13th International Workshop on Advanced Genomics

Tokyo, Japan

June 2019

JOHNS HOPKINS
BIOMEDICAL ENGINEERING

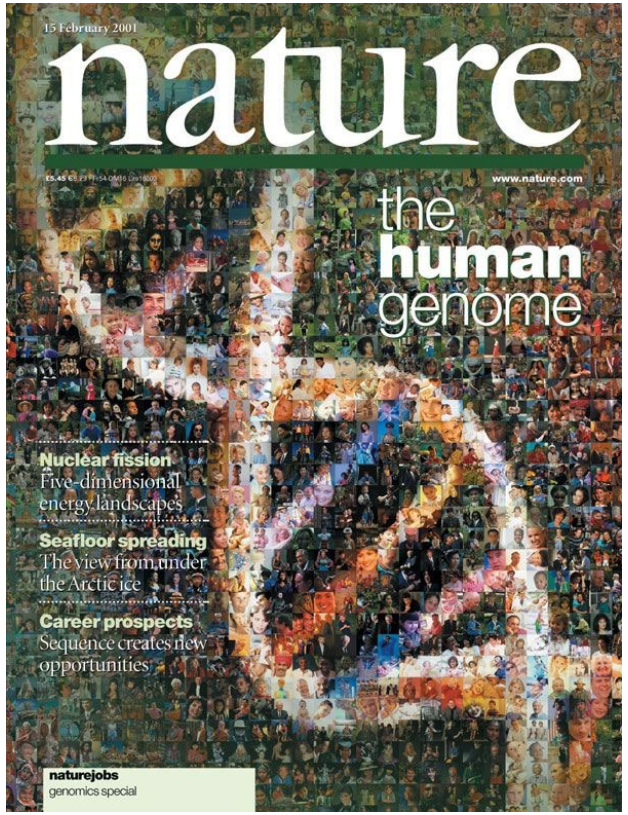
Using nanopore sequencing to interrogate the genome, epigenome and transcriptome

Winston Timp

Department of Biomedical Engineering

Johns Hopkins University

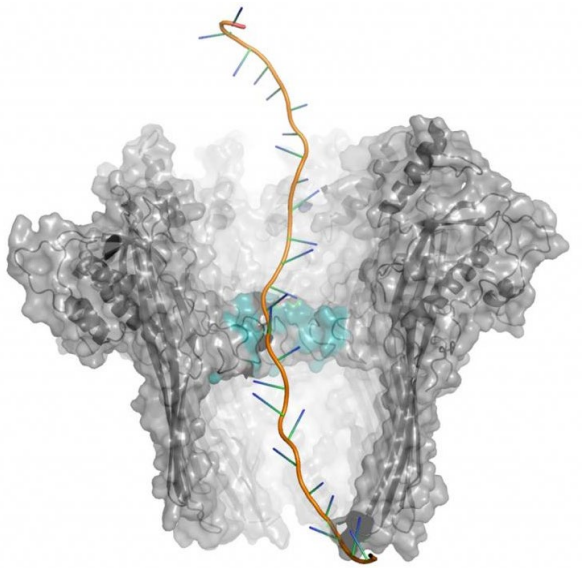
Revolutions in Science: Genomics



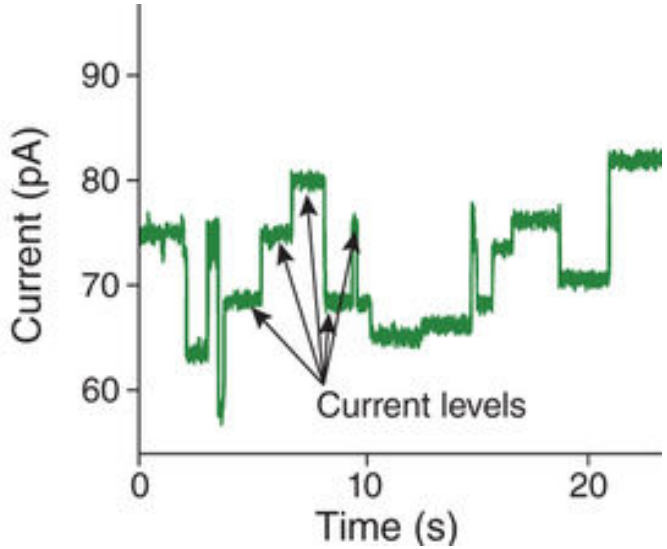
- Draft of the human genome was completed in 2001
- ~3 billion bases in size
- Think about this like the first transistor (1947) – the watershed after which genomic and epigenomic engineering has exploded

Nanopore: Single Molecule Sequencing

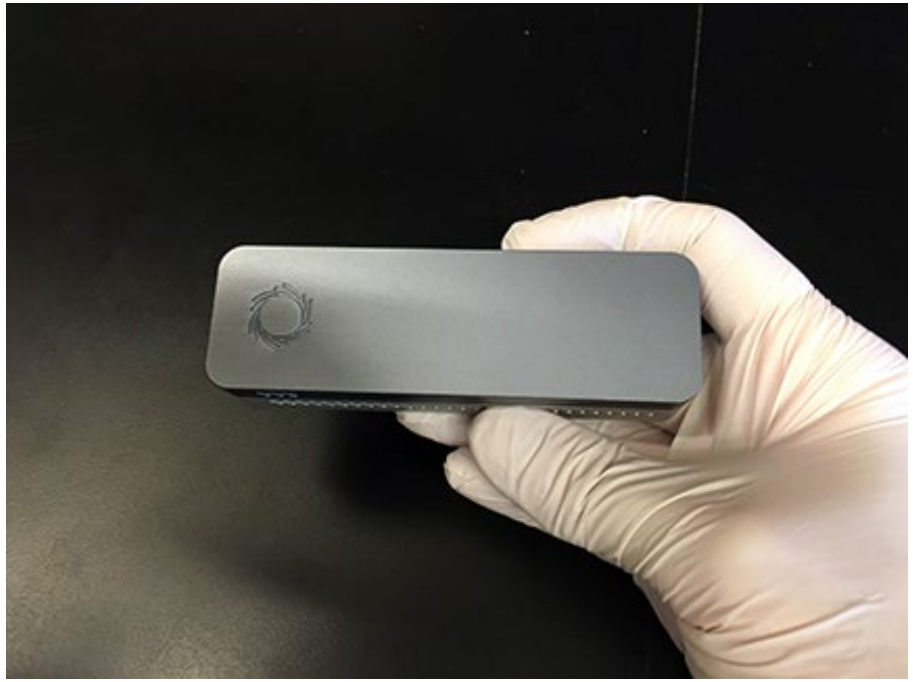
- Oxford Nanopore Technologies, CsgG biological pore
- No theoretical upper limit to sequencing read length, practical limit only in delivering DNA to the pore intact
- Palm sized sequencer
- Sequencing output 5-15Gb



