Long-Read Sequencing Workshop
Jackson Laboratory for Genomic Medicine
Farmington, CT; April 2018

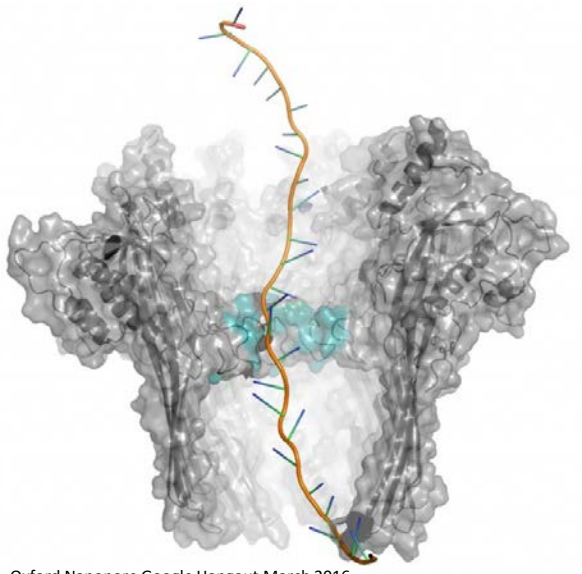
JOHNS HOPKINS
BIOMEDICAL ENGINEERING

Detecting Base Modifications Using Nanopore Sequencing

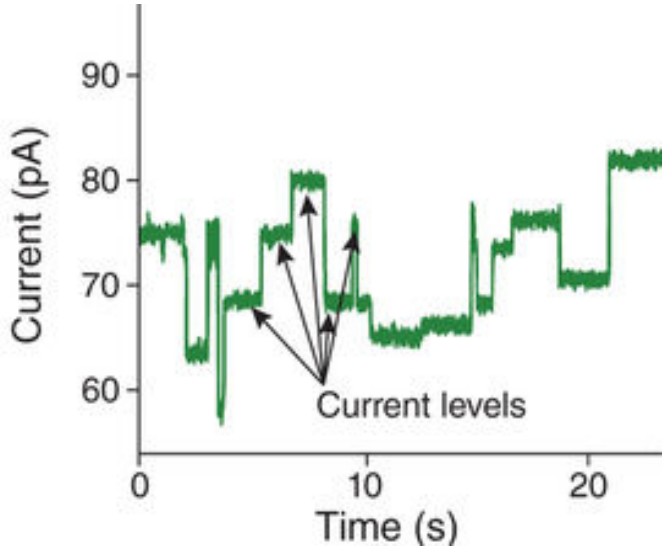
Winston Timp
Department of Biomedical Engineering
Johns Hopkins University

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 5-10Gb



Oxford Nanopore Google Hangout March 2016

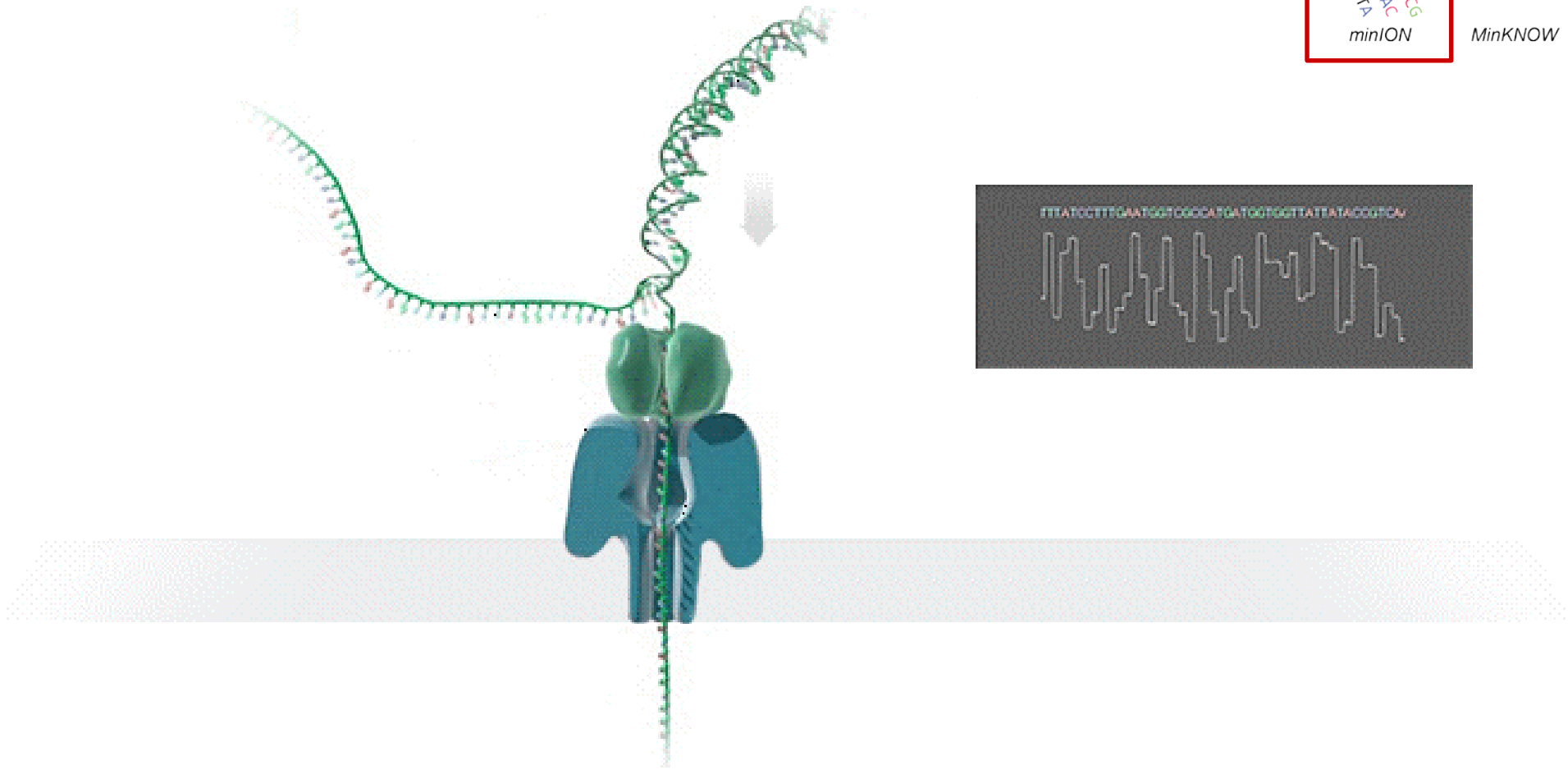


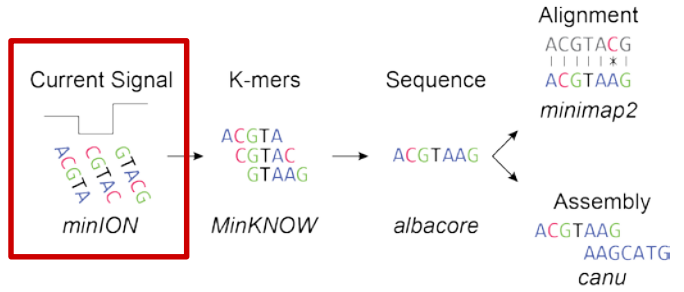
Deamer et al 2016, Nature Biotech

ATCGATCGATAGTAT
TAGATACGACTAGC
GATCAG

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

Sequencing Operation

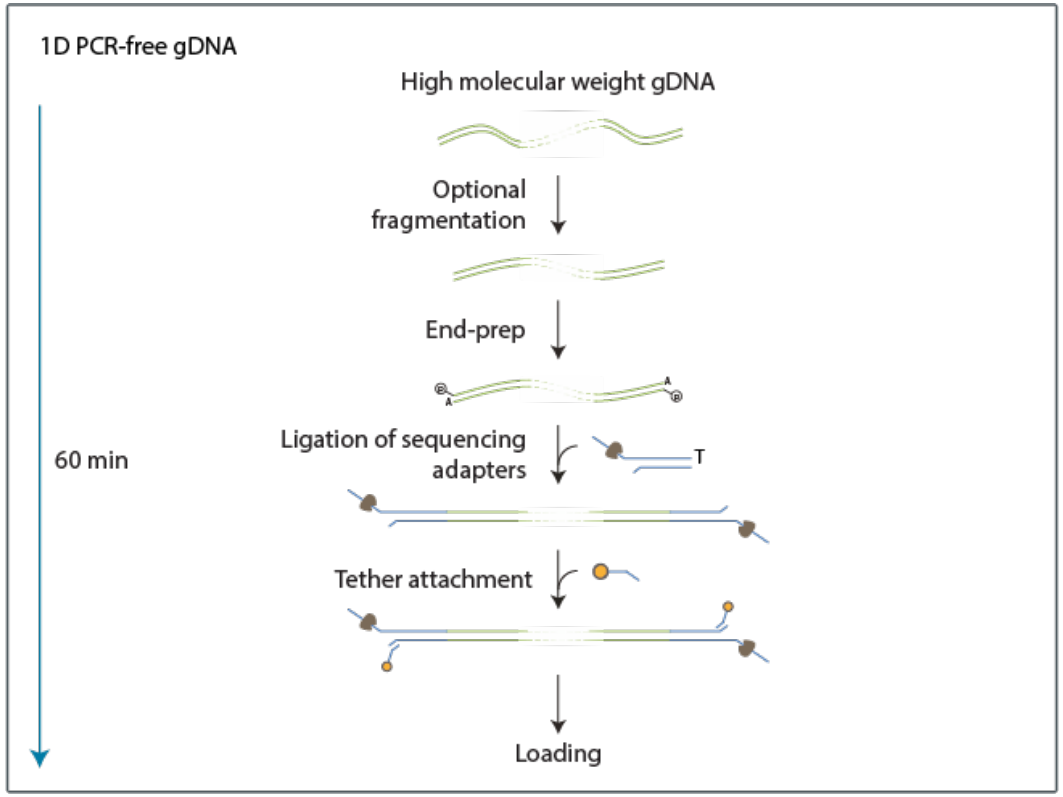


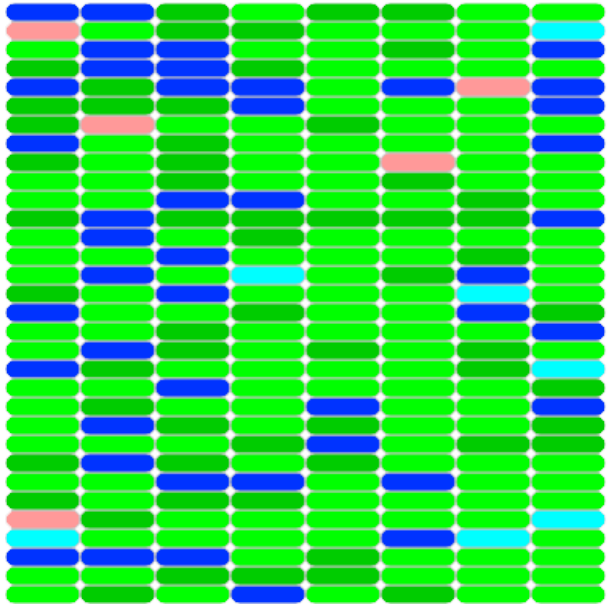


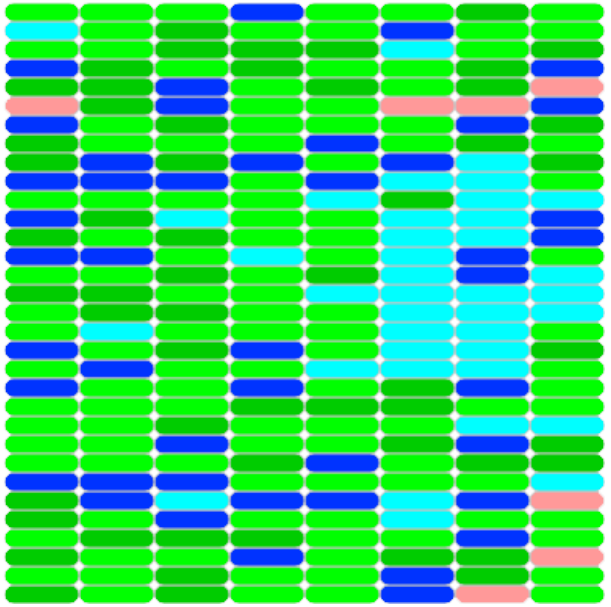
Oxford Nanopore Technologies

- 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

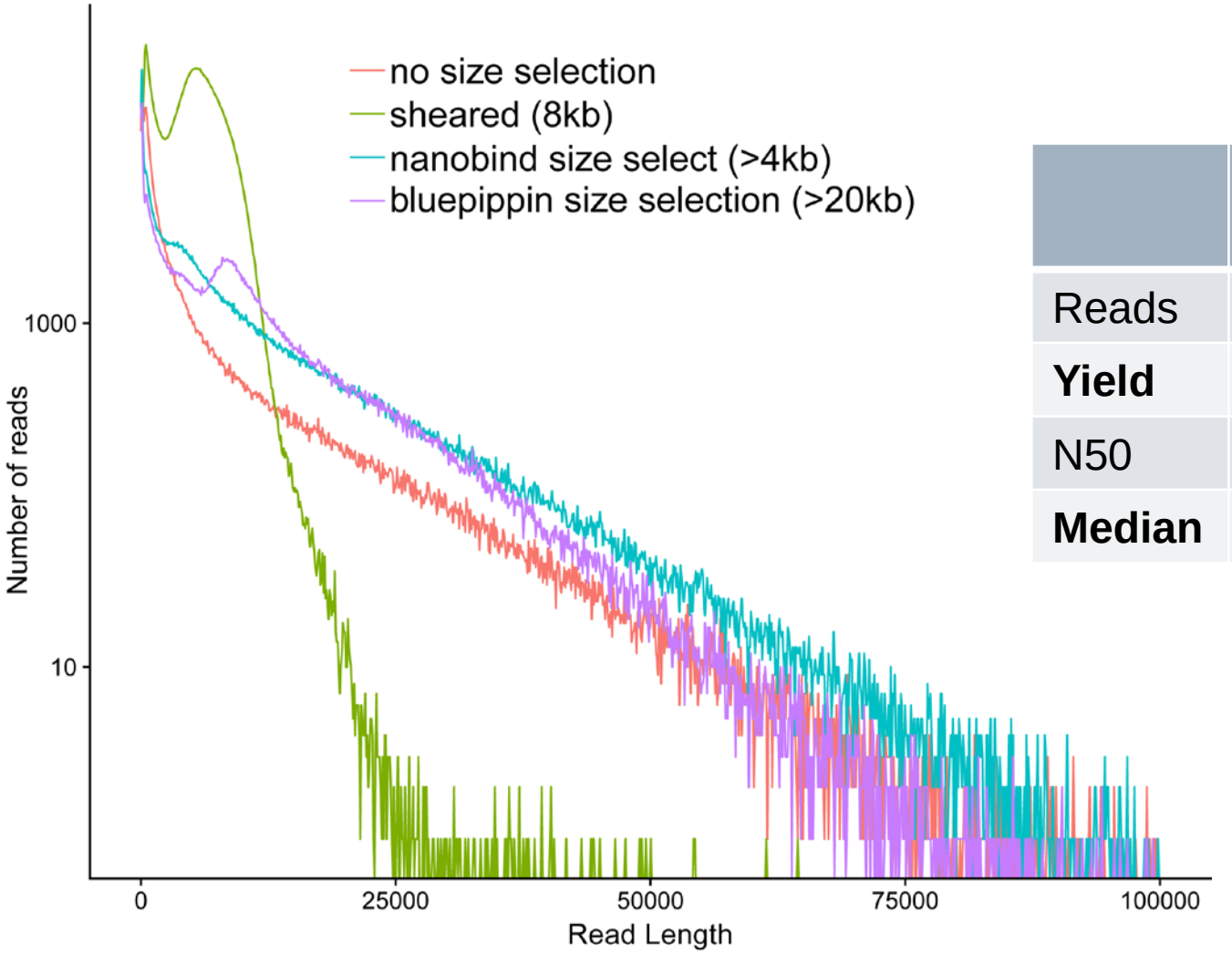






- Library prep is very similar to methods for short-read sequencing
- For DNA shearing we use Covaris gTubes or Diagenode Megaruptor
- After end-repair and A-tailing, leader adapter with motor protein is ligated
- MinION arrays 512 channels (with 4 pores possible per channel) (shown bottom left from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

