Biology of Genomes
Oxford Nanopore Workshop
Cold Spring Harbor Lab; May 2018

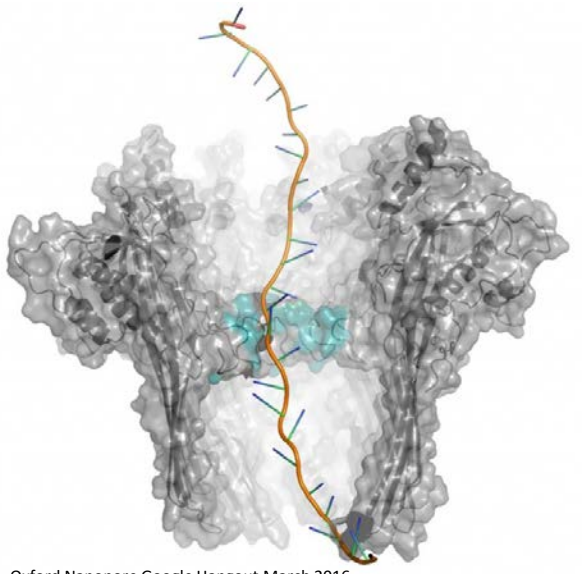
JOHNS HOPKINS
BIOMEDICAL ENGINEERING

Applications of modification detection in nanopore sequencing

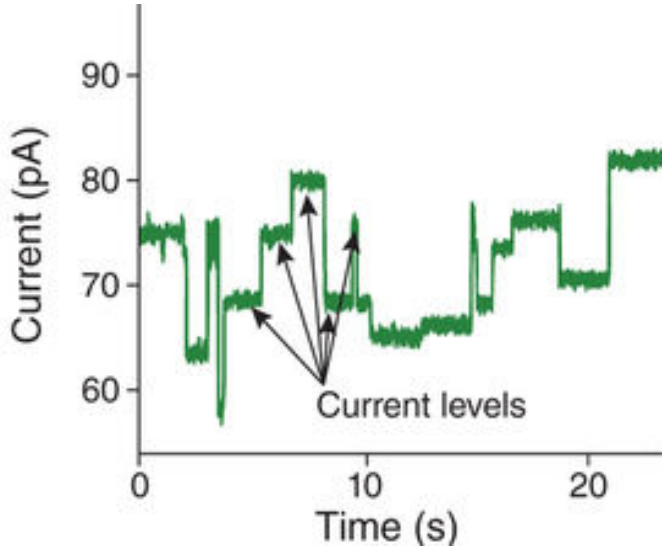
Winston Timp
Department of Biomedical Engineering
Johns Hopkins University

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

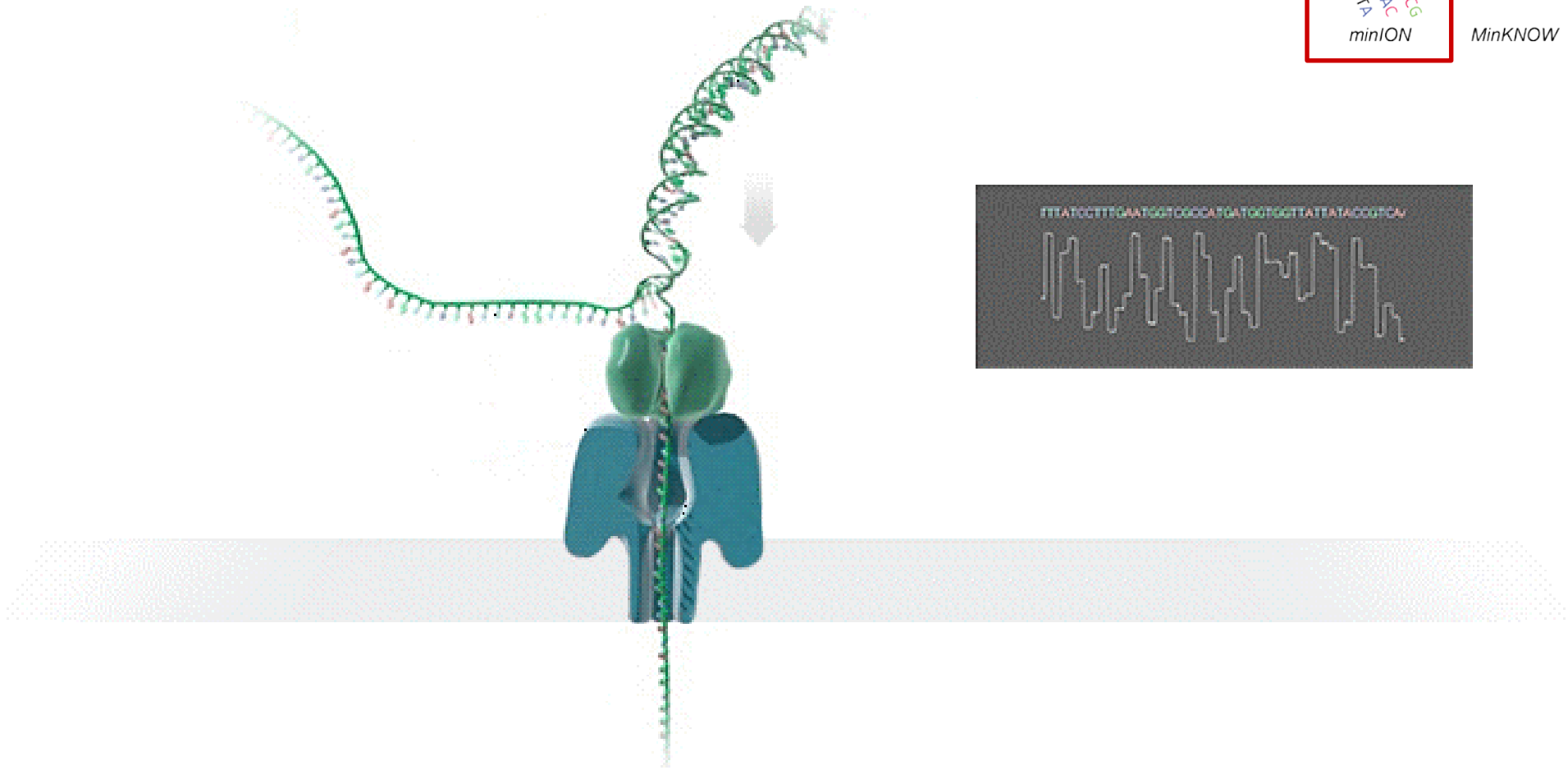


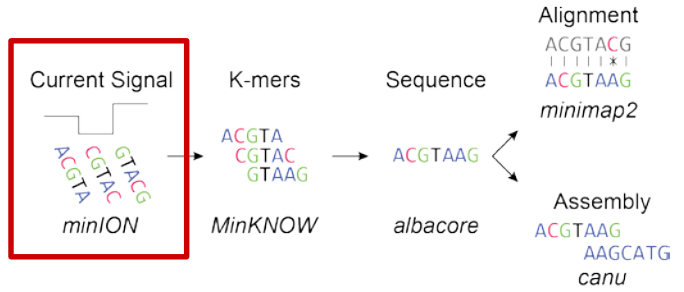
Deamer et al 2016, Nature Biotech

ATCGATCGATAGTAT
TAGATACGACTAGC
GATCAG

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

Sequencing Operation

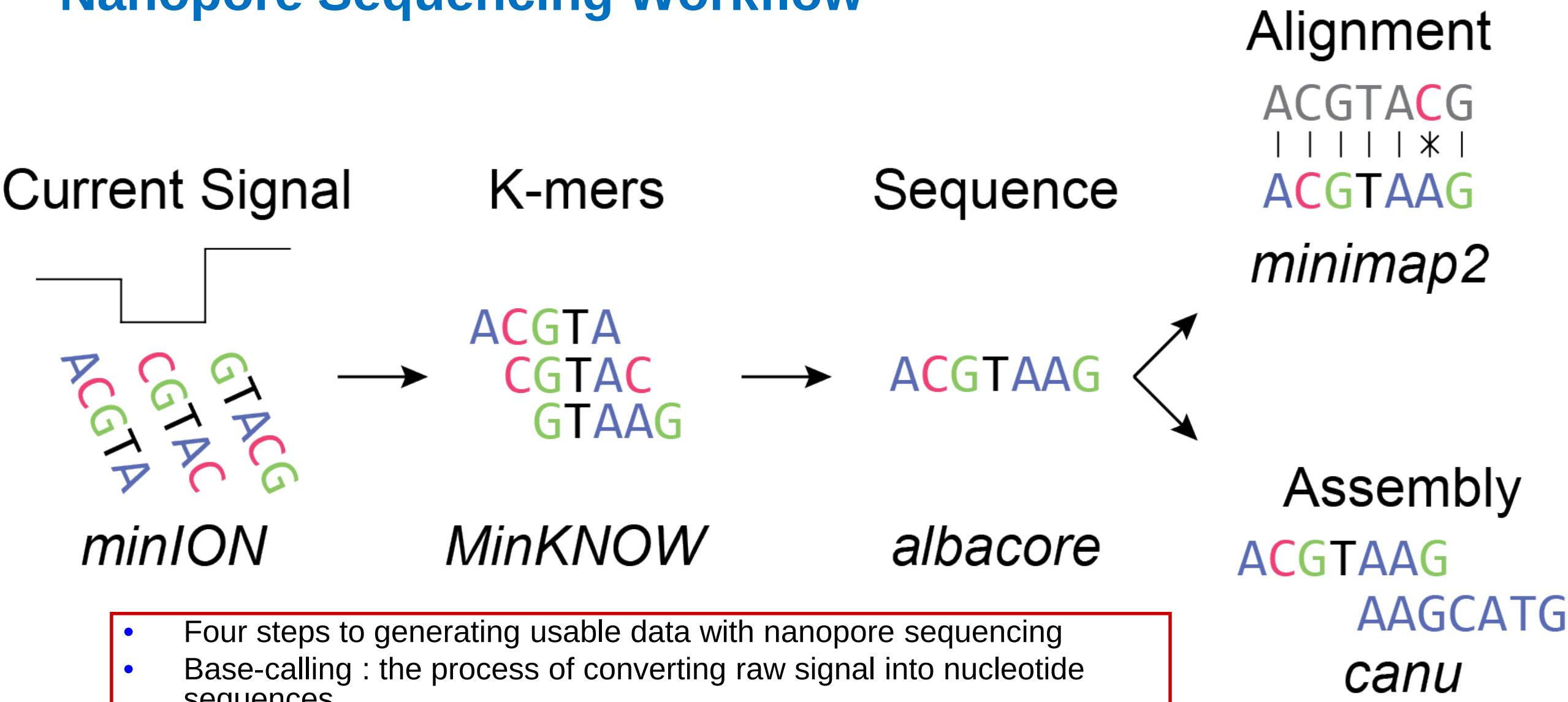




Oxford Nanopore Technologies

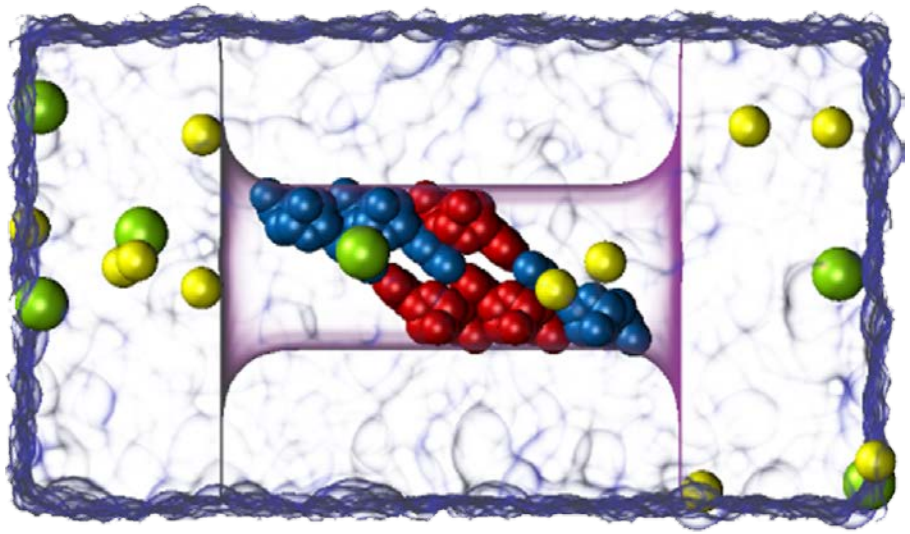
- Protein nanopores on a synthetic polymer
- Multiple base-pairs at a time (“k-mers”)
- Characteristic current signature is converted to nucleotide sequences

Nanopore Sequencing Workflow

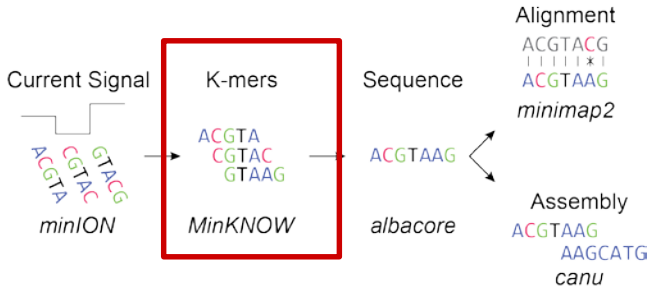


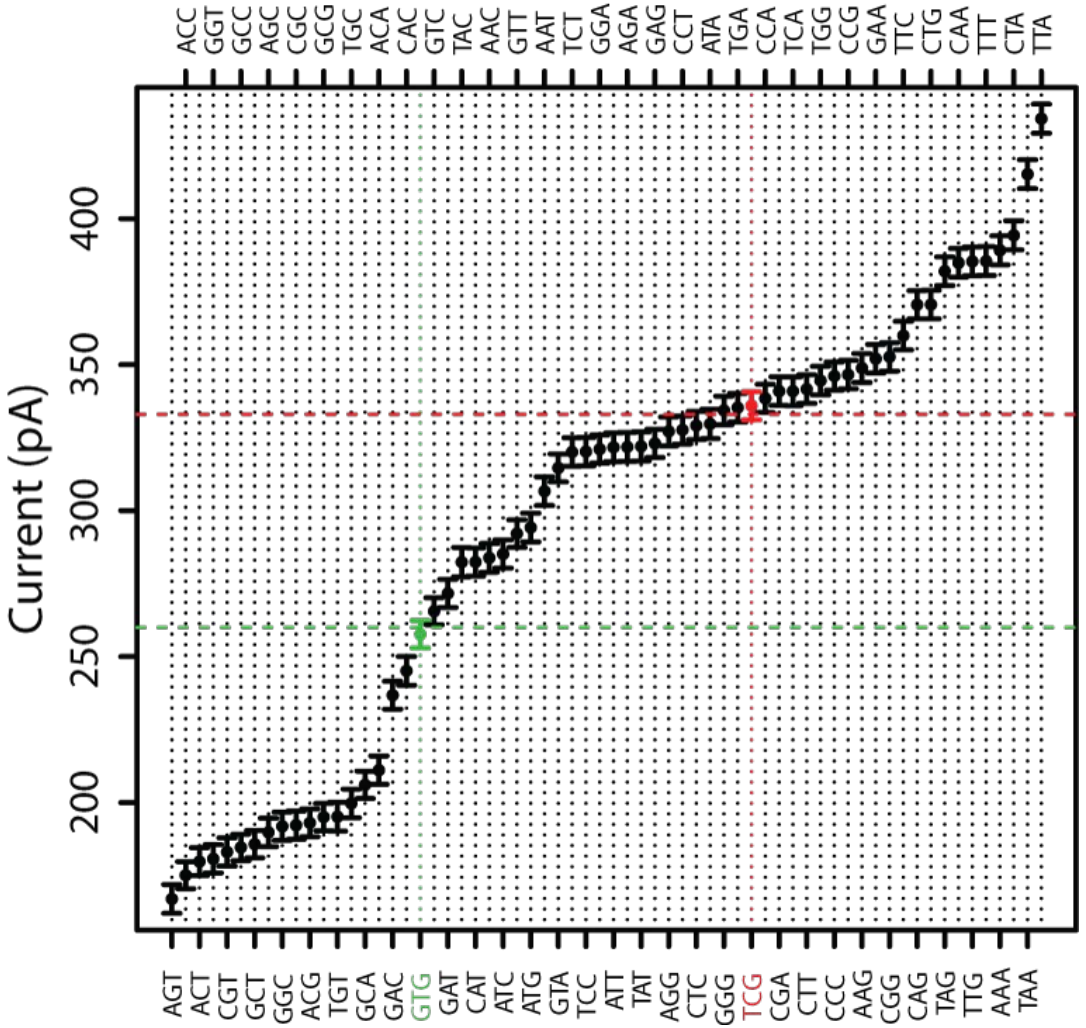
- Four steps to generating usable data with nanopore sequencing
- Base-calling : the process of converting raw signal into nucleotide sequences
- Nanopolish : uses alignment and current signal to improve base-calls