Oxford Nanopore Google Hangout March 2016



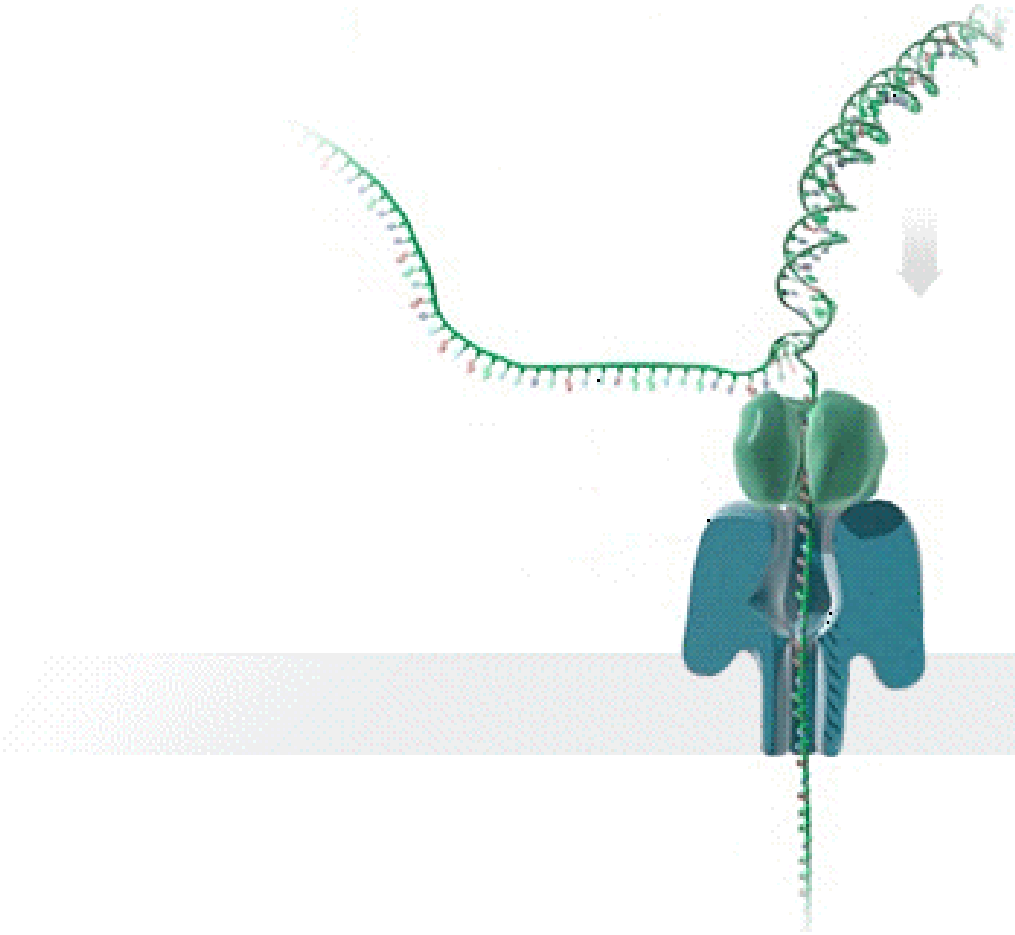
Deamer et al 2016, Nature Biotech

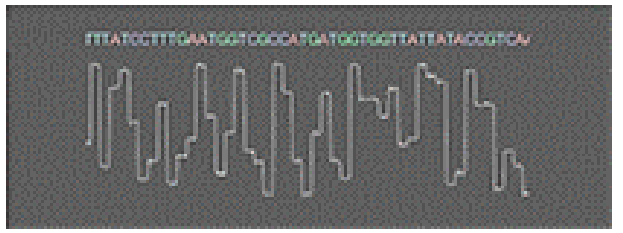


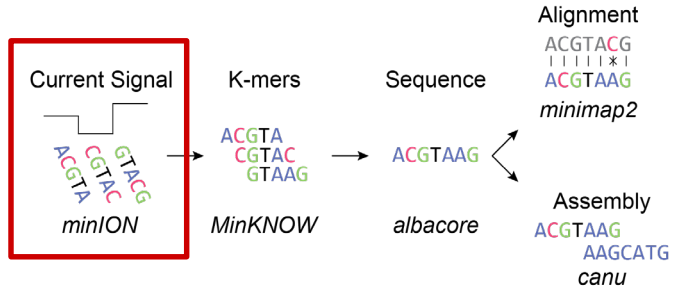
ATCGATCGATAGTAT
TAGATACGACTAGC
GATCAG

Disclosure: Timp has two patents (US Patent 8,748,091; US Patent 8,394,584) licensed to ONT

Sequencing Operation



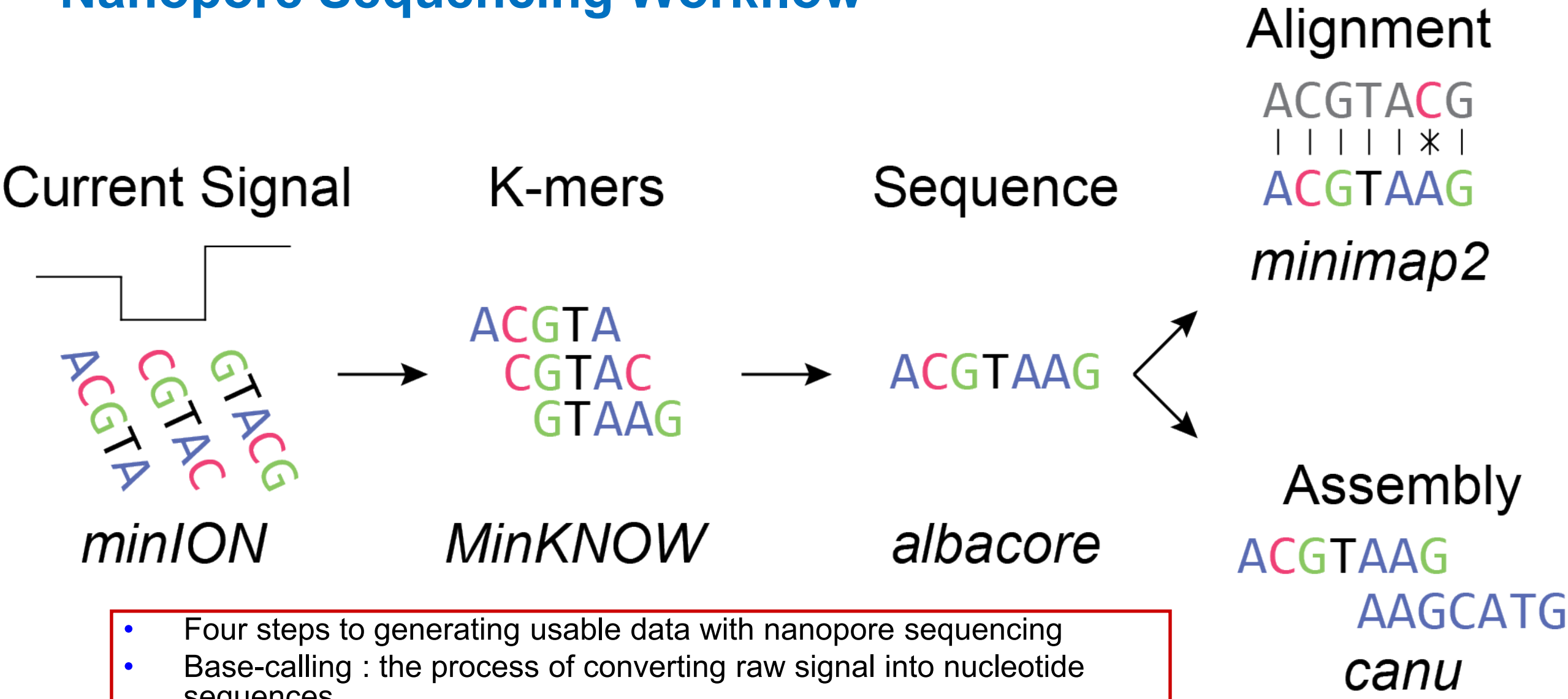




Oxford Nanopore Technologies

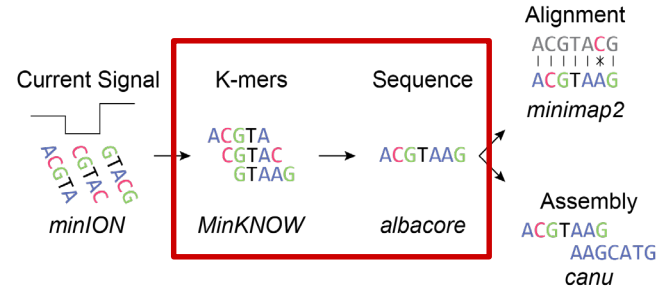
- Protein nanopores on a synthetic polymer
- Multiple base-pairs at a time (“k-mers”)
- Characteristic current signature is converted to nucleotide sequences