Improving Read Lengths: Size selection



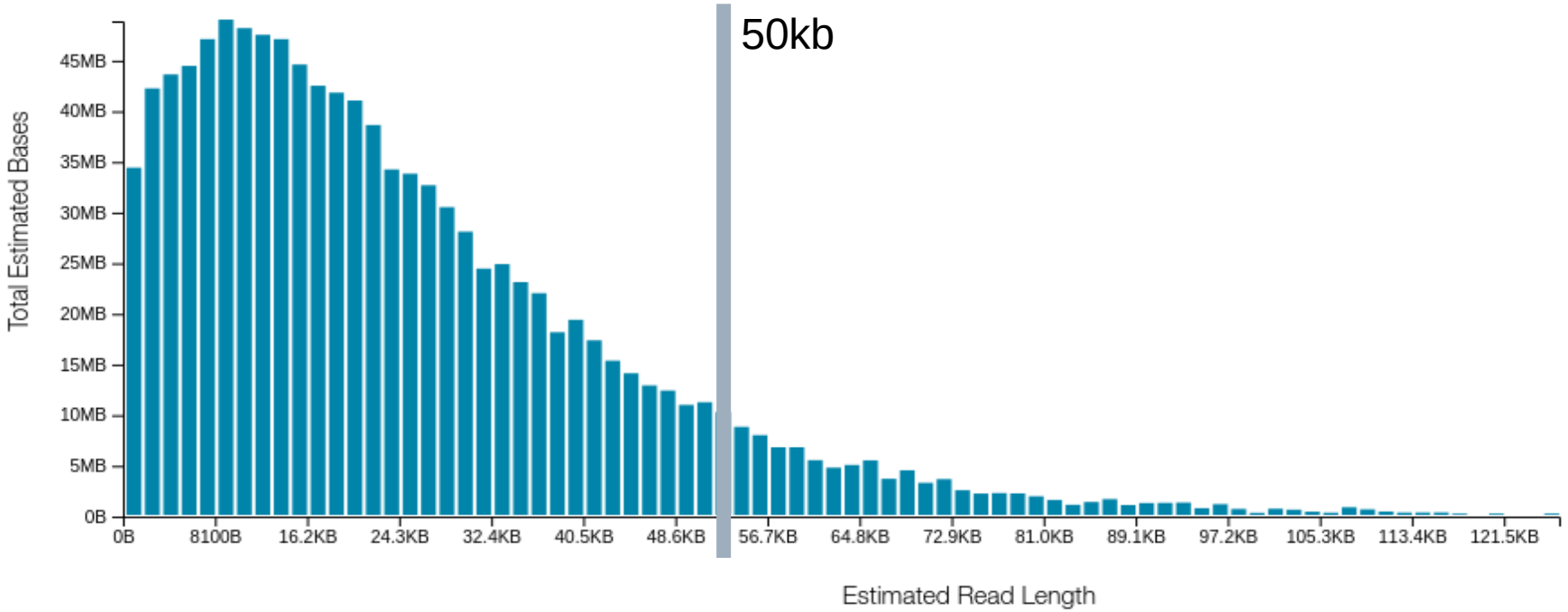
	None	Sheared	Nanobind SS (4kb)	Blue Pippin SS (20kb)
Reads	353k	2060k	400k	435k
Yield	**1.71Gb**	**10.1Gb**	**3.57Gb**	**3.65Gb**
N50	17.3kb	6.6kb	15.7kb	19.0kb
Median	**1.2kb**	**5.1kb**	**6.8kb**	**4.3kb**

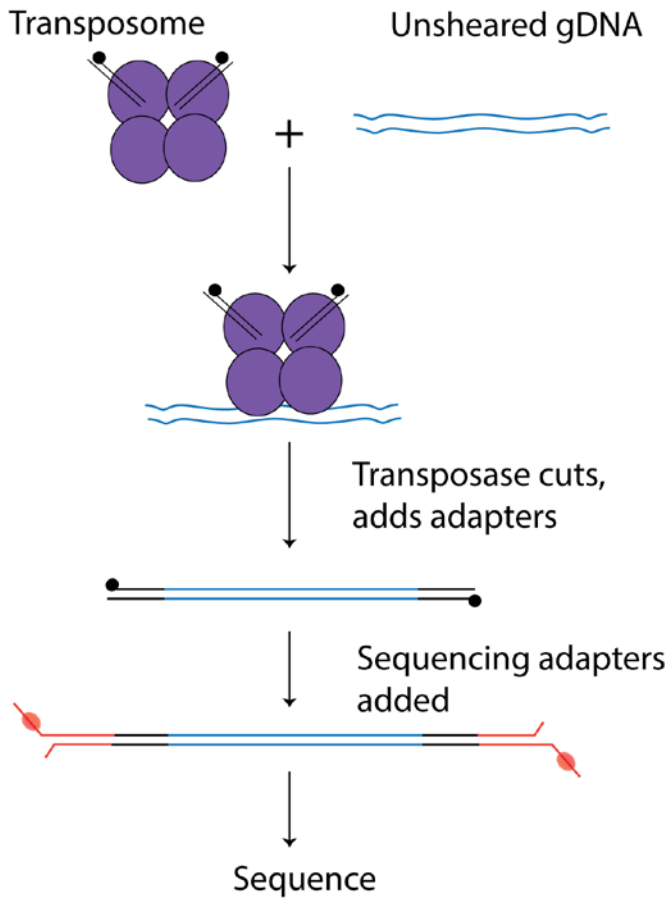
Read length and yield require some optimization and trade-offs

Improving Read Lengths: Rapid kit RAD004

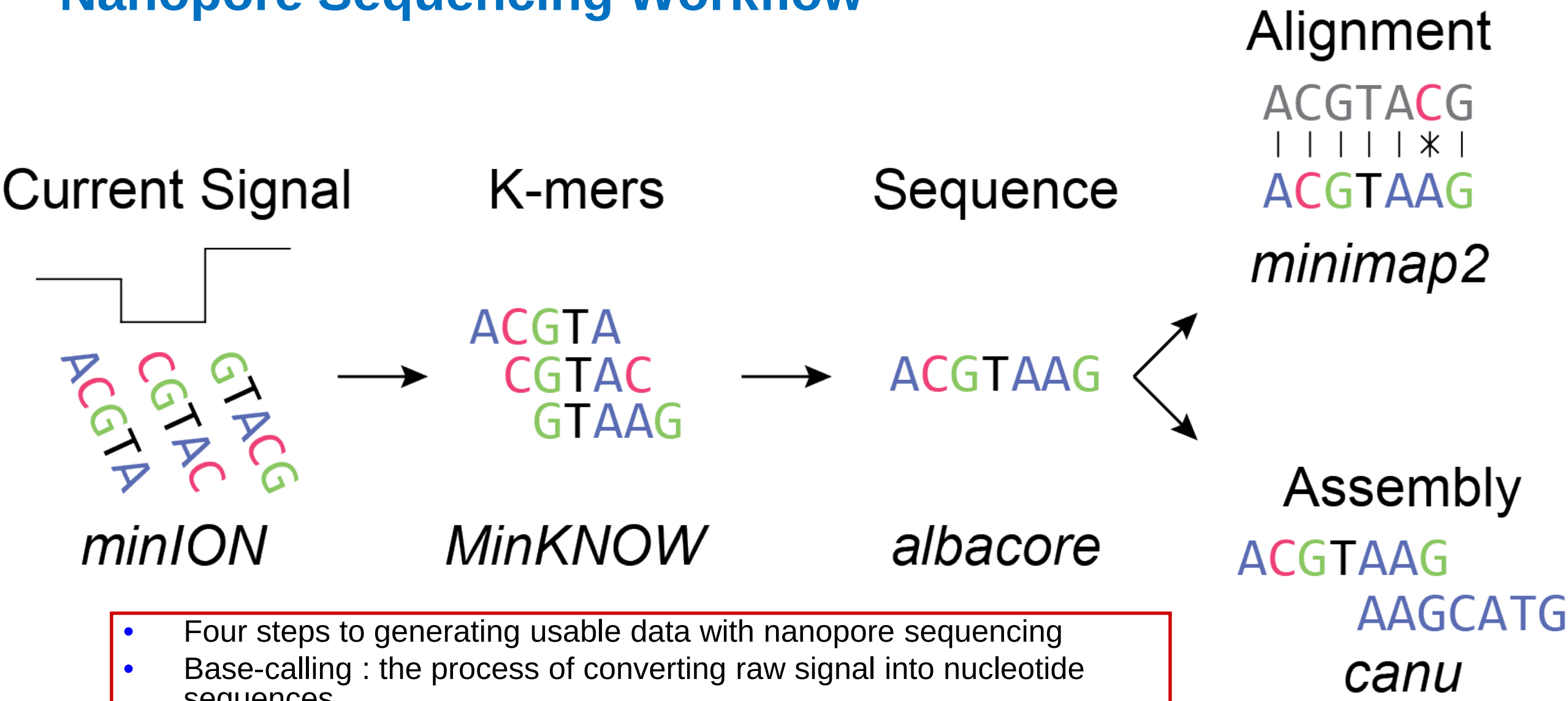
15 minute protocol

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



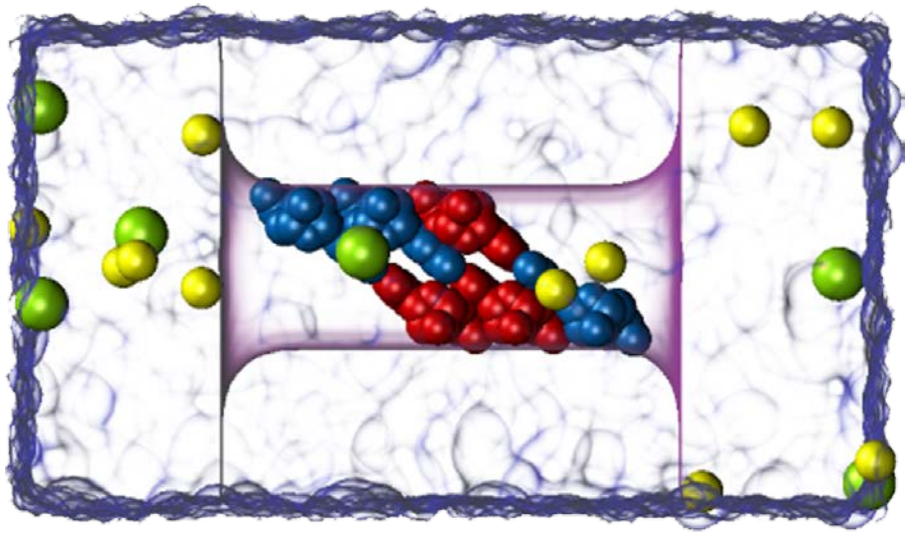


Nanopore Sequencing Workflow

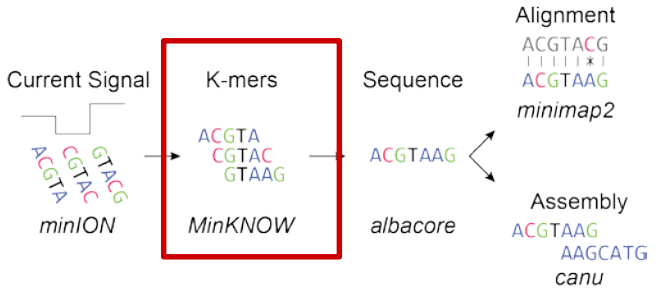


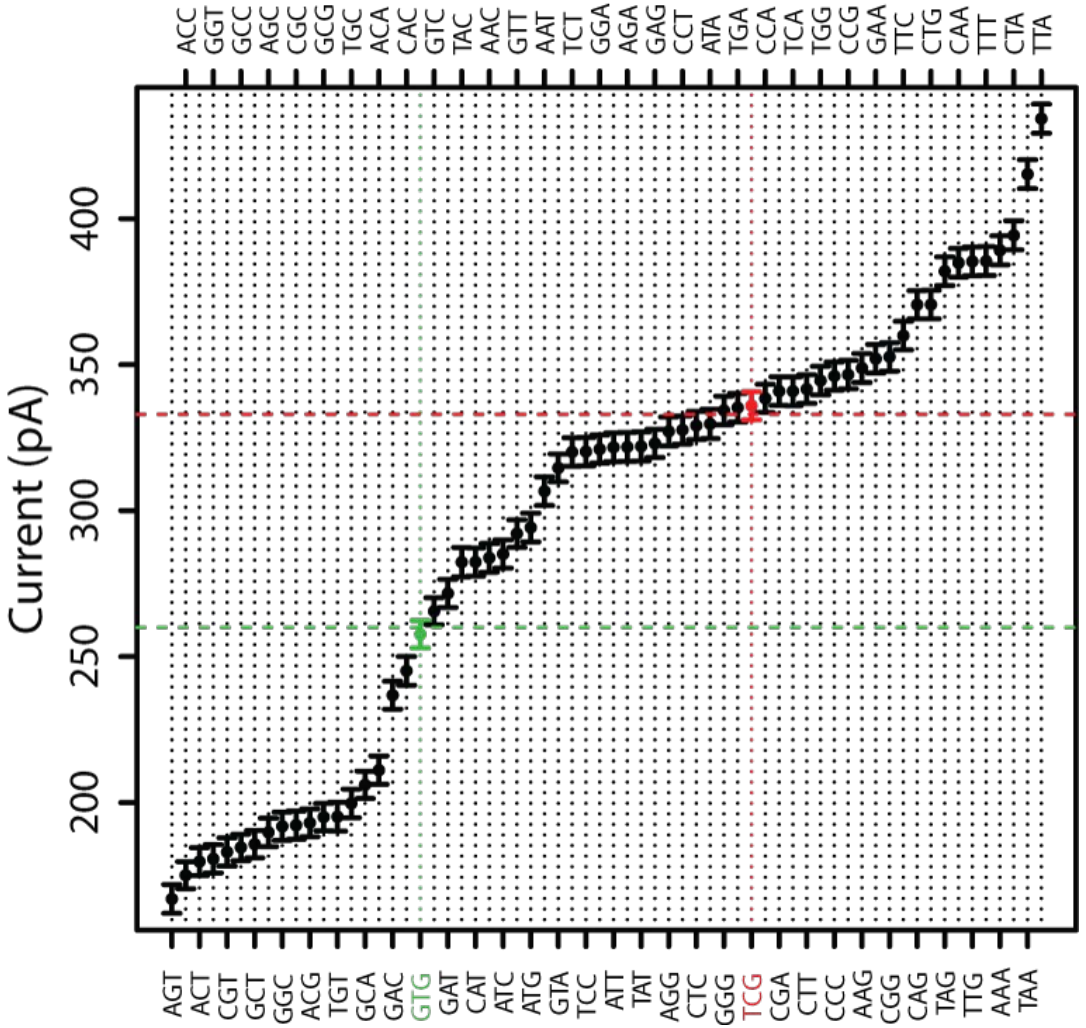
- Four steps to generating usable data with nanopore sequencing
- Base-calling : the process of converting raw signal into nucleotide sequences
- Nanopolish : uses alignment and current signal to improve base-calls

Problems with Nanopore basecalling



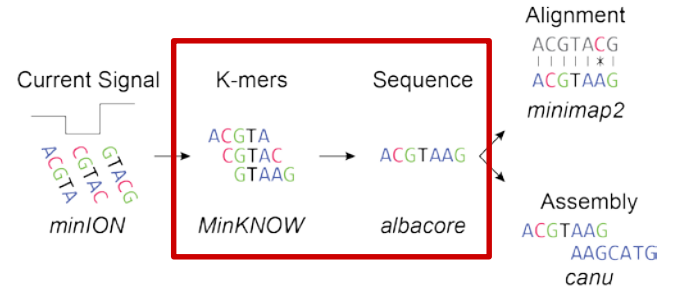
- Multiple bases influence the current passing through the pore.
- Through simulation with Brownian Dynamics, we calculated the contribution from triplets of DNA in a solid-state nanopore - 64 current levels.
- Not all of these different currents are distinguishable

