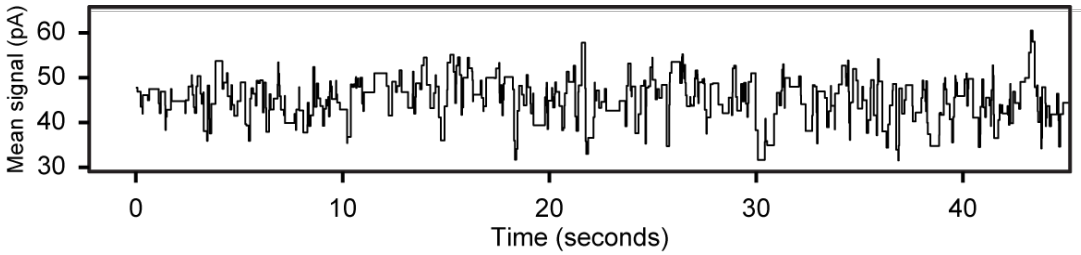
Nanopore: nanopolish methyltrain





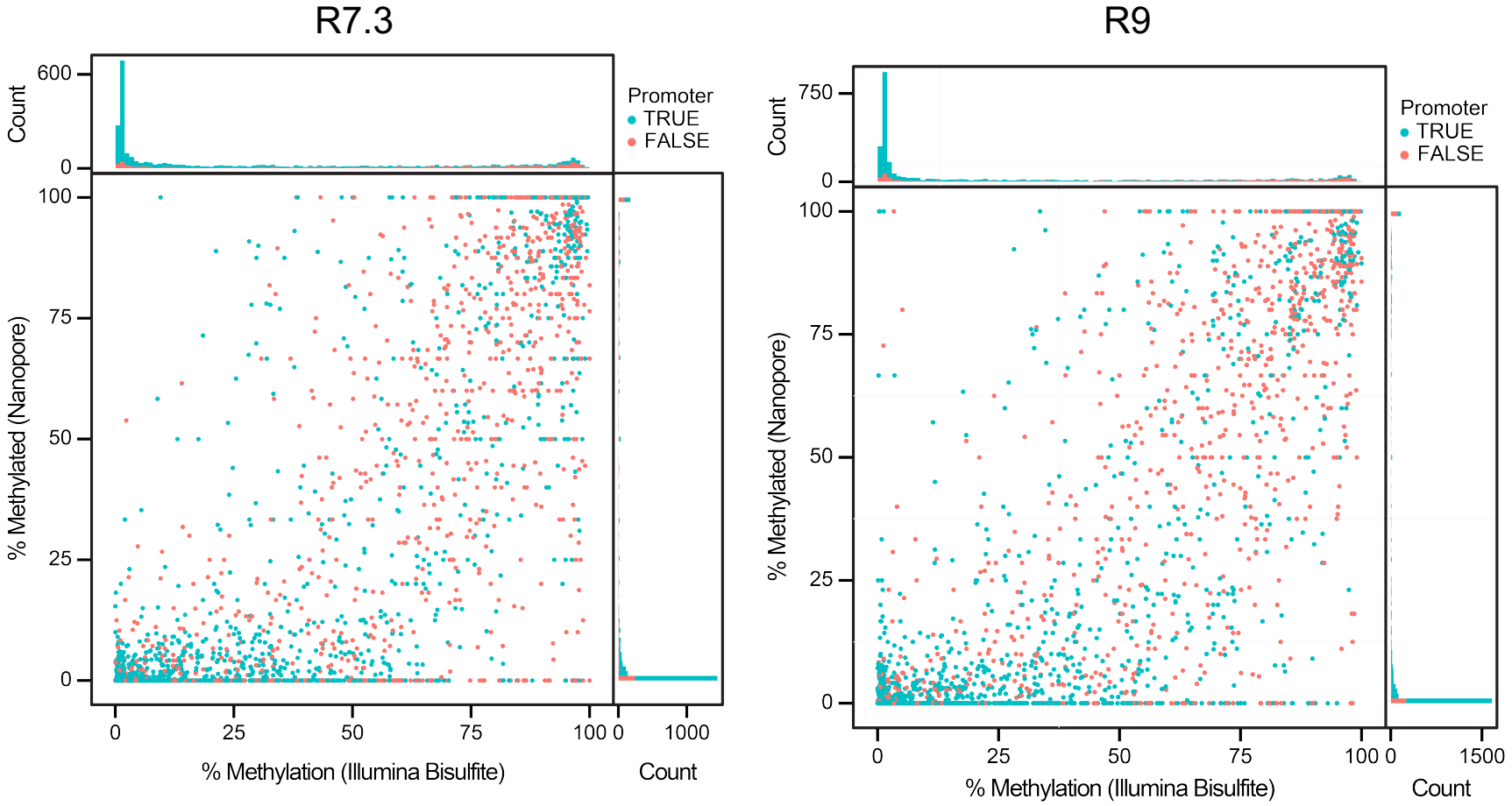
$$\delta_k \prod_t P(I(t) \mid k_t) \times T_{(t-1)(t)}$$