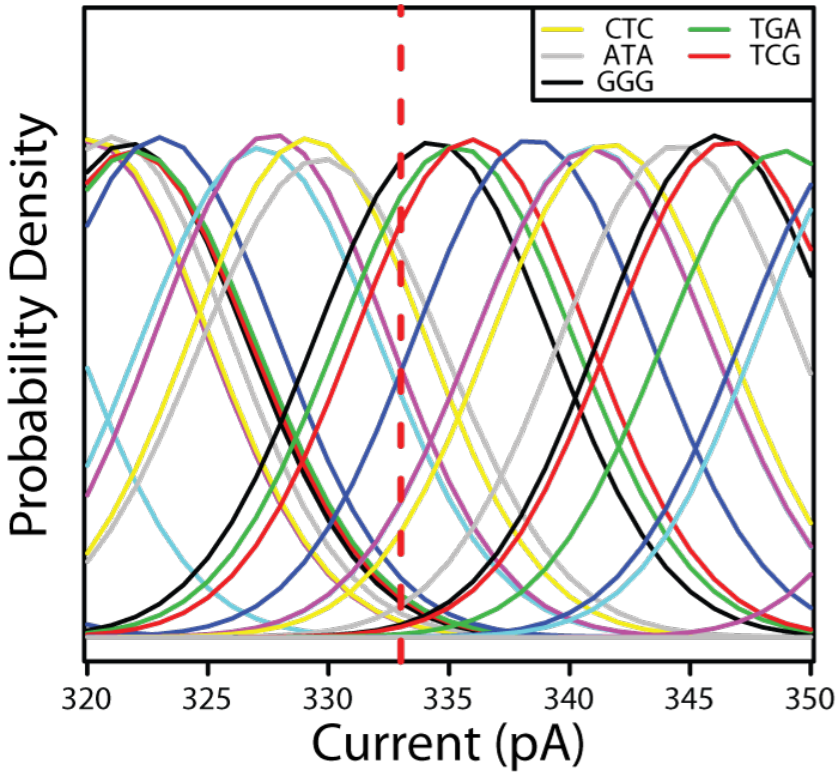
Nanopore Sequencing Workflow

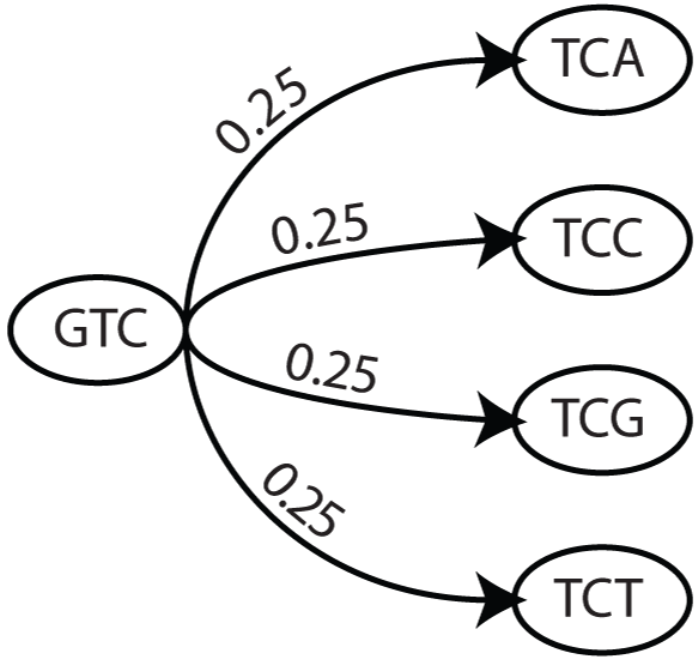


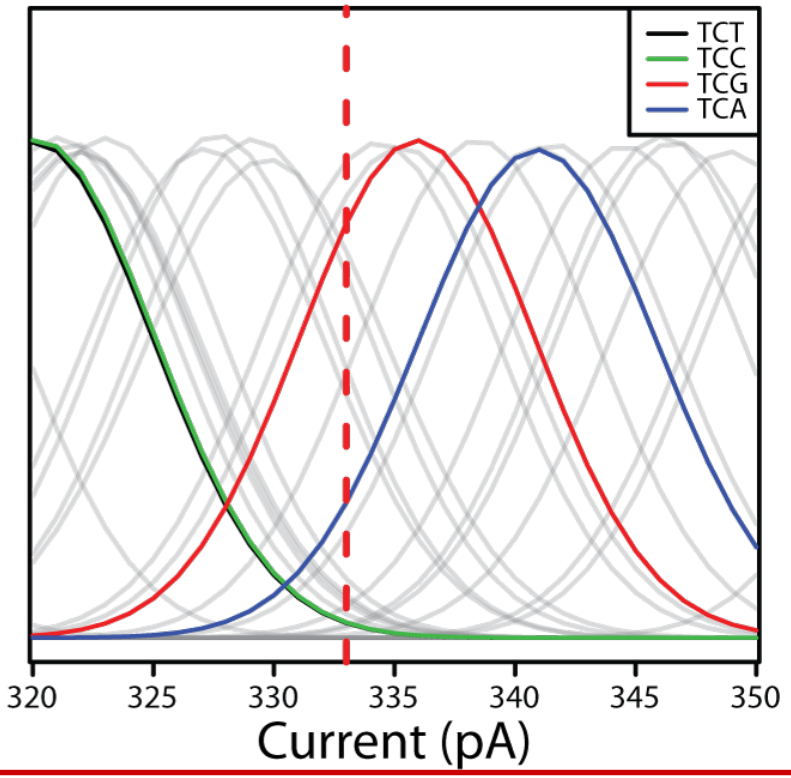
- Four steps to generating usable data with nanopore sequencing
- Base-calling : the process of converting raw signal into nucleotide sequences
- Nanopolish : uses alignment and current signal to improve base-calls