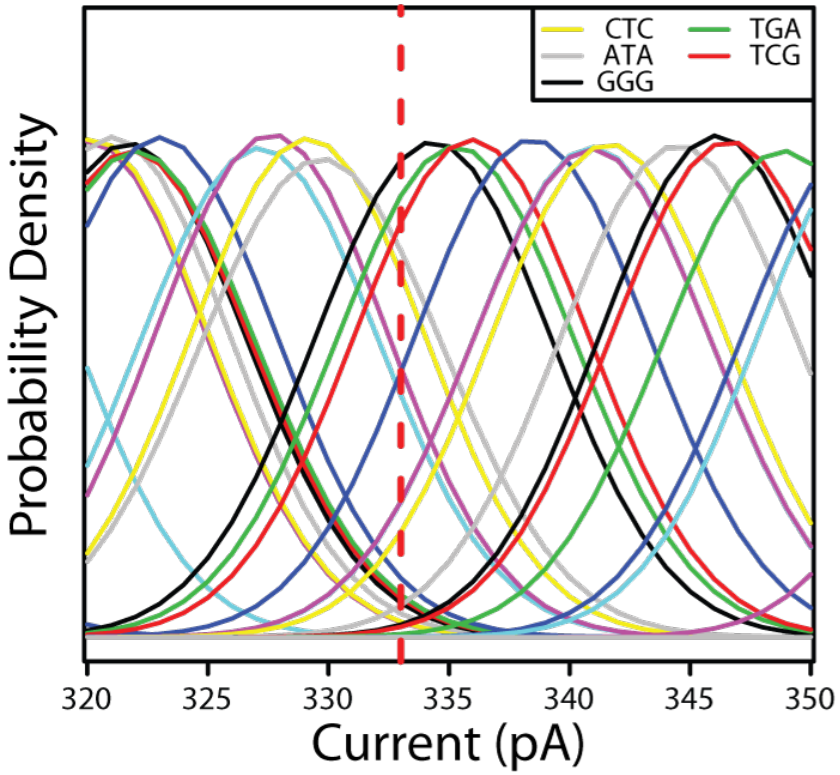


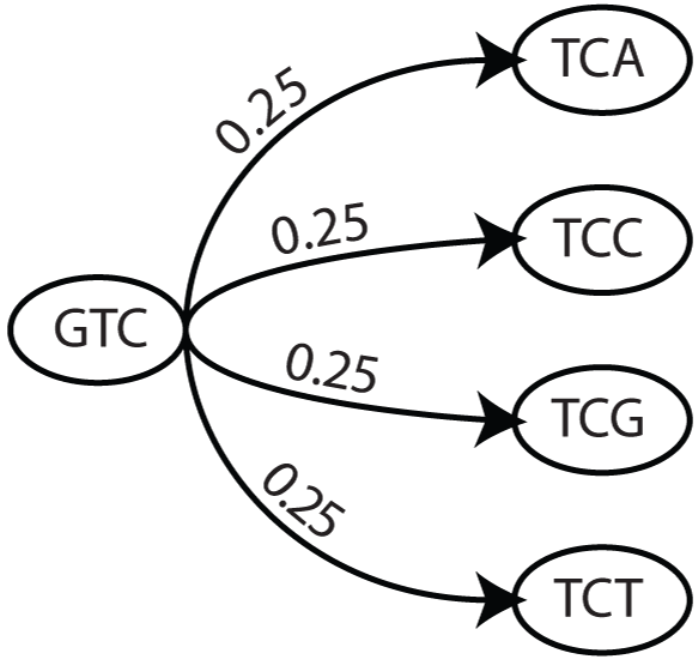


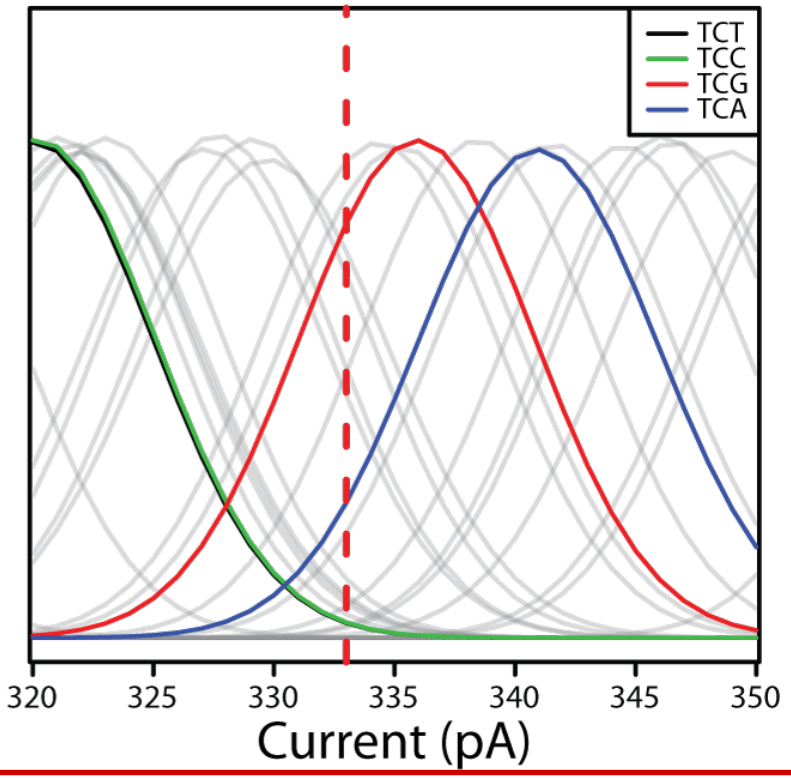
Comer and Aksimentiev *J. of Phys Chem C* 116(5) 3376-3393 (2012)

Prior Information for Decoding







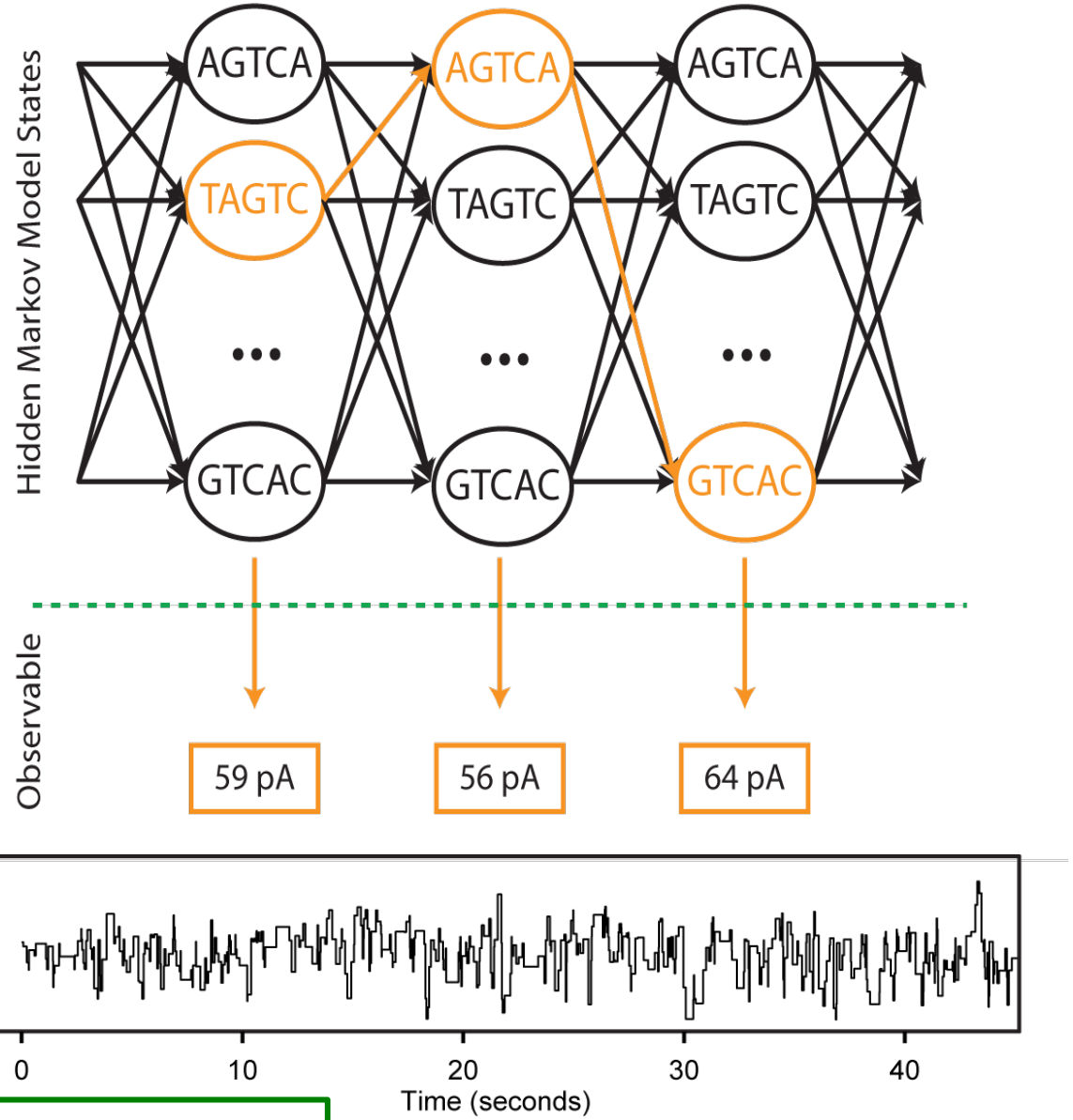


- With no prior information, a given current value may not be called correctly (333pA would be called as GGG)
- If we know the previous triplet, the next triplet is well defined, leaving only four possibilities, resulting in the correct call of TCG

Timp, et al. *Biophys J.* 30, 349-353 (2012)

Nanopore HMM basecalling

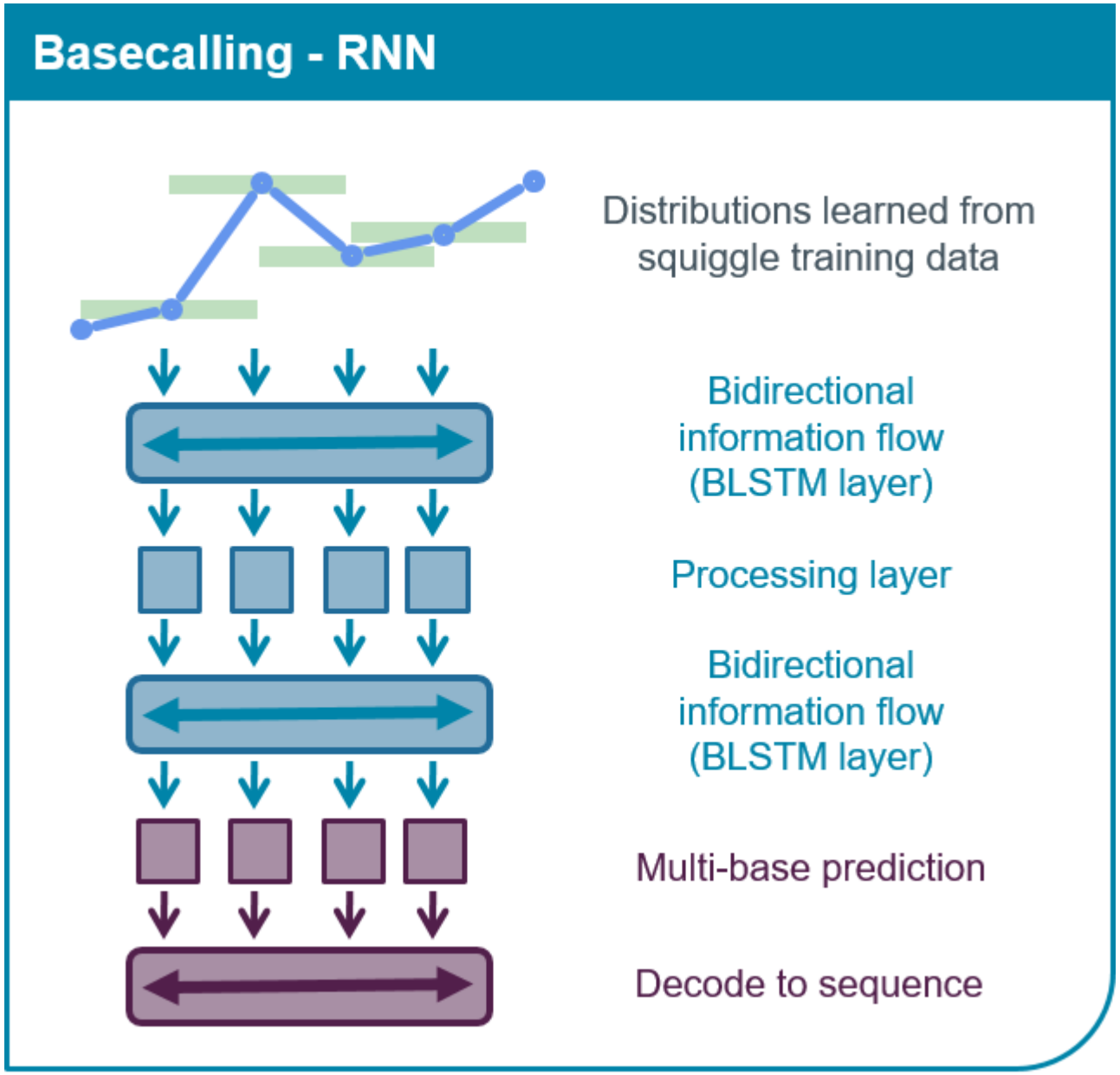
- By using a sequence of observables and maximizing the total joint probability given below, we find the sequence of states.
- This is done using the Viterbi algorithm – which grows, finding the most likely path for each step, saving the probabilities, to avoid recalculation.
- 1st generation basecallers from Oxford used a HMM for basecalling similar to the one detailed in our Biophysical paper
- Transistion probablility matrix for oxford seems to allow for a 0, 1 (most common), 2, or 5 (reset) move.
- We think that Oxford trained its basecalling model on unmethylated lambda



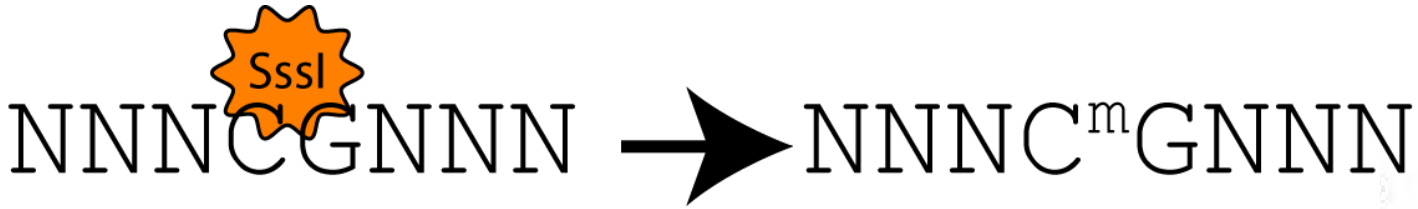
$$\delta_k \prod_t P(I(t) \mid k_t) \times T_{(t-1)(t)}$$

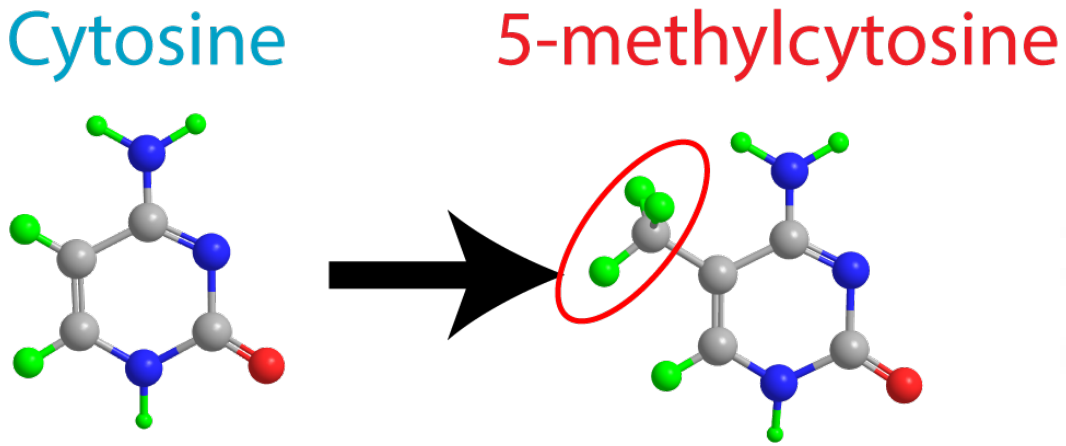
Basecalling shifting to RNN

- Recently (over the past year) there has been a shift to neural network based basecalling
- A *recurrent* neural network is still one with memory, that has a dependence on past computations
- Specifically two layers of Bidirectional Long Short Term Memory (BLSTM)
- These still require the same “training” data to learn what current distributions correspond to which k-mers – and the results are still k-mer based, as multiple bases still influence the current.