$$\frac{P(\mathcal{D}|S_m)}{P(\mathcal{D}|S_r)}$$



- Standard basecalling does not take into account the 5th base – mC.
- Using nanopolish, after alignment we return to the signal and asses the logliklihood ratio: $\log(\frac{P(D|S_m)}{P(D|S_r)})$ where S_m is the probability methylated for a given observable current D and S_r the probability unmethylated
- We then use 2.5 as the threshold for methylated and -2.5 for unmethylated.

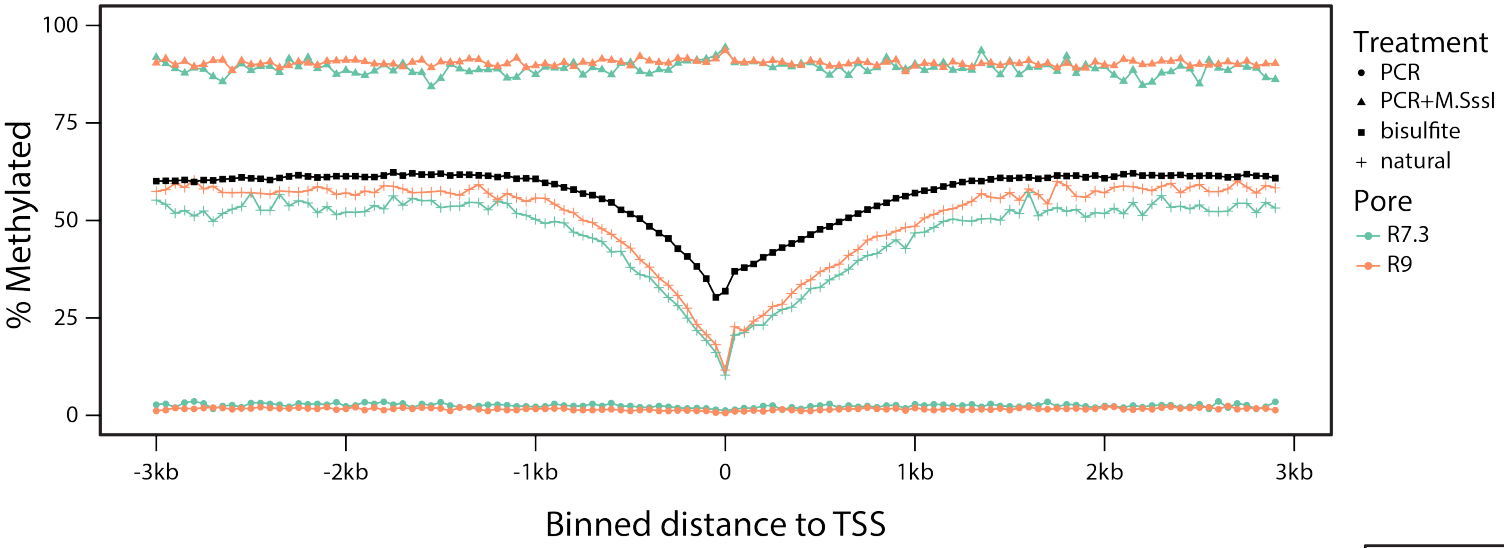
NA12878 Methylation



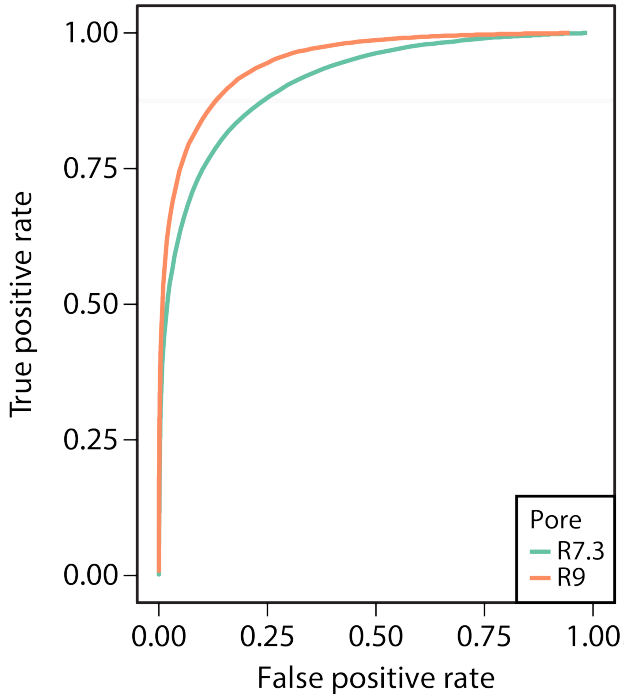
Simpson, Workman, Nature Methods (2017)

- NA12878 (lymphoblast) gDNA: Illumina WGBS on X-axis (24X coverage) (SRA: GSM1002650) vs. R7.3 (0.02X) or R9 (0.13X) nanopore sequencing.
- Correlation of 0.83 (R7.3) and 0.84 (R9) – most gene promoters unmethylated

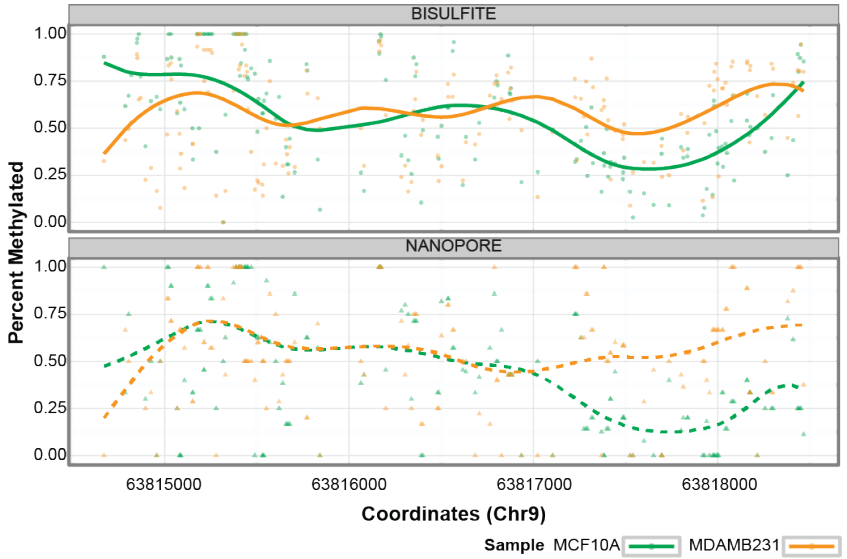
Methylation at Transcription Start Sites