Problems with Nanopore basecalling



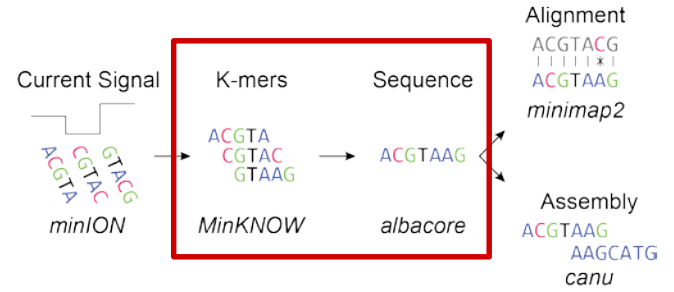
- Multiple bases influence the current passing through the pore.
- Through simulation with Brownian Dynamics, we calculated the contribution from triplets of DNA in a solid-state nanopore - 64 current levels.
- Not all of these different currents are distinguishable

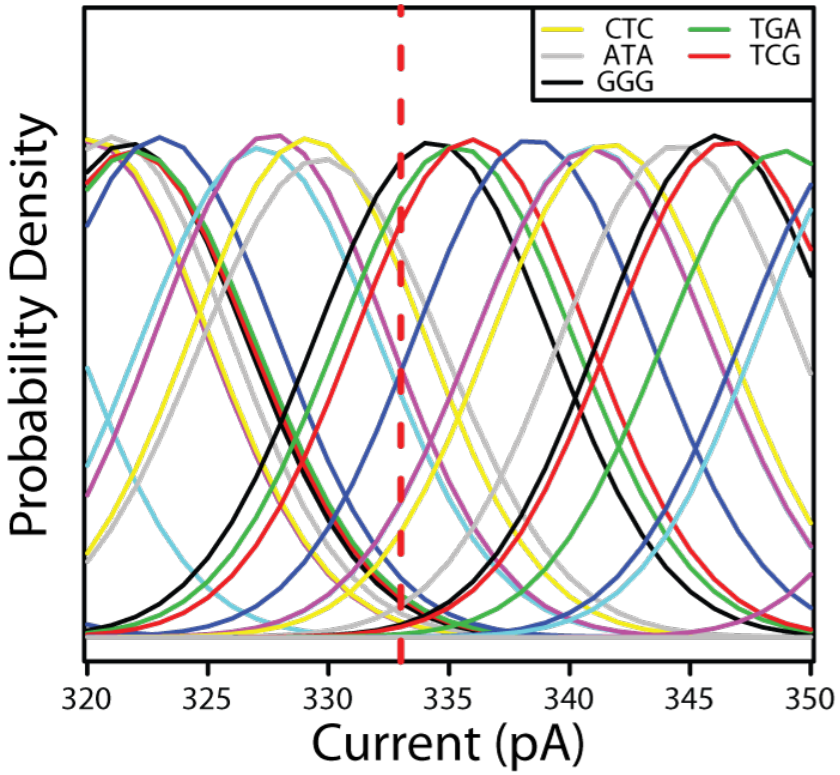


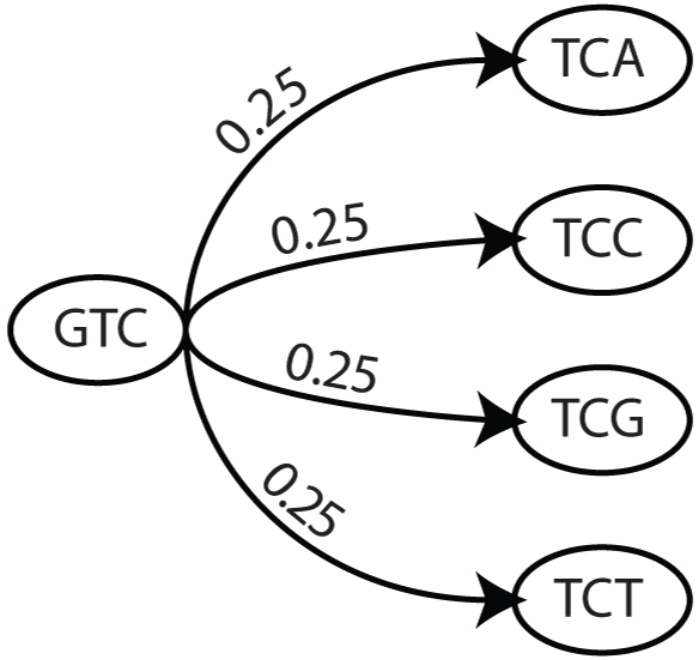


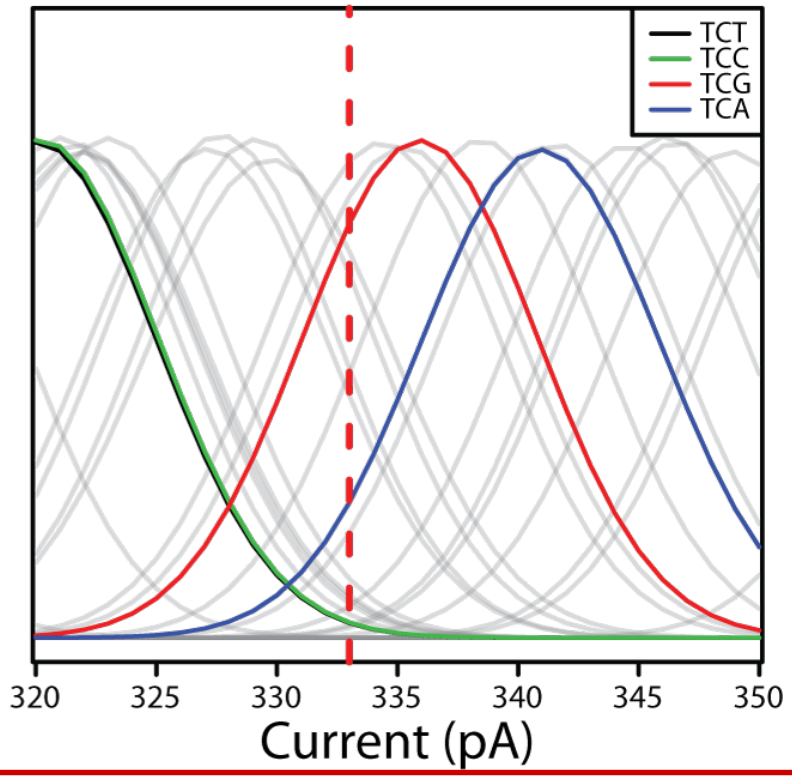
Comer and Aksimentiev *J. of Phys Chem C* 116(5) 3376-3393 (2012)

Prior Information for Decoding







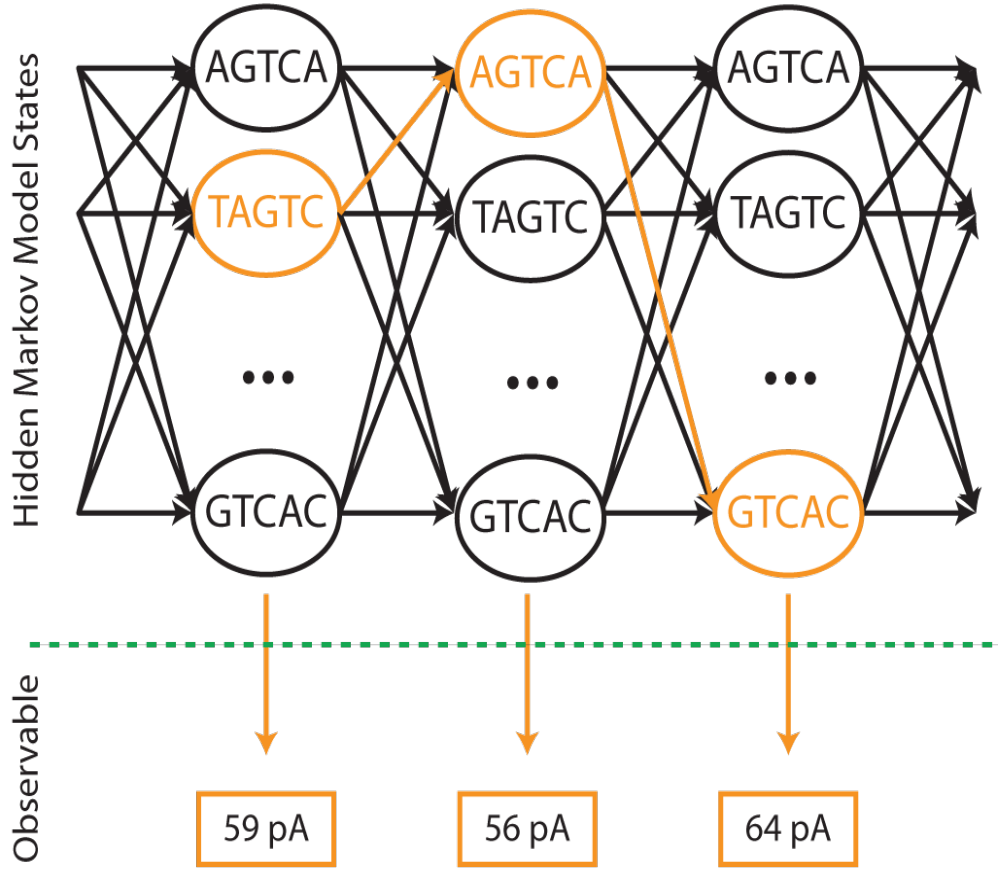


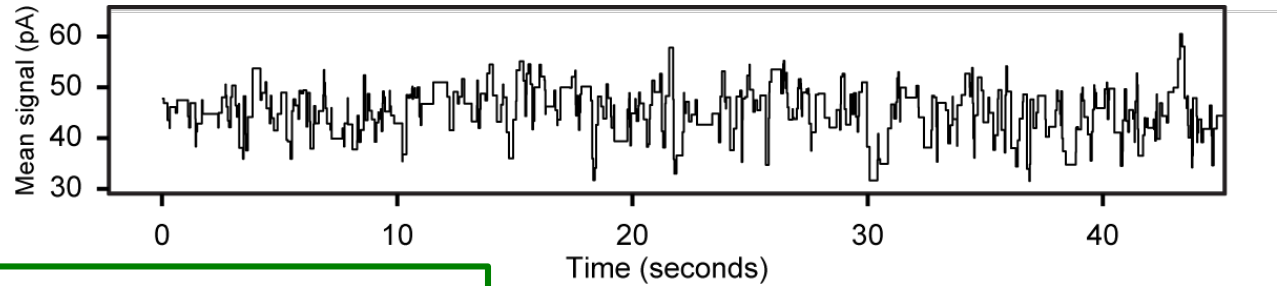
- With no prior information, a given current value may not be called correctly (333pA would be called as GGG)
- If we know the previous triplet, the next triplet is well defined, leaving only four possibilities, resulting in the correct call of TCG

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
- Transition probability 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

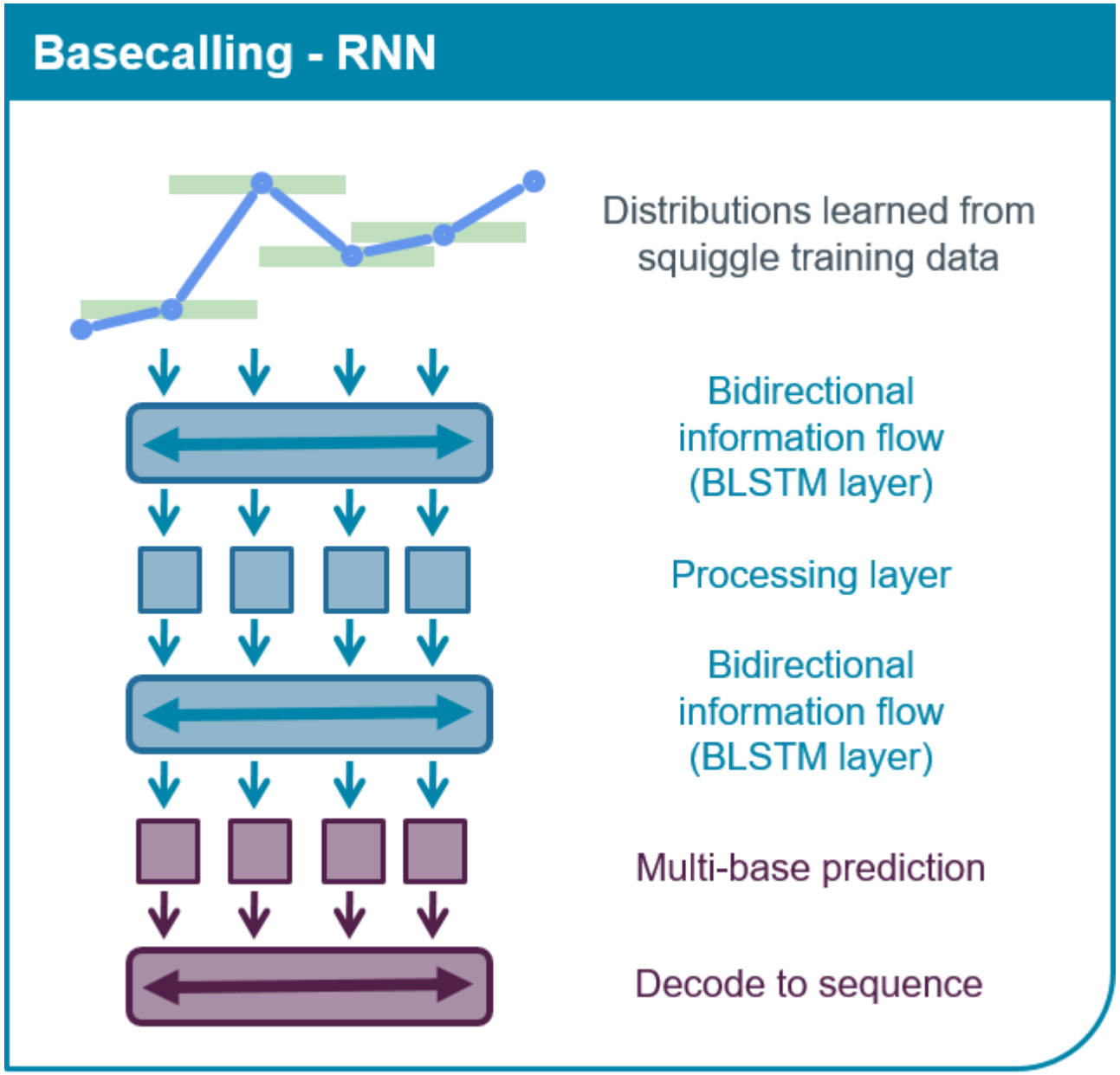




$$\delta_k \prod_t P(I(t) \mid k_t) \times T_{(t-1)(t)}$$

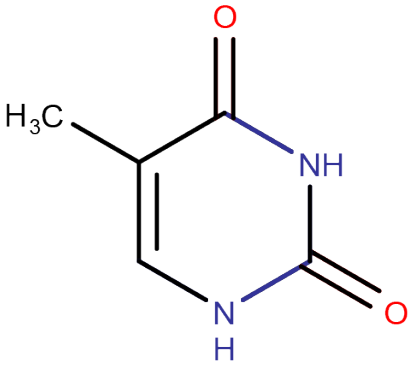
Basecalling shifting to RNN

- 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.

