NHGRI Advanced DNA Sequencing Technology Development

Grantee Meeting

Boston, MA; May 2017

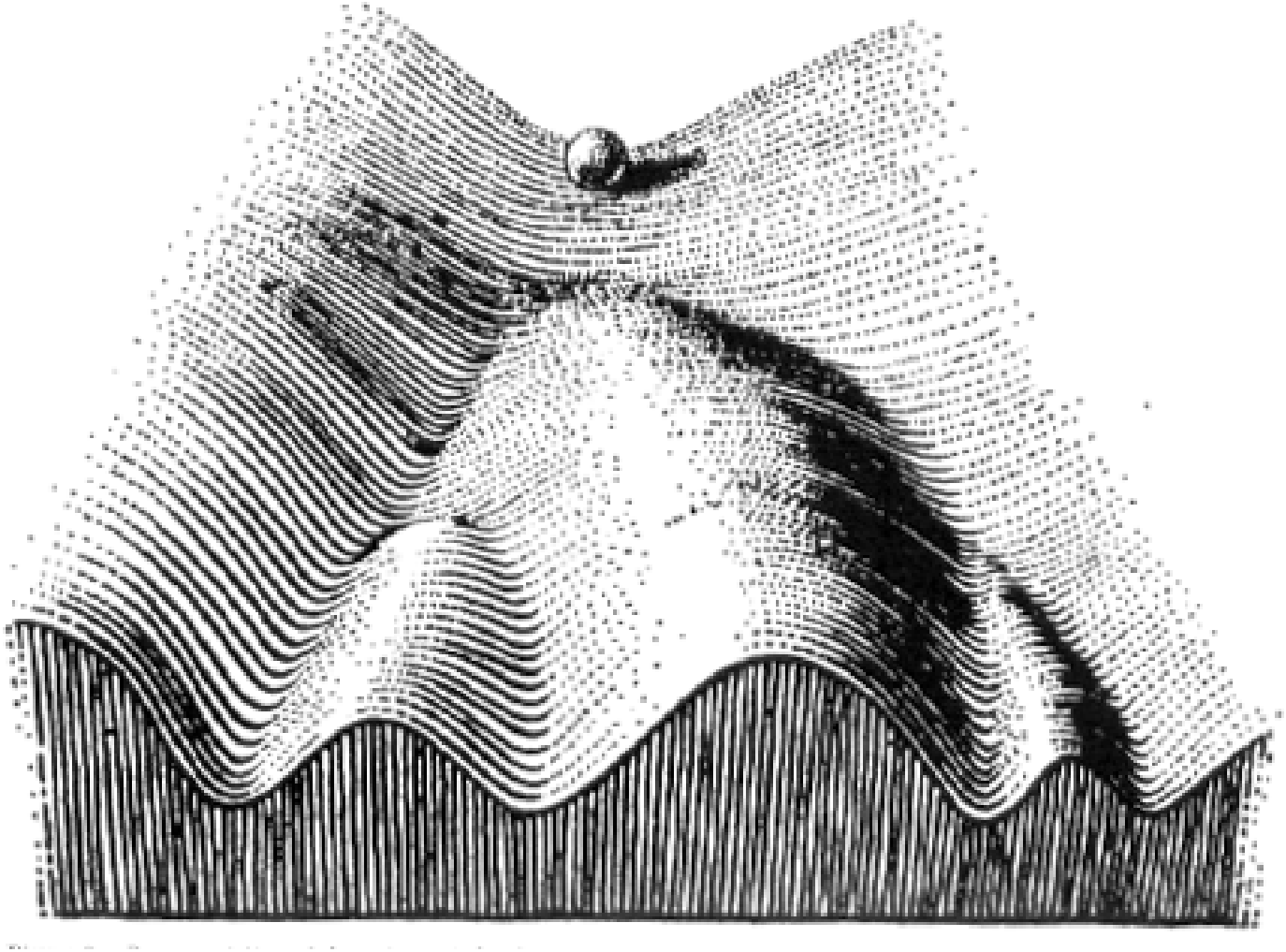
JOHNS HOPKINS
BIOMEDICAL ENGINEERING

Detecting nucleotide modifications using nanopore sequencing

Winston Timp
Department of Biomedical Engineering
Johns Hopkins University

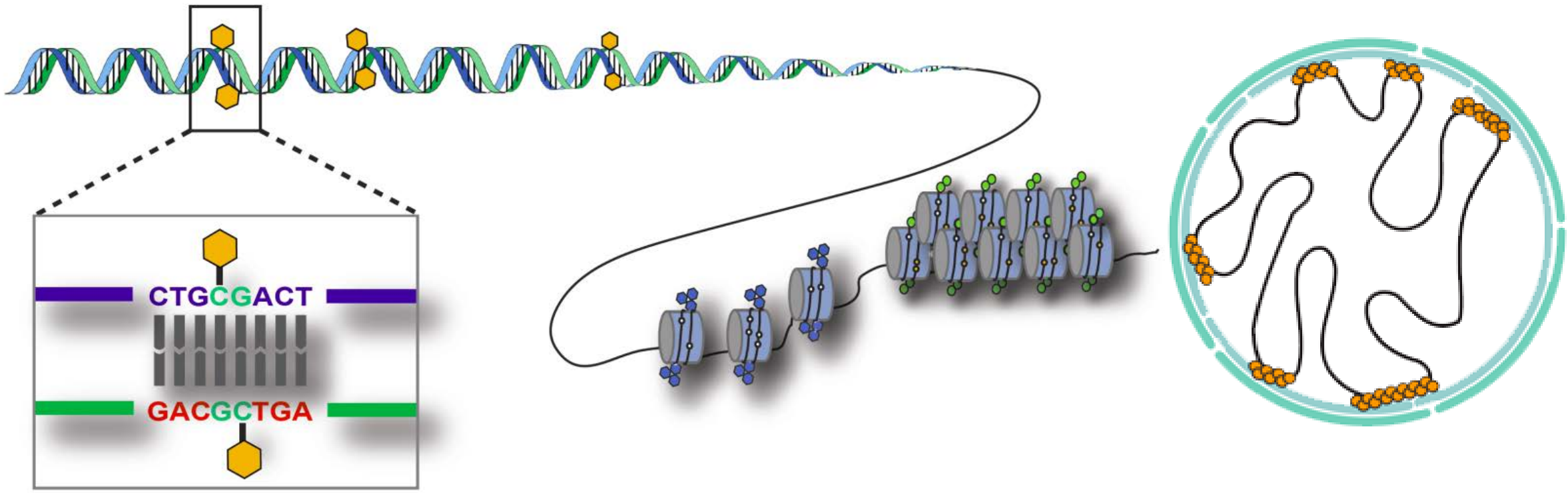
Epigenetics: Historical

- The *historical* definition of epigenetics, by Waddington, is how **genotype** interacts with the environment to create **phenotype**. The analogy is the rolling ball in the image to the right – the genotype has set up the hills, environment nudges the ball into one valley or another, then the ball is, to mix metaphor, lineage committed.
- Which valley the ball rolls into is not predetermined, but a stochastic behavior.



Waddington, 1940; Raser et al. Science (2005)

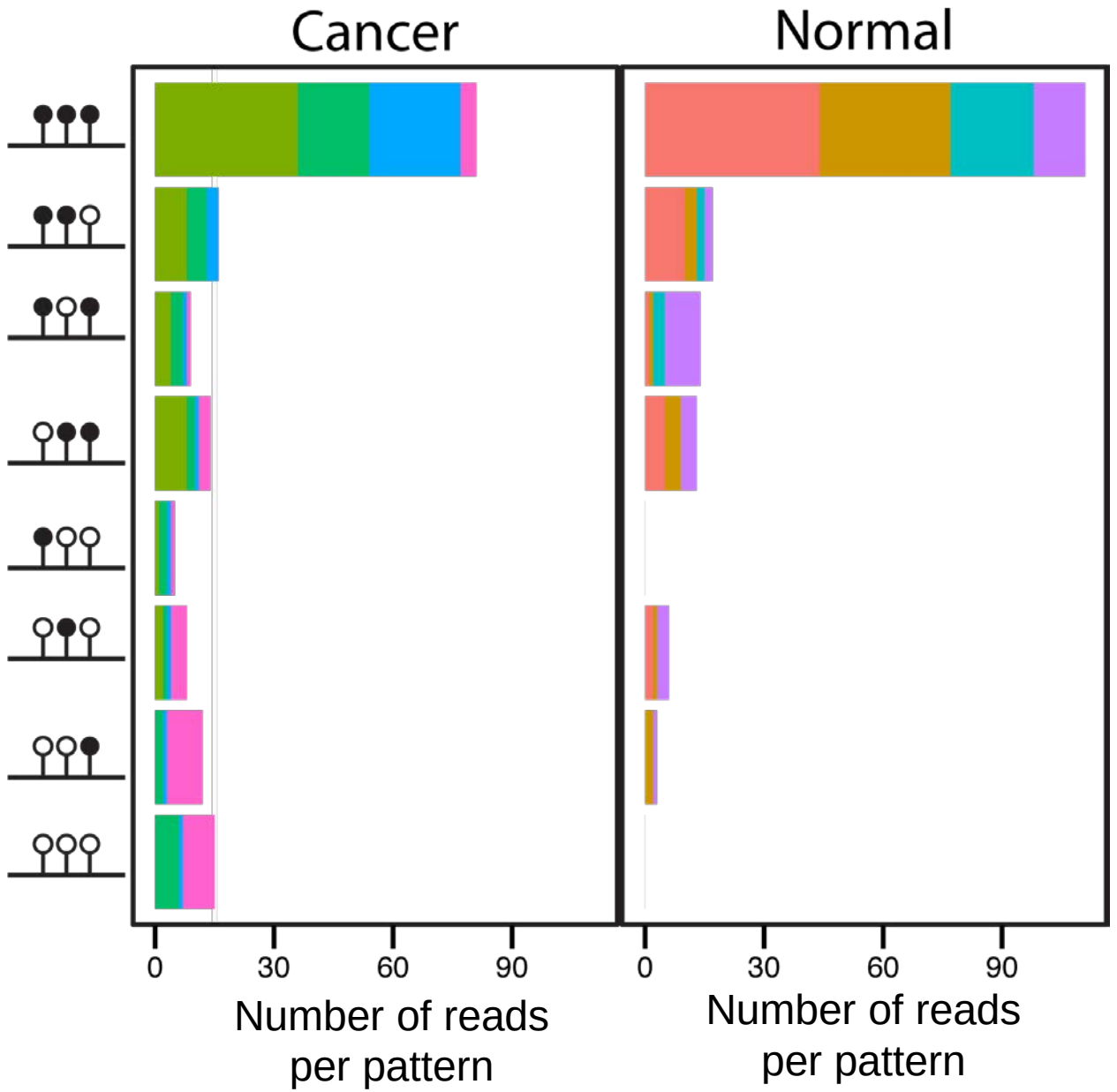
Epigenetics: Modern



- 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**

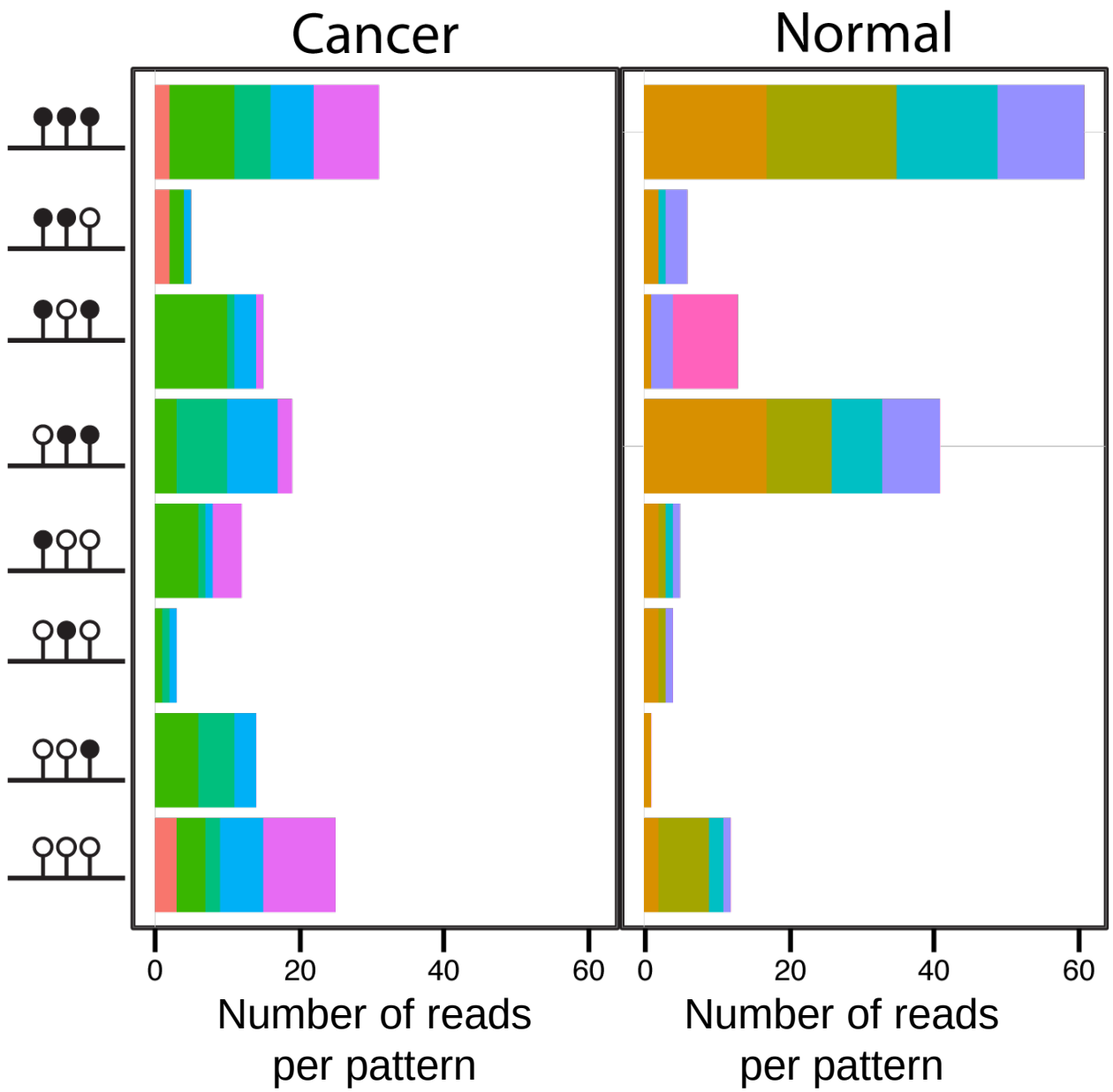
Single Read Methylation: Distribution

- We performed hybridization capture, then Illumina bisulfite sequencing on 6 paired colon cancer and normal samples.
- We then examined methylation patterns *within reads* and looked at the distribution in normal vs. cancer samples.
- Colors in the stacked bar graph represent different sequenced samples.
- Areas are clusters in regions which show significantly different methylation levels (t-test).