Prior Information for Decoding





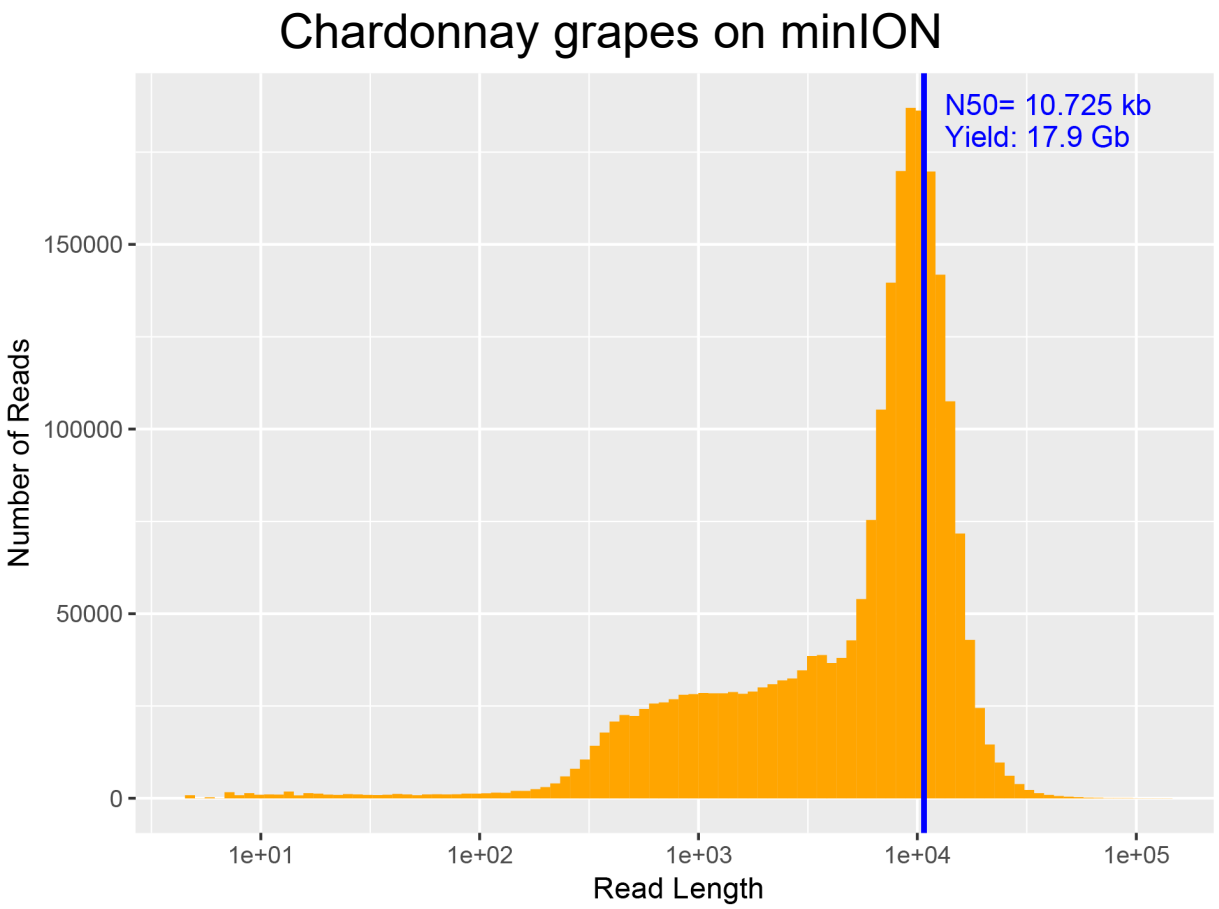


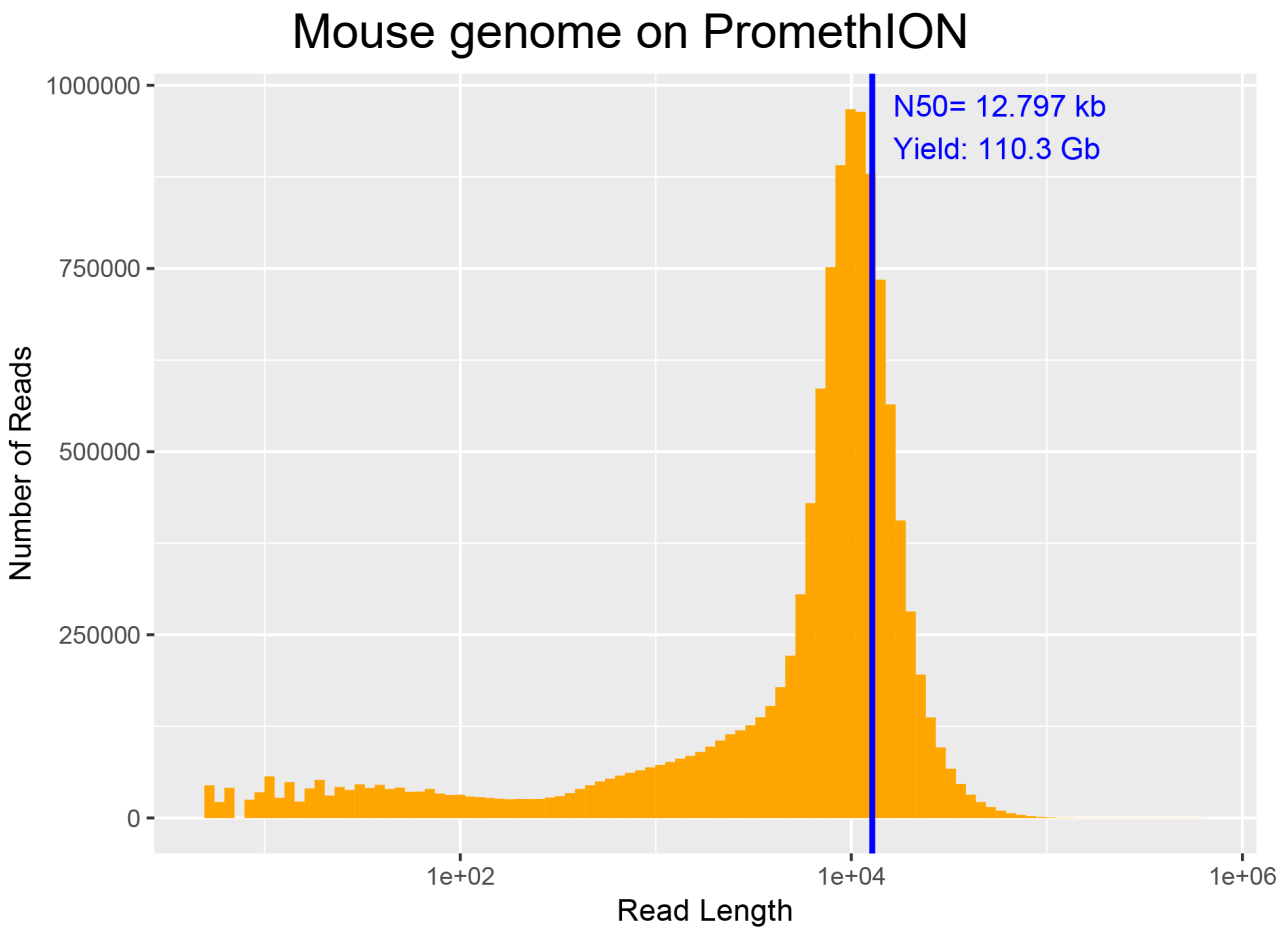


- With no prior information, a given current value may not be called correctly (333pA would be called as GGG)
- If we know the previous triplet, the next triplet is well defined, leaving only four possibilities, resulting in the correct call of TCG

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

Improving Yield

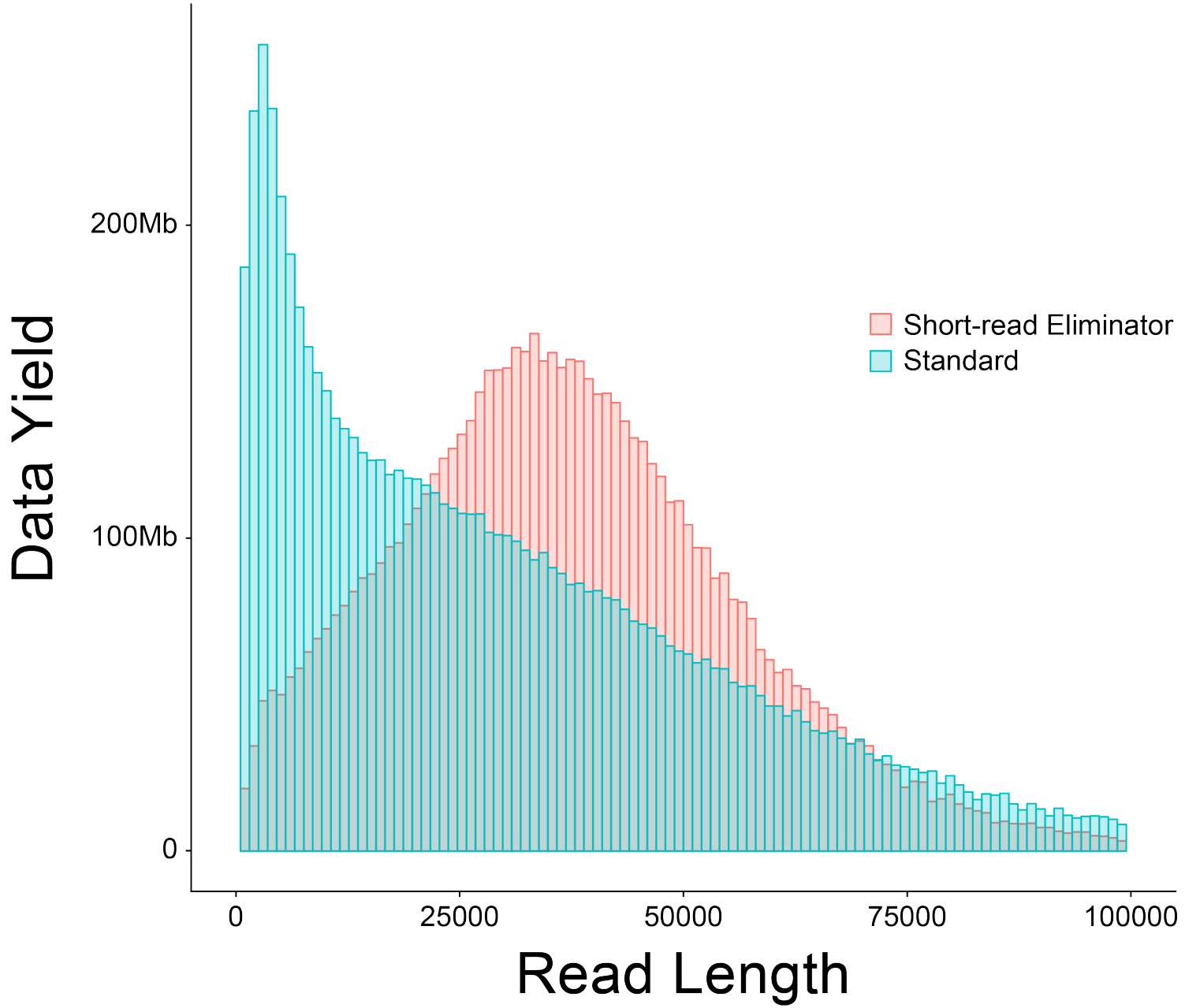




Yield is continually improving; costs for good runs are now commensurate with costs (in raw yield) for Illumina sequencing (~$25/Gb)