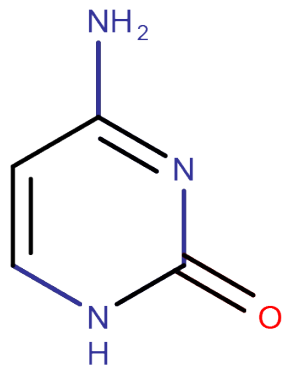


Oxford Nanopore

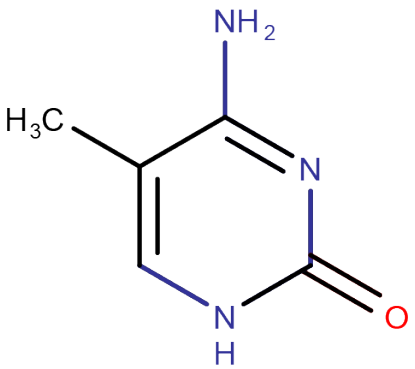
Nanopore Sequencing of Modifications



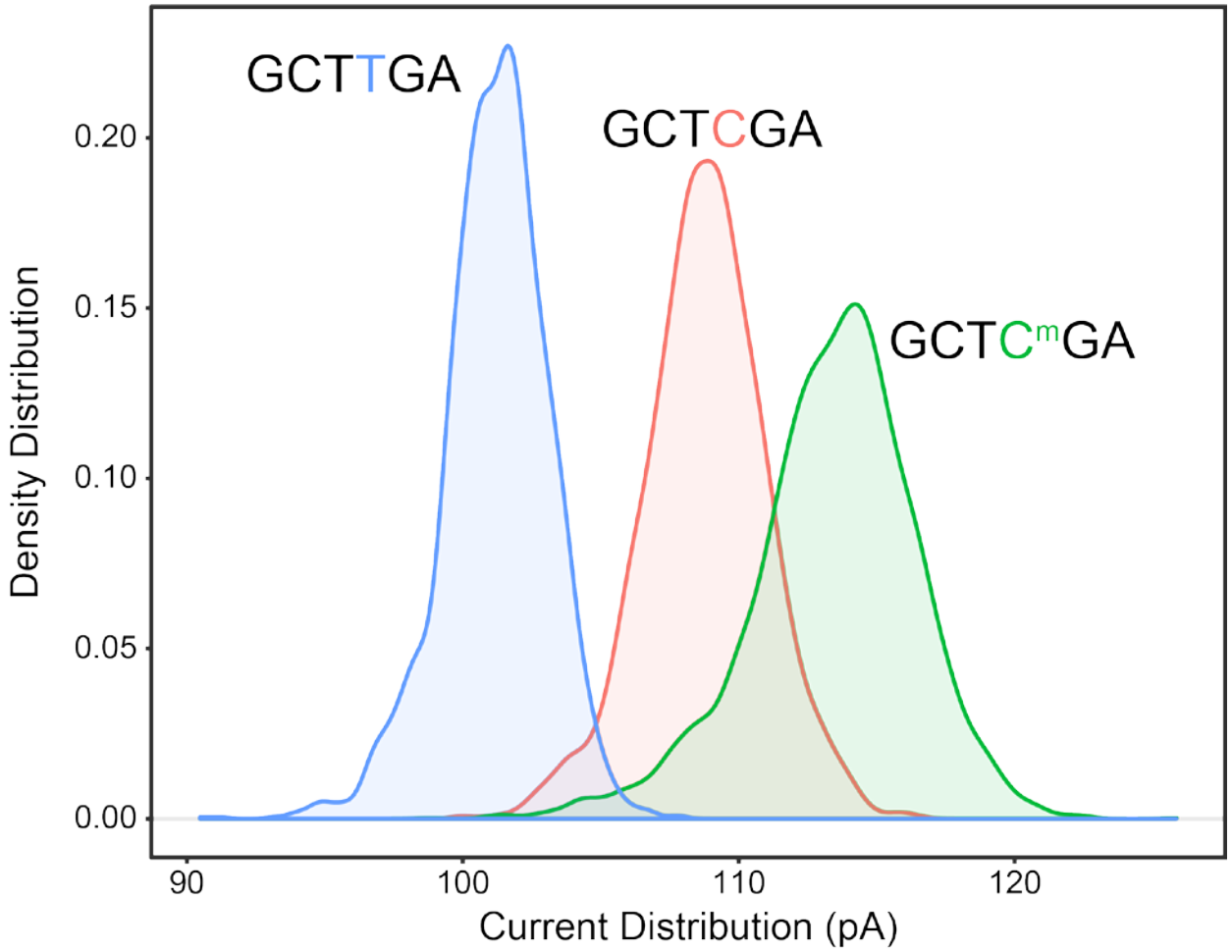
Thymine



Cytosine



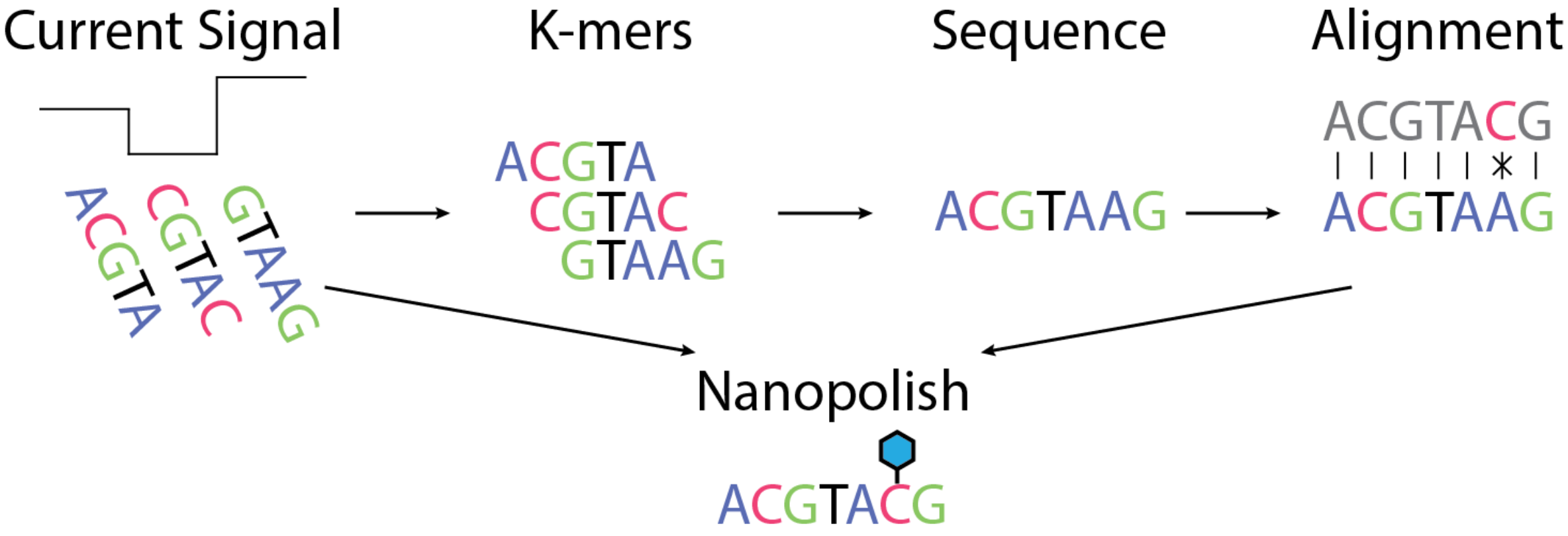
5-methylcytosine



- To generate methylated samples, we treat unmethylated DNA (PCR amplified E. Coli gDNA) with M. SssI methyltransferase
- Distributions of observed current for GCT[T/C/mC]GA demonstrate the type of signal between methylated and unmethylated k-mers