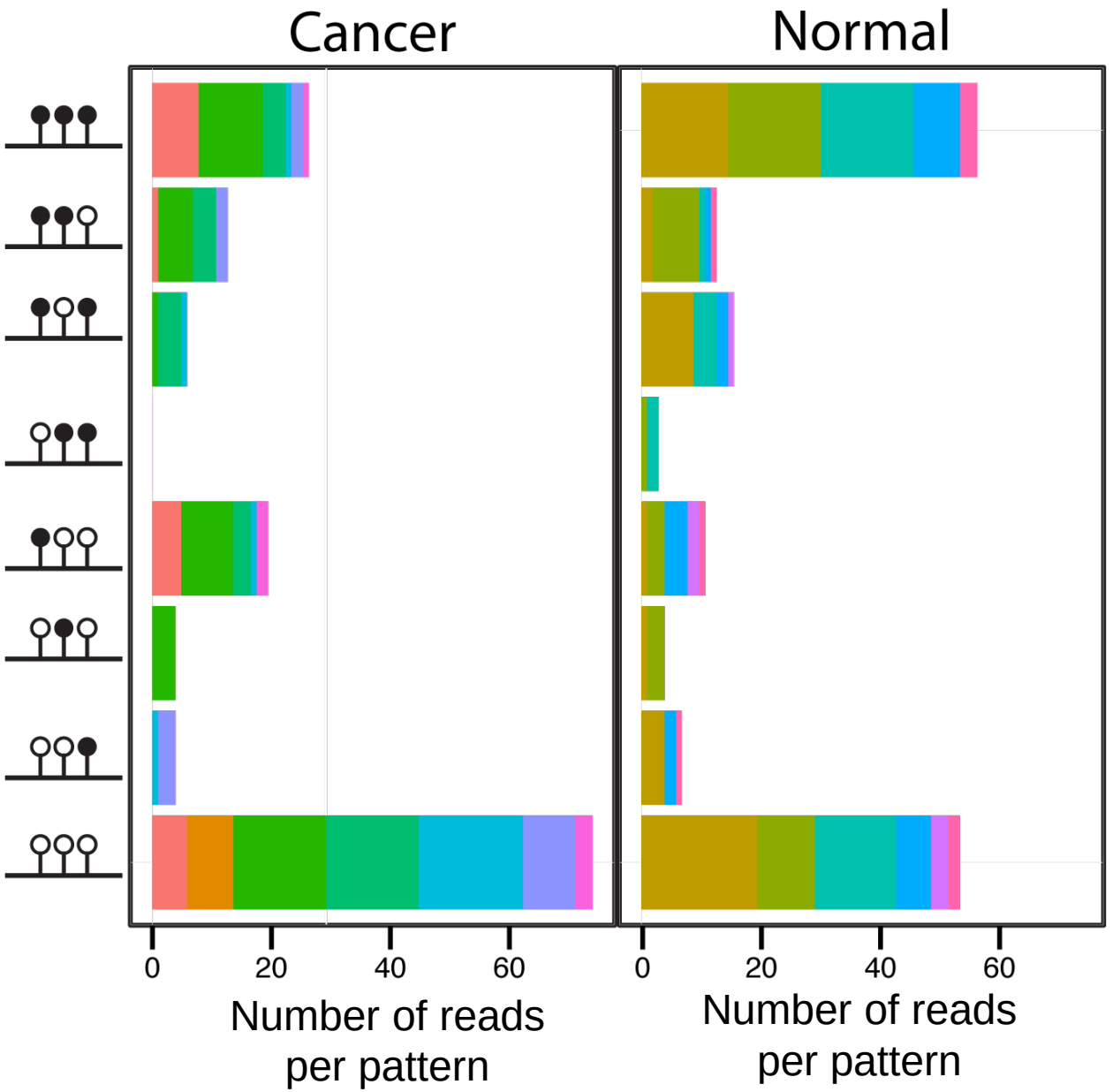
Single Read Methylation: Distribution

- We performed hybridization capture, then Illumina bisulfite sequencing on 6 paired colon cancer and normal samples.
- We then examined methylation patterns *within reads* and looked at the distribution in normal vs. cancer samples.
- Colors in the stacked bar graph represent different sequenced samples.
- Areas are clusters in regions which show significantly different methylation levels (t-test).

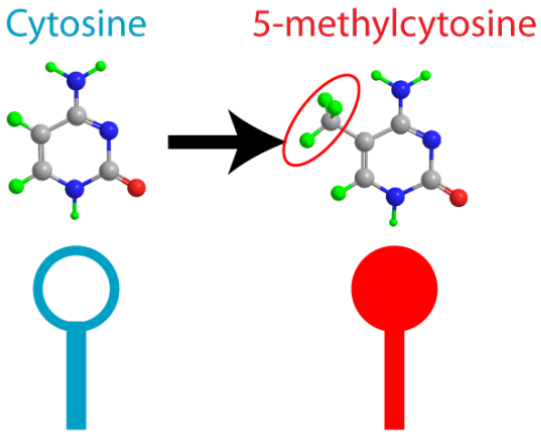


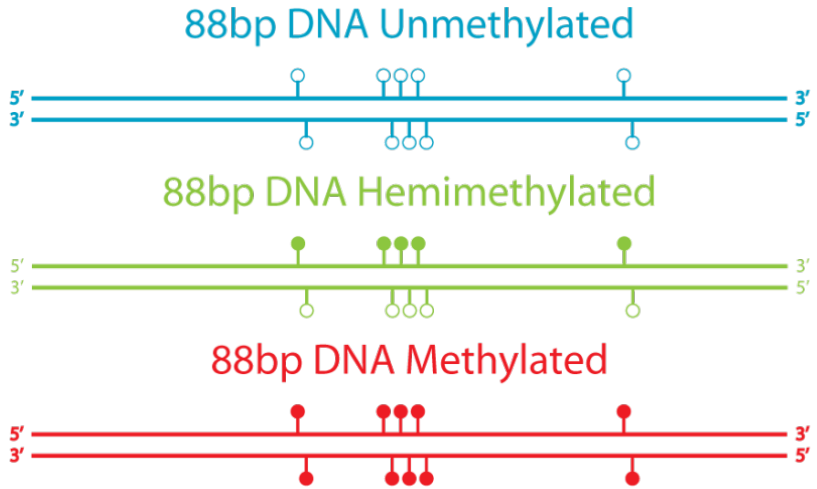
Single Read Methylation: Distribution

- We performed hybridization capture, then Illumina bisulfite sequencing on 6 paired colon cancer and normal samples.
- We then examined methylation patterns *within reads* and looked at the distribution in normal vs. cancer samples.
- Colors in the stacked bar graph represent different sequenced samples.
- Areas are clusters in regions which show significantly different methylation levels (t-test).

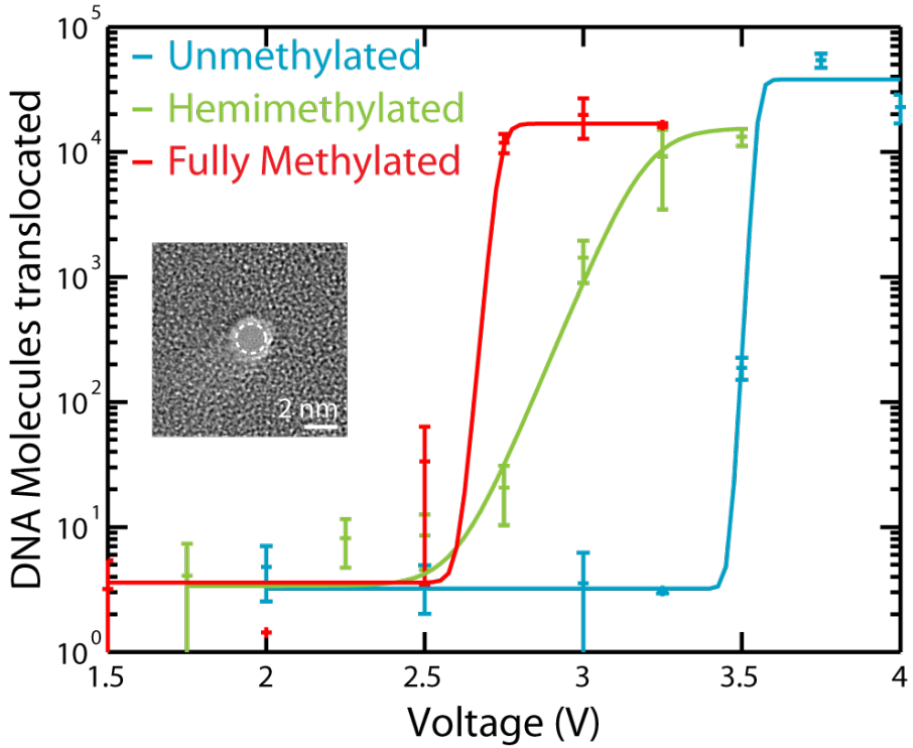


Nanopore: Methylation



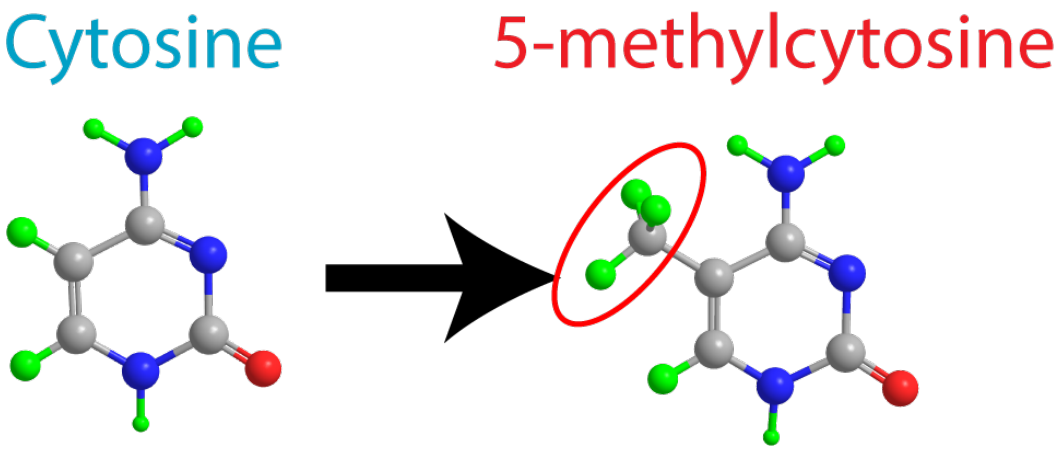


- Cytosine methylation increases DNA stiffness, decreasing ***stretching force***
- This may allow for filtering of methylated versus unmethylated DNA using a pore

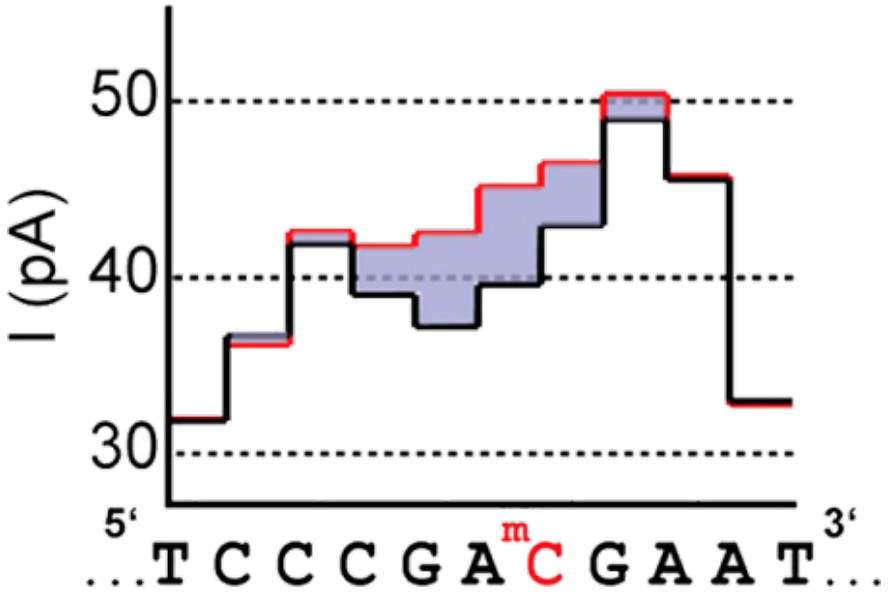


Mirsaidov, et al. *Biophys J.* (2009)

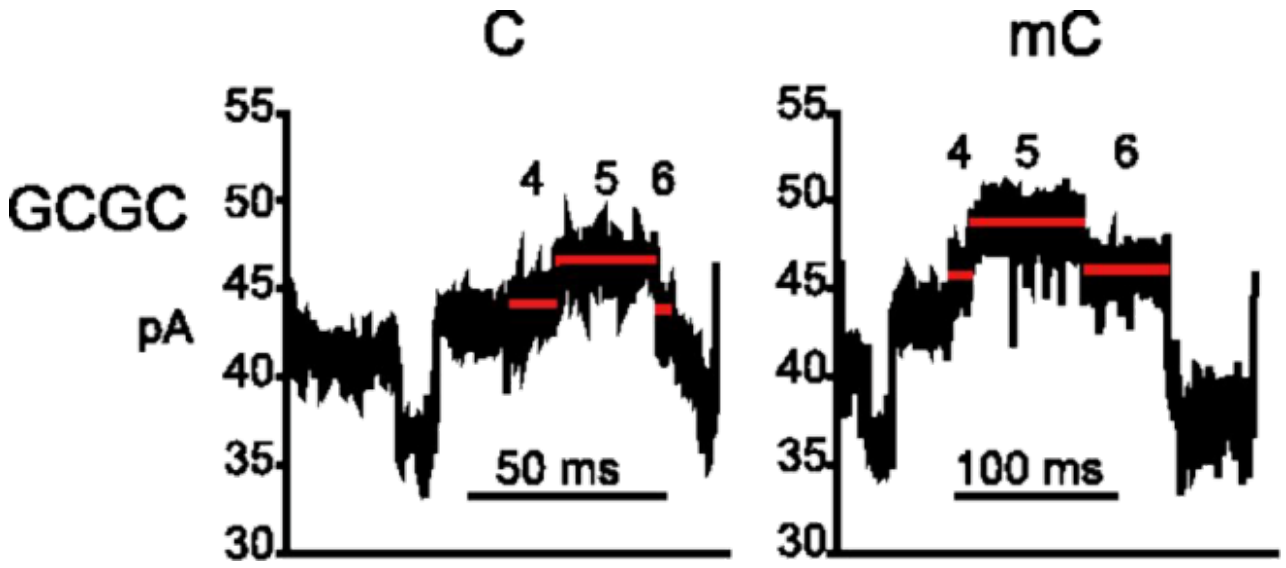
Nanopore: Methylation



- Differences between methylated and unmethylated cytosine have been detected using nanopores.
- Methylation state can be called with 90% accuracy.
- We have implemented a classifier for mC on using Oxford Nanopore signals.