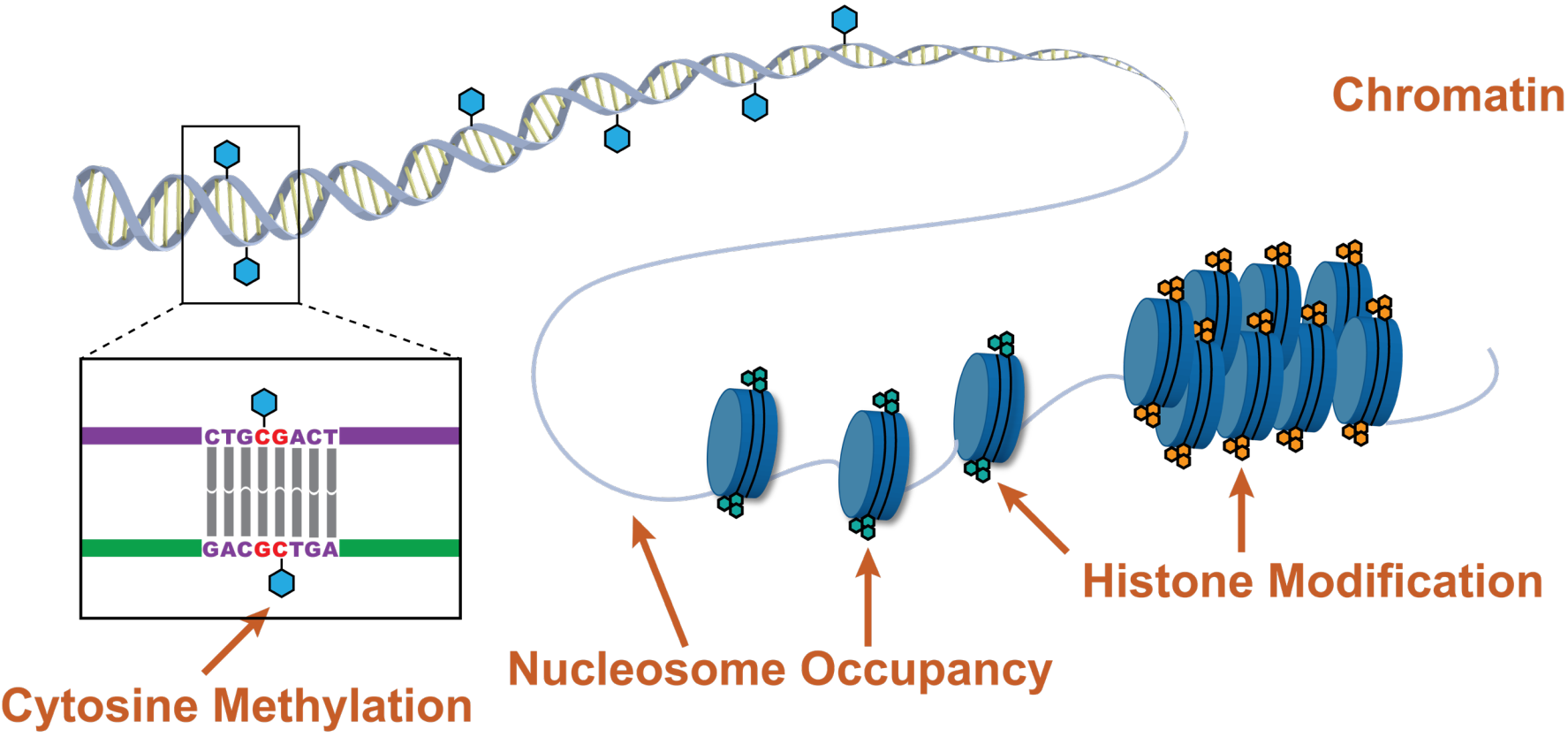
Read Length

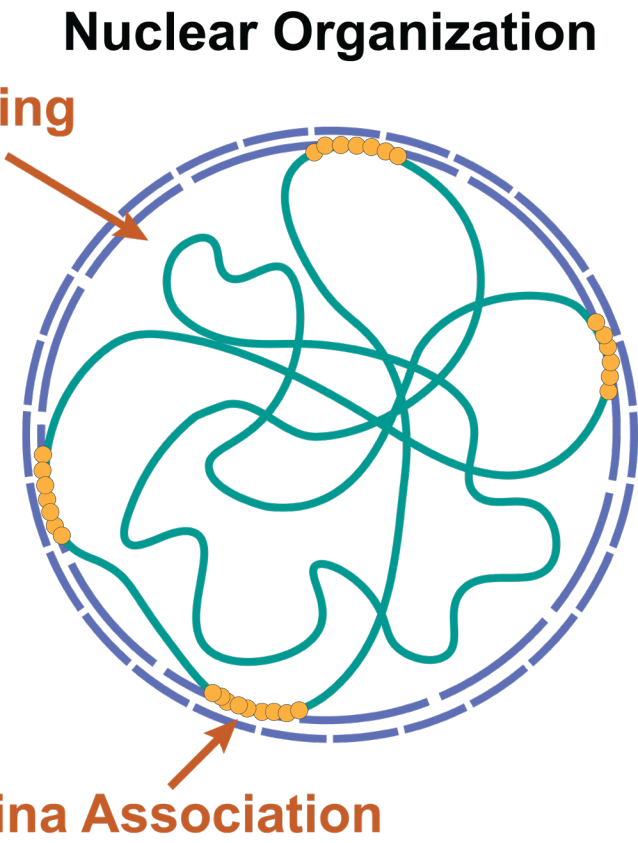
- Working with Circulomics we have been trying to get the read length up
- Using a size selection with their Nanobind material, read N50 can be substantially improved
- There is still room for improvement often still difficult to get both high yield and high read length



circulomics

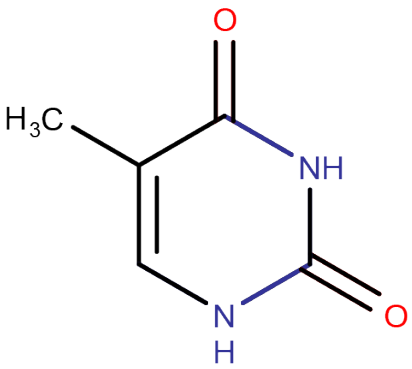
Epigenetics: Modern



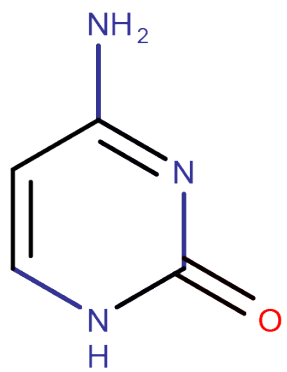


- Modern Definition of epigenetics involves heritable changes other than genetic sequence, e.g., positive feedback, high order structure, chromatin organization, histone modifications, DNA methylation.
- An analogy to a computer system:
  - DNA Sequence = Hardware
  - User input = Environment
  - Systems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