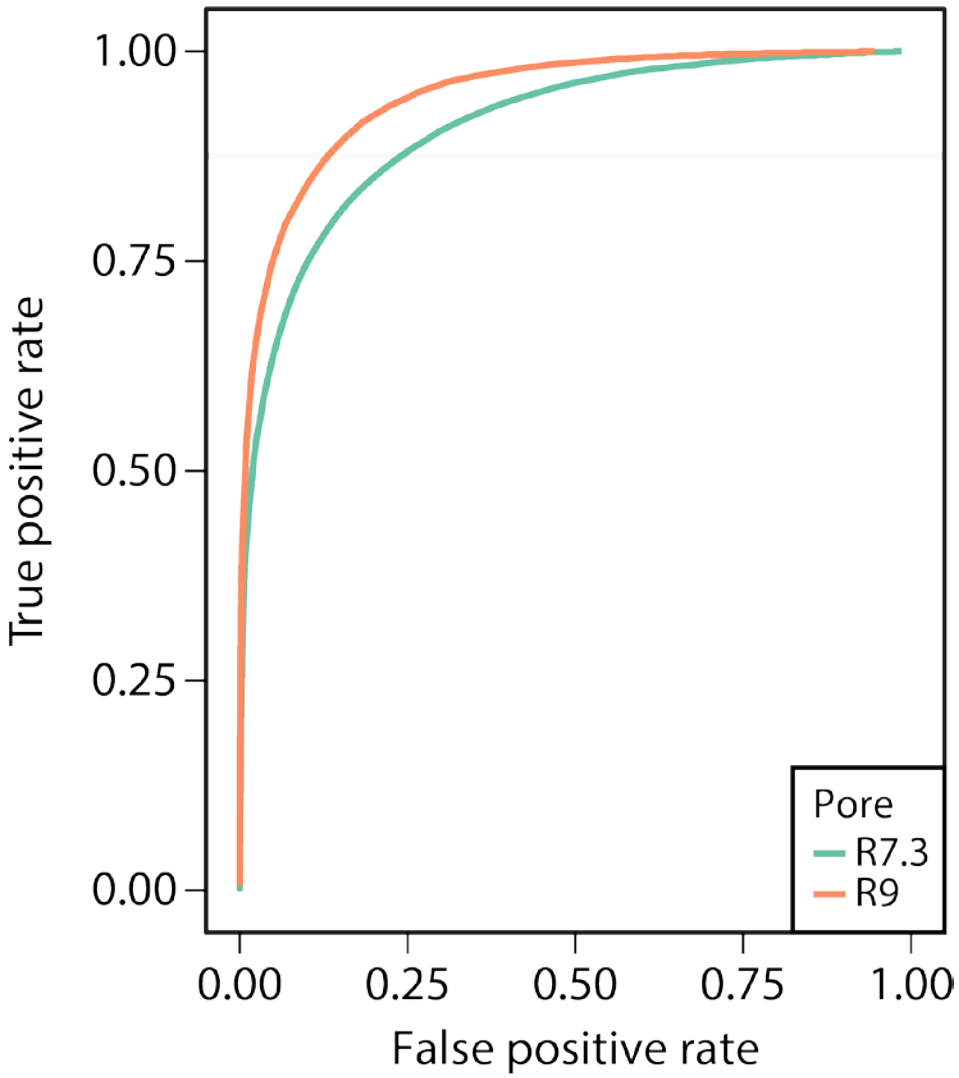
Simpson et. al *Nature* Methods 2017

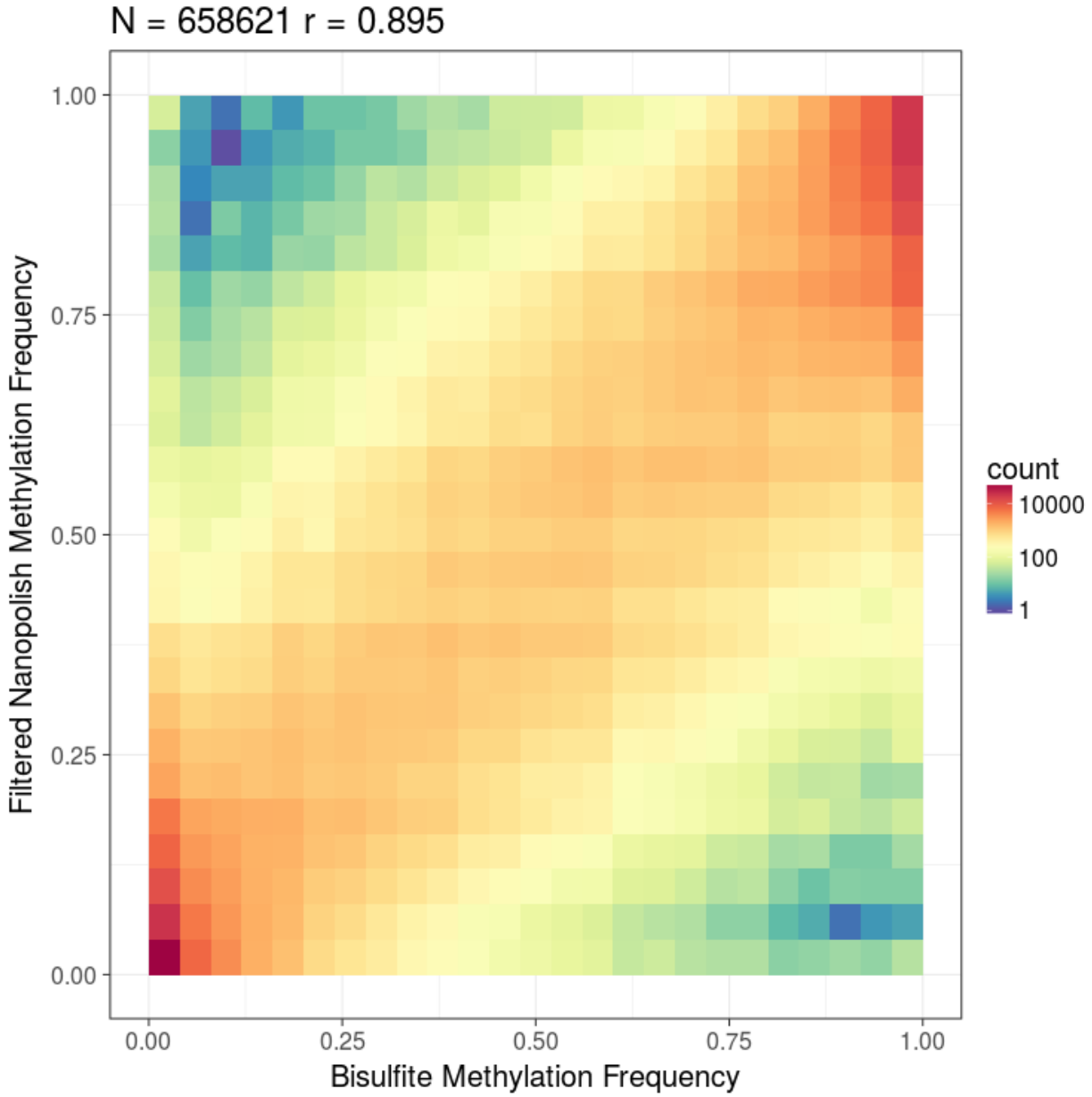
Nanopore: nanopolish methyltrain



- 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

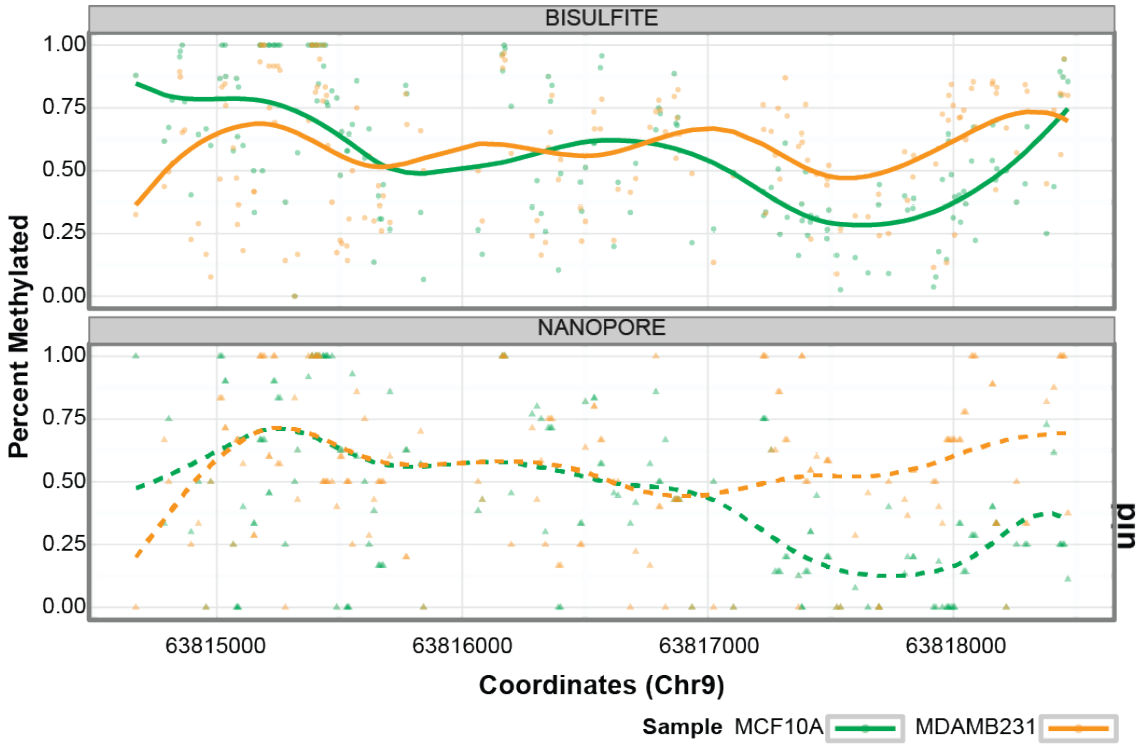


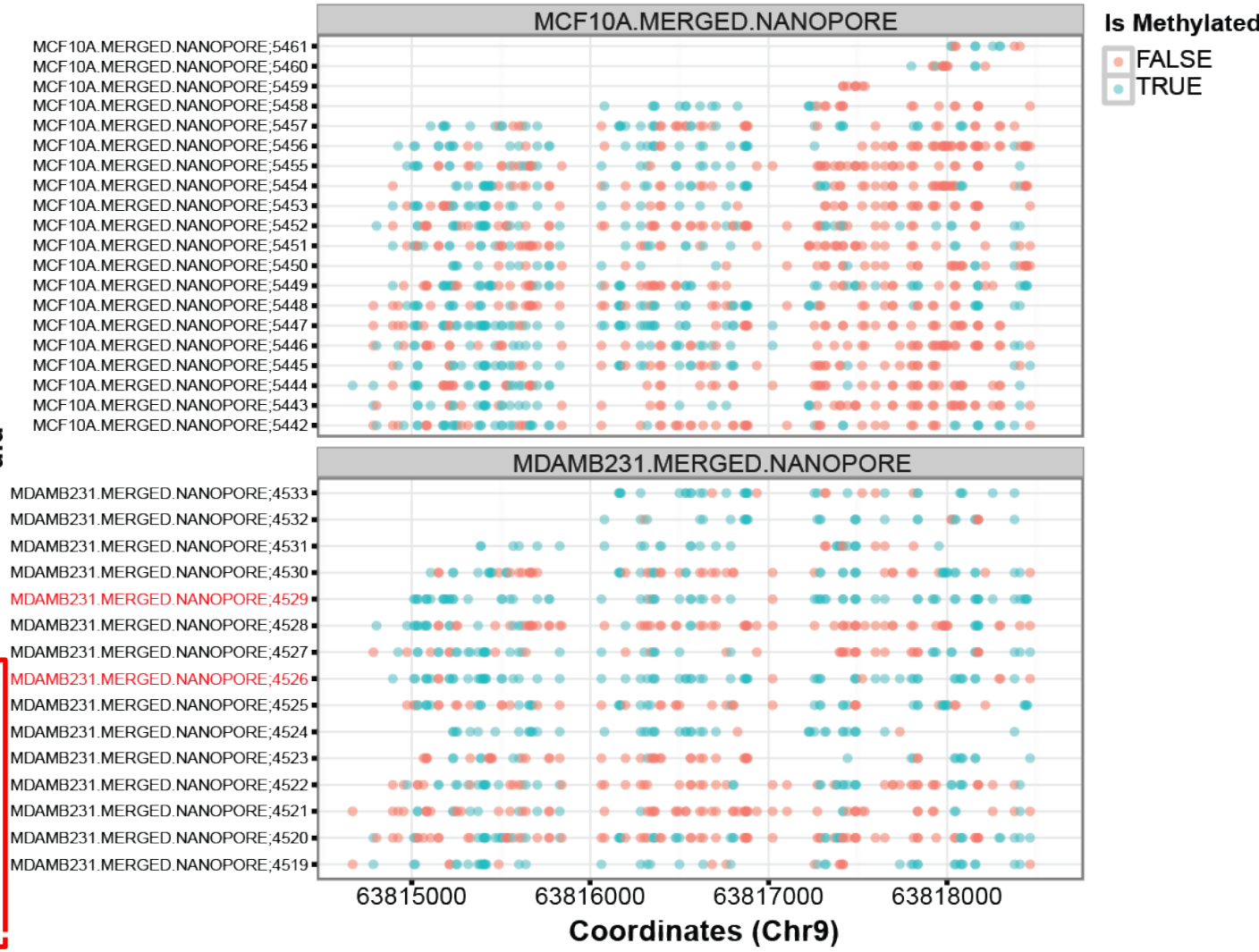


R9 calculates methylation 94% accurate at 77% of sites
NA12878 data shows .895 correlation with bisulfite

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

Cancer-Normal Comparison



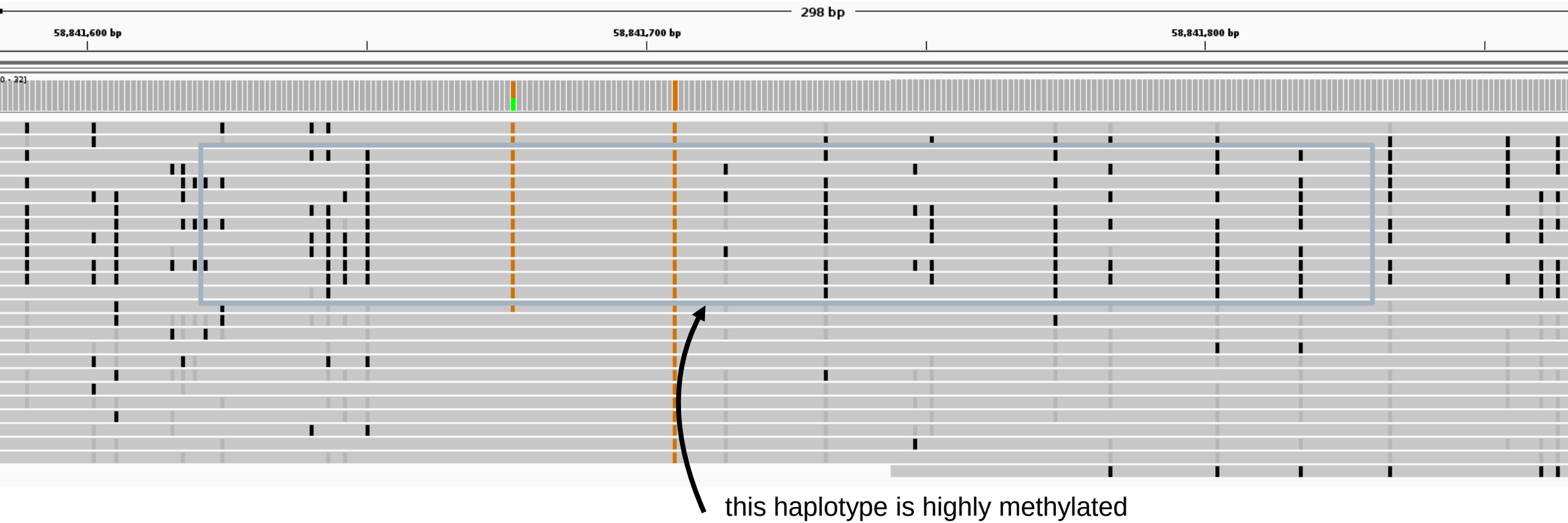


- Reduced representation method:12.5Mb of the genome (3.5-6kb size selection)
- We sequenced this fraction on nanopore and bisulfite Illumina seq
- Long reads measure *phased* methylation

Simpson, Workman, Nature Methods (2017)

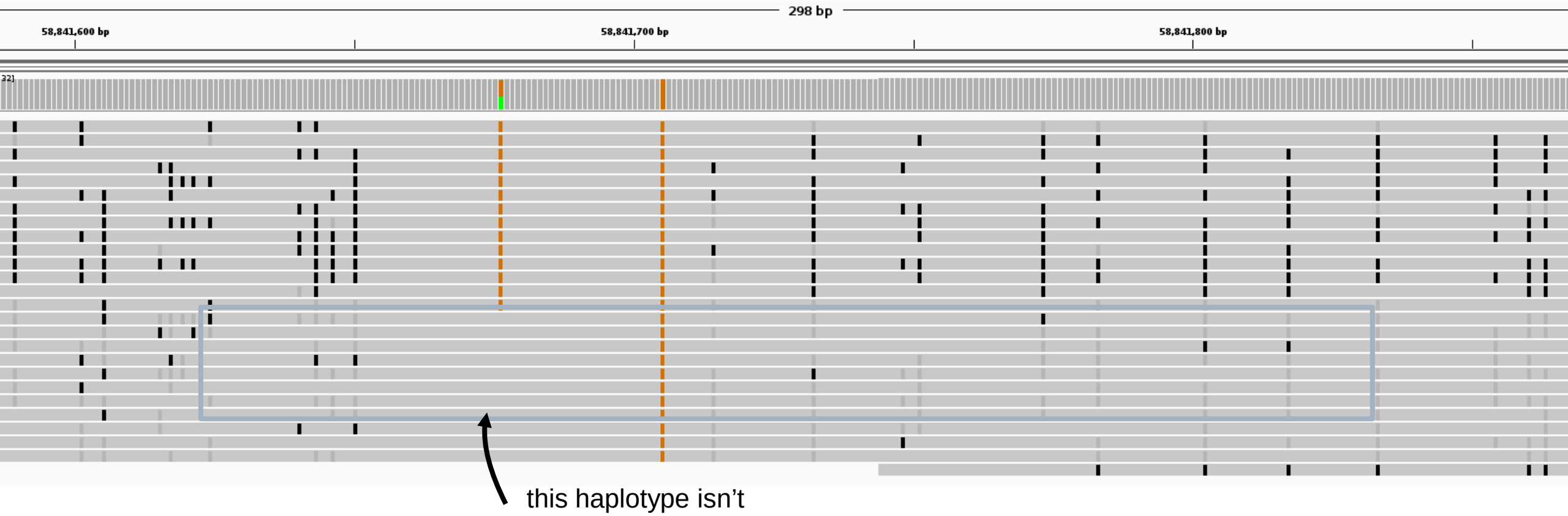
Haplotype-Phased Methylation

nanopolish has experimental support for phasing methylation patterns

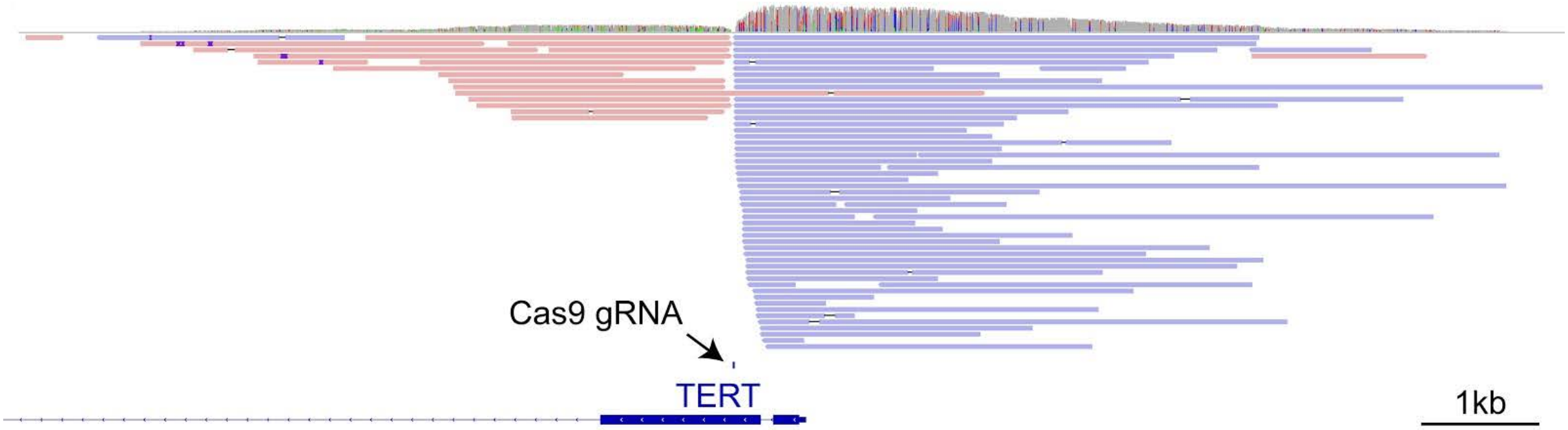


Haplotype-Phased Methylation

nanopolish has experimental support for phasing methylation patterns



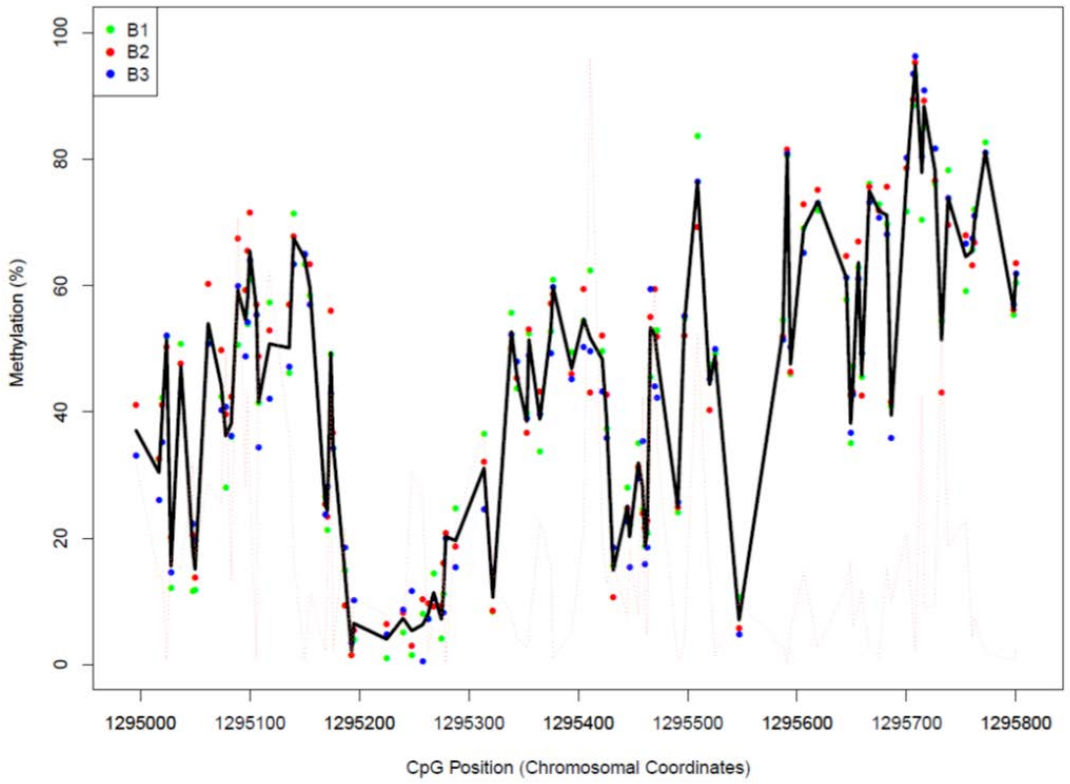
Cas9 Enrichment around target



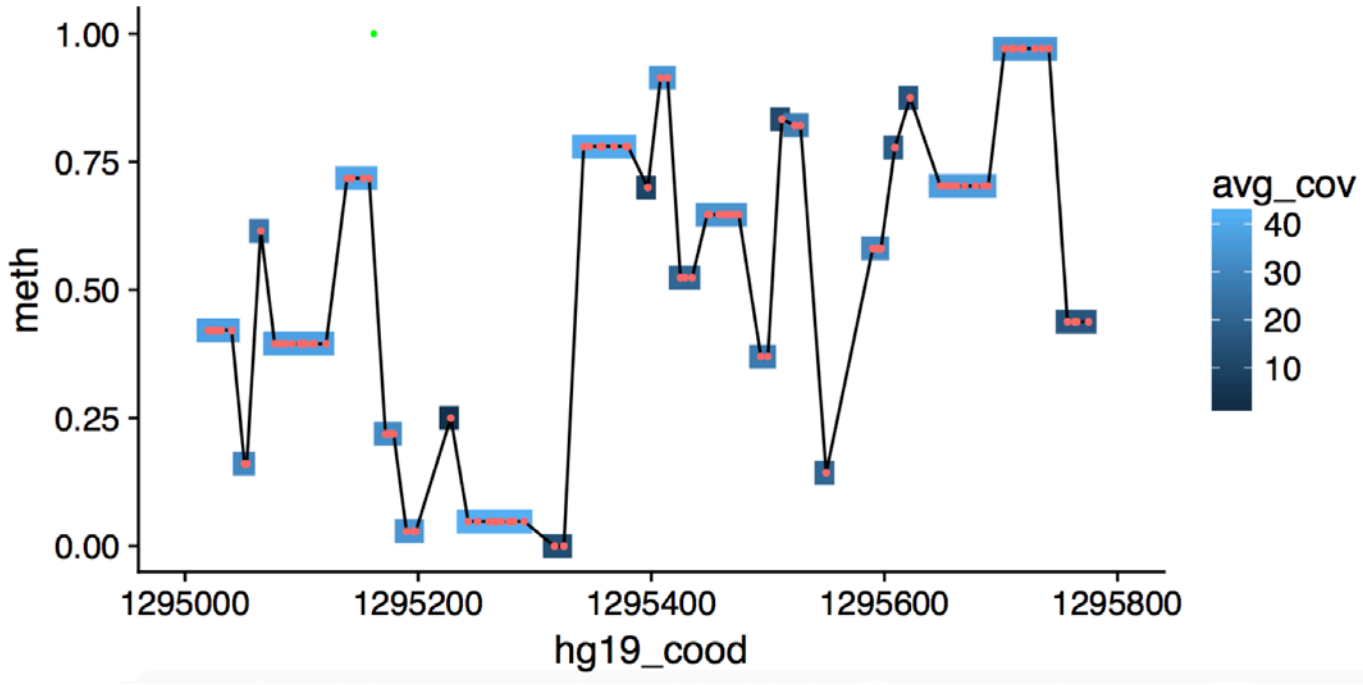
- Capture around the hTERT promoter, region with aberrant methylation in many cancers
- gDNA source from a BCPAP thyroid cancer cell line (poorly differentiated papillary thyroid carcinoma
- Hard to amplify with bisulfite PCR because of high CG-density, required many iterations of primer design