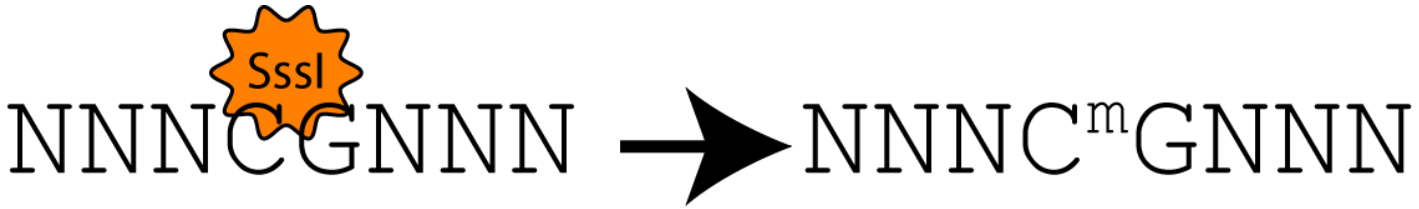


Laszlo, et al. *PNAS* (2013)

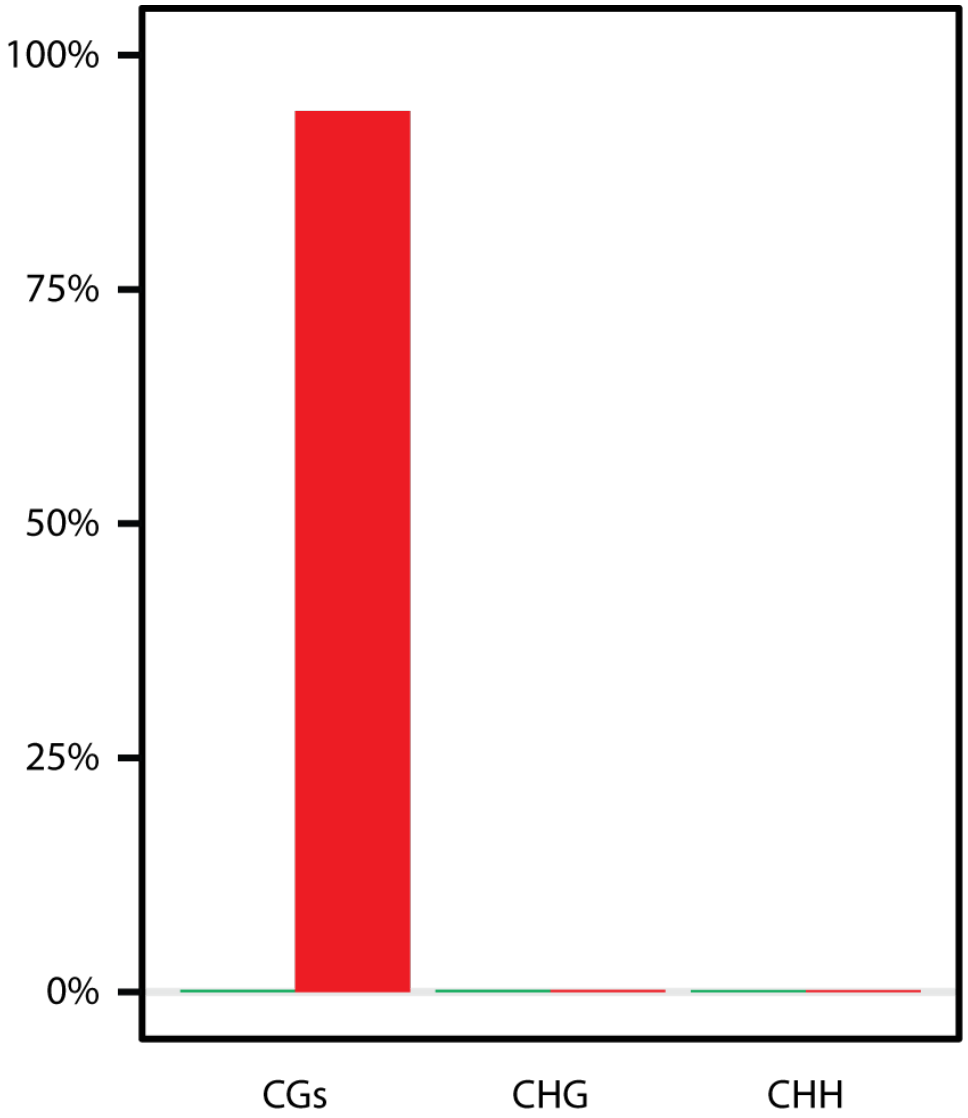


Schreiber, et al. *PNAS.* (2013)

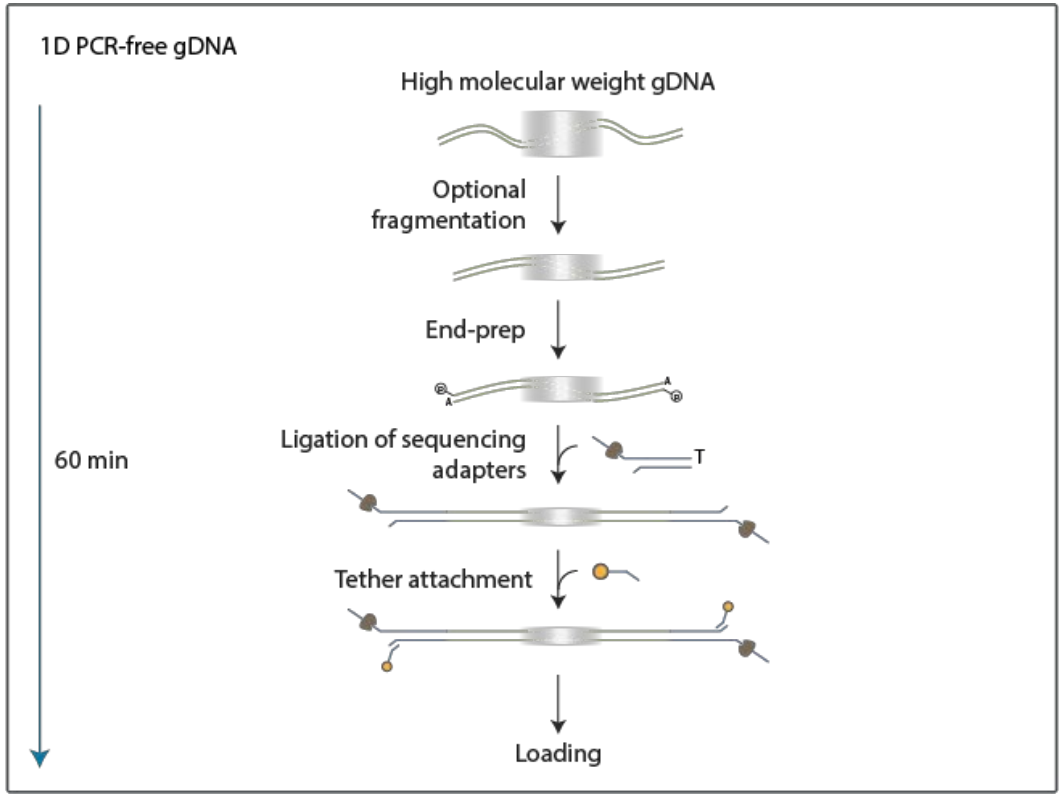
Generation of methylated Samples

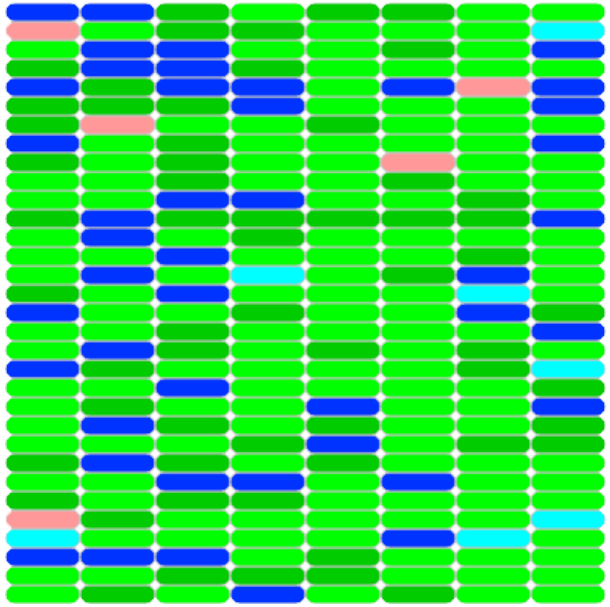


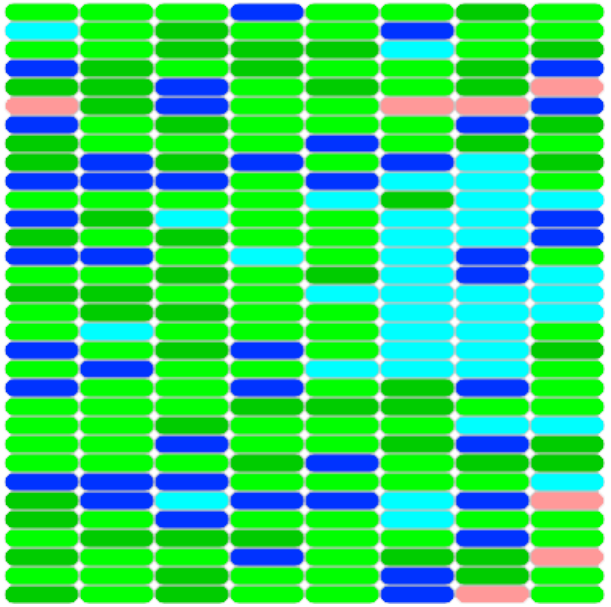
- To generated methylated samples, we treat unmethylated DNA (lambda, dam-/dcm- E. Coli, PCR product) with M. SssI methyltransferase
- We confirm the CG specific methylation using Illumina bisulfite sequencing of the sample – pictured right is methylation in different contexts E. Coli dataset treated with M. SssI (red) versus untreated (green)