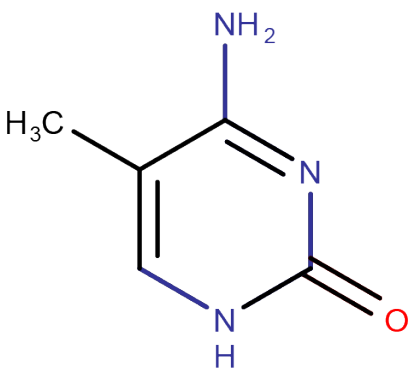
Nanopore Sequencing of Modifications



Thymine

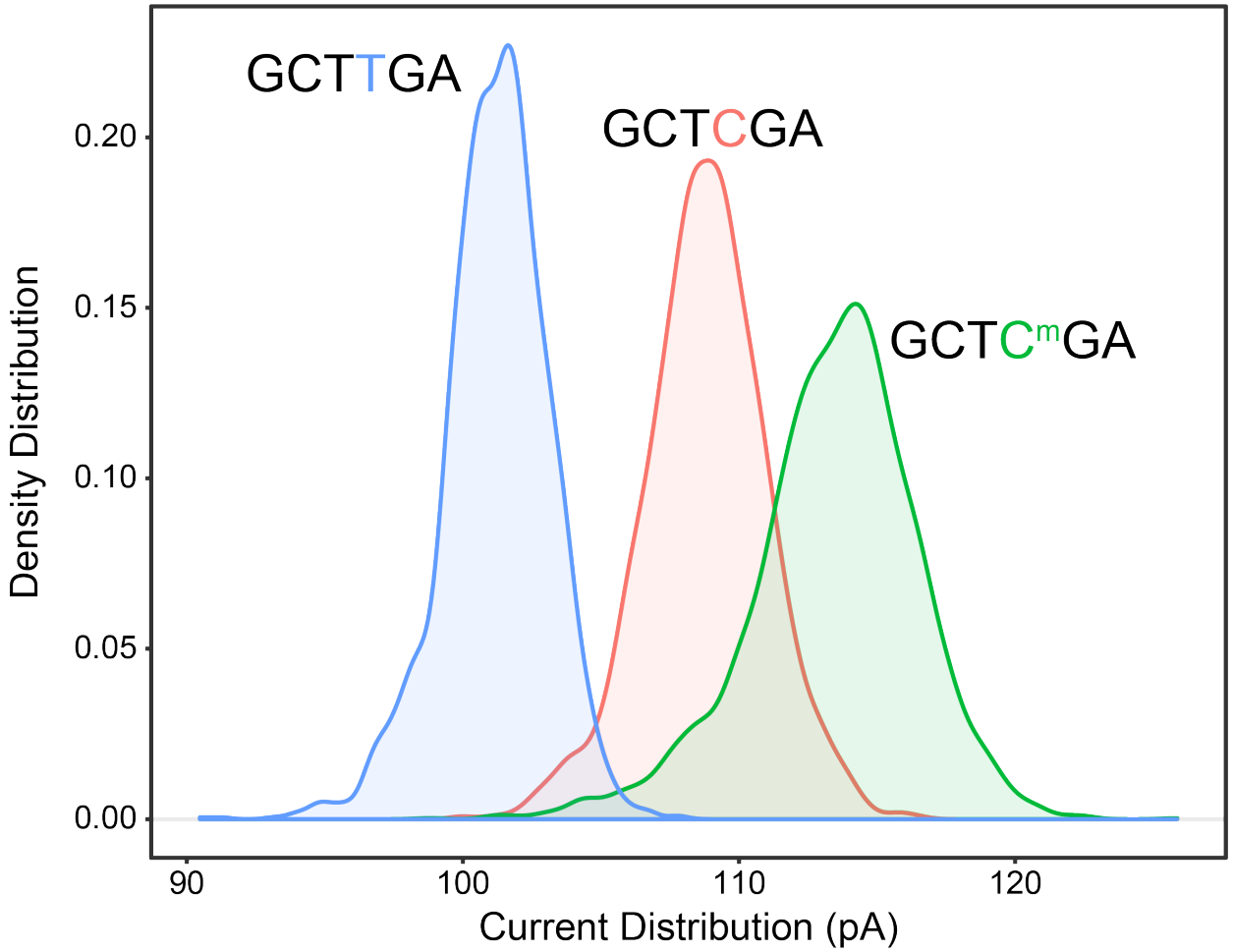


Cytosine



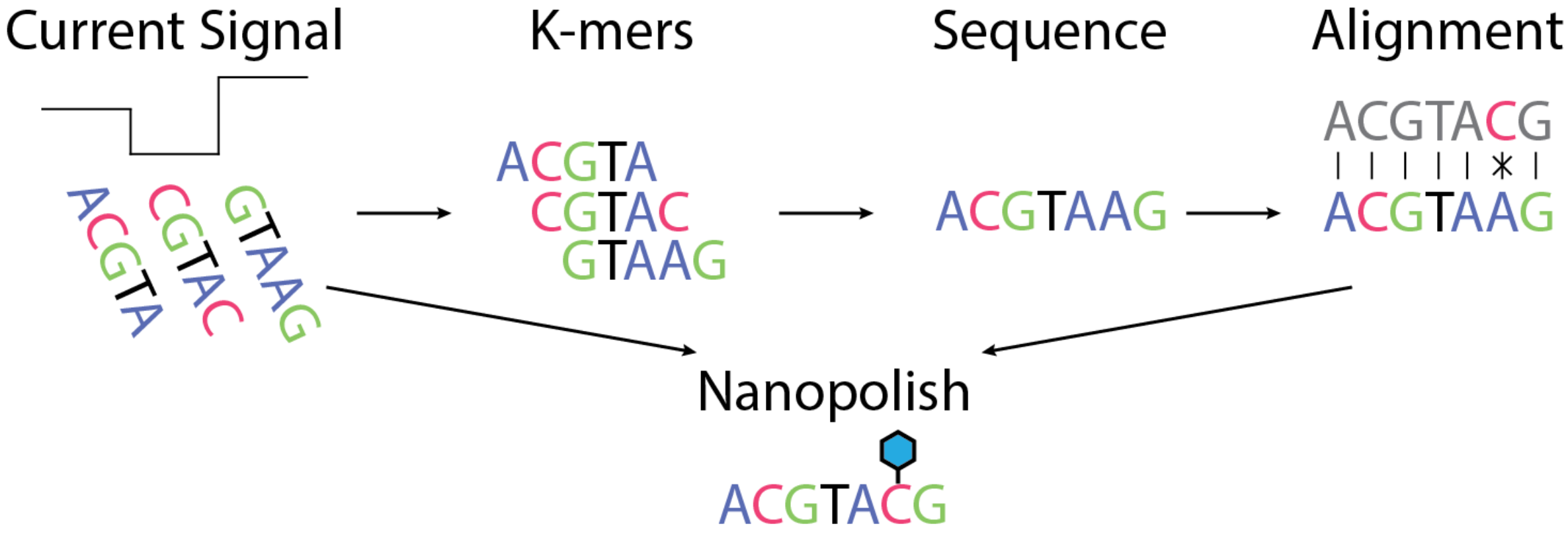
5-methylcytosine

- 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



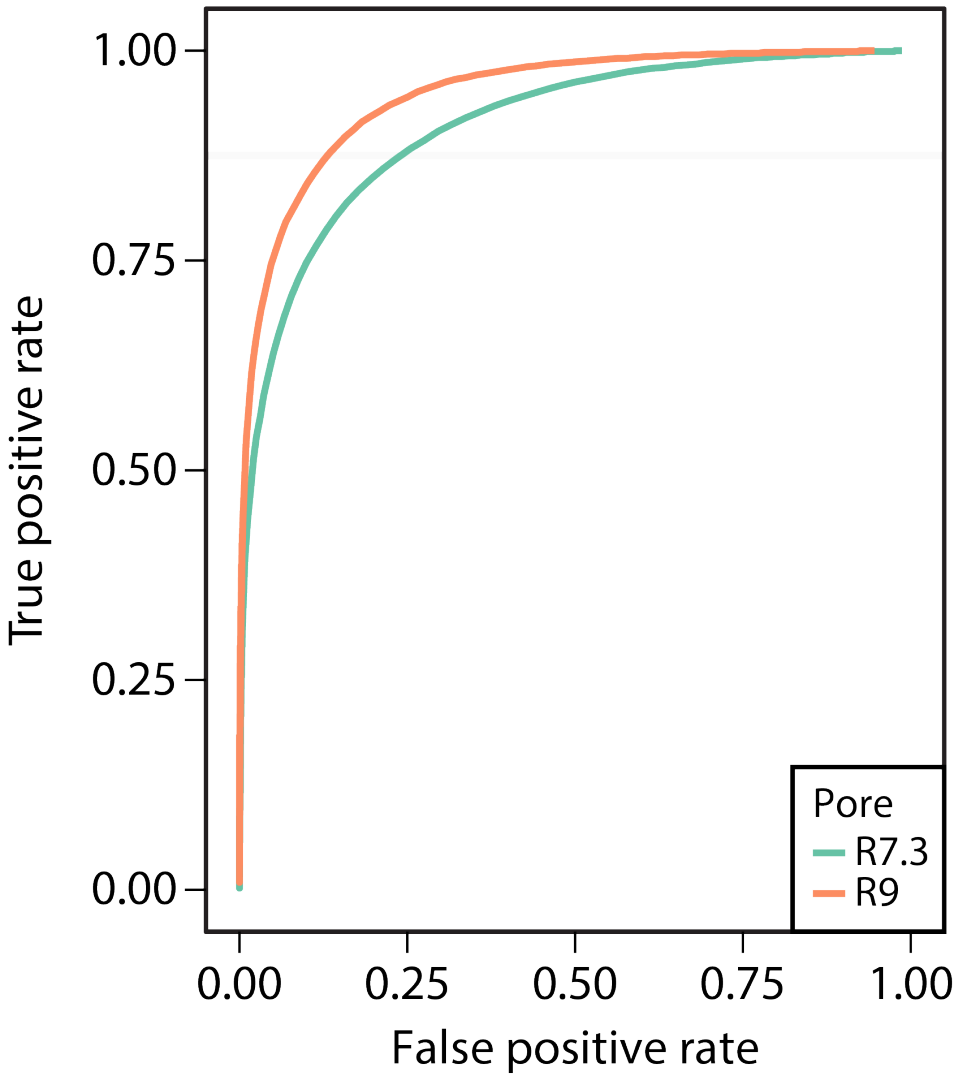