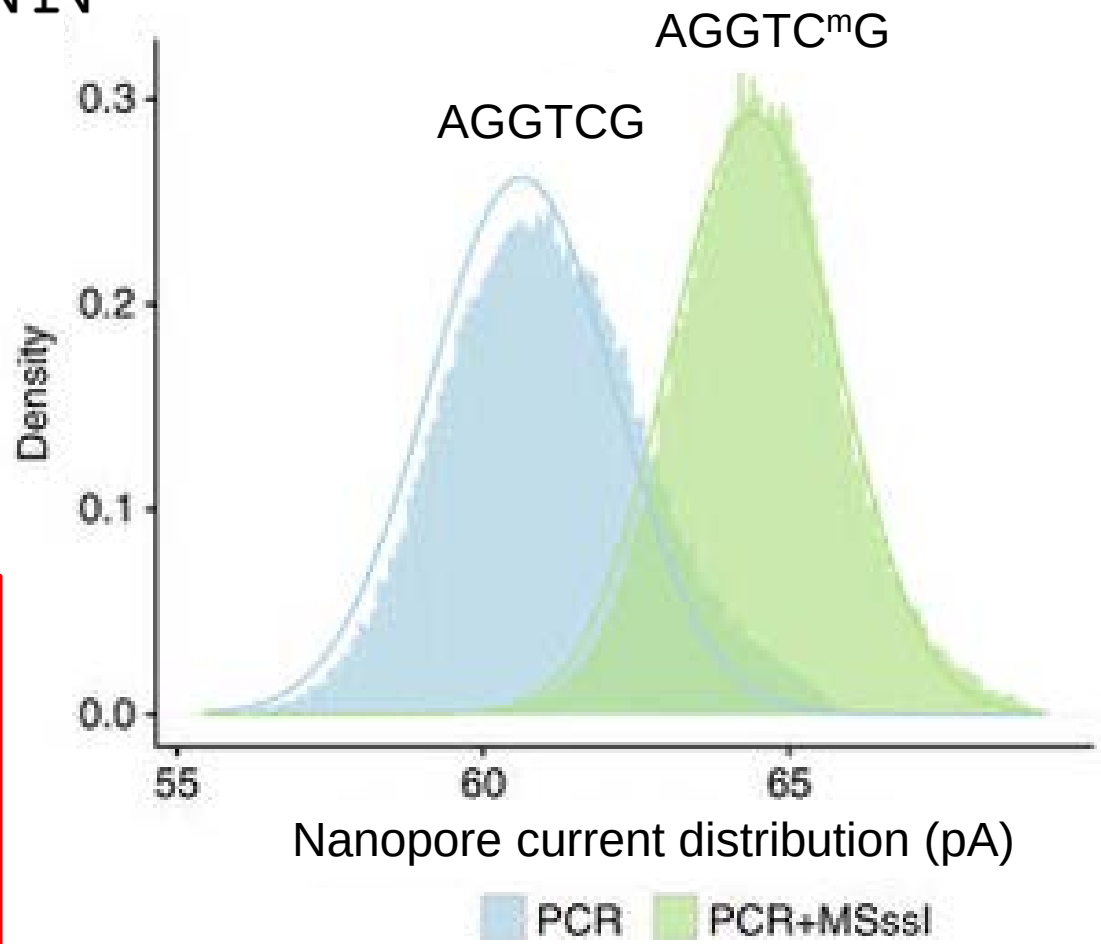


Nanopore Sequencing in Epigenomics



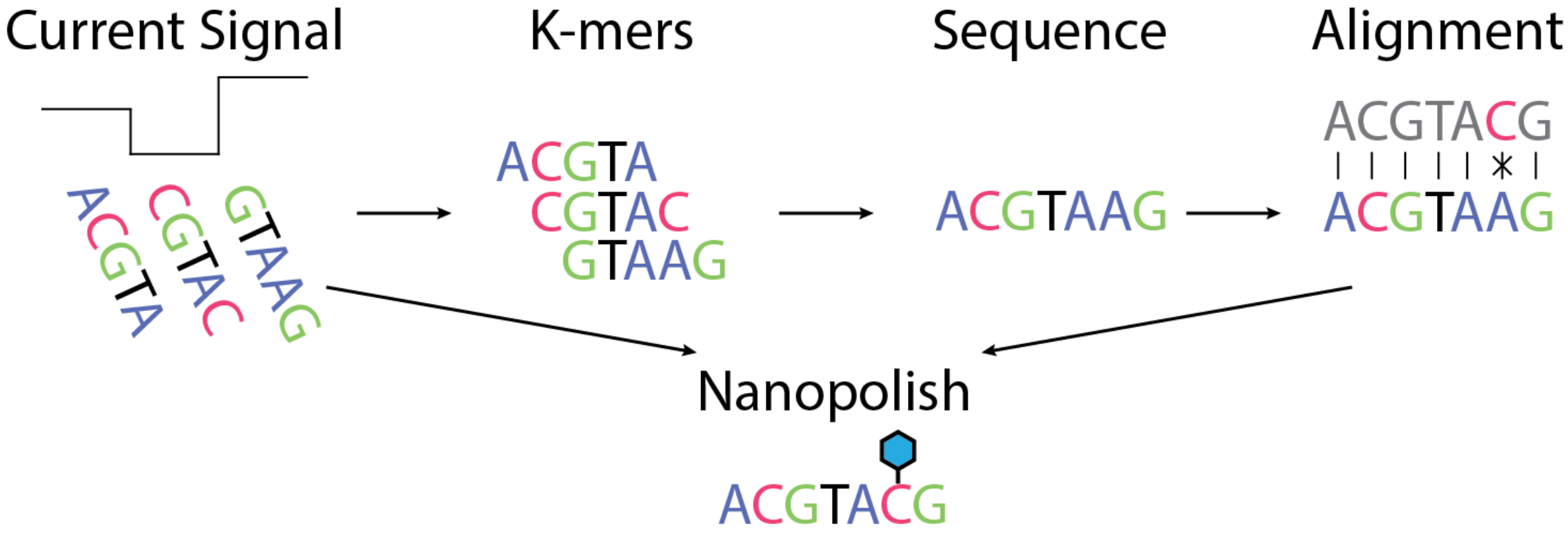


- To generate methylated samples, we treat unmethylated DNA (PCR amplified E. Coli gDNA) with M. SssI methyltransferase
- Distributions of observed current for AGGTCG demonstrate the type of signal between methylated and unmethylated k-mers



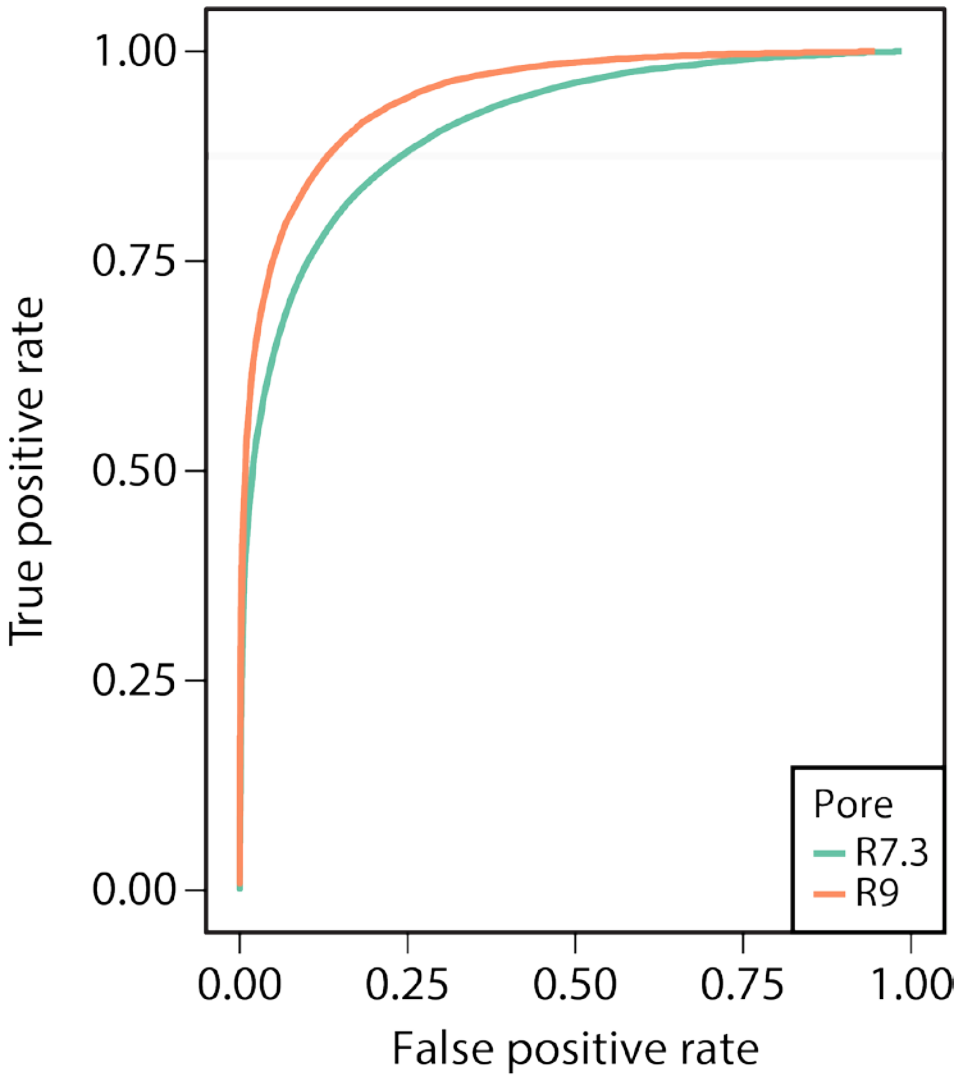
Simpson et. al *Nature* Methods 2017

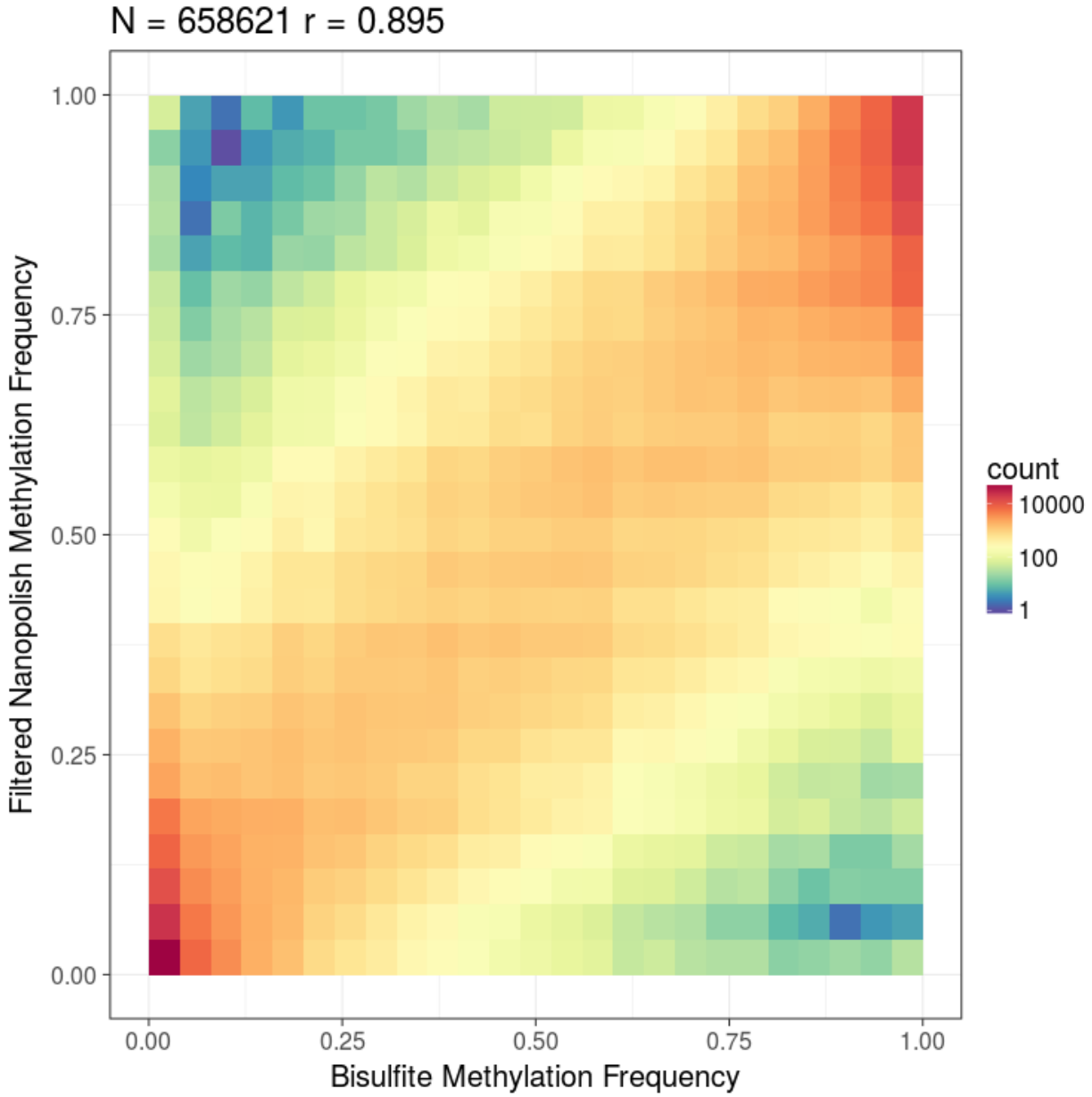
Nanopore: nanopolish methyltrain



- 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, and threshold for >2.5 as methylated; <2.5 as unmethylated