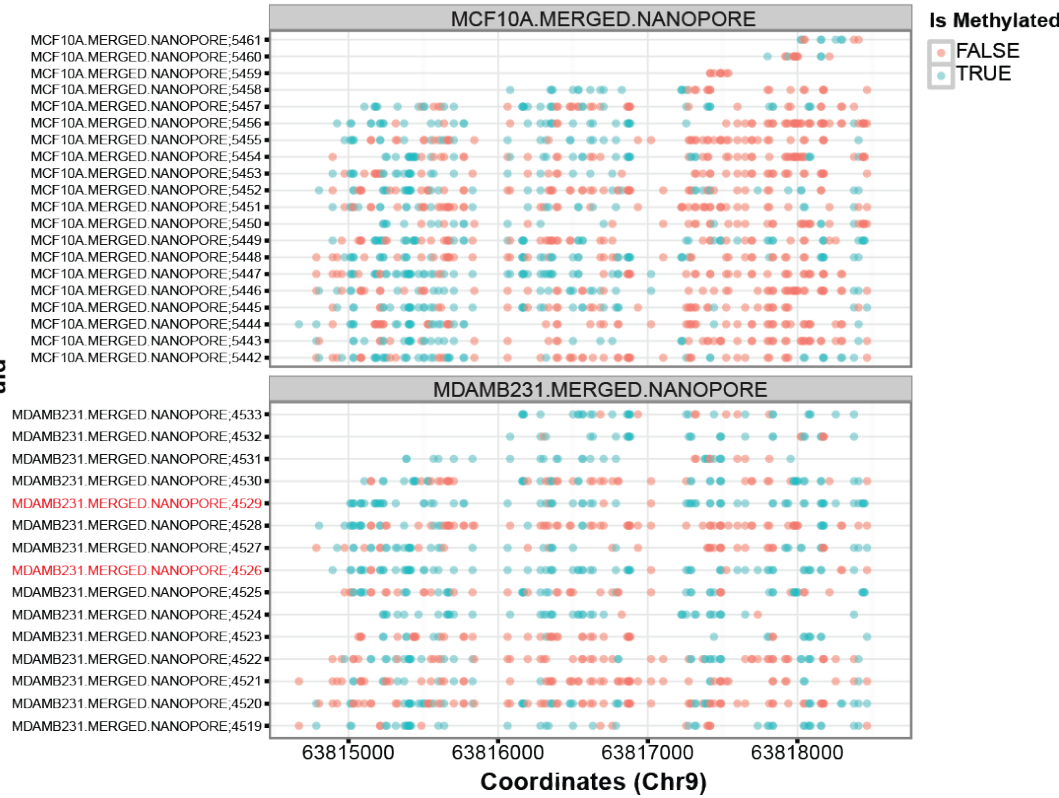


- On human genomic samples:
- Binning methylation levels vs. distance to TSS sites, compared to bisulfite data (NA12878).
- We also generated completely methylated (M.SssI treated; ~95% meth) and unmethylated – used to generate the ROC curve (right)
- R7.3 91% accurate at 68% of sites
- R9 94% accurate at 77% of sites