Nanopore Library Prep





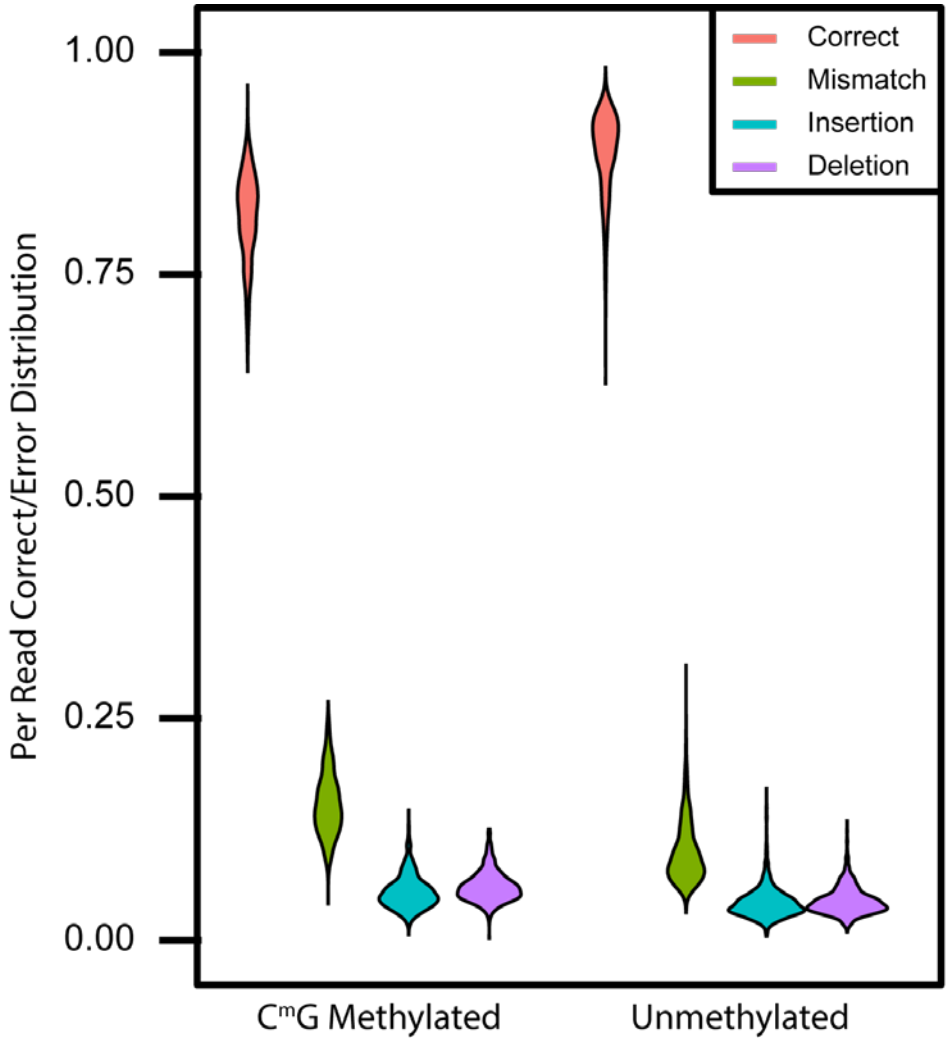


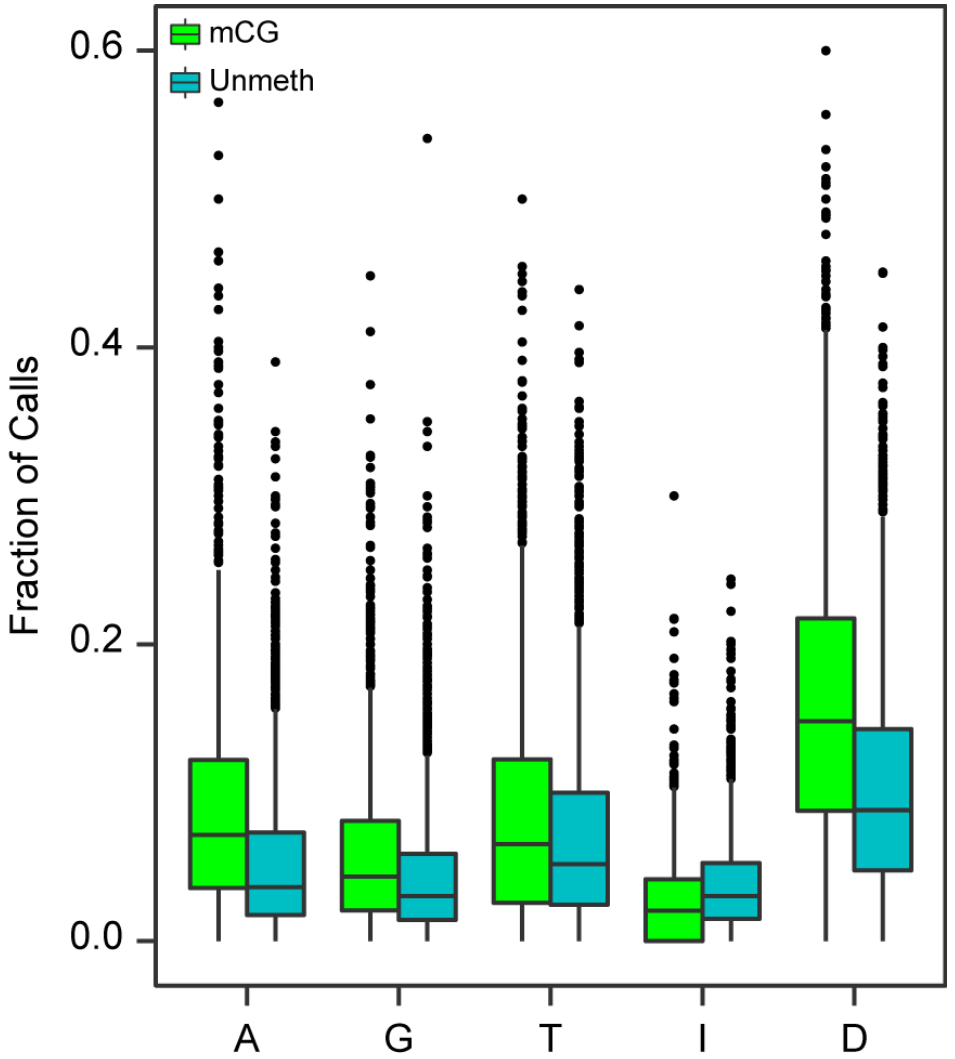
- Library prep is very similar to methods for short-read sequencing
- For DNA shearing we used Covaris gTubes
- After end-repair and A-tailing, leader adapter with motor protein is ligated
- MinION arrays 512 channels (with 4 pores possible per channel) (shown bottom left from running software); dark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

Nanopore: Methylated Error

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

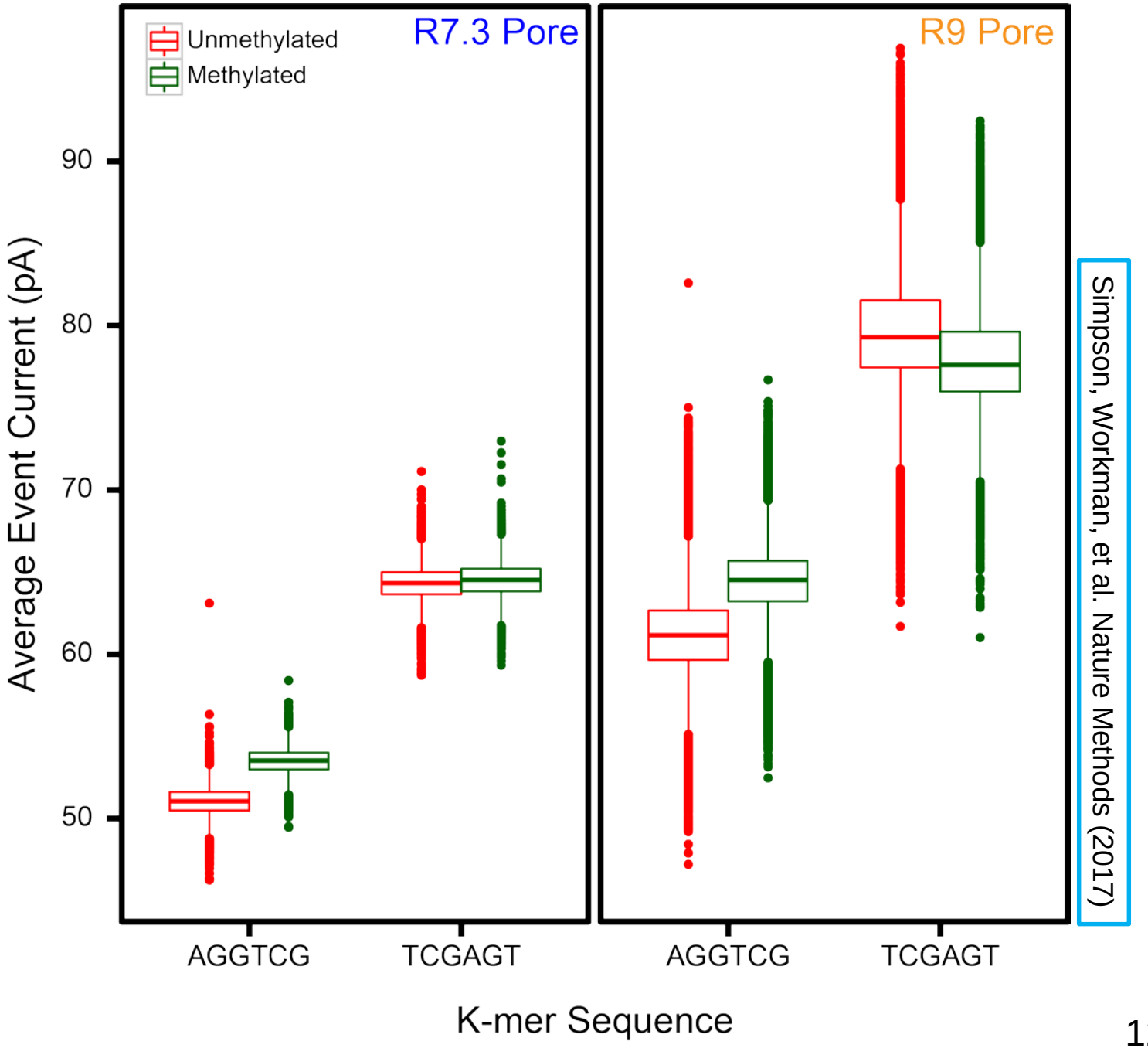
- We sequenced samples with either SssI or no treatment.
- Notably, mismatch error rate is higher on methylated samples than unmethylated, though indel rate seems mostly unchanged
- At CG locations, there is not a clear alternative basecalling, though error is higher; Ts are not substituted for methylated Cs





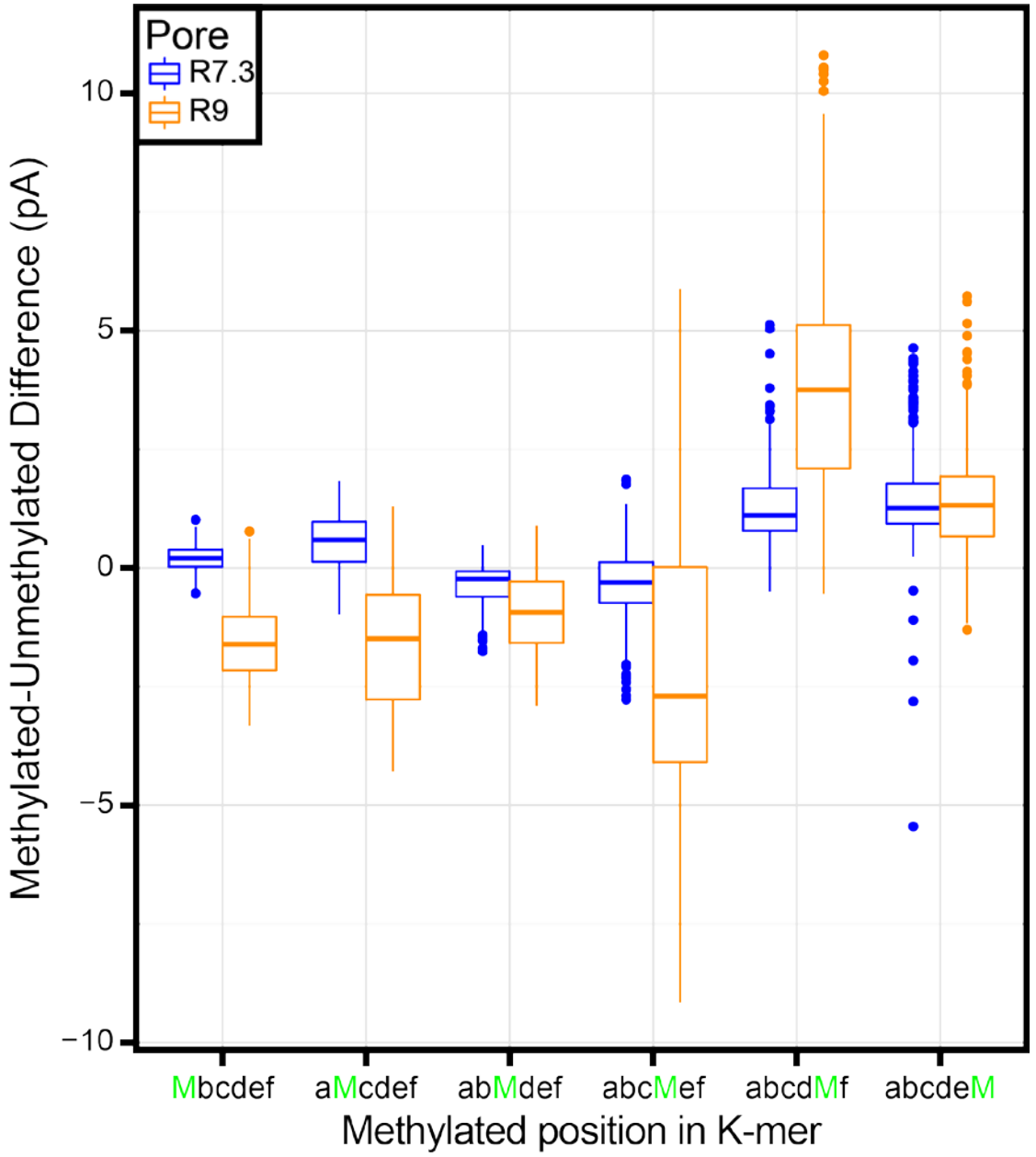
Emission Probabilities

- 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.



Distance of methylation effect

- We looked at the difference in current levels dependent on the position of the methylated base – plotted are the current differences for R7.3(blue) and R9 pores(orange).
- Signal seems again stronger but more variable for R9 pores than R7.3
- Methylation can either reduce current or increase it.
- Some positions are more sensitive to methylation than others.



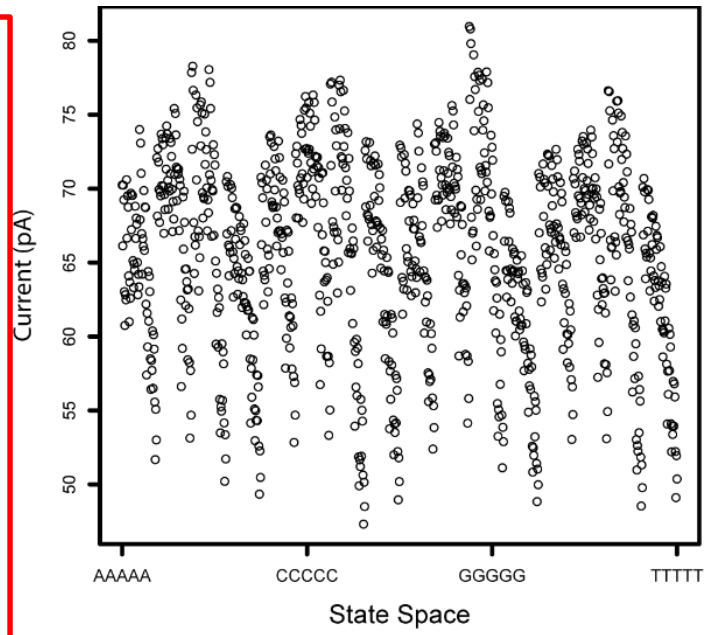
Simpson, Workman, *Nature Methods* (2017)

Nanopore: nanopolish methyltrain

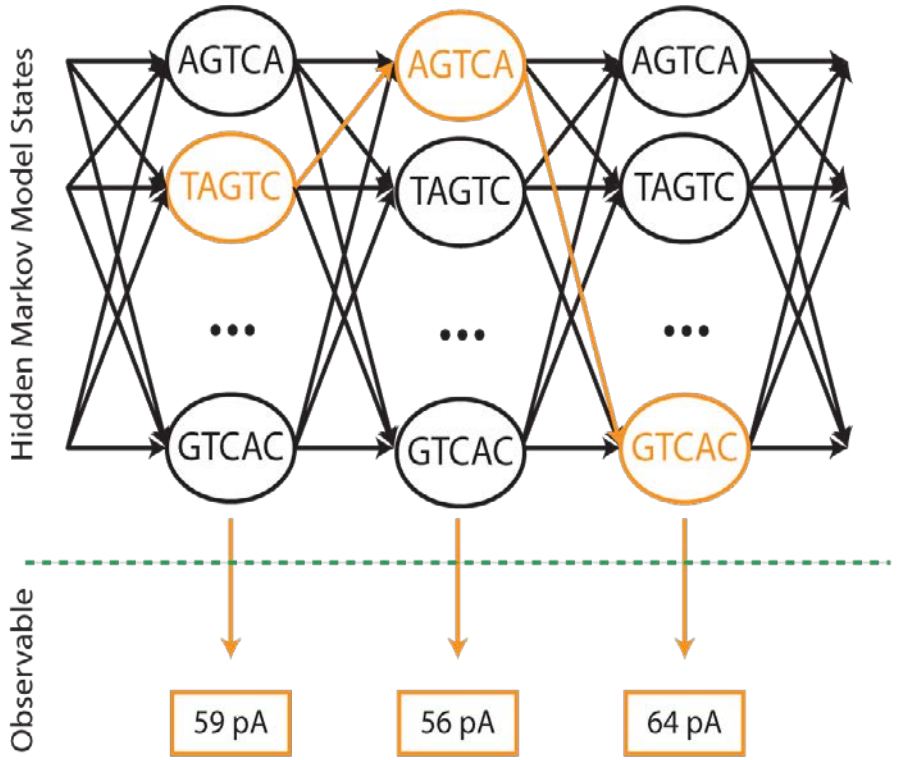
- Multiple bases influence the current passing through the pore.
- Current basecallers use a neural-network based methodology to call bases.
- We currently use a HMM based classifier to call methylation
- With *nanopolish* we can call the probability:

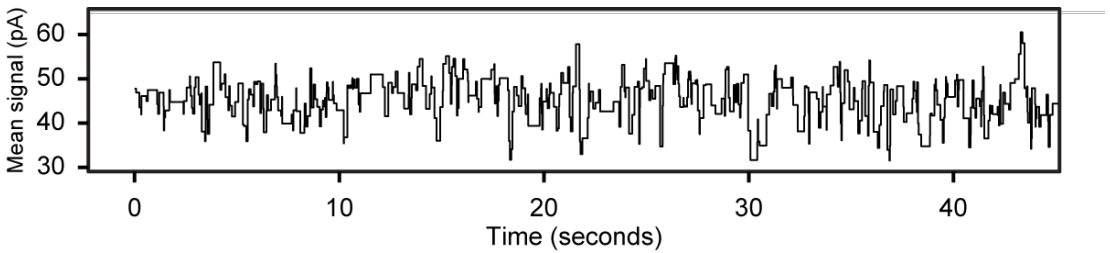
$$\frac{P(\mathcal{D}|S_m)}{P(\mathcal{D}|S_r)}$$

- Where S_m is the probability methylated for a given observable D and S_r the probability unmethylated
- We then take the log of this likelihood ratio.



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

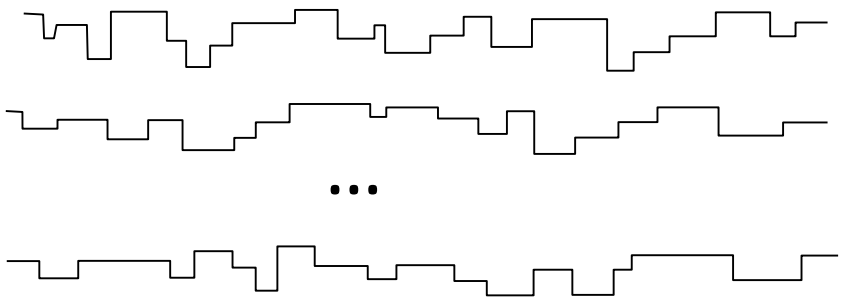




Nanopolish tools

github.com/jts/nanopolish

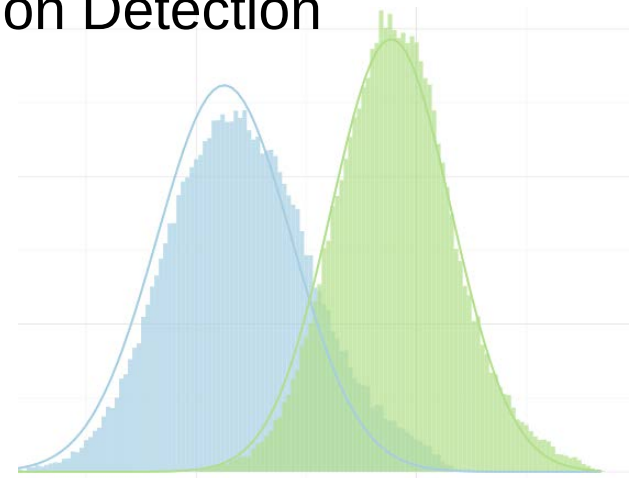
Consensus Calling



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

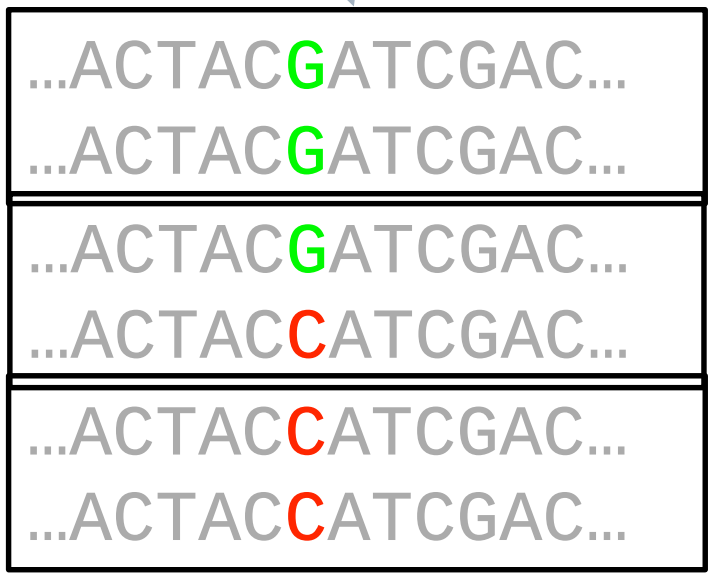


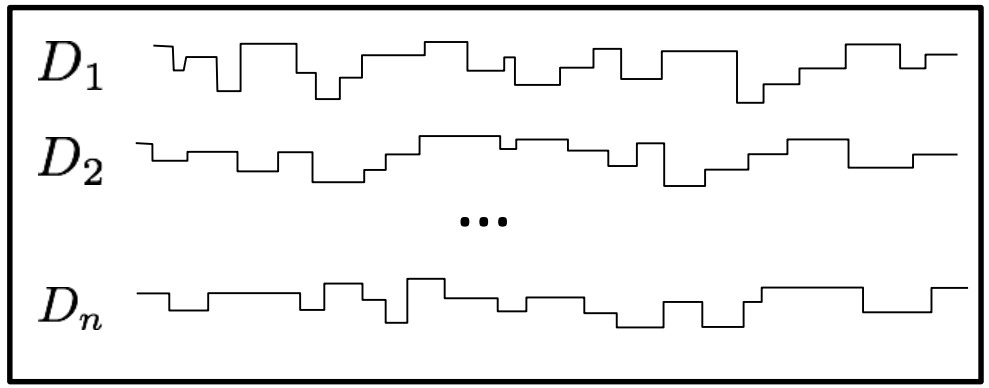
Read Phasing

TAGAAGATATCATGTATAGTACGAT
TAGAAGATATCATG
TAGCAGATATCATGTATATTACGAT
CATGTATATTACGAT

Nanopolish SNP Calling and Genotyping

input: pairs of haplotypes





hidden Markov model

-332
-300
-350

0/1

output: genotype call

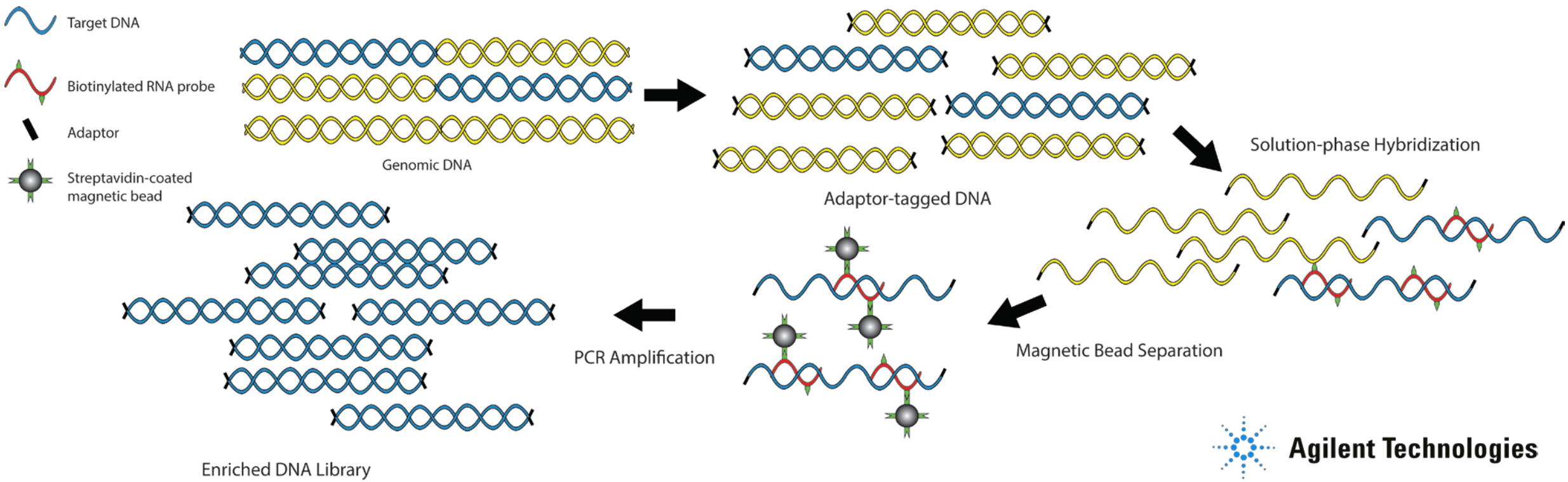
Human Genotyping Results