Nanopolish Methylation

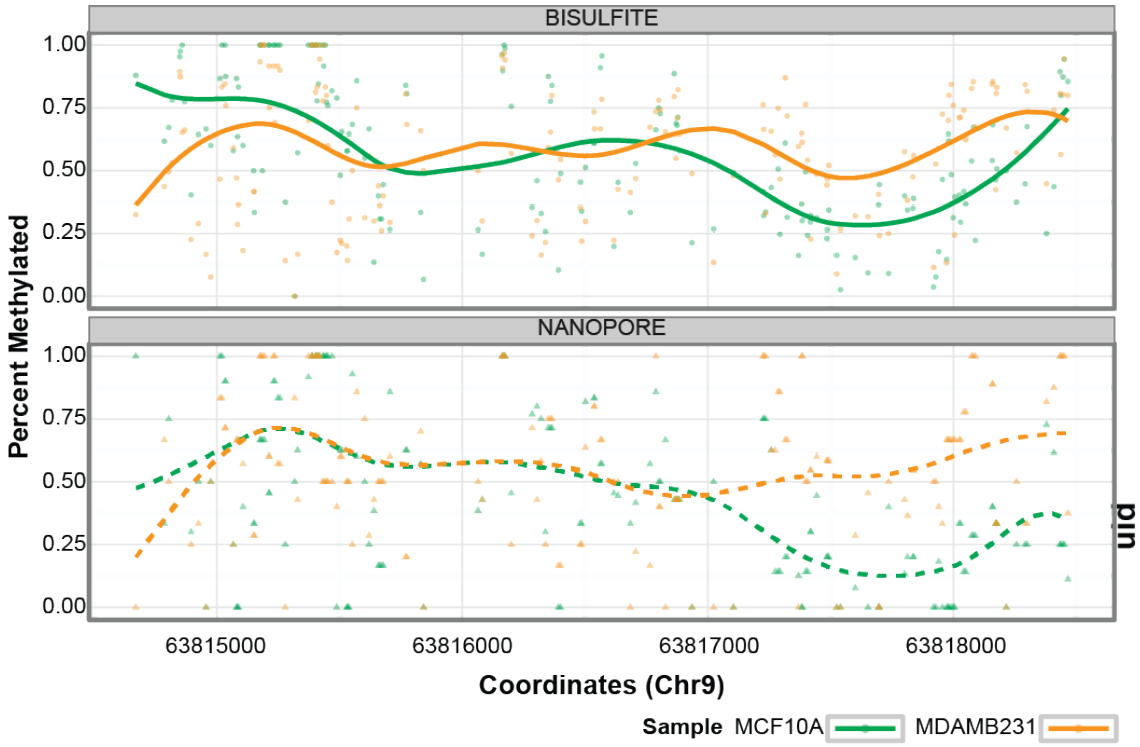


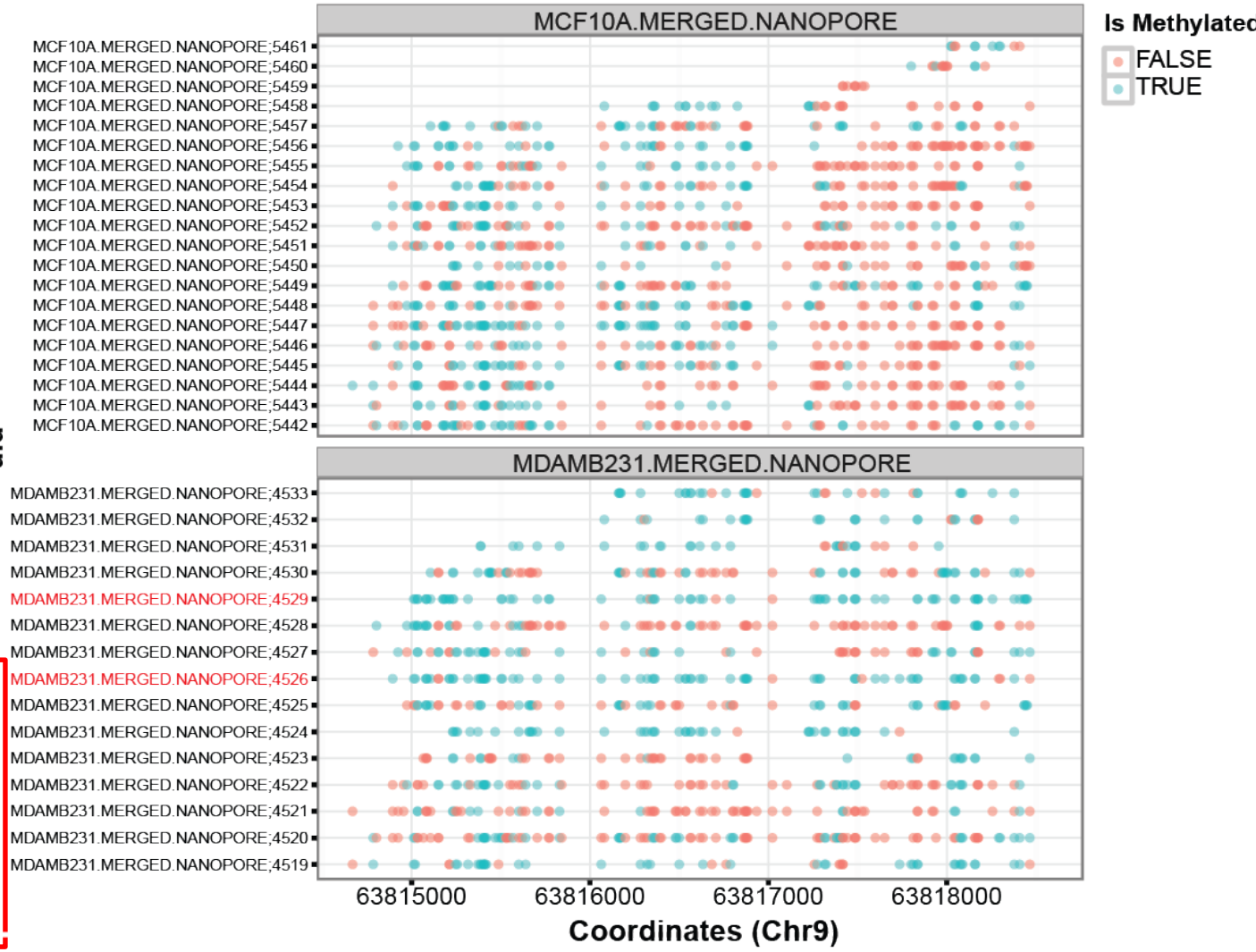


R9 calculates methylation 94% accurate at 77% of sites
NA12878 data shows .895 correlation with bisulfite

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

Cancer-Normal Comparison



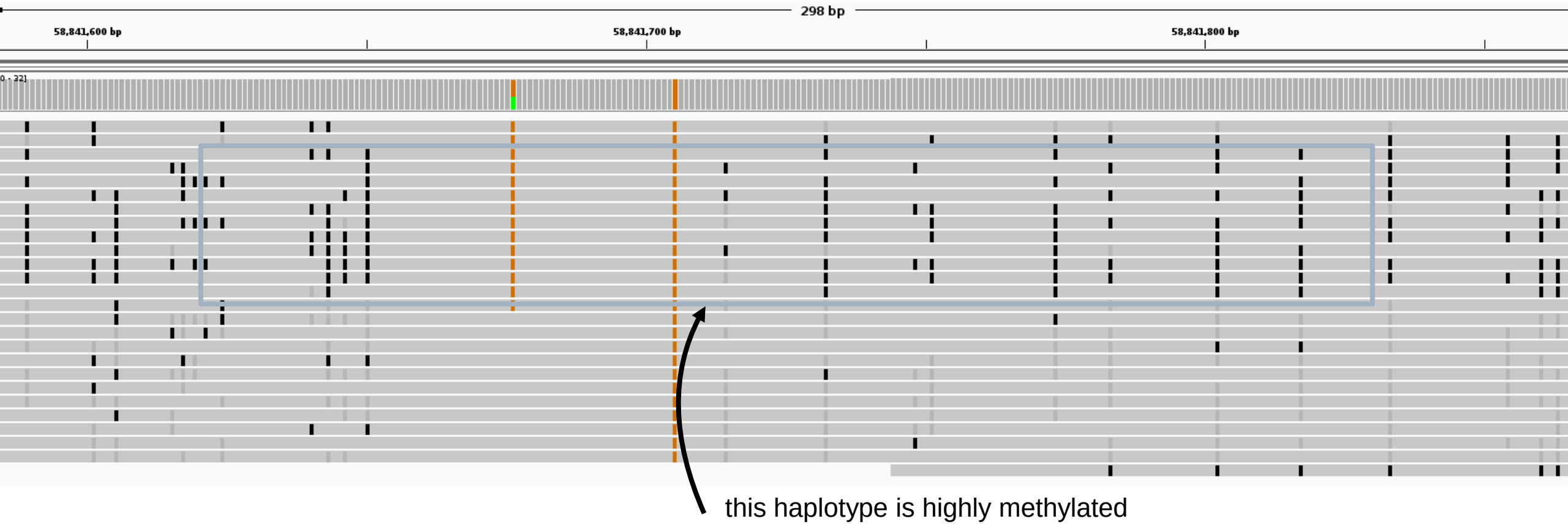


- Reduced representation method:12.5Mb of the genome (3.5-6kb size selection)
- We sequenced this fraction on nanopore and bisulfite Illumina seq
- Long reads measure *phased* methylation

Simpson, Workman, Nature Methods (2017)

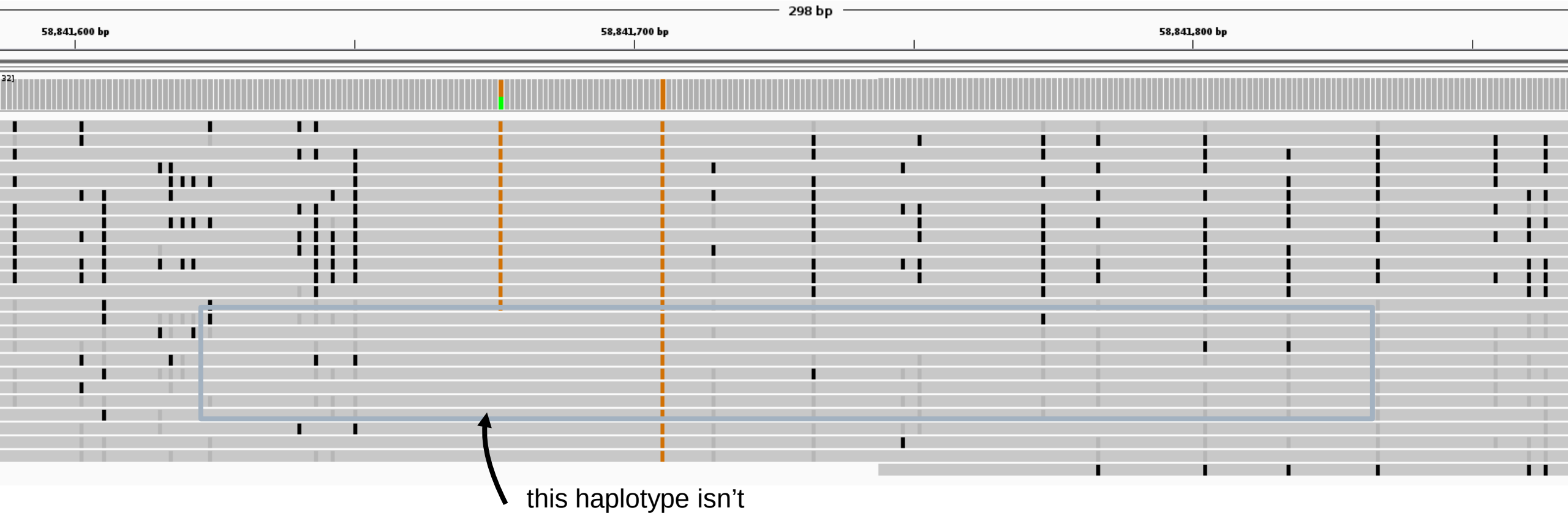
Haplotype-Phased Methylation

nanopolish has experimental support for phasing methylation patterns

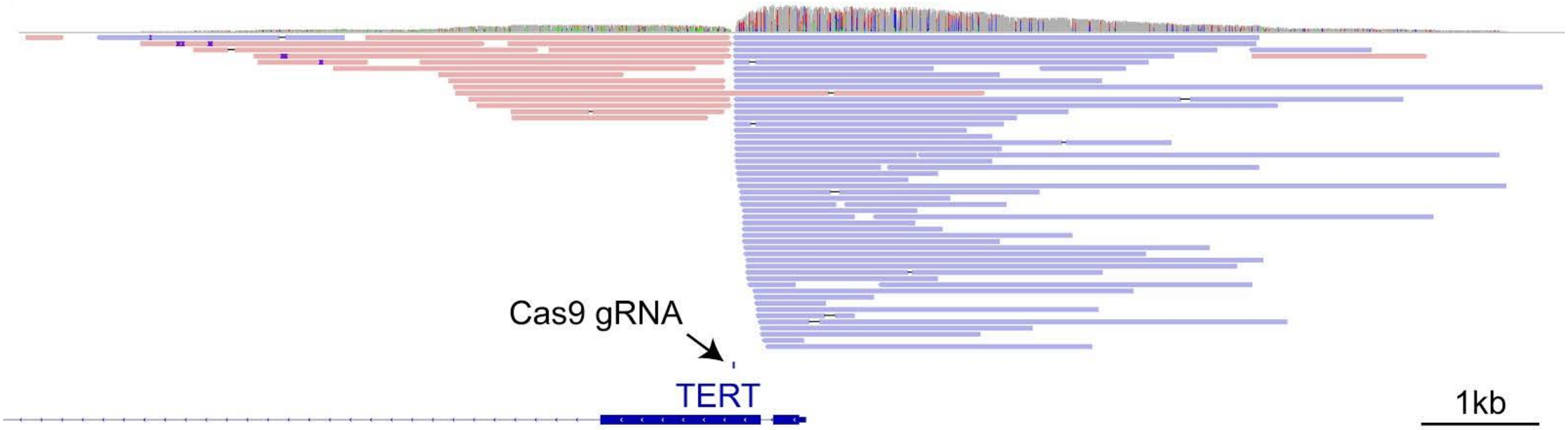


Haplotype-Phased Methylation

nanopolish has experimental support for phasing methylation patterns



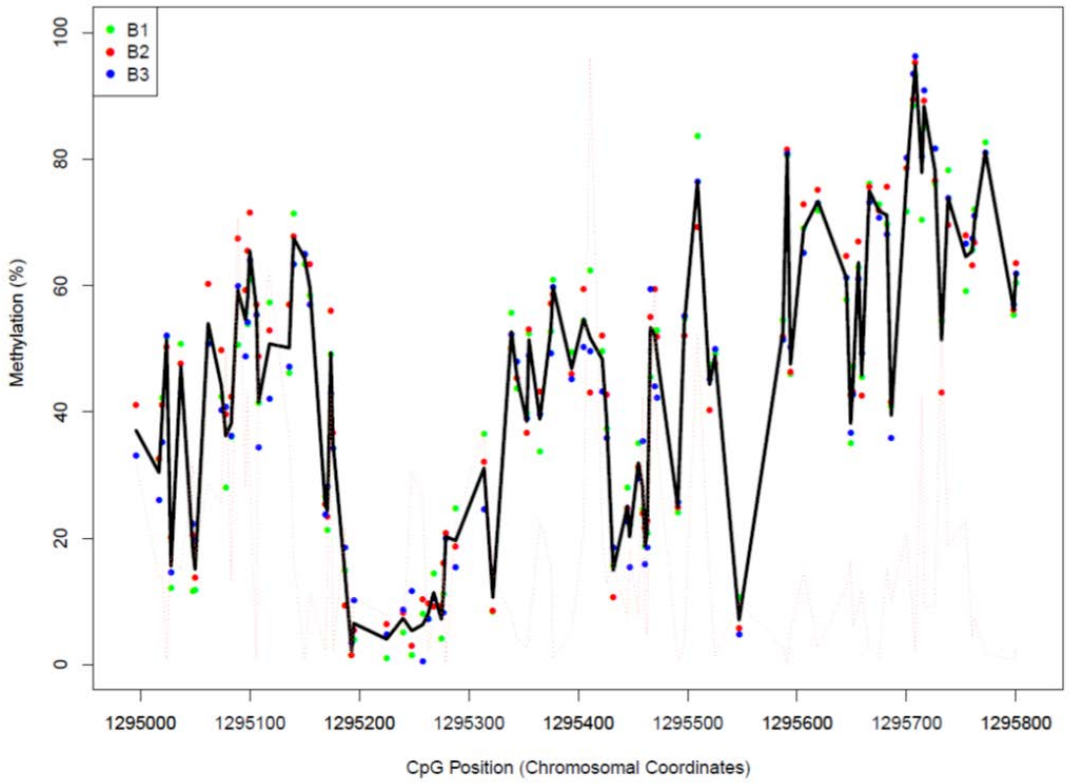
Cas9 Enrichment around target



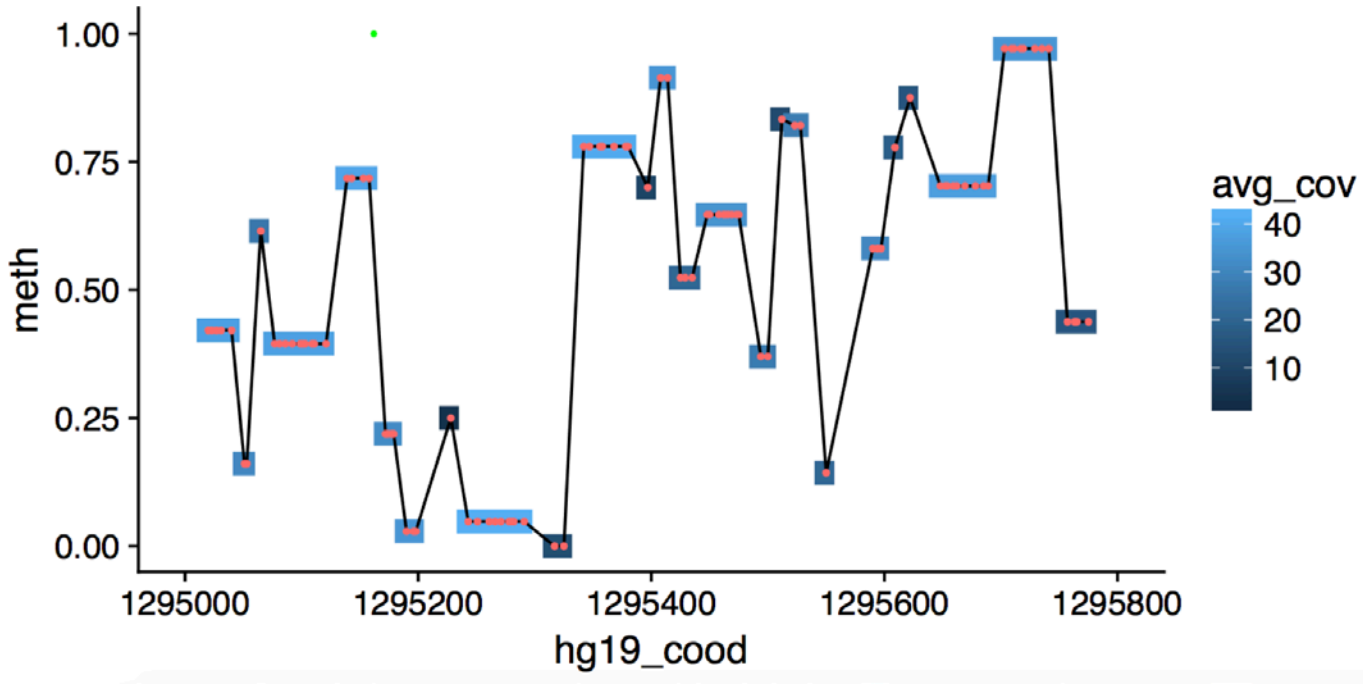
- Capture around the hTERT promoter, region with aberrant methylation in many cancers
- gDNA source from a BCPAP thyroid cancer cell line (poorly differentiated papillary thyroid carcinoma
- Hard to amplify with bisulfite PCR because of high CG-density, required many iterations of primer design