Methylation compare of capture/bisulfite

illumina



Nanopore

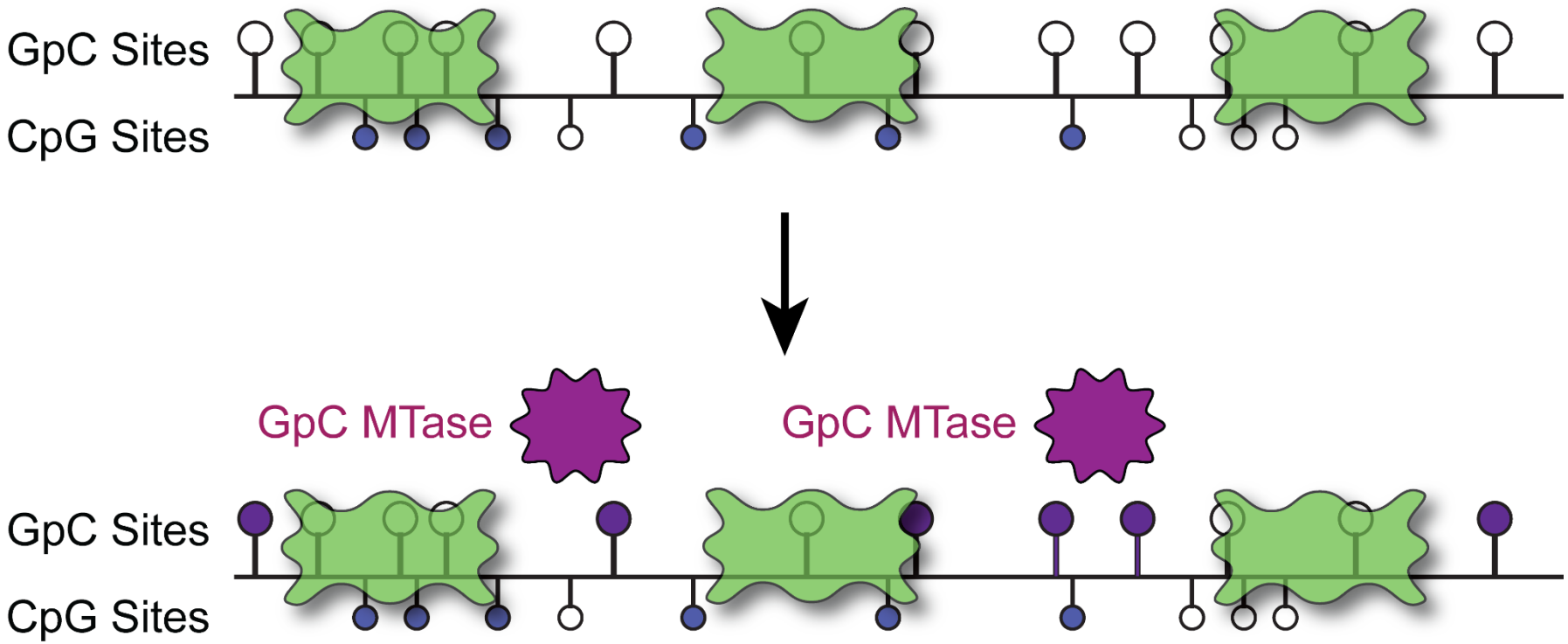


Preliminary data indicates methylation patterns largely concordant between bisulfite and nanopore

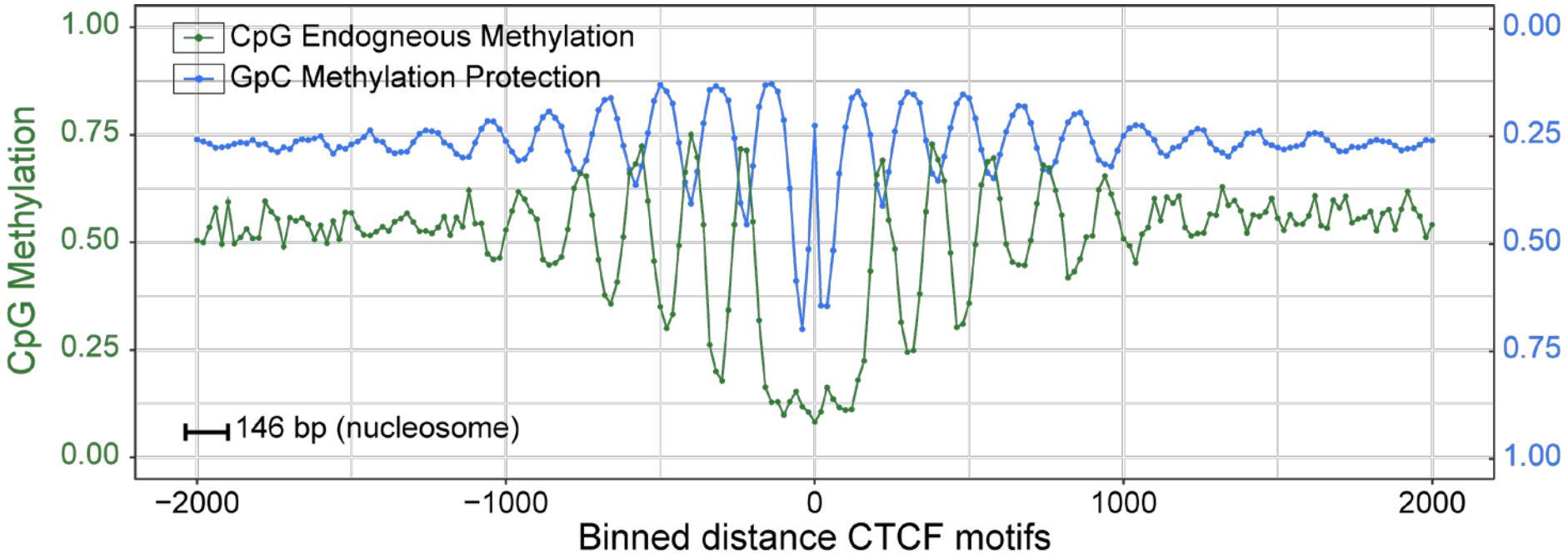
NanoNOMe: Chromatin Accessibility with Nanopore

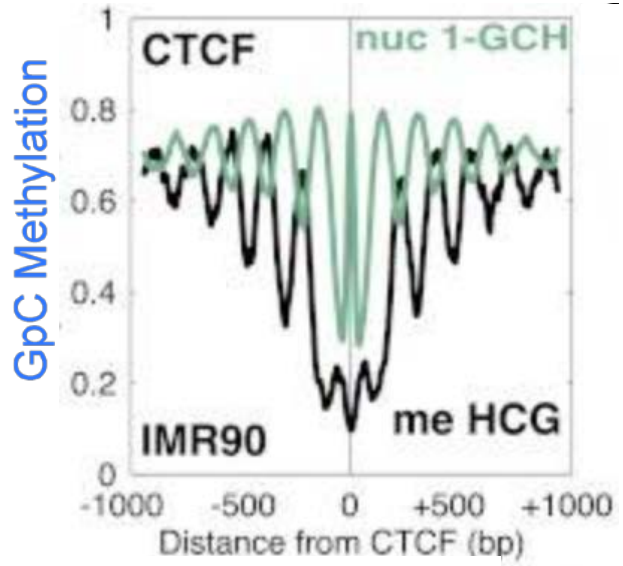
- NOMe-seq : Nucleosome Ocupancy and Methylome sequencing (Kelly et. al. *Genome Res.* 2012)

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



NanoNOMe – Validation with GM12878



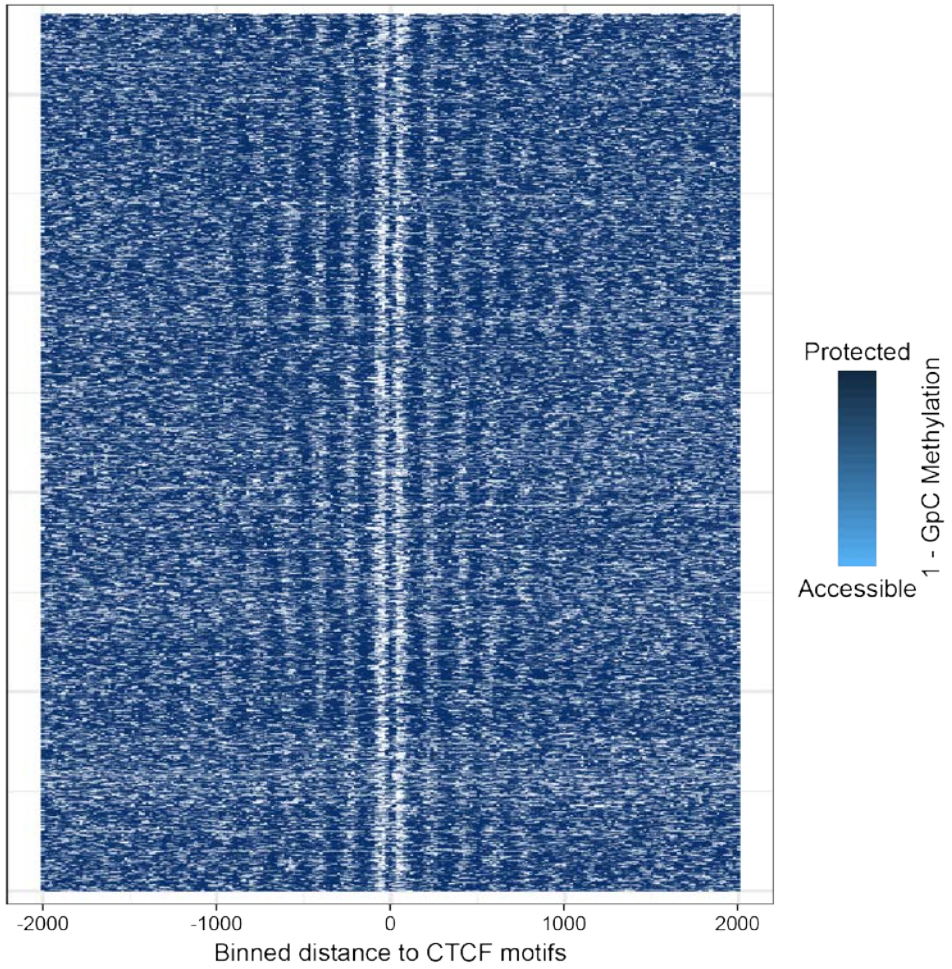


Kelly, et. al. *Genome Res.* 2012

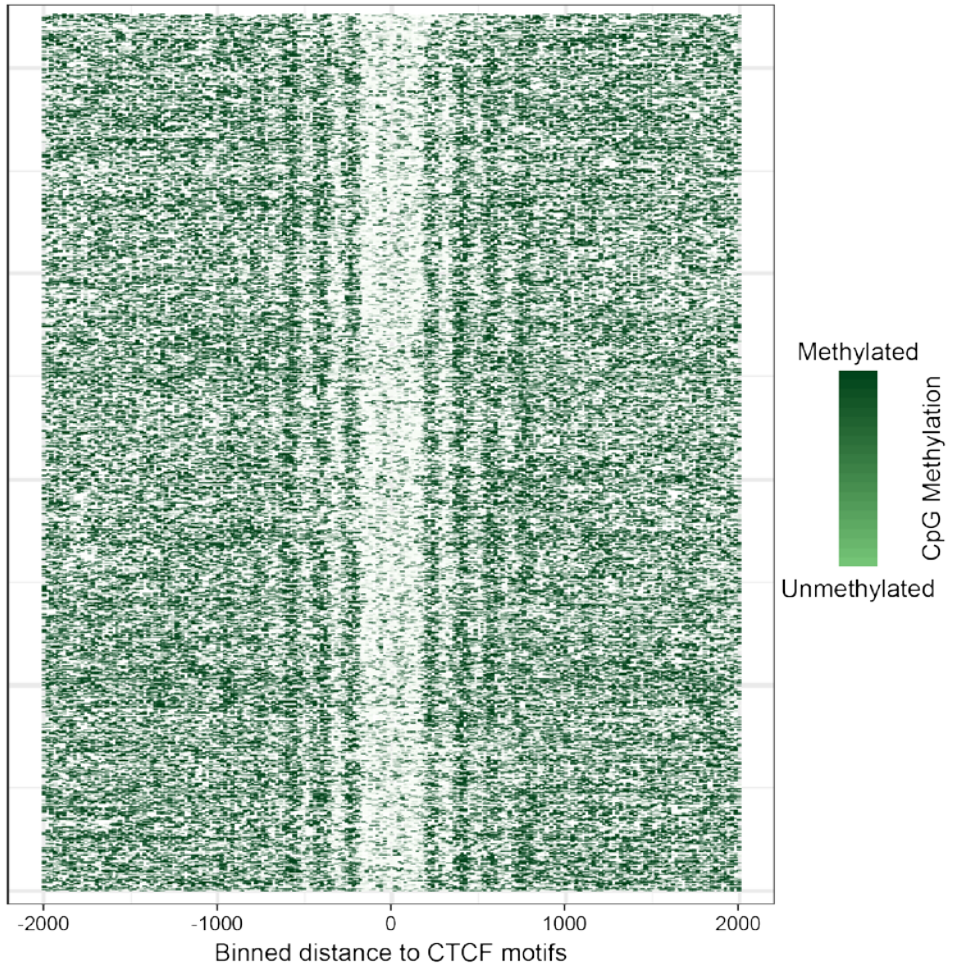
- CpG methylation and open chromatin correlated
- Chromatin states around CTCF agrees with NOMe-seq

NanoNOMe – Validation with GM12878

Chromatin Protection (1-GpC)