		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 *bioRxiv* 128835

Solution-phase Hybridization Capture

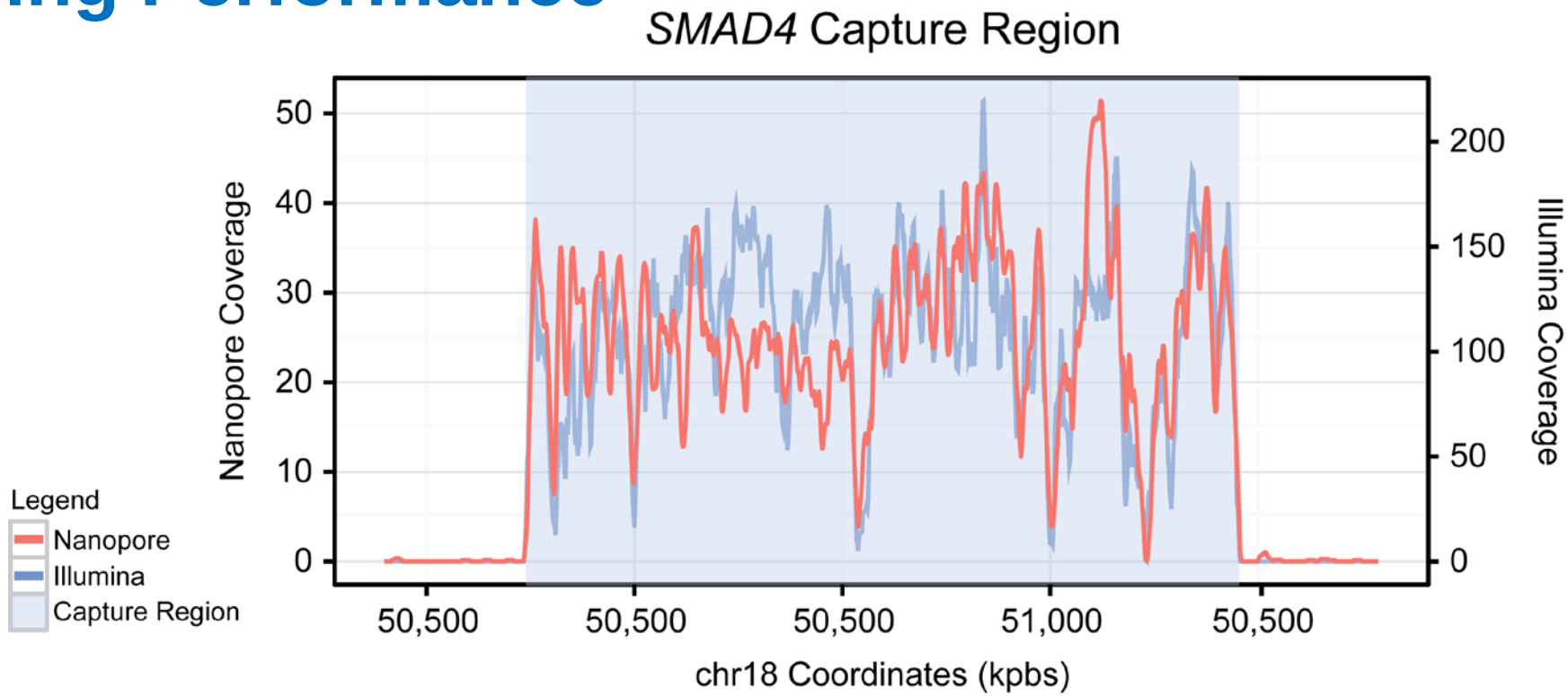


Agilent SureSelectXT Targeted Sequencing System

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

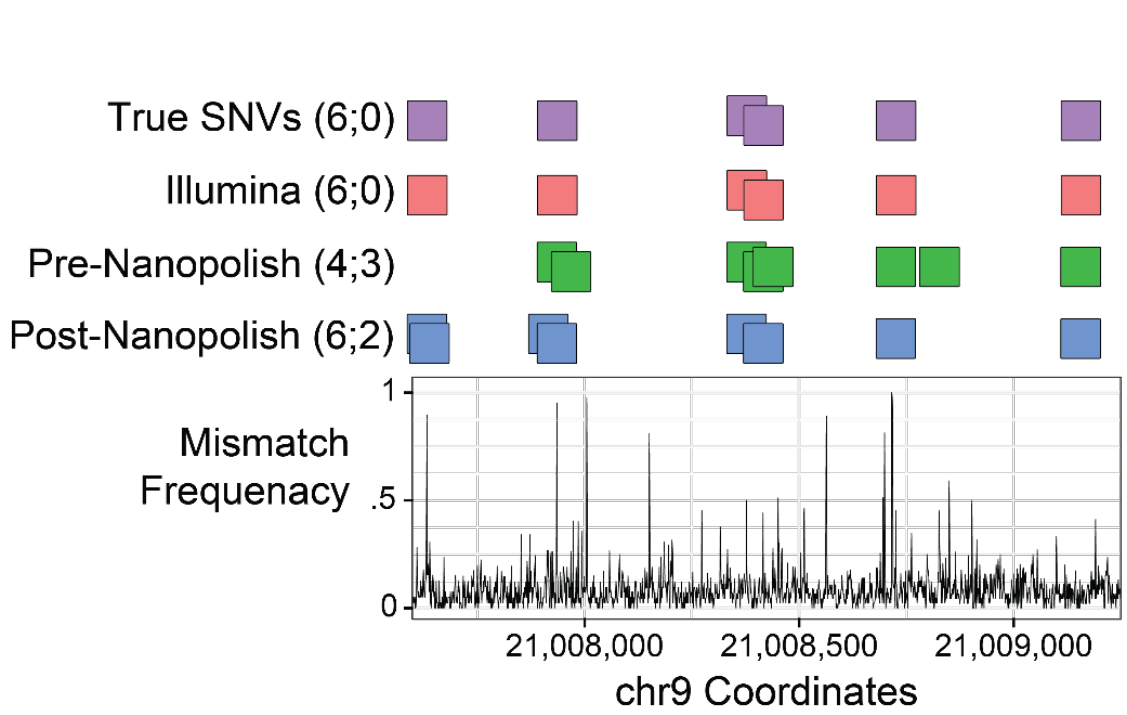
Targeted Sequencing Performance

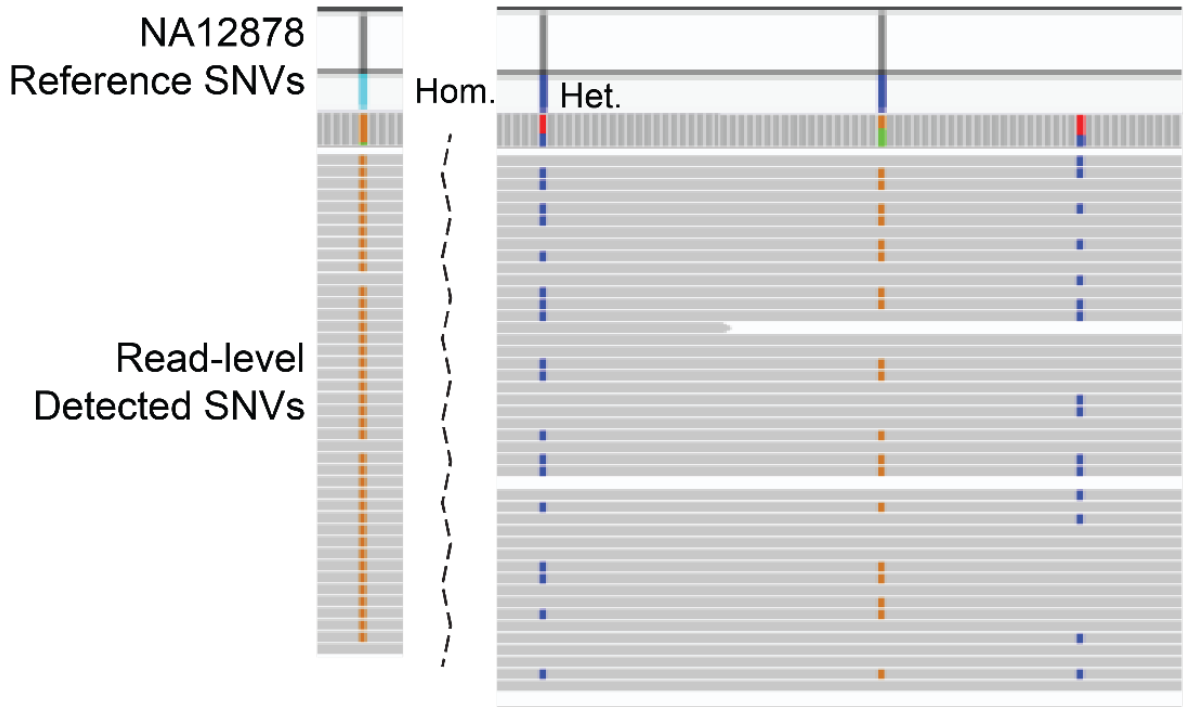
- Control : NA12878 lymphoblast
- Sample : PDAC from Eshleman lab
- Illumina short-read targeted sequencing for comparison
- > 300-fold enrichment
- > 20X average coverage
- Agilent App Note: https://goo.gl/8V2Fei



	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

Single Nucleotide Variation Detection





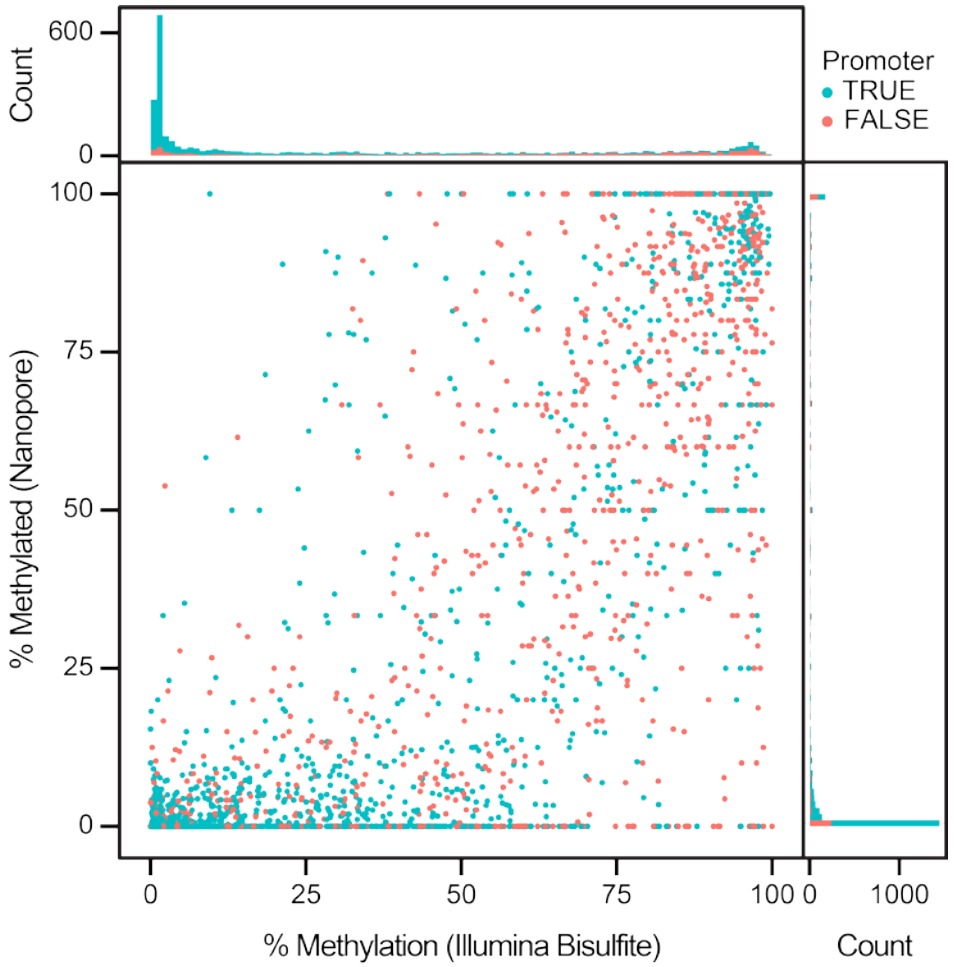
Phased SNV analysis is possible with coverage from targeted sequencing

	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%**

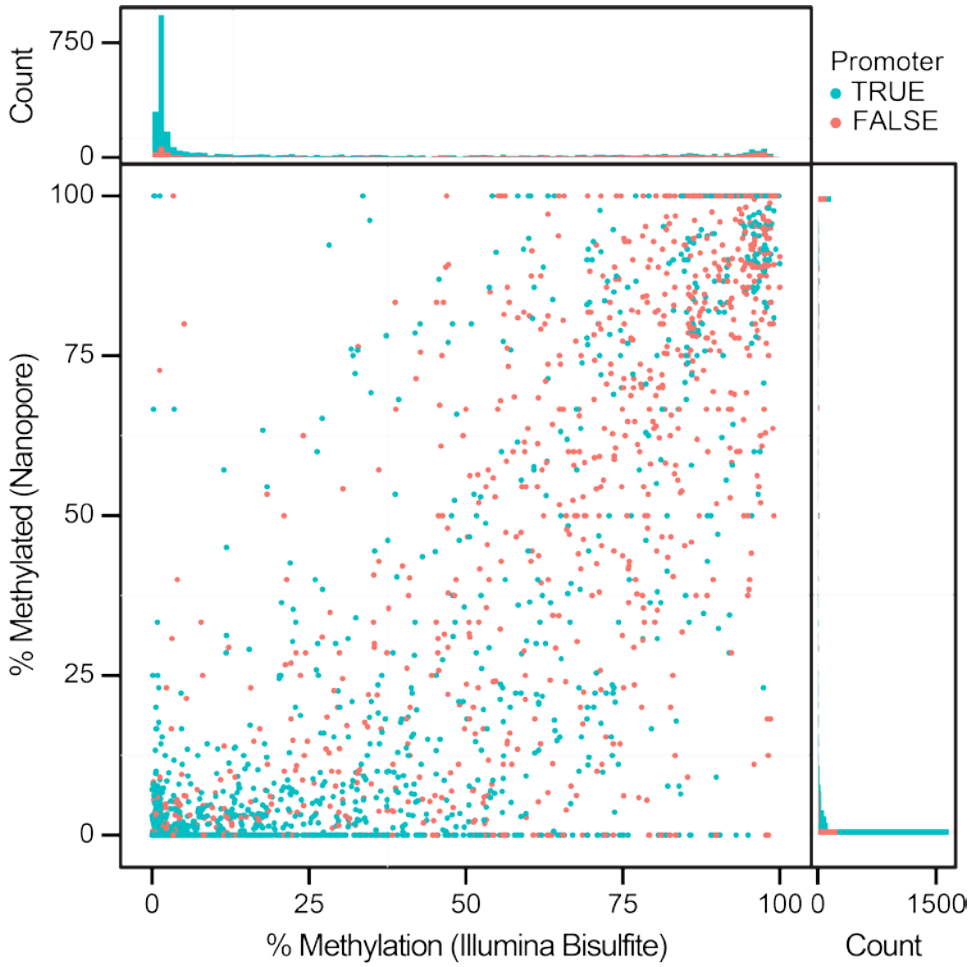
Number of True SNVs: 3587(Eberle,et al. bioRxiv, 2016)