Methylation compare of capture/bisulfite

illumina



Nanopore

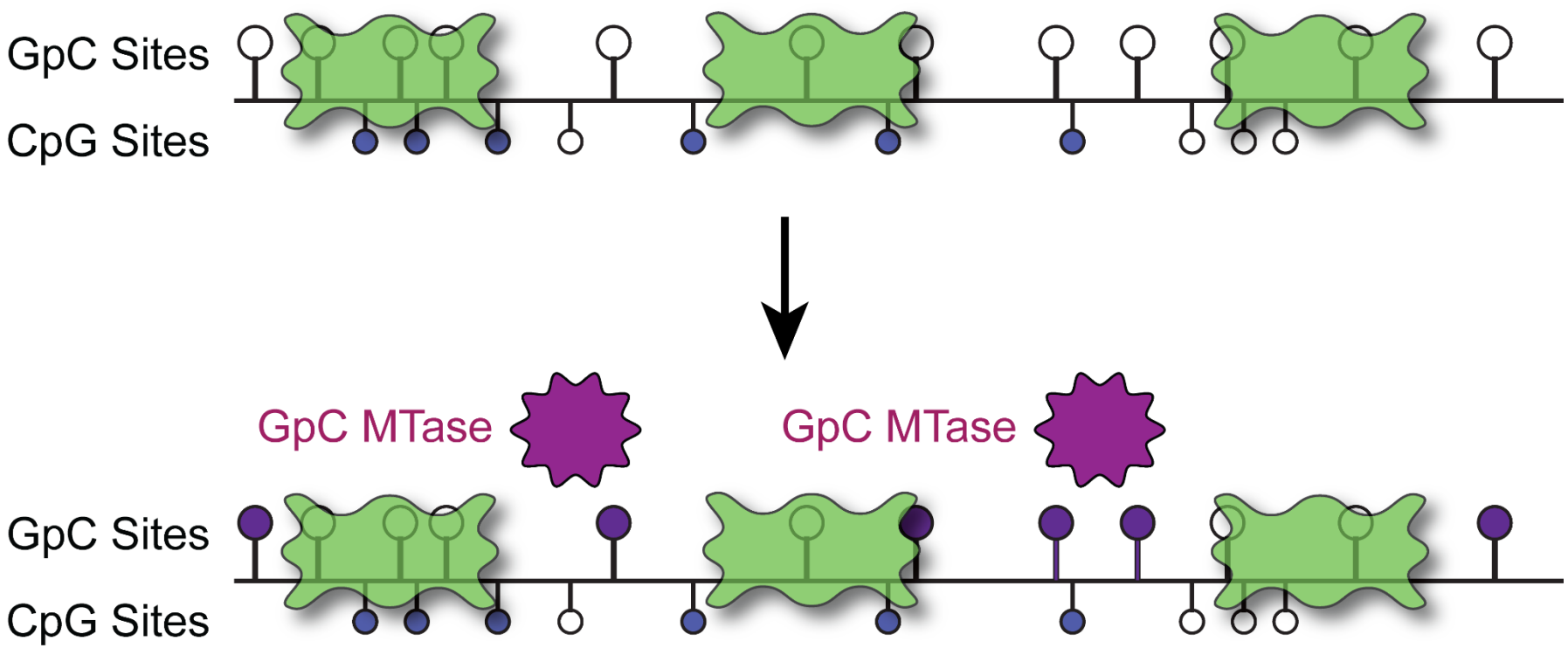


Preliminary data indicates methylation patterns largely concordant between bisulfite and nanopore

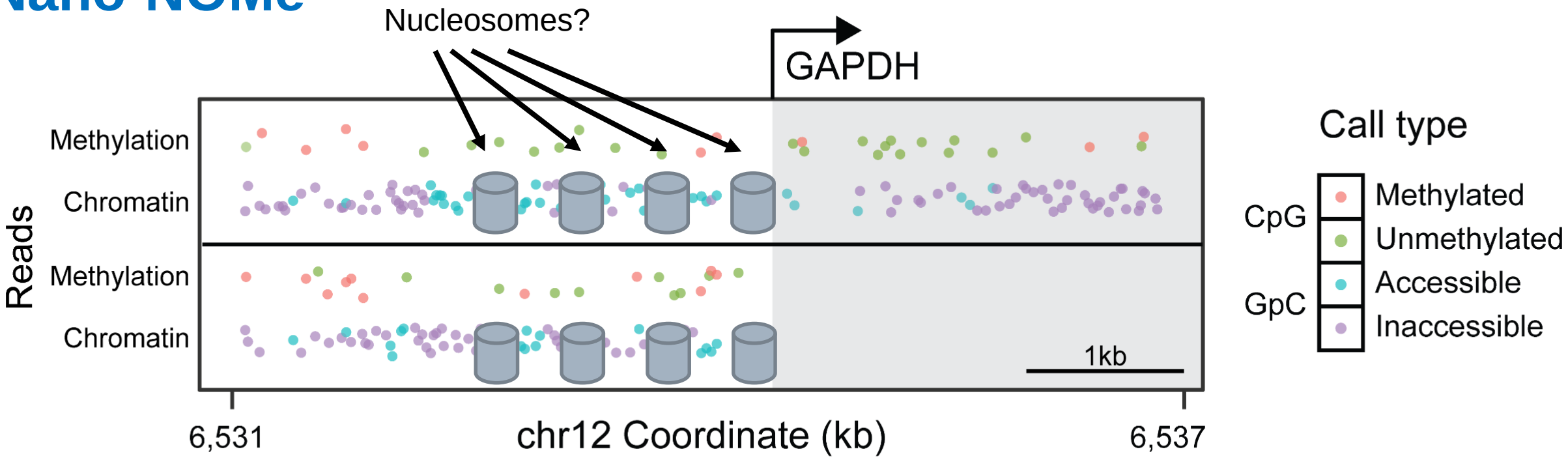
NanoNOMe: Chromatin Accessibility with Nanopore

- NOMe-seq : Nucleosome Ocupancy and Methylome sequencing (Kelly et. al. *Genome Res.* 2012)

Simultaneously measures DNA methylation (CpG) and nucleosome occupancy (GpC)



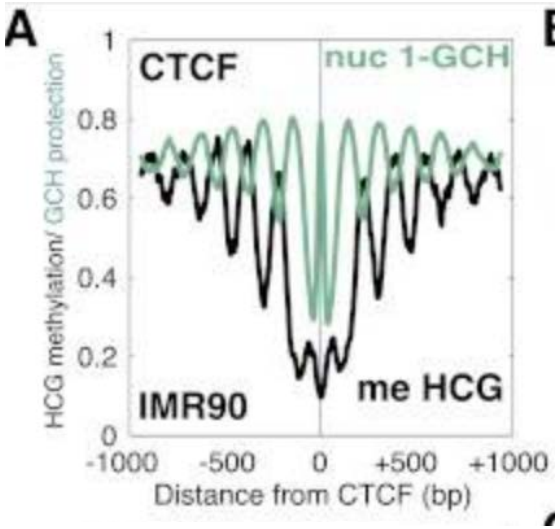
Nano-NOMe



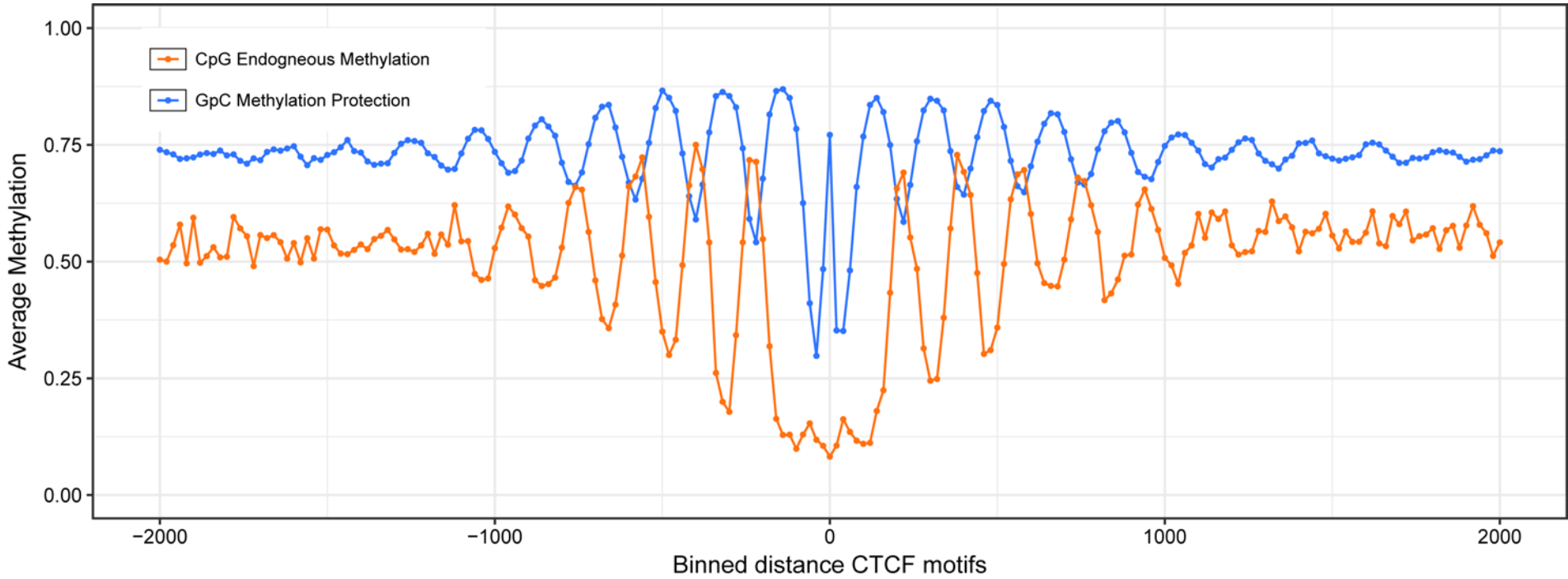
Read-level analysis of nucleosome occupancy **along with** DNA methylation

- MDA-MB-231 GAPDH : house-keeping gene
- promoter : Unmethylated / Open chromatin

Nano-NOMe - Results

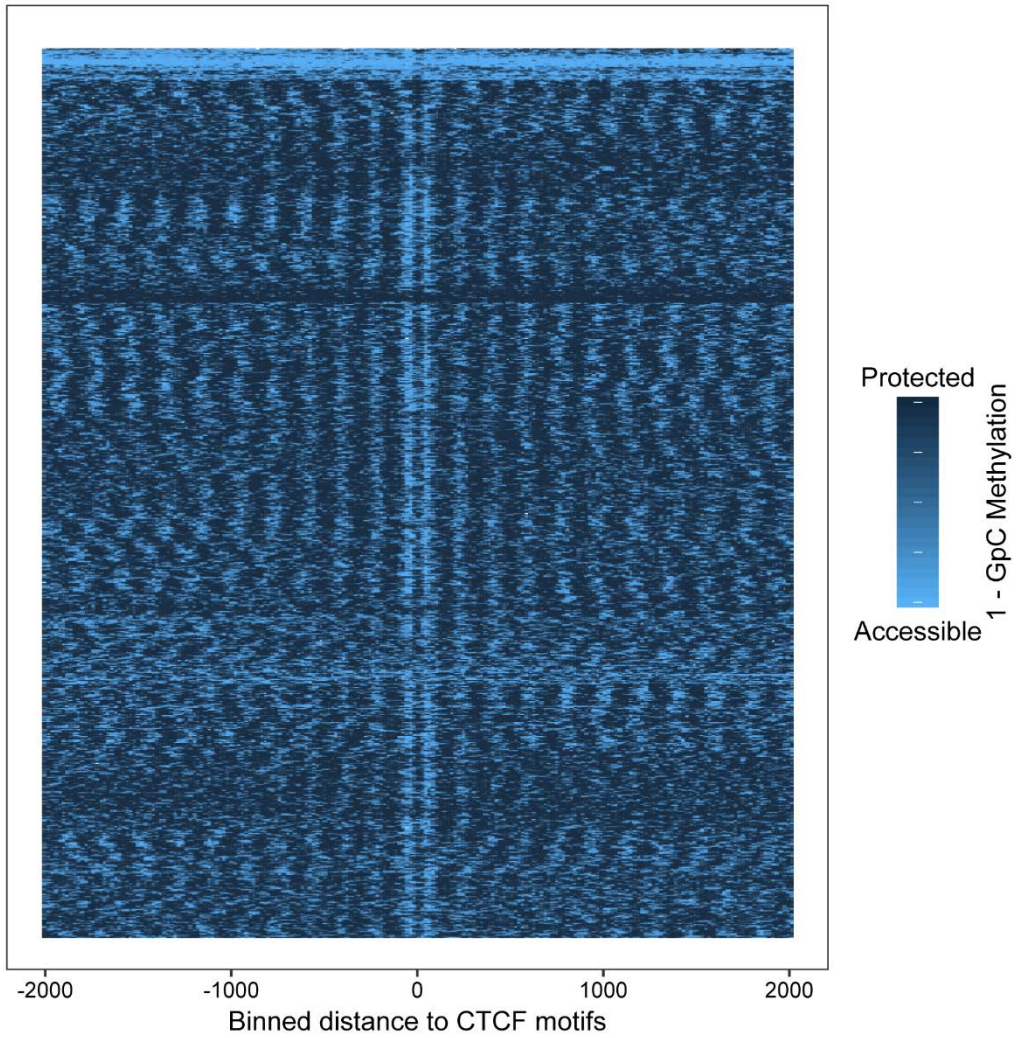


Kelly, et. al. *Genome Res.* 2012



- Genome-wide cumulative methylation profile shows comparable chromatin states in CTCF motif