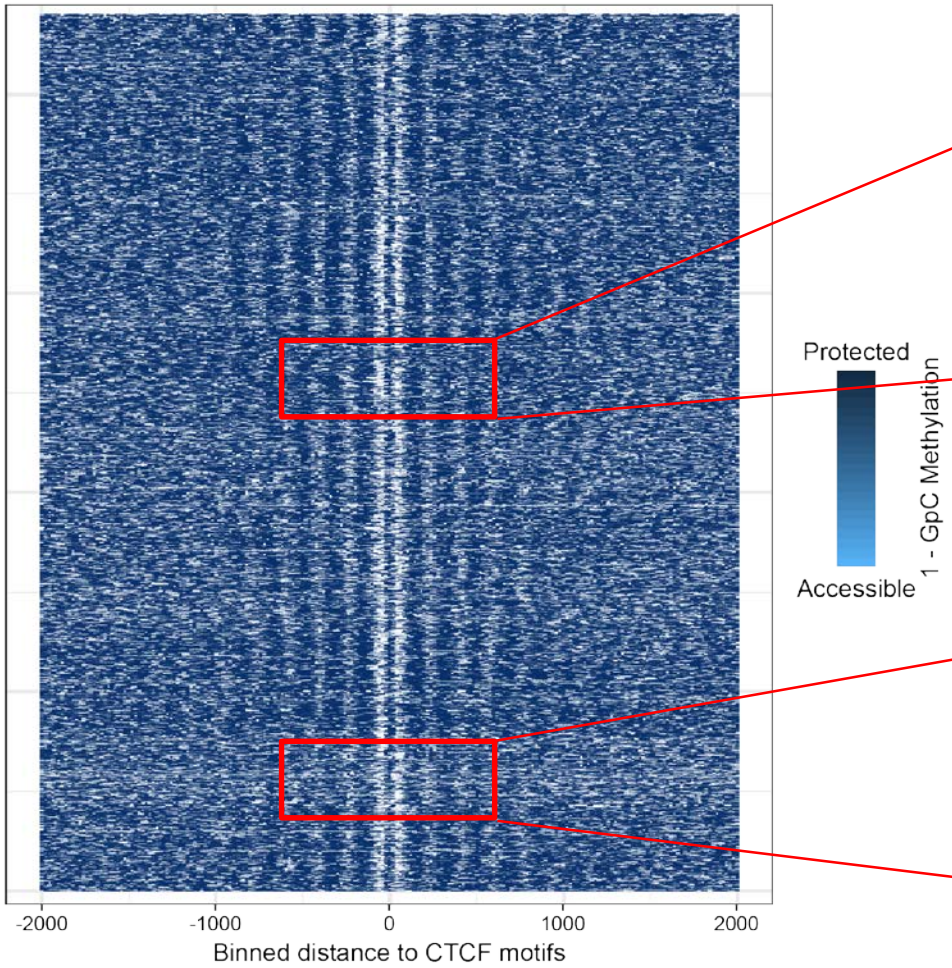


Endogenous Methylation (CpG)

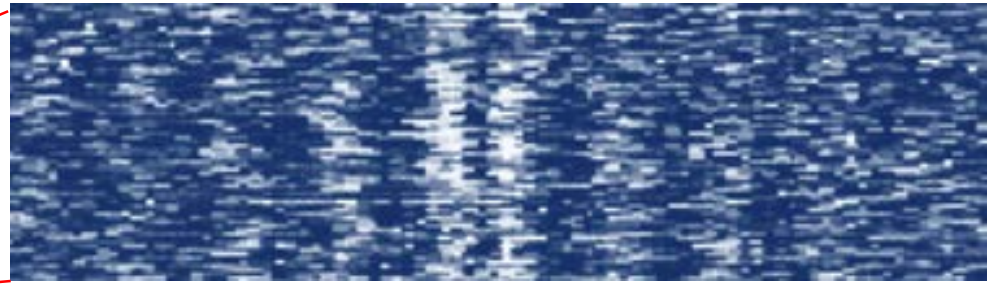


NanoNOMe – Validation with GM12878

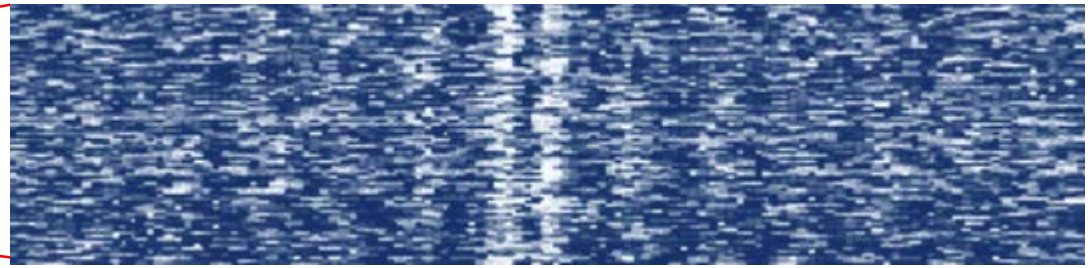
Chromatin Protection (1-GpC)



Shift in nucleosome positioning

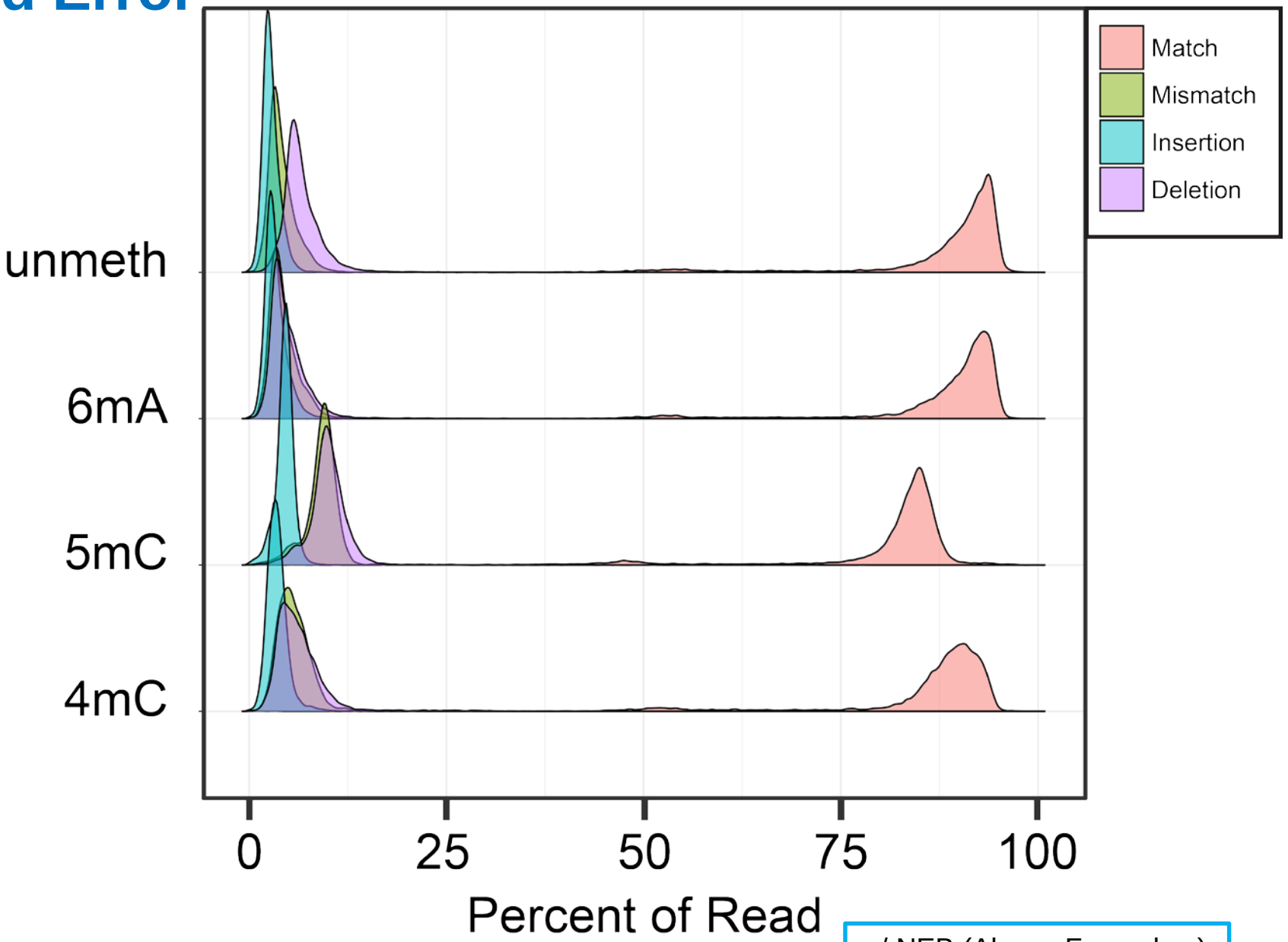


Loss of homogeneity



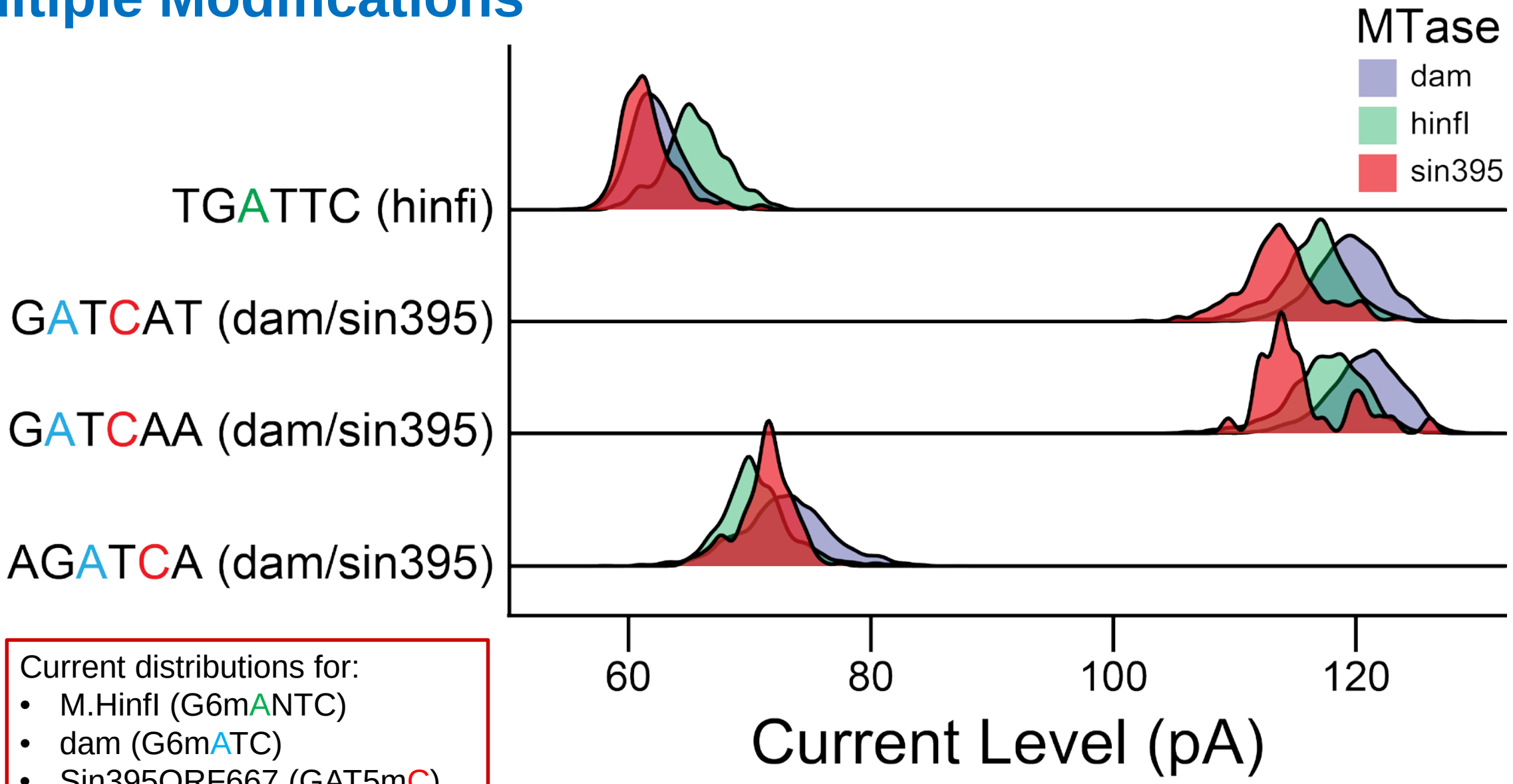
Nanopore: Methylated Error

- 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)

Multiple Modifications

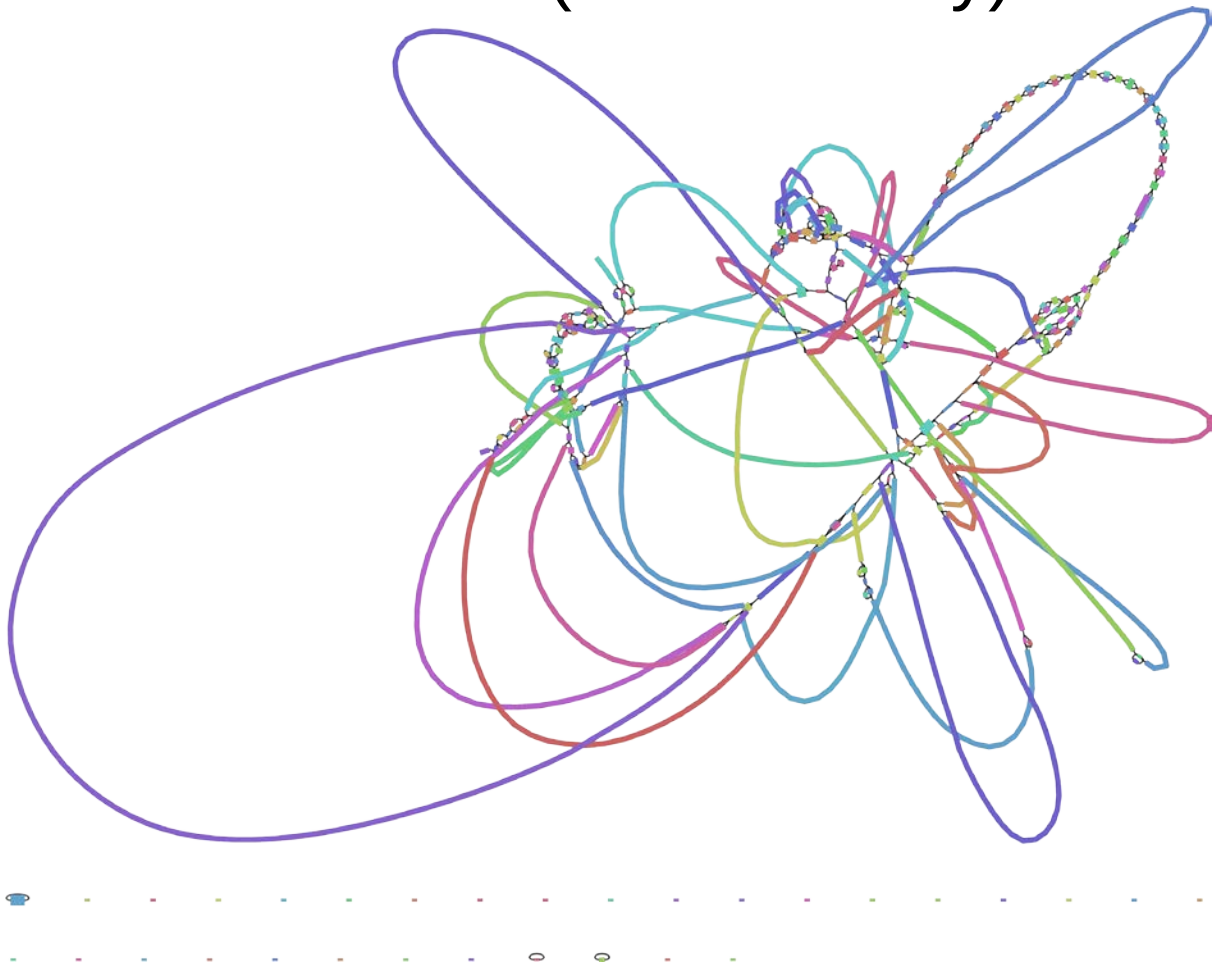


Current distributions for:
- M.HinfI (G6mANTC)
- dam (G6mATC)
- Sin395ORF667 (GAT5mC)