Simpson et. al *Nature* Methods 2017

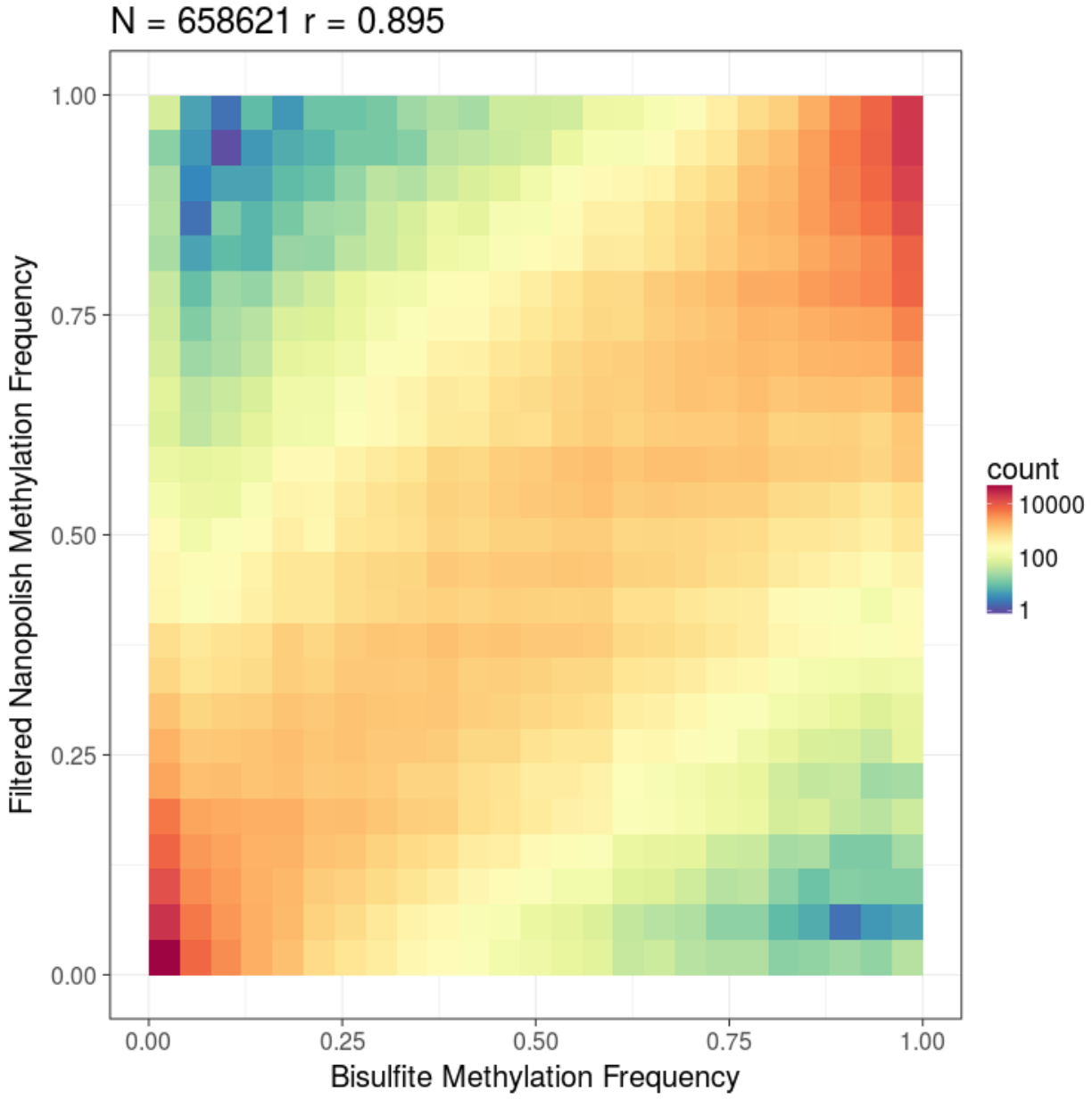
Nanopore: nanopolish methyltrain



- 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





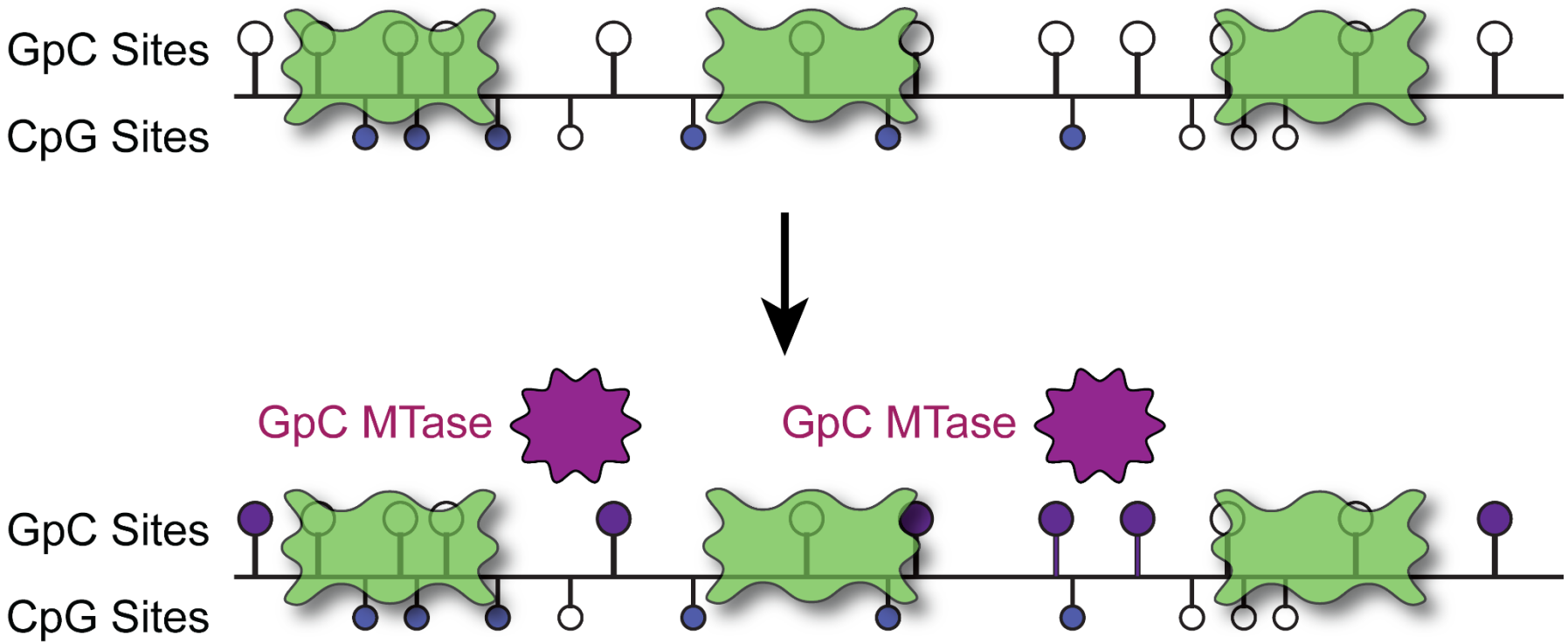
R9 calculates methylation 94% accurate at 77% of sites
NA12878 data shows .895 correlation with bisulfite