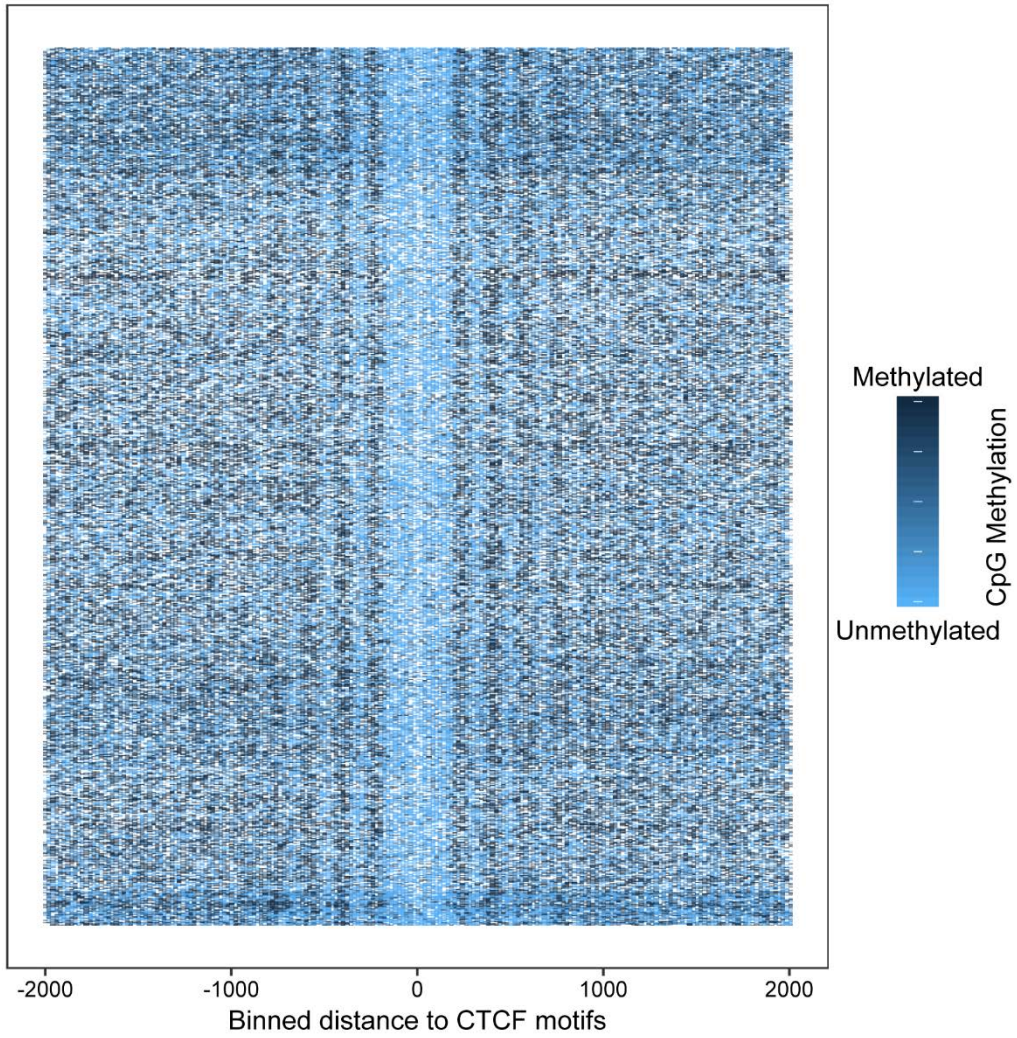
Nano-NOMe - Results



Heatmap of methylation reads that span 4kb region surrounding any CTCF site

- Nucleosome positioning is visible at single-read resolution

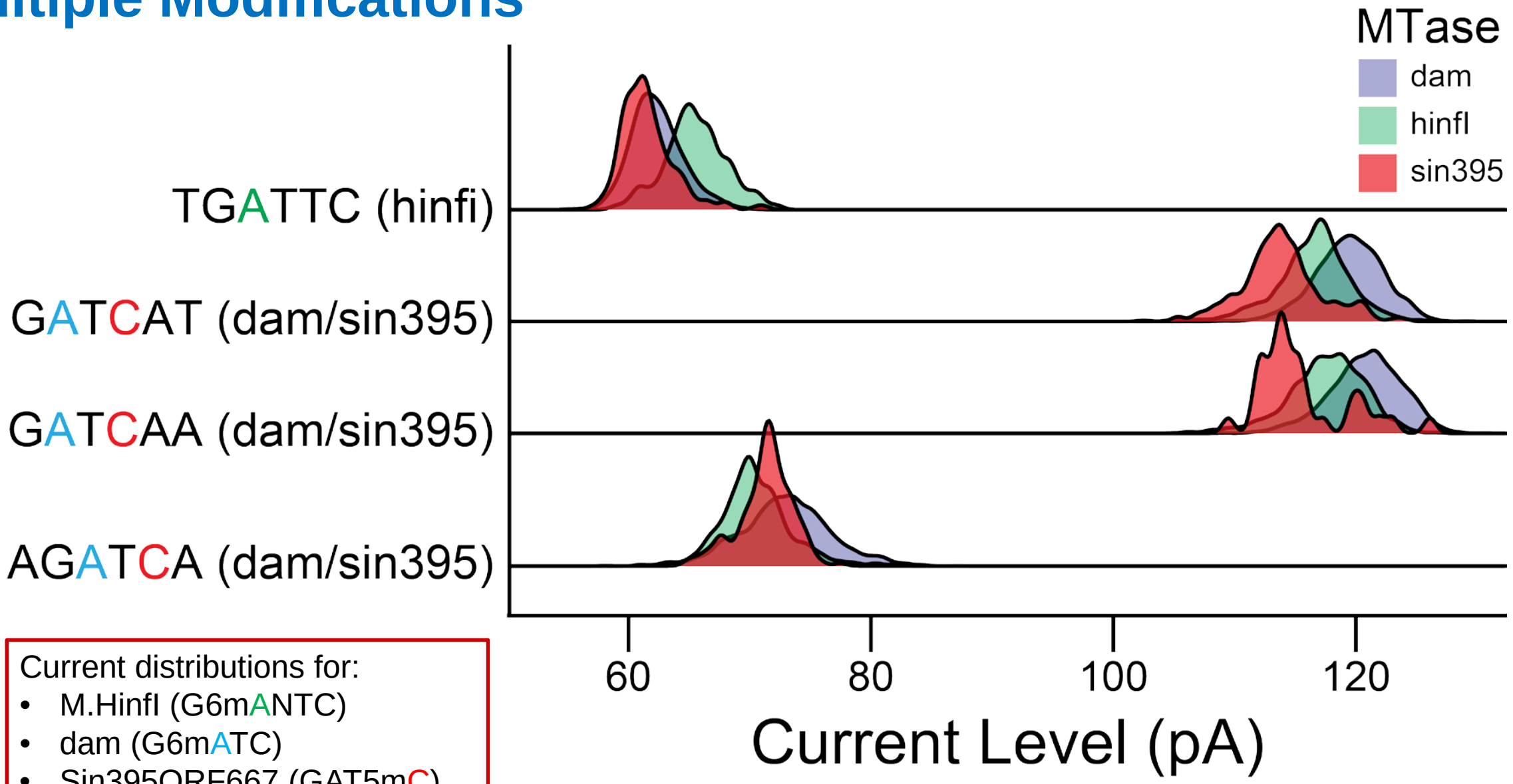
Nano-NOMe - Results



Heatmap of methylation reads that span 4kb region surrounding any CTCF site

- Nucleosome positioning is visible at single-read resolution
- Methylation pattern is also visible

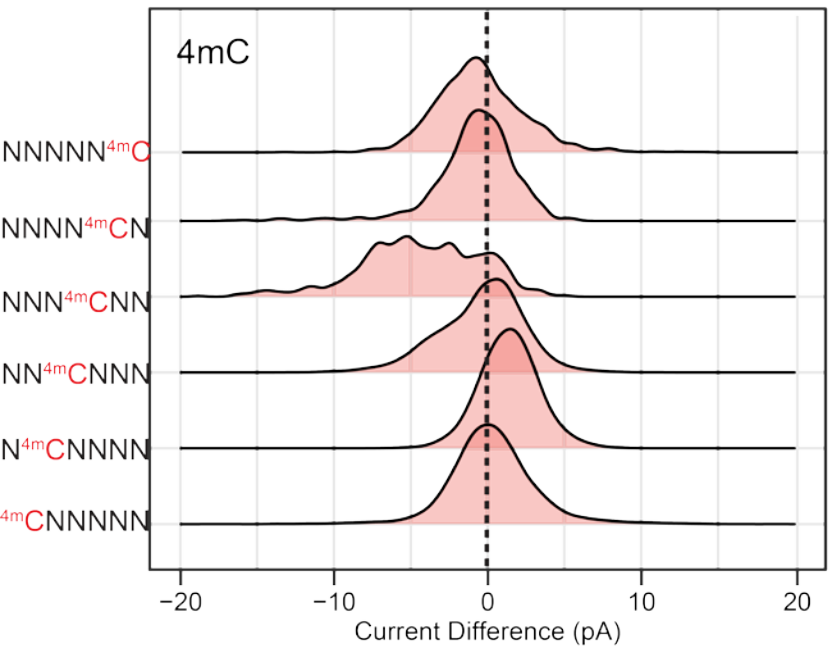
Multiple Modifications

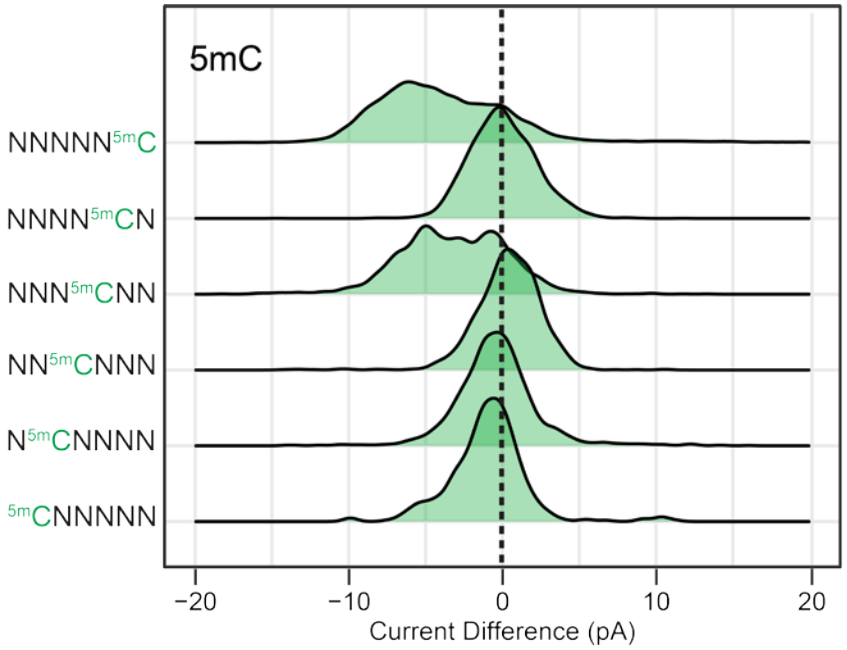


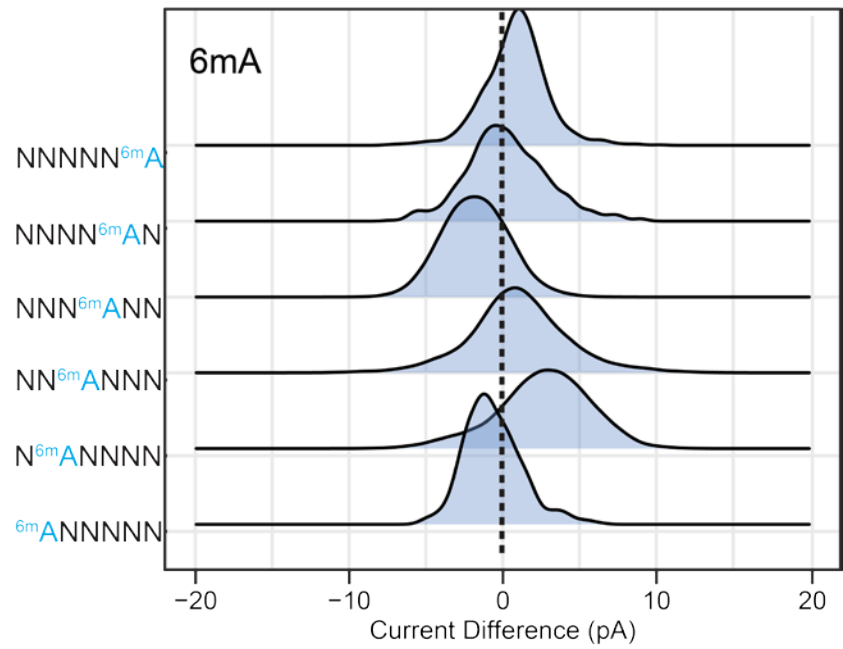
Current distributions for:
- M.HinfI (G6mANTC)
- dam (G6mATC)
- Sin395ORF667 (GAT5mC)

w/ Alexey Fomenkov (NEB)

4-methylcytosine and N6-methyladenine





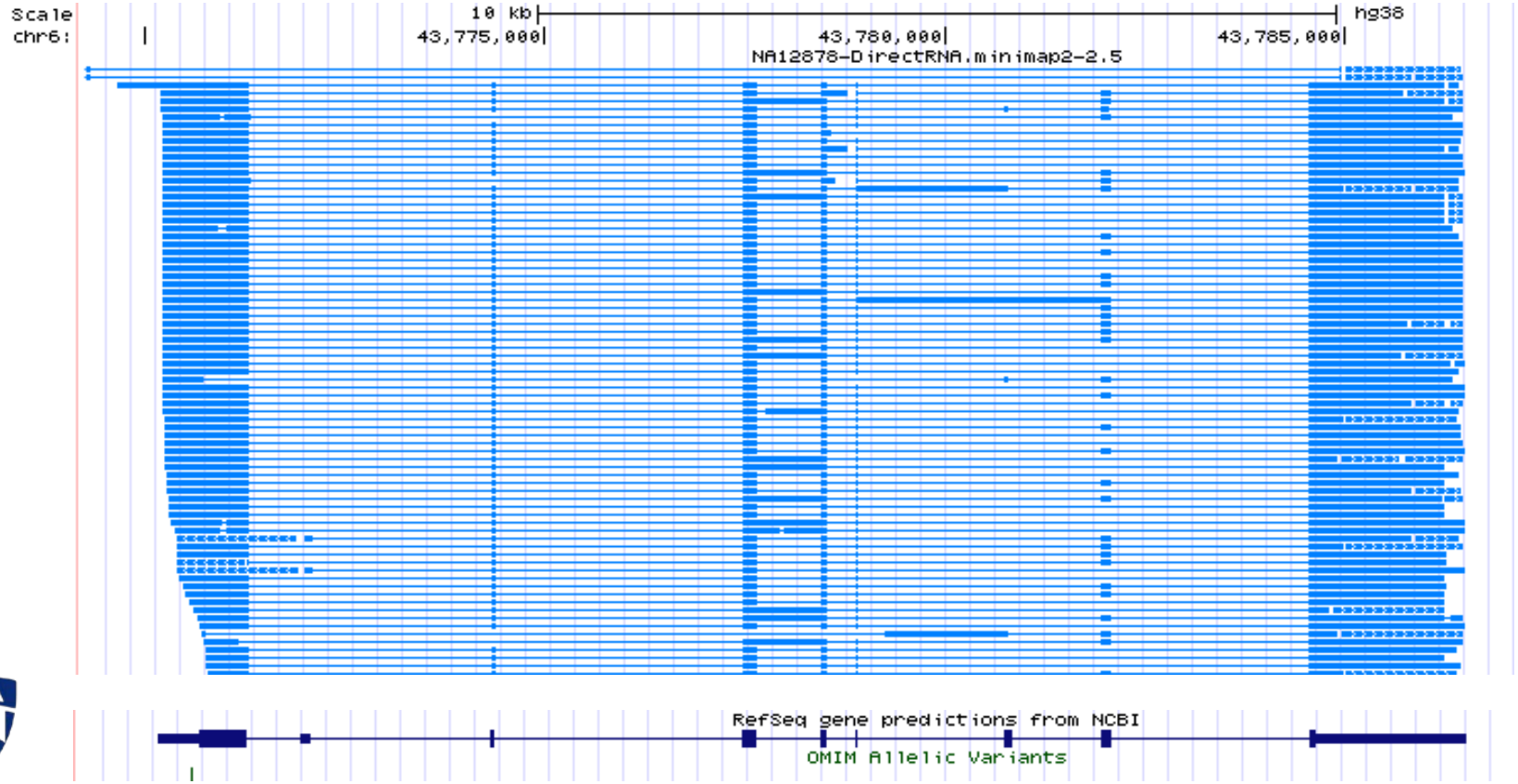


Initial work using the methyltransferases from NEB demonstrates that we can see signal from several different methylation marks (4-mC, 5-mC, N6-mA)

w/ Alexey Fomenkov (NEB)