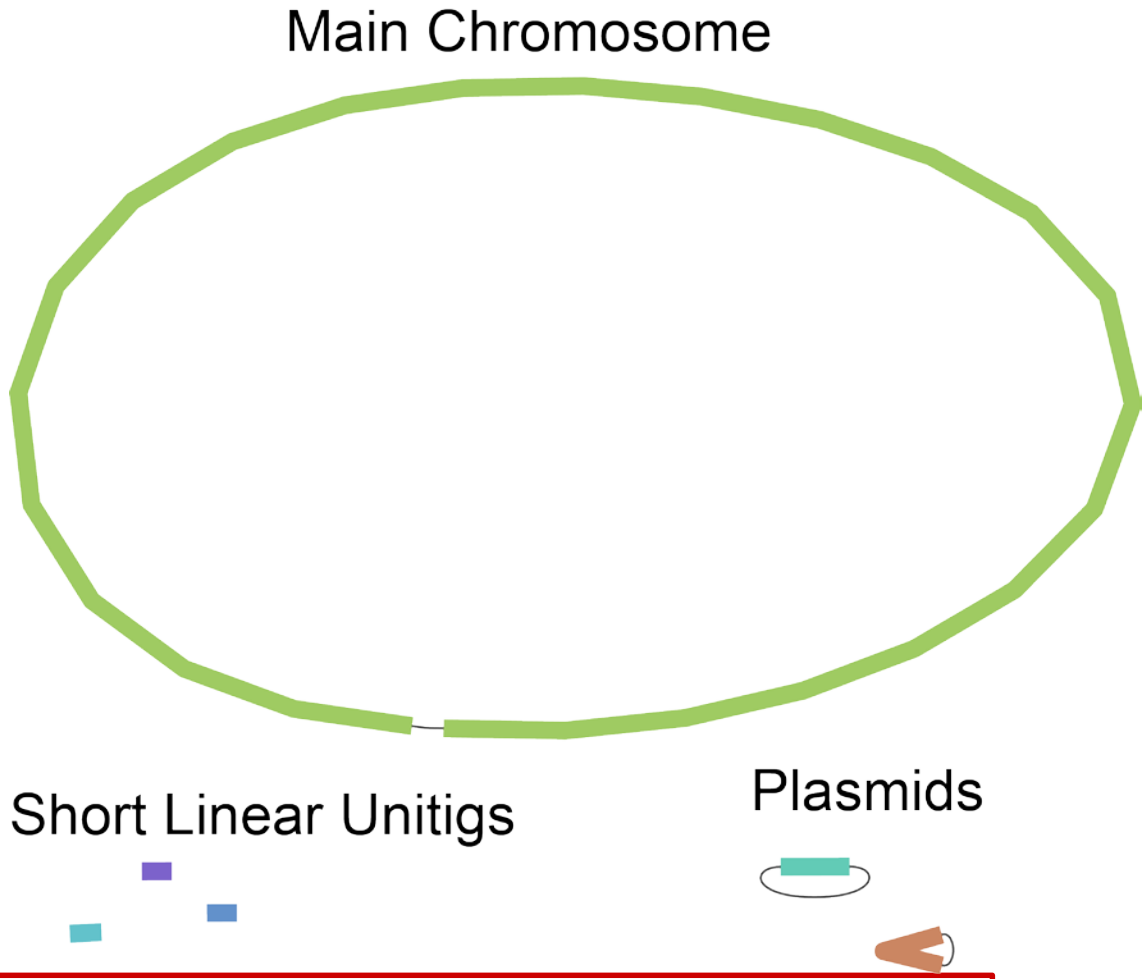
w/ Alexey Fomenkov (NEB)

Assemblies

SPAdes (Illumina only)



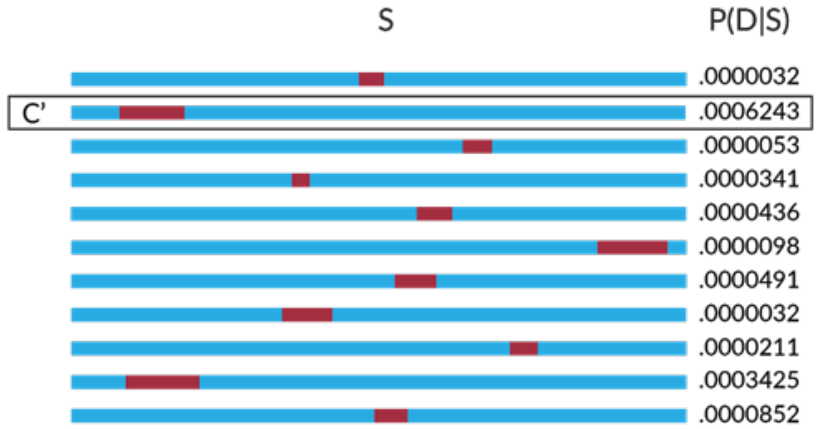
Canu (Nanopore only)



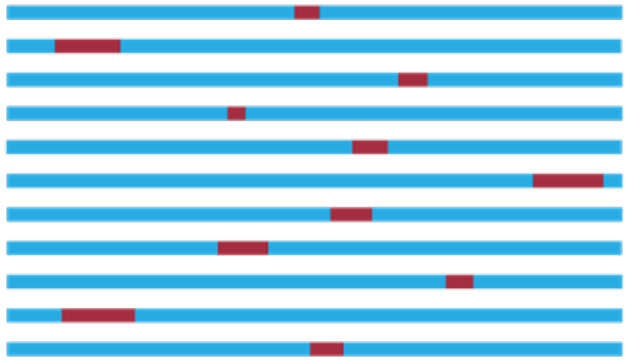
Long reads **really** help in getting complete assemblies – full single contig chromosomes and plasmids identified cleanly.

Nanopolish

Set the best S as the new C, and repeat

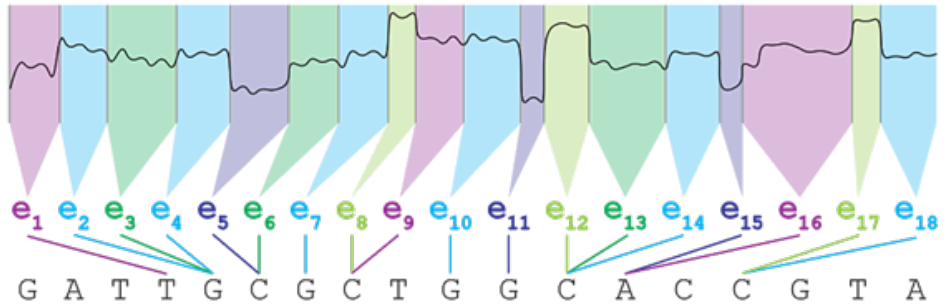


Based on non-matching stretches of sequence between the assembly and the reads, generate a list of candidate improvements to C, called S

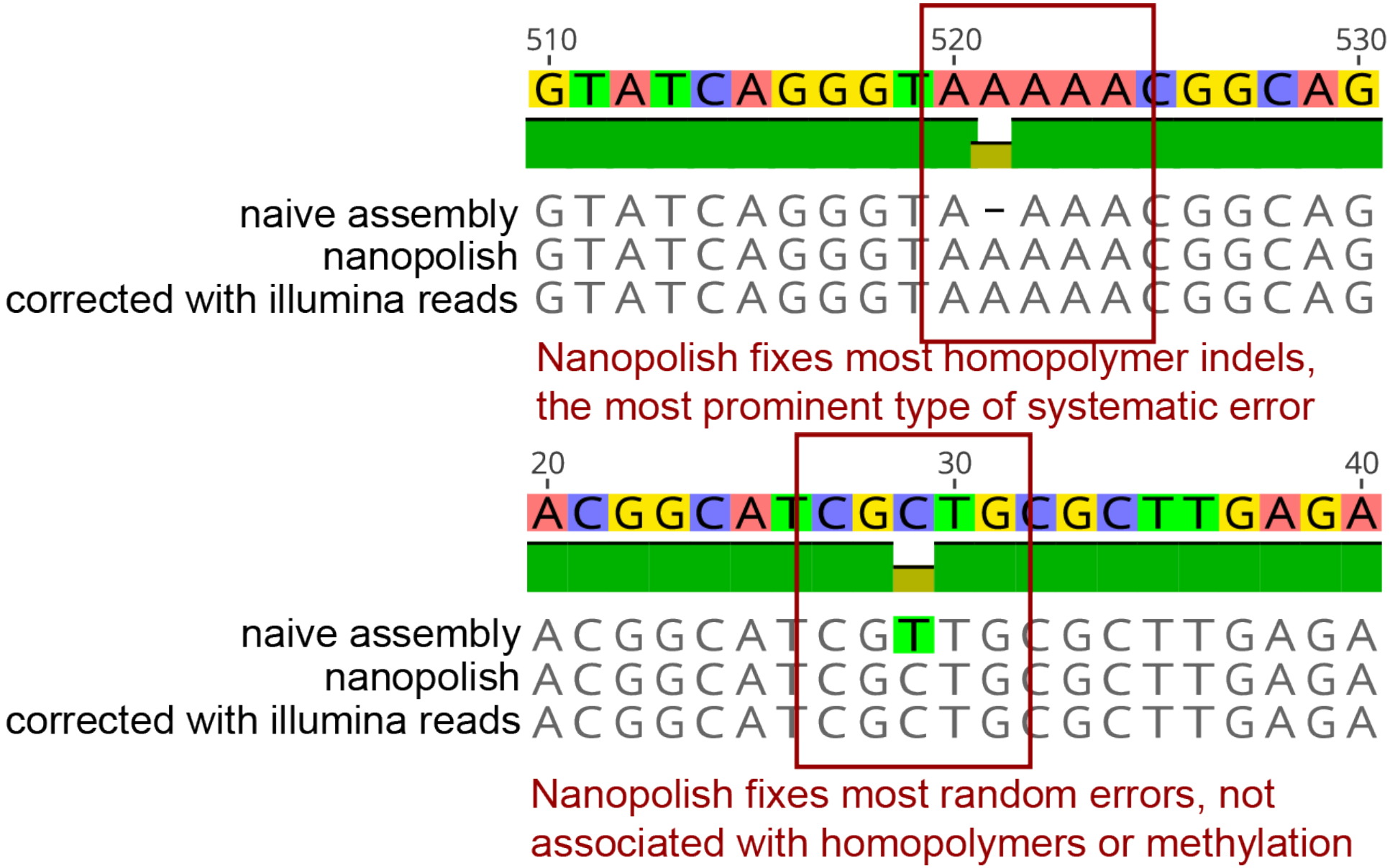


Use a Hidden Markov Model to align the electrical data from the reads to each S, and compute the probability of observing the event data given S, P(D|S)

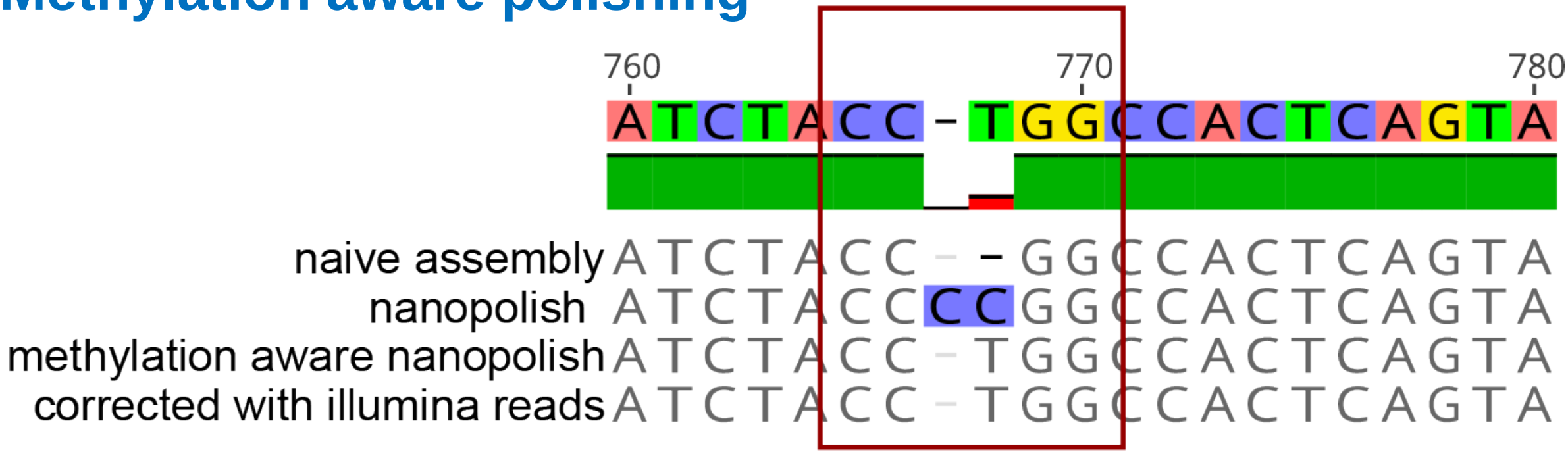
Choose the S that maximizes the probability of observing the event data



Assembly Using Signal to polish



Methylation aware polishing

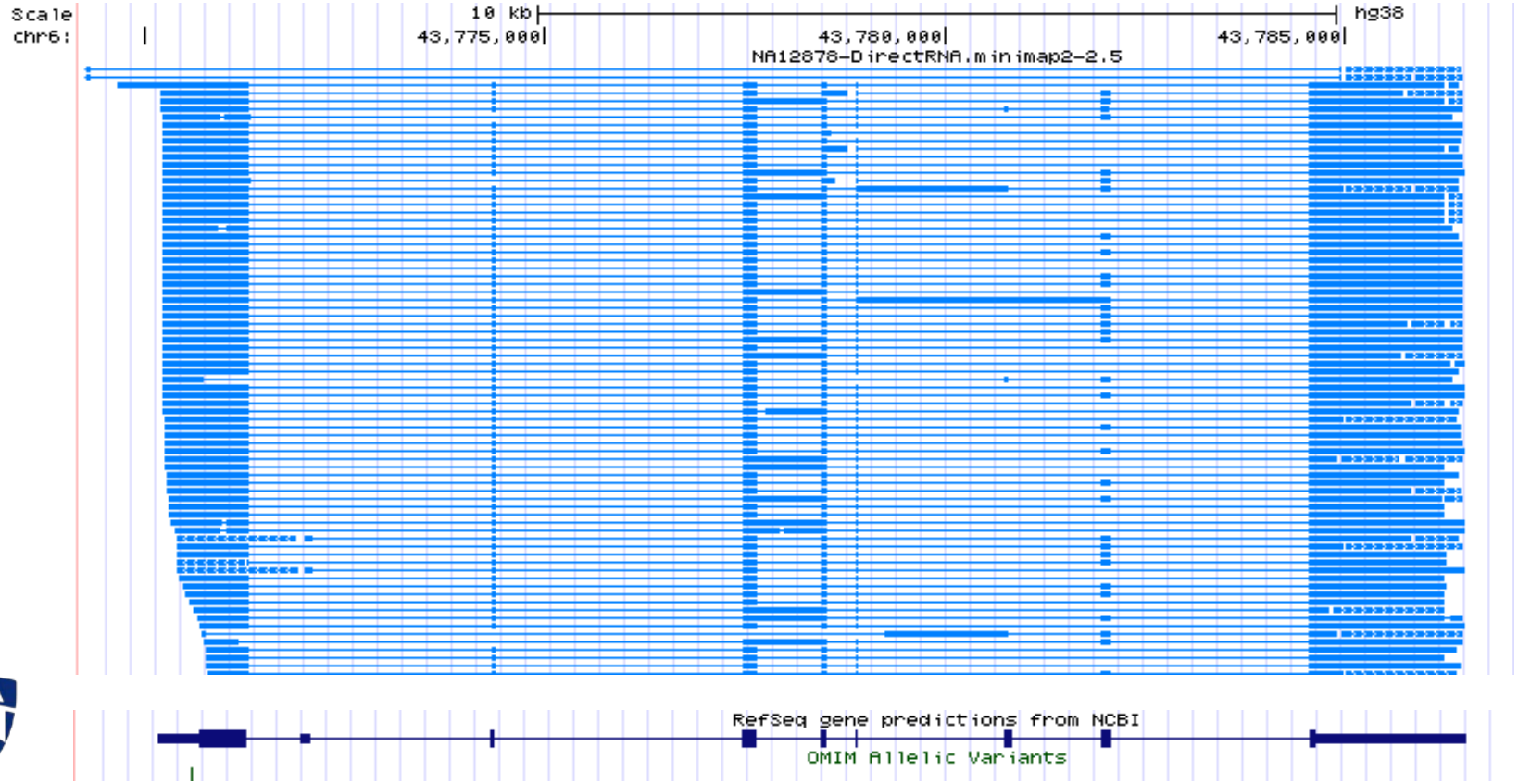


models need to be trained on methylated k-mers in order to correct MTase motifs (dcm MTase= **CCWGG**)