NA12878 Methylation

R7.3



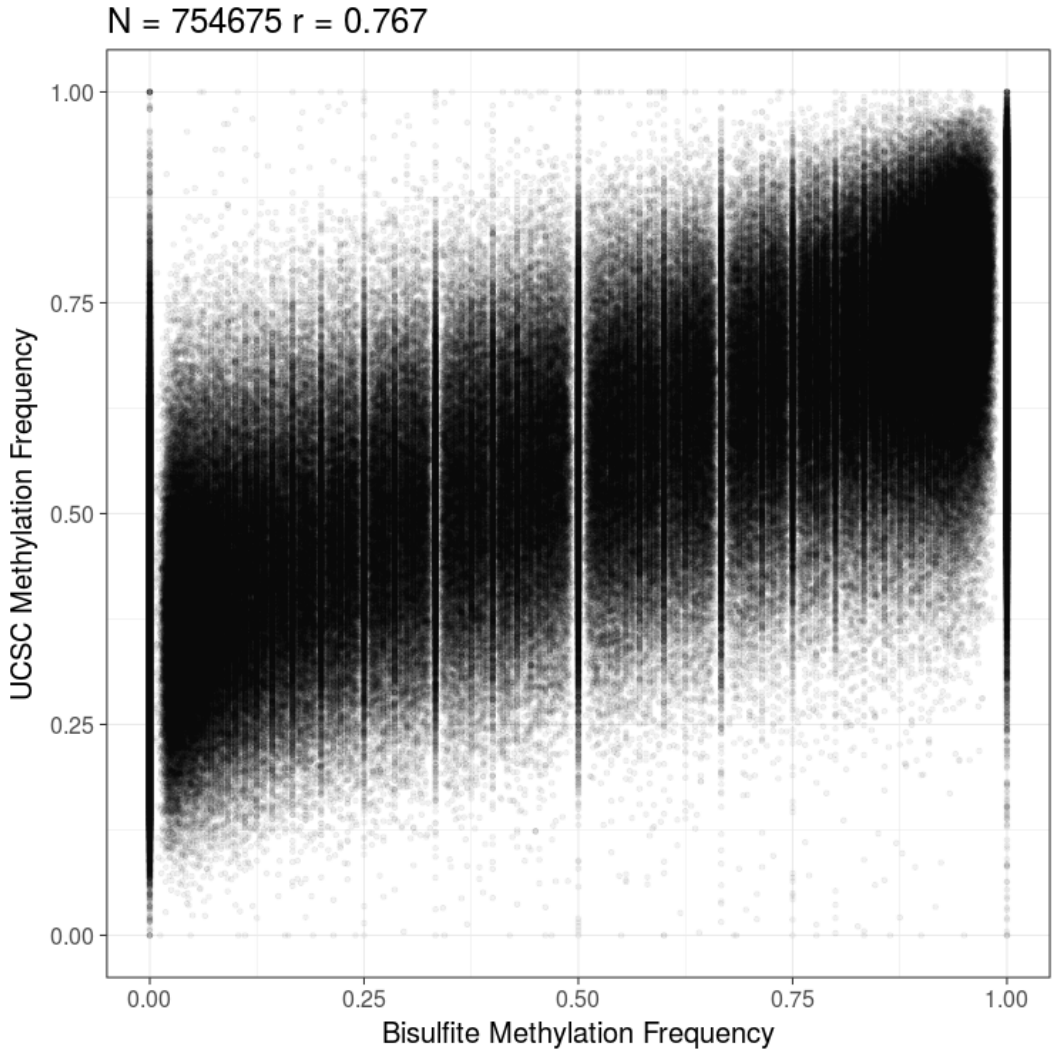
R9

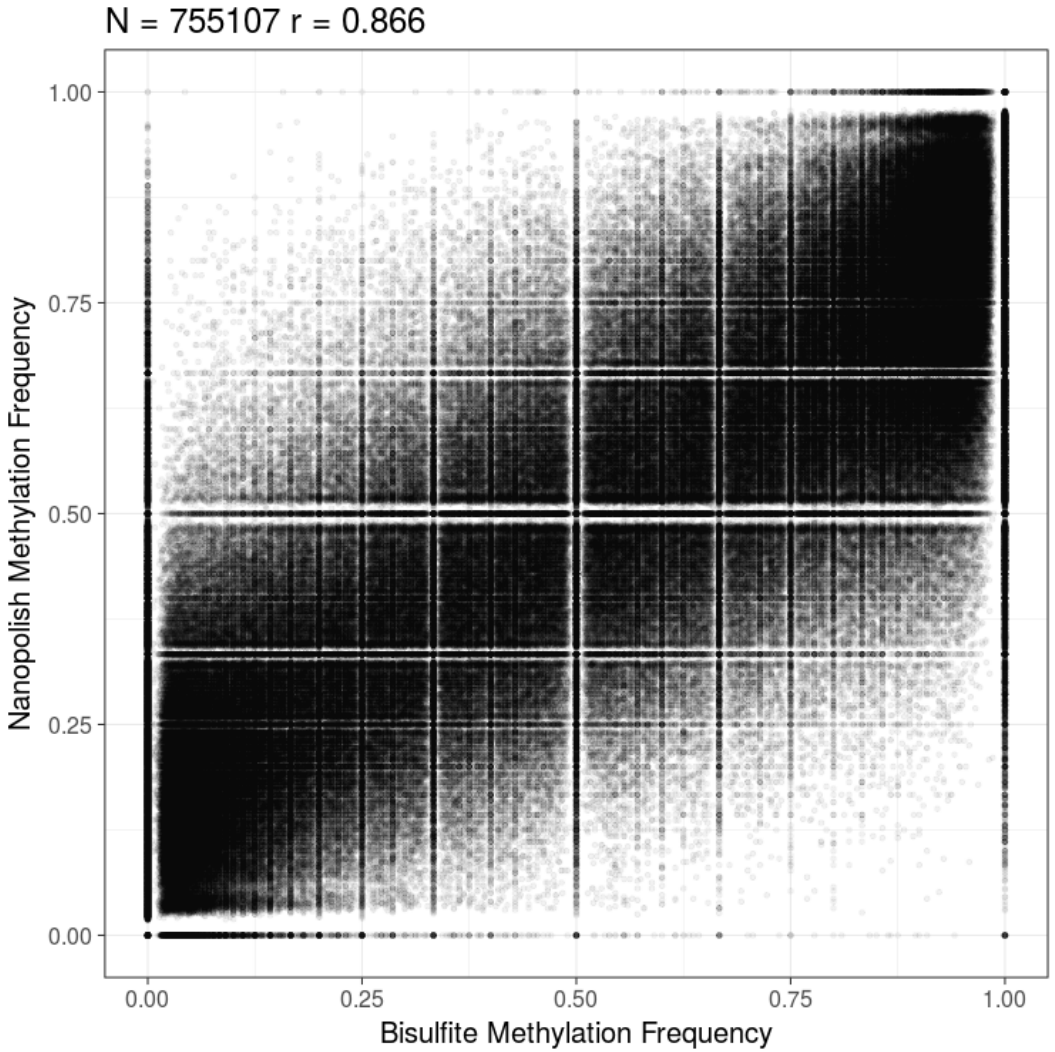


Simpson, Workman, *Nature Methods* (2017)

- 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

Nanopolish Methylation

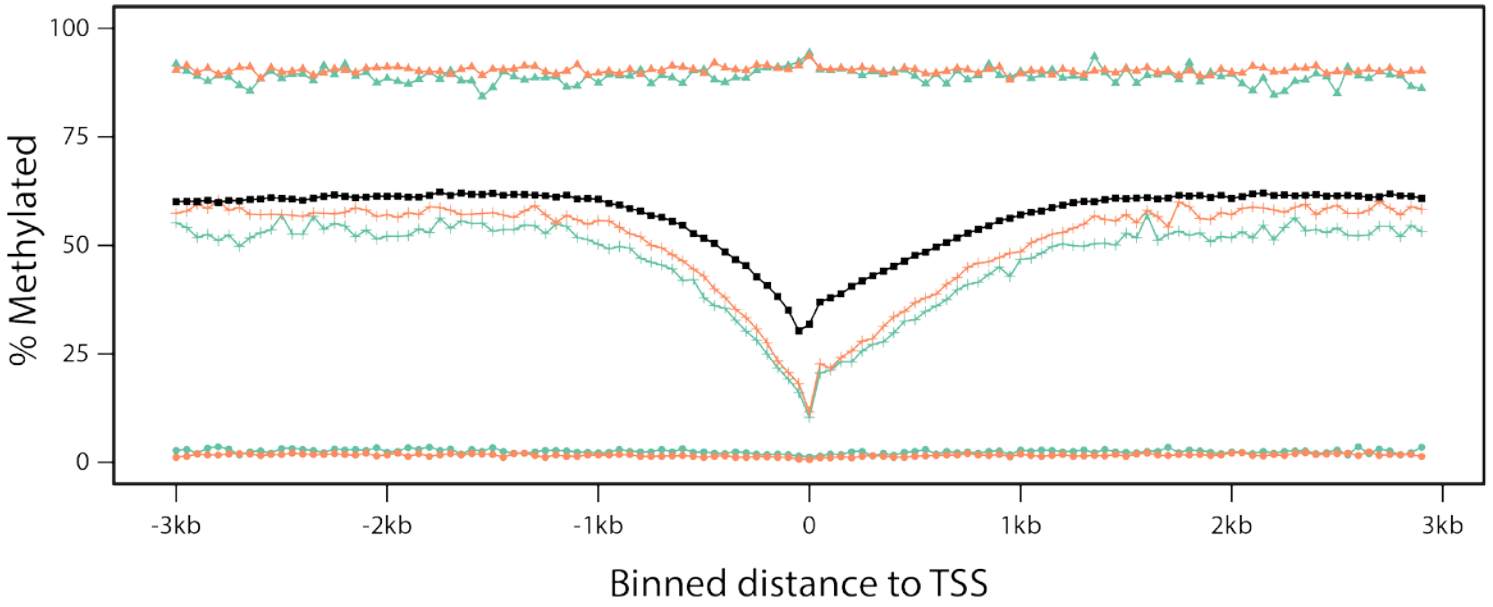


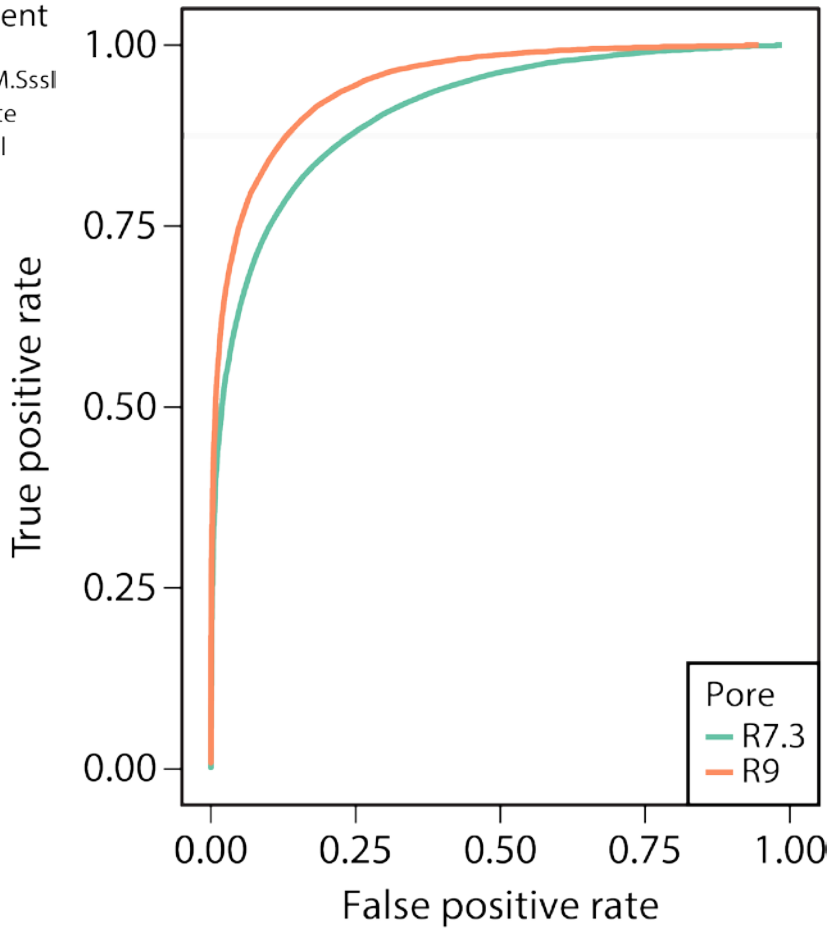


Methylation comparison between bisulfite and nanopore calling
In whole genome

Jain et al *bioRxiv* 128835

Binned Methylation vs. Transcription Start Sites

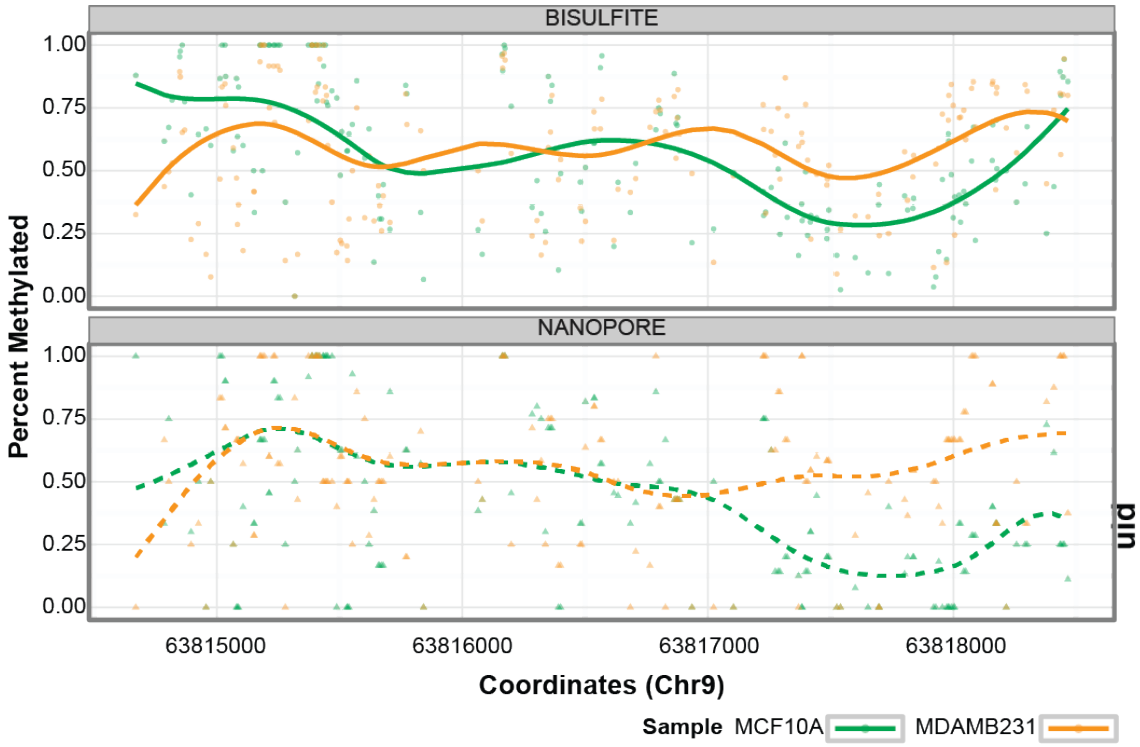


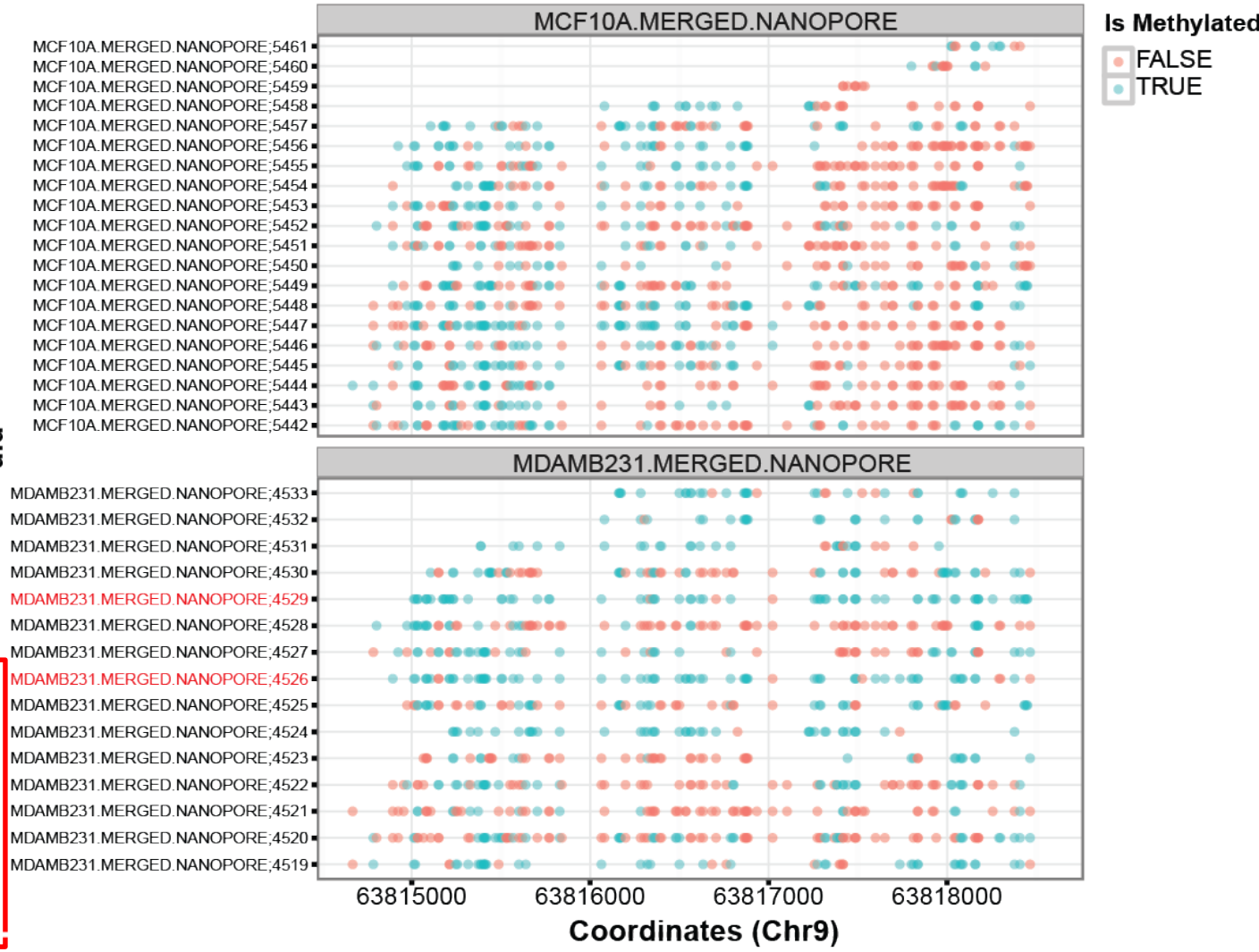


Simpson, Workman, *Nature Methods* (2017)