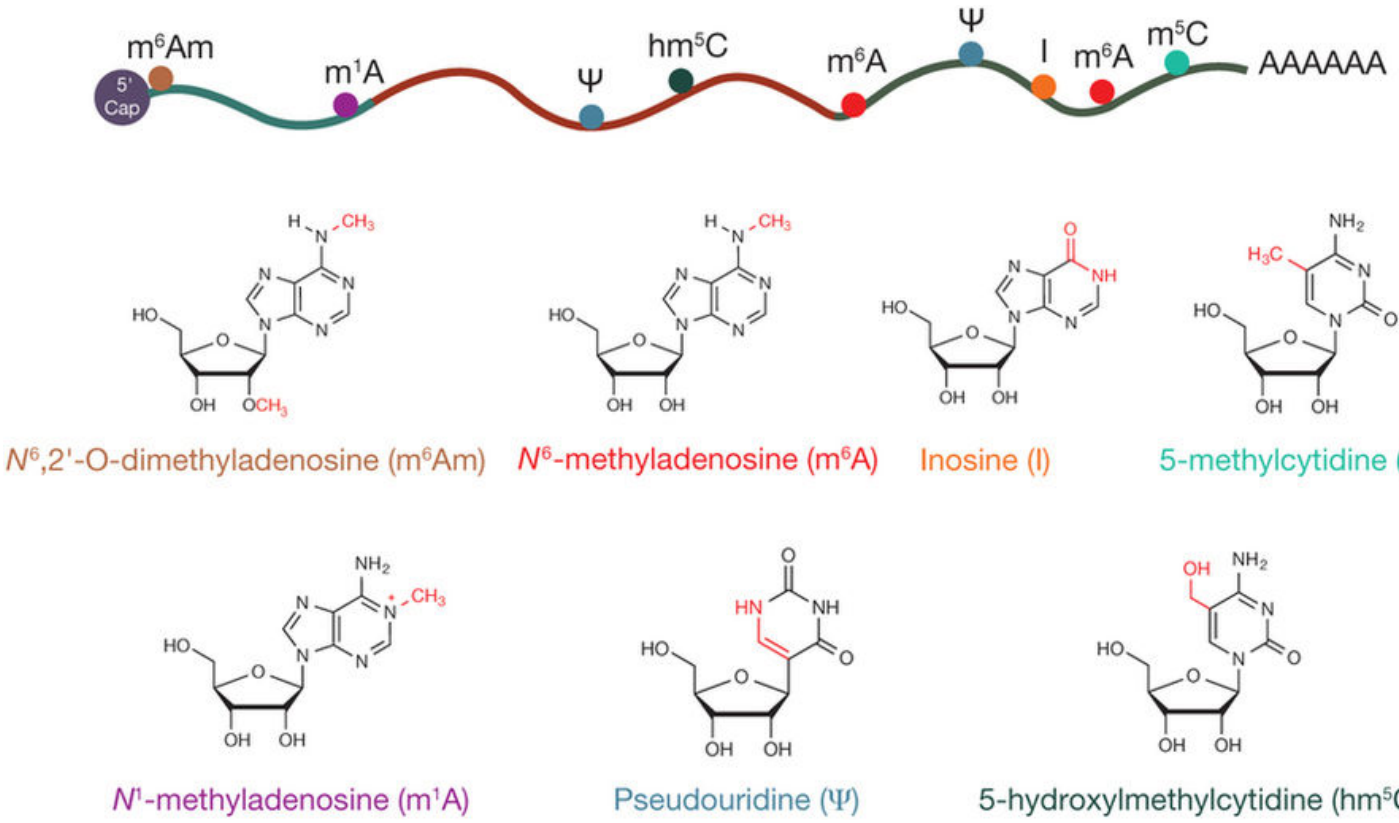
NA12878 RNA Consortium

- 13M dRNA reads (30 flowcells); 24M cDNA reads (12 flowcells)
  - Assess ability to sequence full-length isoforms
  - Quantify bias introduced by RT-PCR
  - Poly-A tail length
  - **RNA modifications?**



Direct RNA Sequencing

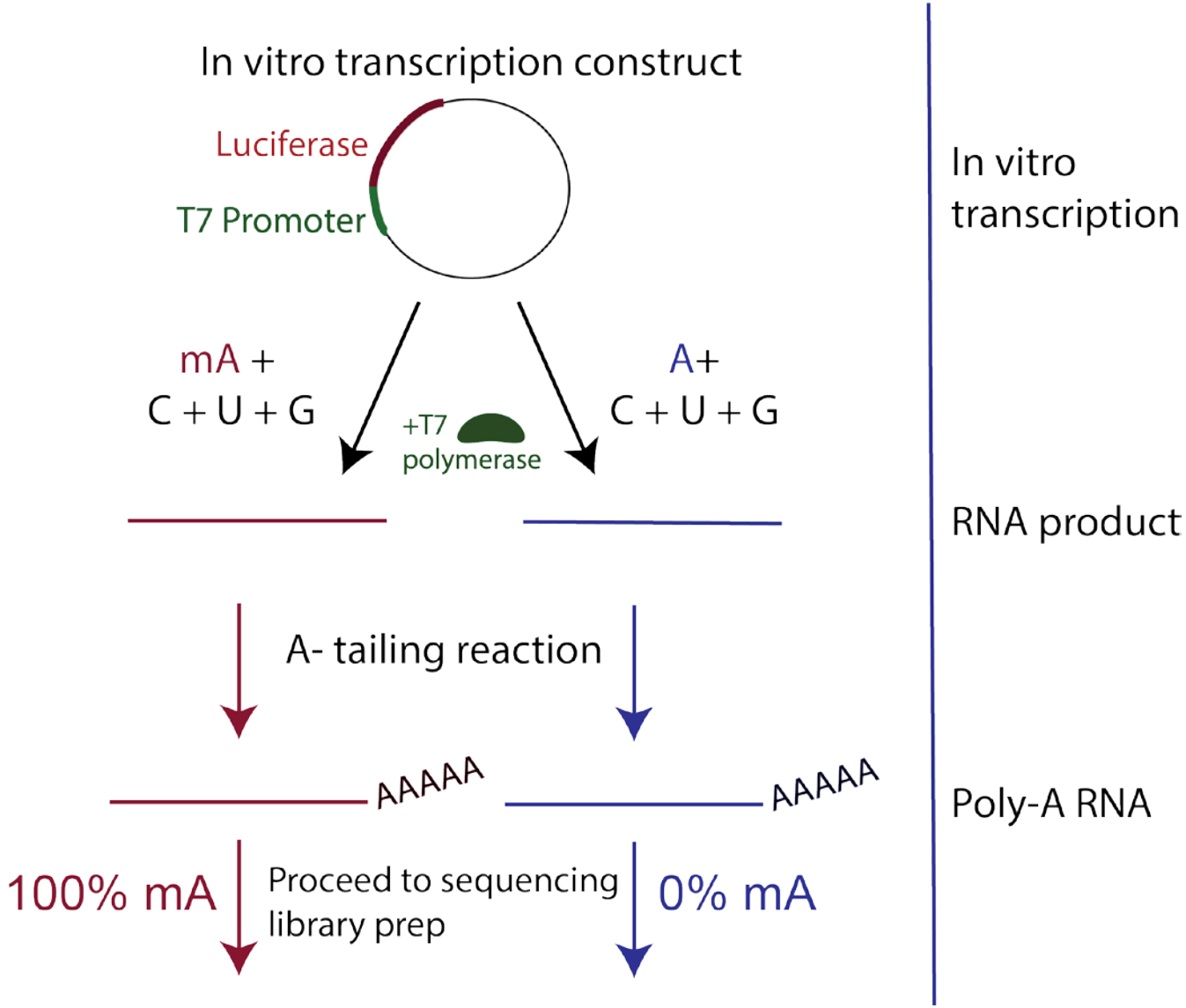
- We can use this to understand RNA modifications – the **epitranscriptome**
- Other methods are challenging – either inefficient, or lack resolution, and always only one modification at a time



Li, Xiong, Yi, Nature Methods (2017)

Detection of RNA modifications with modIVT

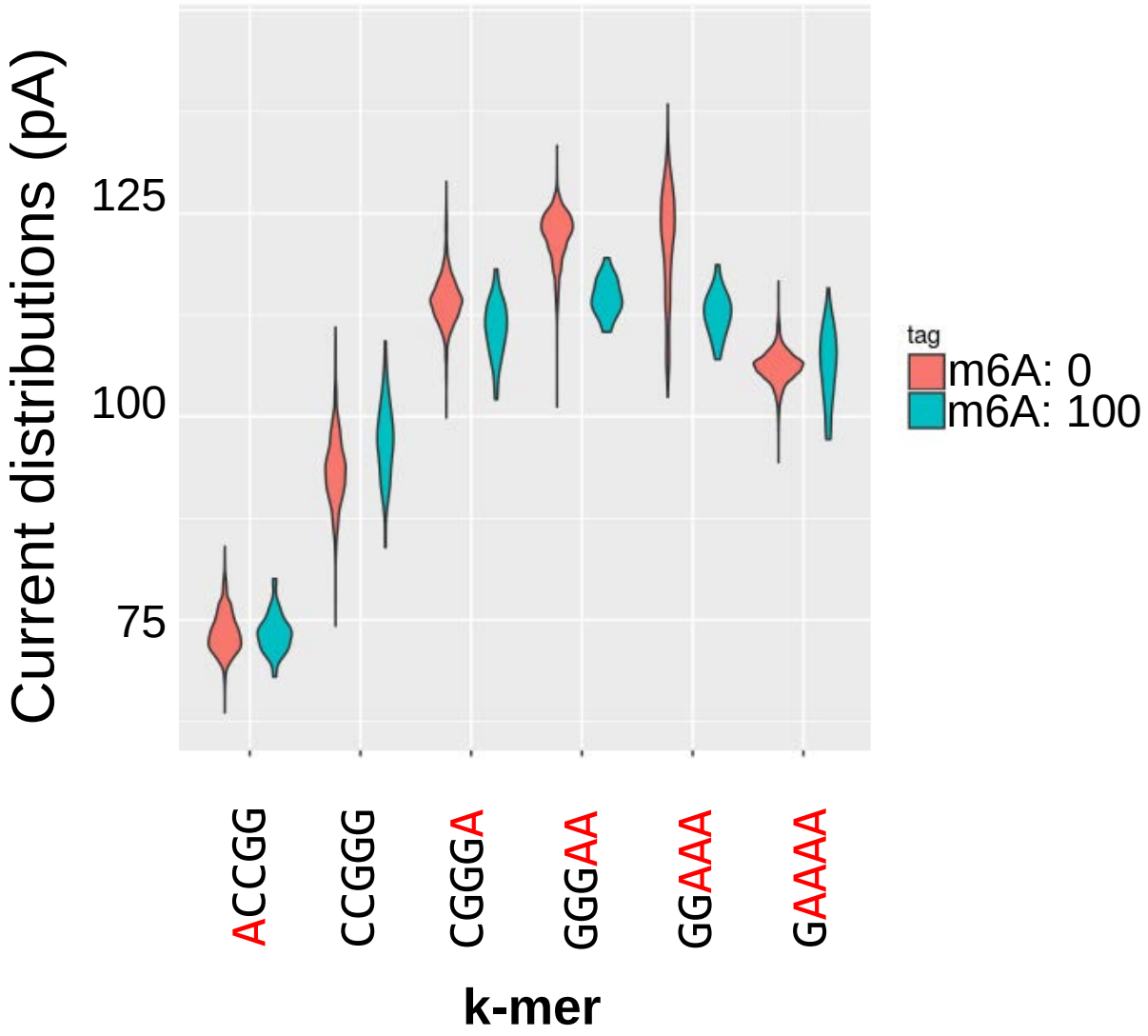
- IVT based RNA synthesis allows incorporation of labeled nucleotides
- All or none reaction right now, T7 has a strong preference for the unmodified nucleotides, making mixtures hard



NEB HiScribe™ T7 High Yield RNA Synthesis kit

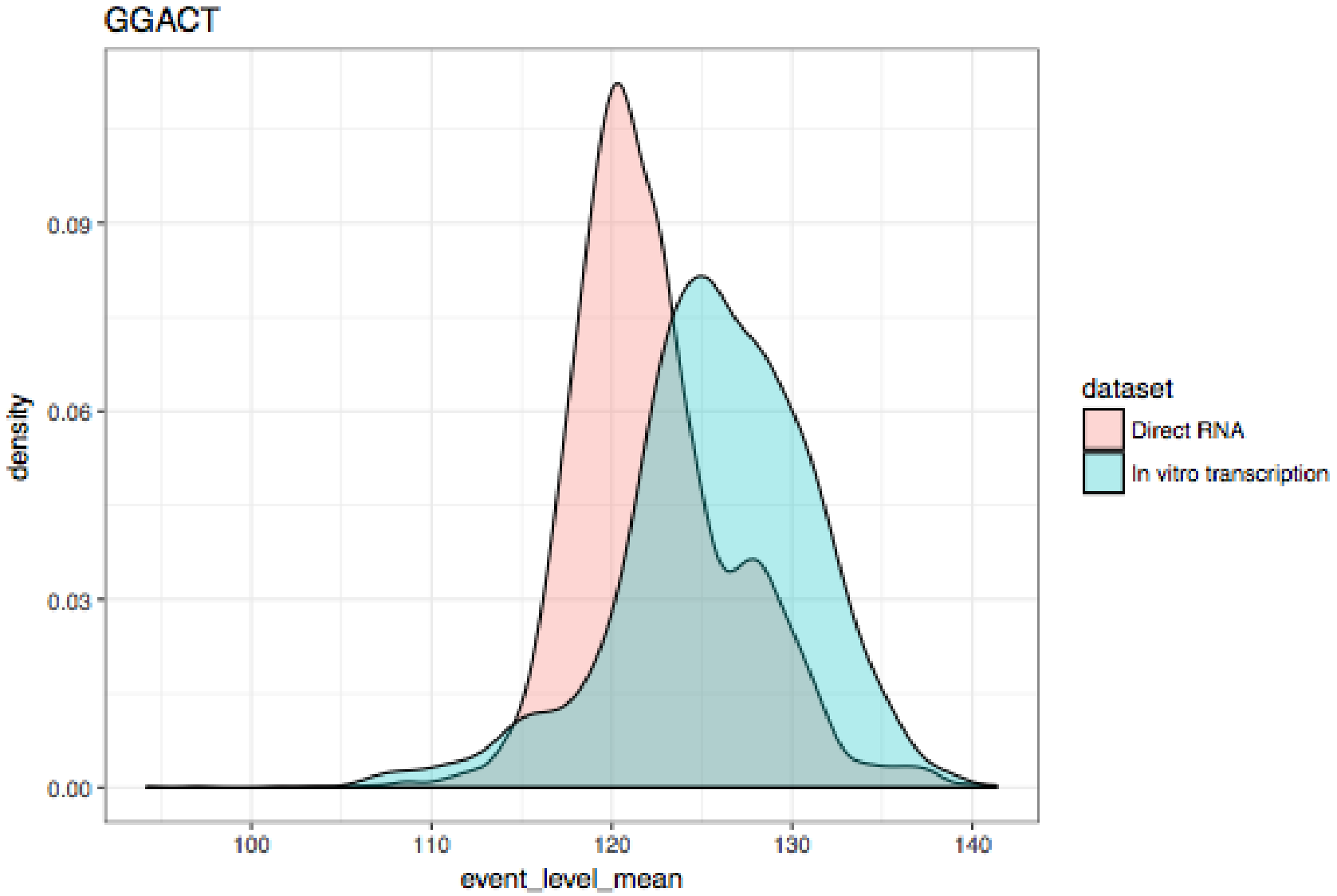
Detection of RNA modifications with modIVT

- From Luciferase we can already see strong signal depending on context
- Using nanopolish eventalign, we can extract the distribution of current values along the RNA strand



Exploring the dRNA for m6A

- Eukaryotic elongation factor 2 has a METTL3 motif GGACU (m6A writer) in the mRNA sequence
- Has been shown to have m6A via IP-seq methods (Meyer et al Cell 2012)
- Compared dRNA data with IVT'd dRNA signal



Unpublished, preliminary data - NA12878 RNA Consortium

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.
- Modifications to the primary DNA sequence (e.g. cytosine methylation) can be detected directly using nanopores
- Exogenous labeling allows simultaneous detection of chromatin and methylation state using nanopore sequencing
- Preliminary data from direct RNA sequencing suggests we can also see *RNA* modifications

Acknowledgments

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

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

- Alexey Fomenkov, PhD

Redwood Genome Project (Neale)

National Human Genome Research Institute
1R01HG009190-01A1

Nanopore RNA Consortia
- UCSC (Akeson, Brooks)
- UBC (Snutch, Tyson)
- OICR (Simpson)
- JHU (Timp)
- Nottingham (Loose)
- Birmingham (Loman)

Looking for Postdocs to come to