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

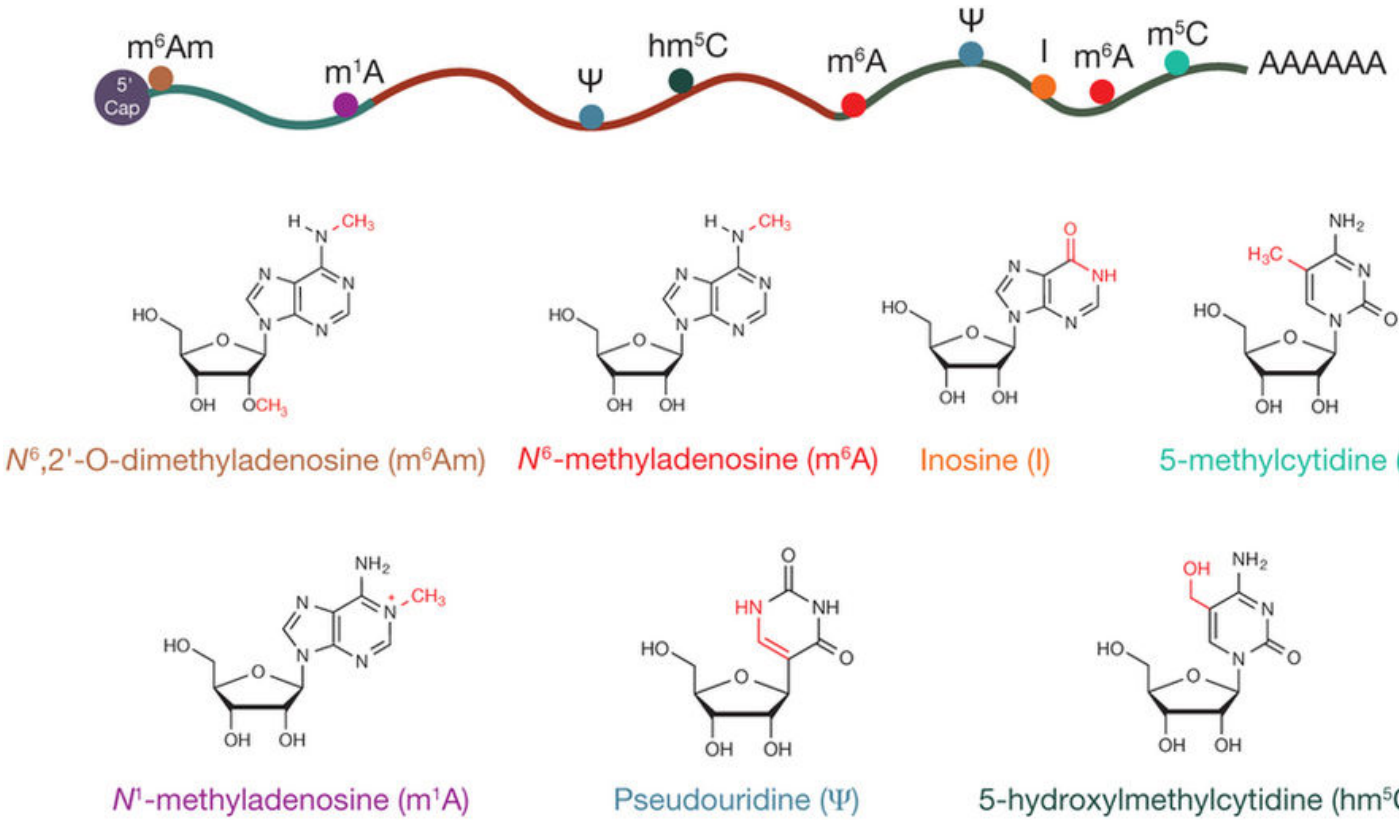
NA12878 RNA Consortium

- 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

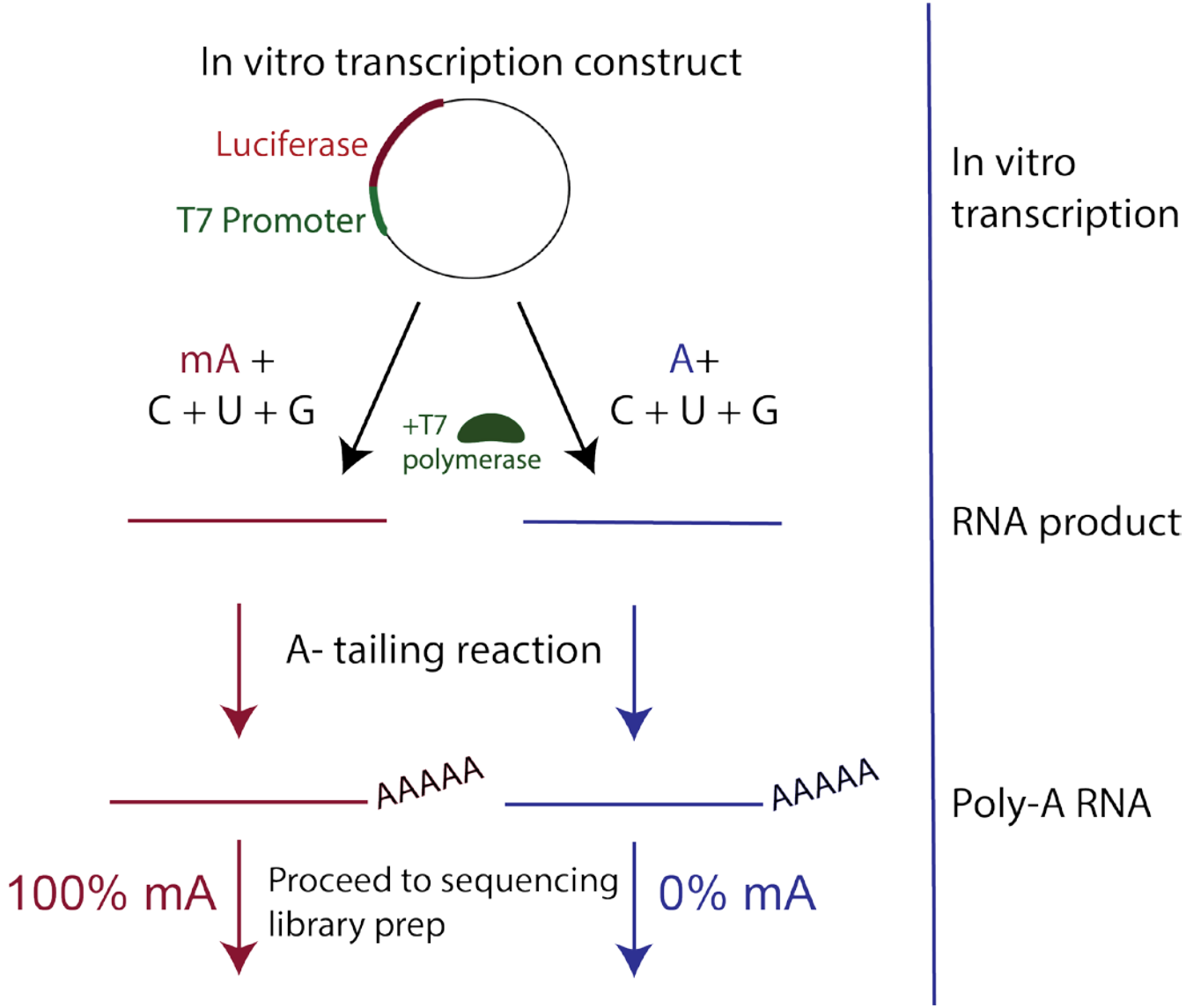
- We can use this to understand RNA modifications – the **epitranscriptome**
- Other methods are challenging – either inefficient, or lack resolution, and always only one modification at a time



Li, Xiong, Yi, Nature Methods (2017)

Detection of RNA modifications with modIVT

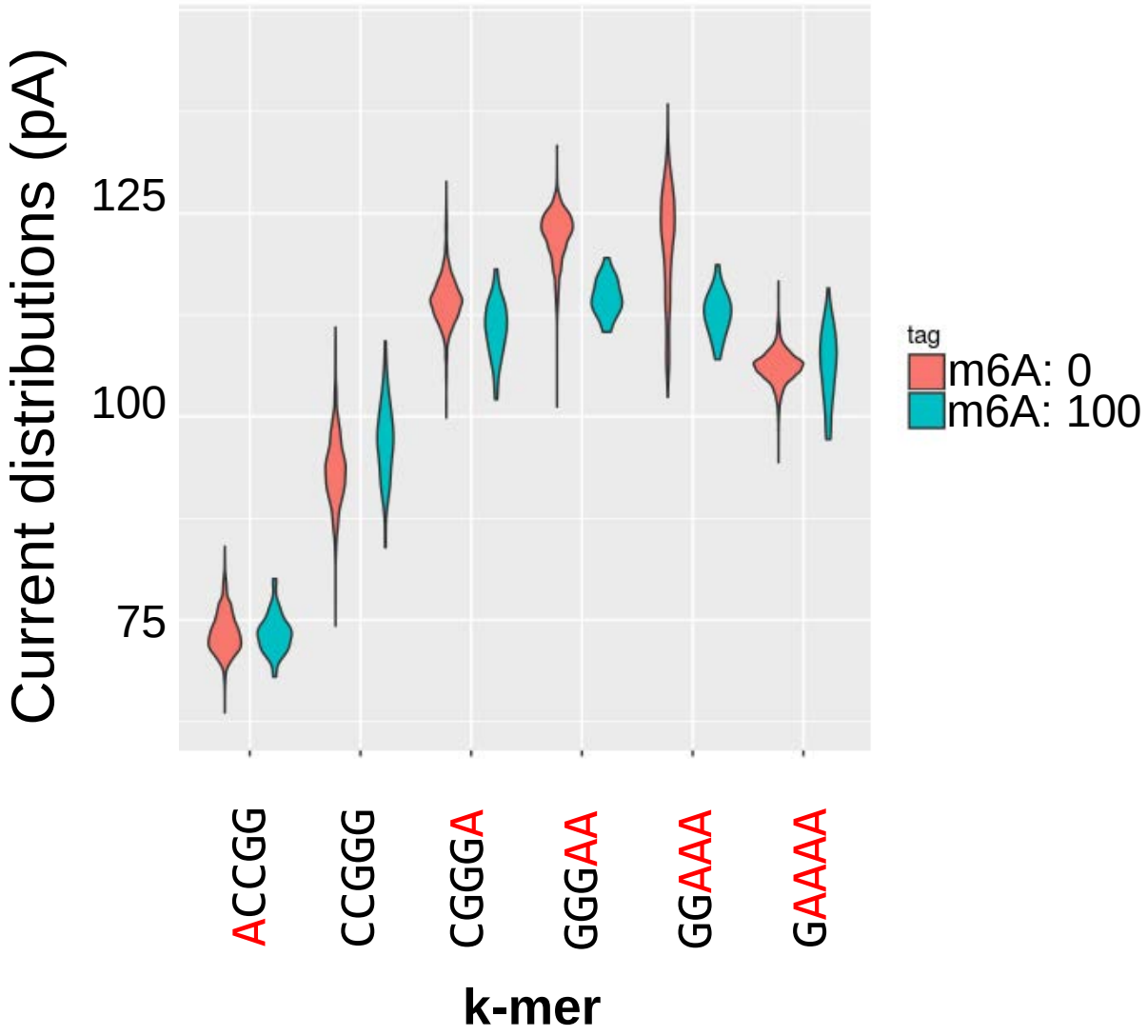
- IVT based RNA synthesis allows incorporation of labeled nucleotides
- All or none reaction right now, T7 has a strong preference for the unmodified nucleotides, making mixtures hard



NEB HiScribe™ T7 High Yield RNA Synthesis kit

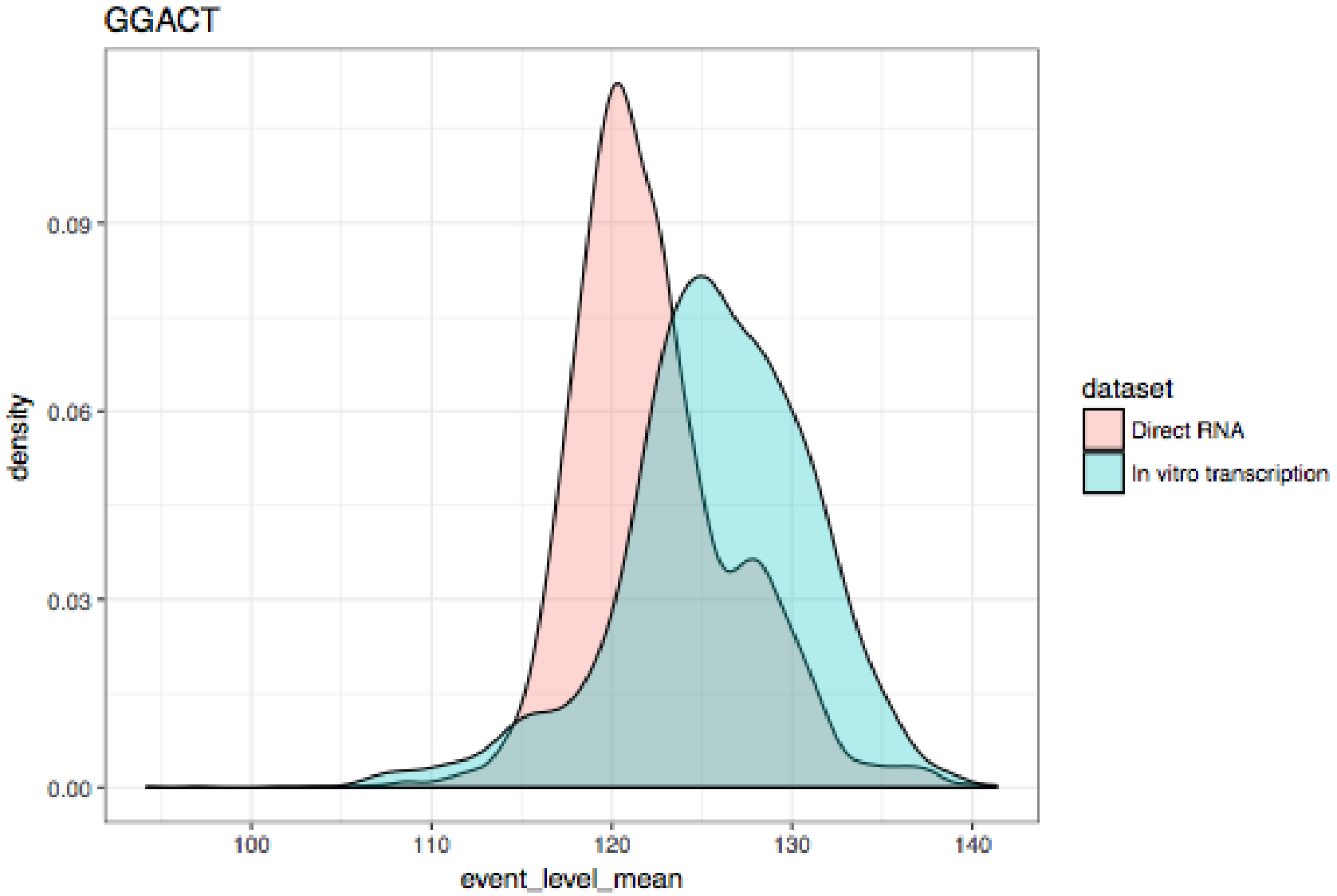
Detection of RNA modifications with modIVT

- 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

- **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

- Alexey Fomenkov, PhD

National Human Genome Research Institute
1R01HG009190-01A1

National Institute of Allergy and Infectious Diseases
1R21AI130608-01 (Simner)

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

Looking for Postdocs!!