Cancer-Normal Comparison

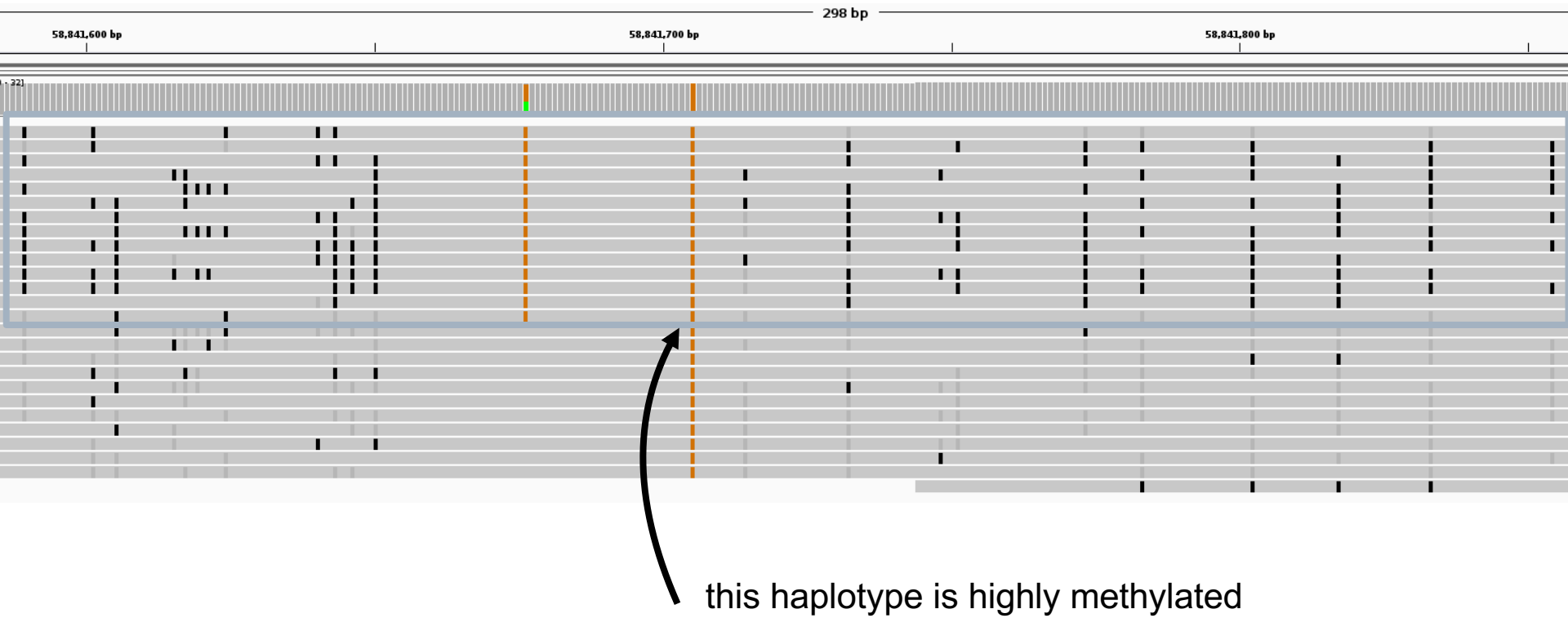




- Reduced representation method:12.5Mb of the genome (3.5-6kb size selection)
- We sequenced this fraction on nanopore and bisulfite Illumina seq
- Long reads measure *phased* methylation