- 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



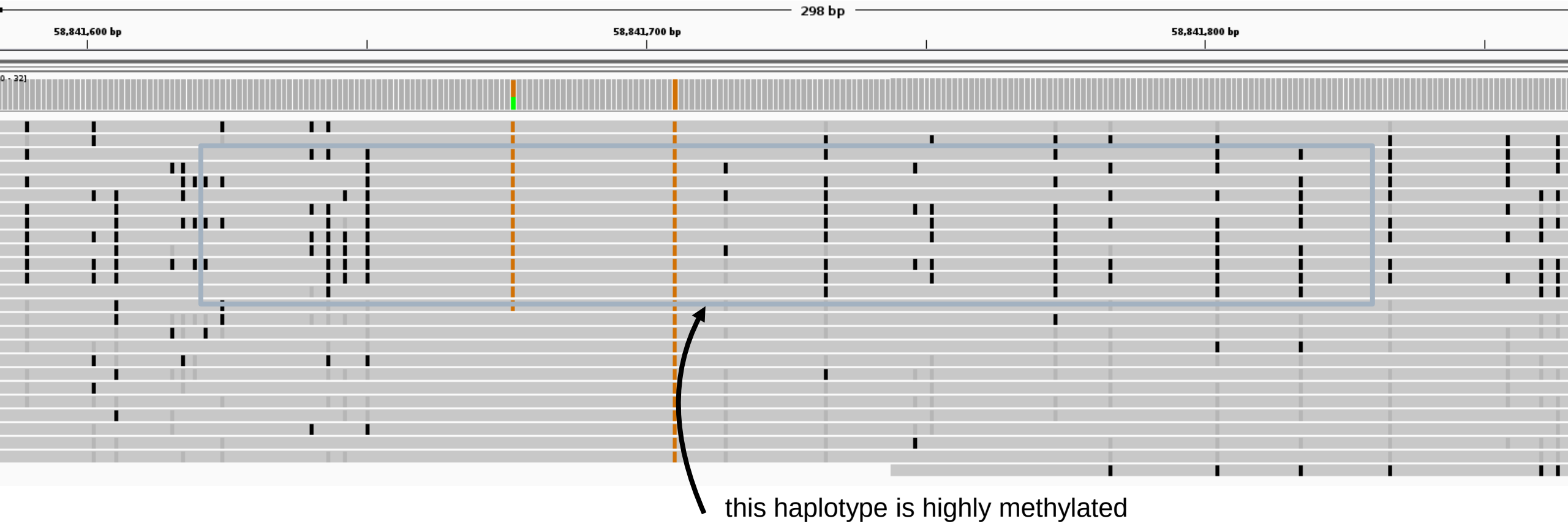


- Reduced representation method:12.5Mb of the genome (3.5-6kb size selection)
- We sequenced this fraction on nanopore and bisulfite Illumina seq
- Long reads measure *phased* methylation

Simpson, Workman, Nature Methods (2017)

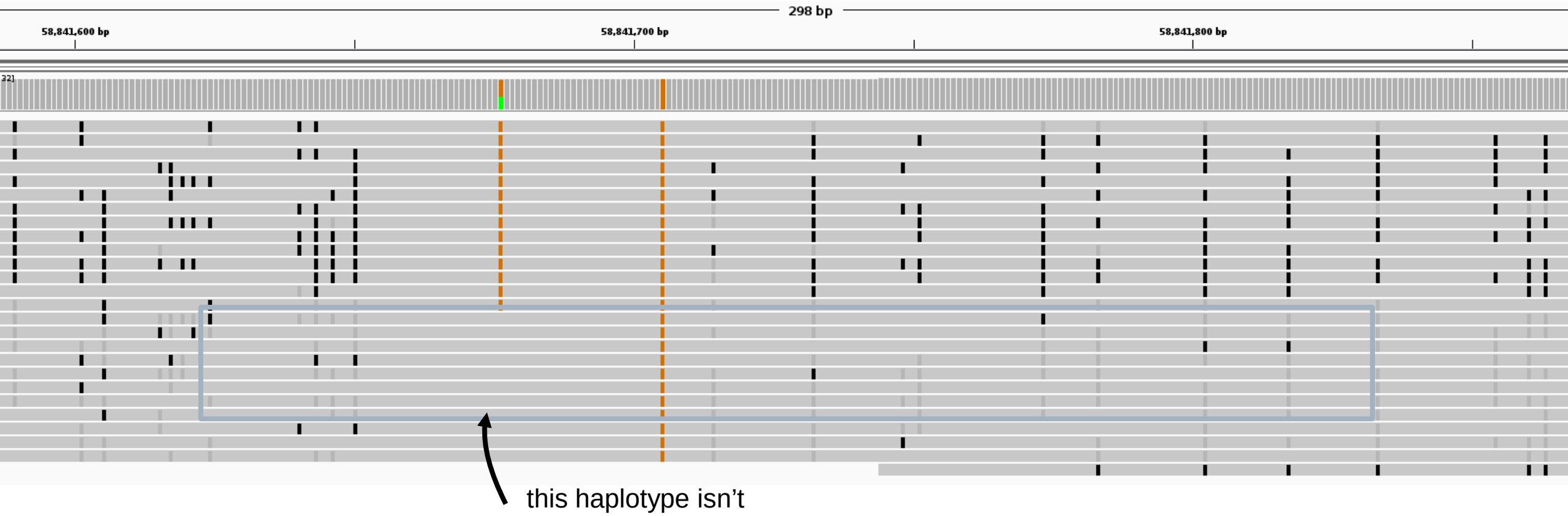
Haplotype-Phased Methylation

nanopolish has experimental support for phasing methylation patterns

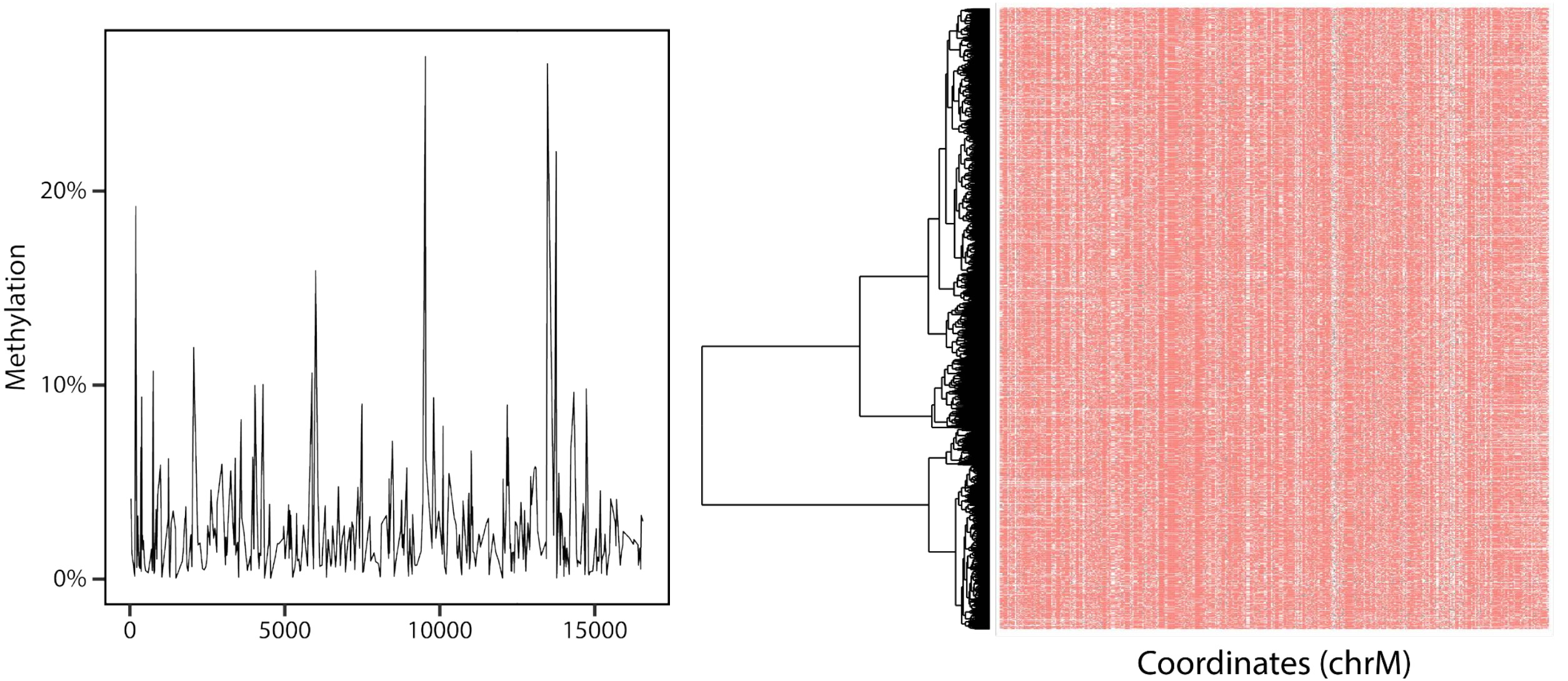


Haplotype-Phased Methylation

nanopolish has experimental support for phasing methylation patterns



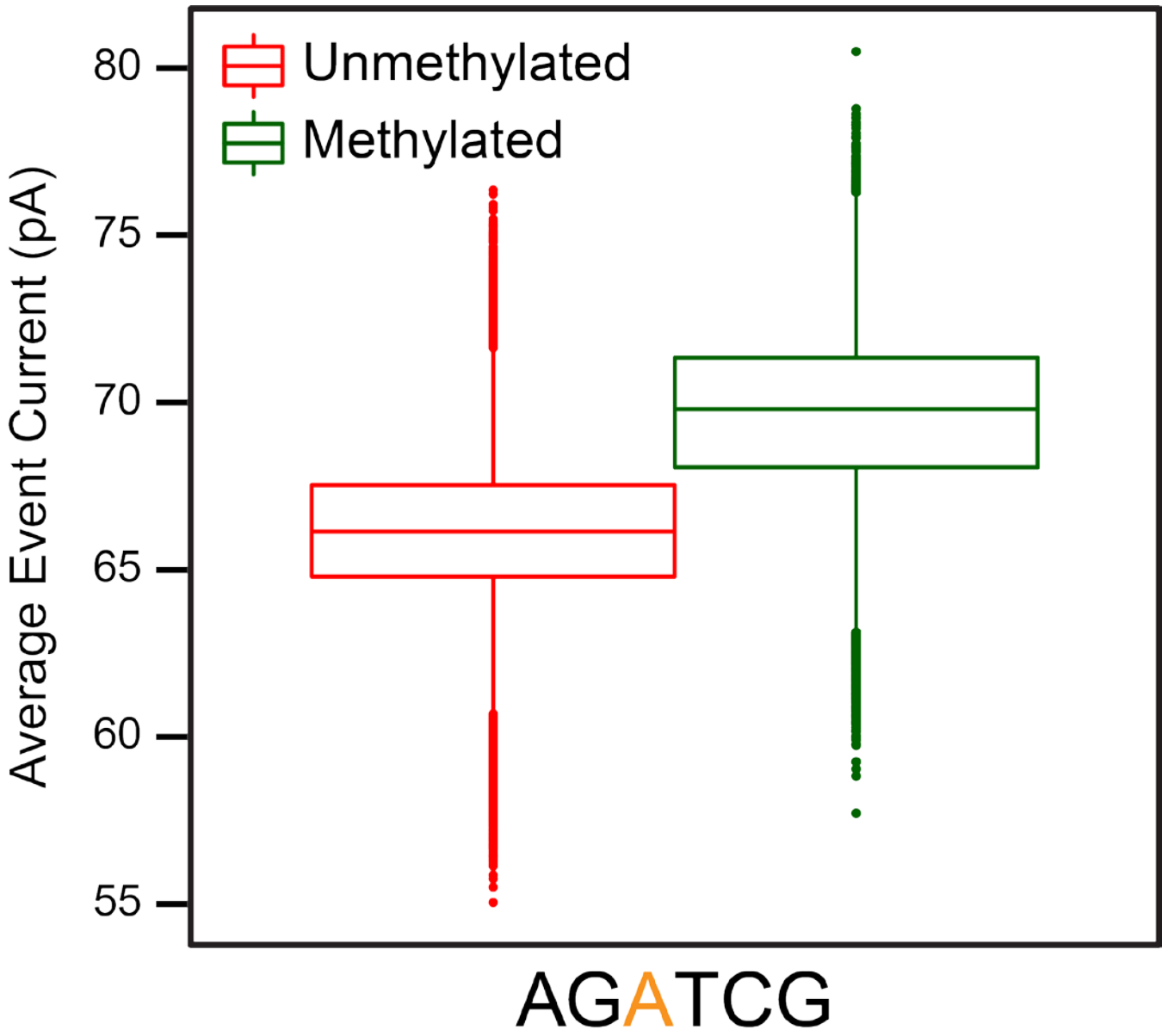
Mitochondrial methylation/clustering



Preliminary data clustering mitochondrial CG methylation from MCF10A

Future Work

- 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
- Exploring replacing the core hidden Markov model with a neural network to capture more of the signal



Acknowledgments

- **Timp Lab – Johns Hopkins University**
- Winston Timp, PhD
- Rachael Workman, MS
- Stephanie Hao, MS
- Yunfan Fan
- Isac Lee
- Amy Vandiver, MD, 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

Agilent Technologies
Josh Wang, PhD
Jonathan Levine, PhD

National Human Genome Research Institute
1R01HG009190-01A1