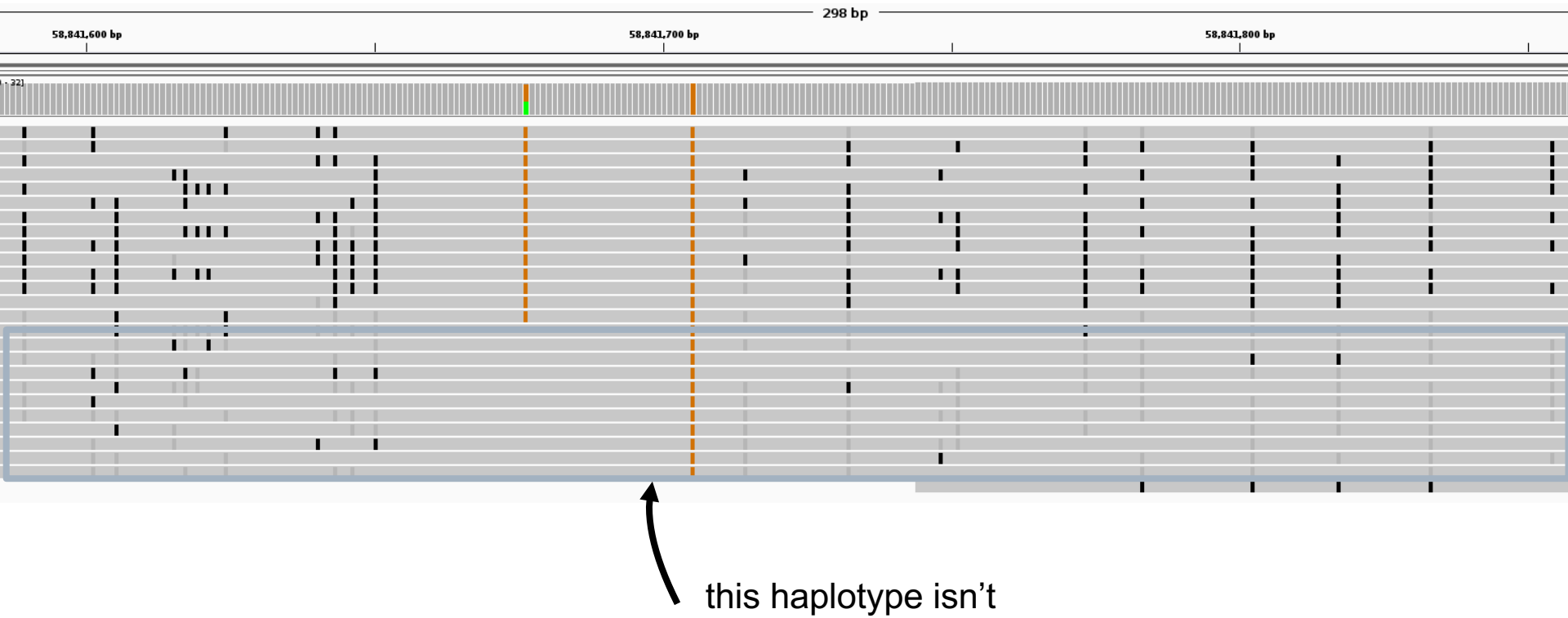
Haplotype-Phased Methylation

nanopolish has experimental support for phasing methylation patterns



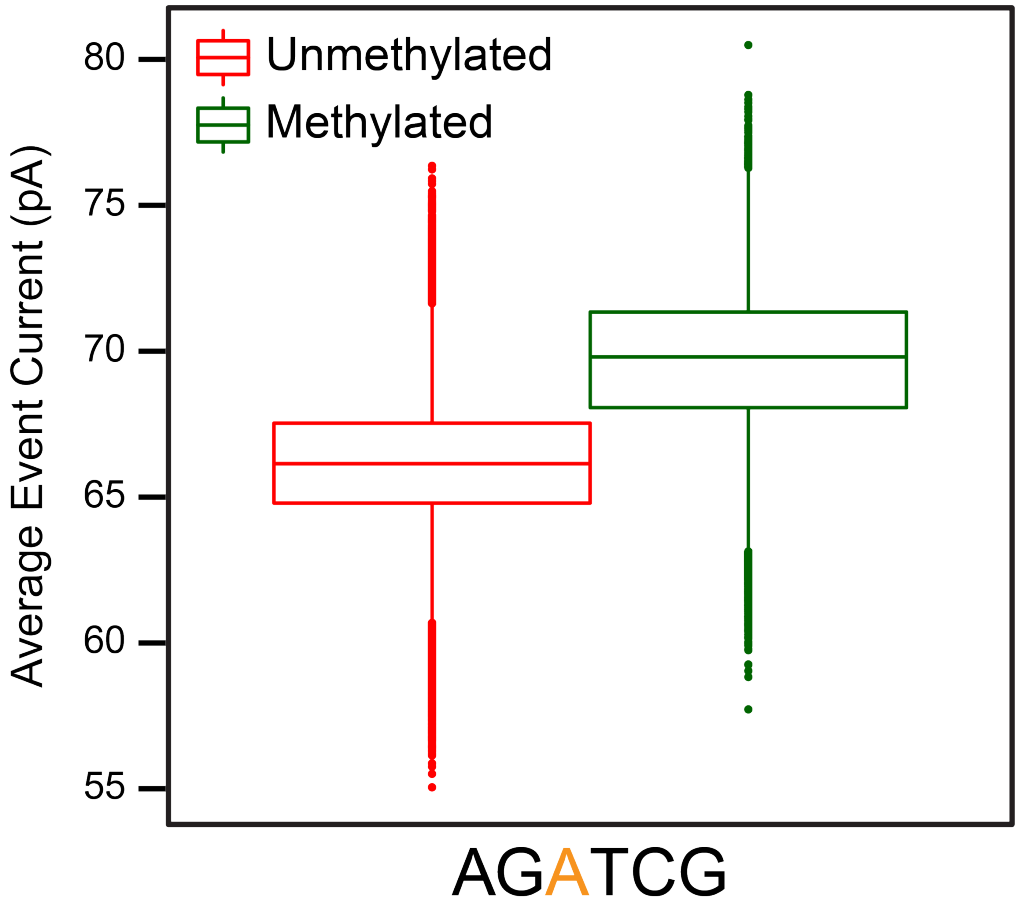
Haplotype-Phased Methylation

nanopolish has experimental support for phasing methylation patterns



Next Steps

- Expand to non-CpG methylation
- Expand to non 5-methylcytosine methylation
  - Strong signal for N6-methyladenine
- Apply to clinical samples
- Exogenous labeling of DNA and readout



Acknowledgments

- **Timp Lab – Johns Hopkins University**
- Winston Timp
- Rachael Workman
- Stephanie Hao
- Yunfan Fan
- Isac Lee

Agilent Technologies
Josh Wang, PhD
Jonathan Levine, PhD

- **Eshleman Lab – Johns Hopkins School of Medicine**
- James Eshleman, MD, PhD
- Alexis Norris, PhD

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

Extra Slides