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

NanoNOMe: Chromatin Accessibility with Nanopore

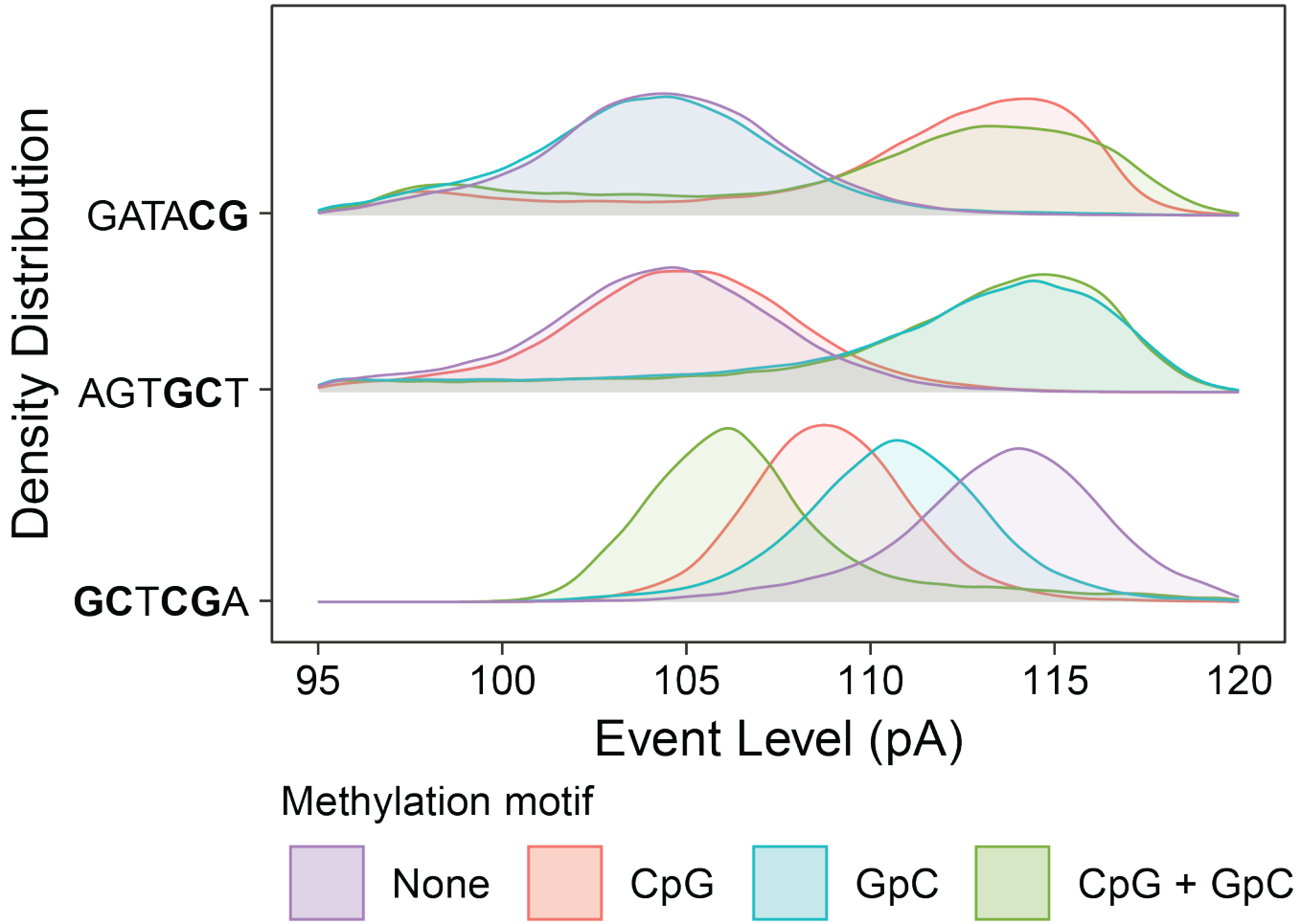
- NOMe-seq : Nucleosome Ocupancy and Methylome sequencing (Kelly et. al. *Genome Res.* 2012)

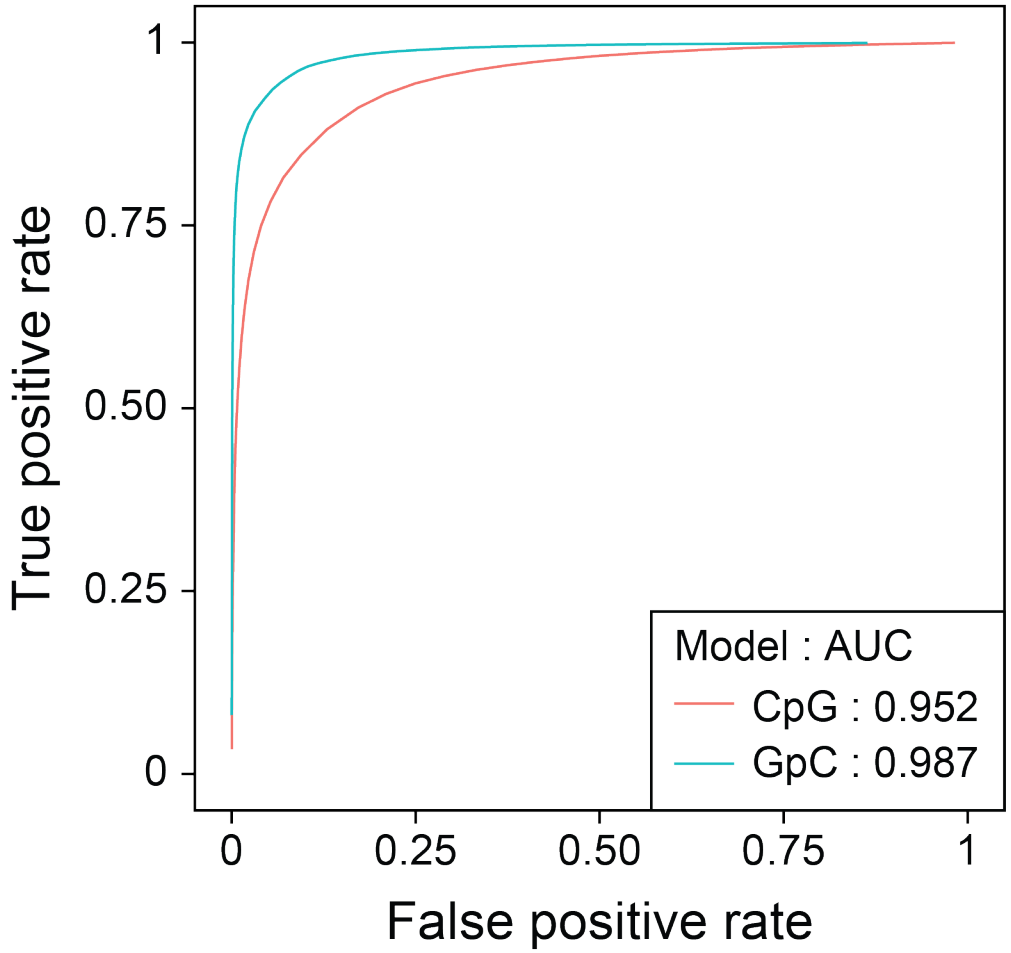
Simultaneously measures DNA methylation (CpG) and nucleosome occupancy (GpC)



Lee et al. bioRxiv 2018

Nanonome Signal

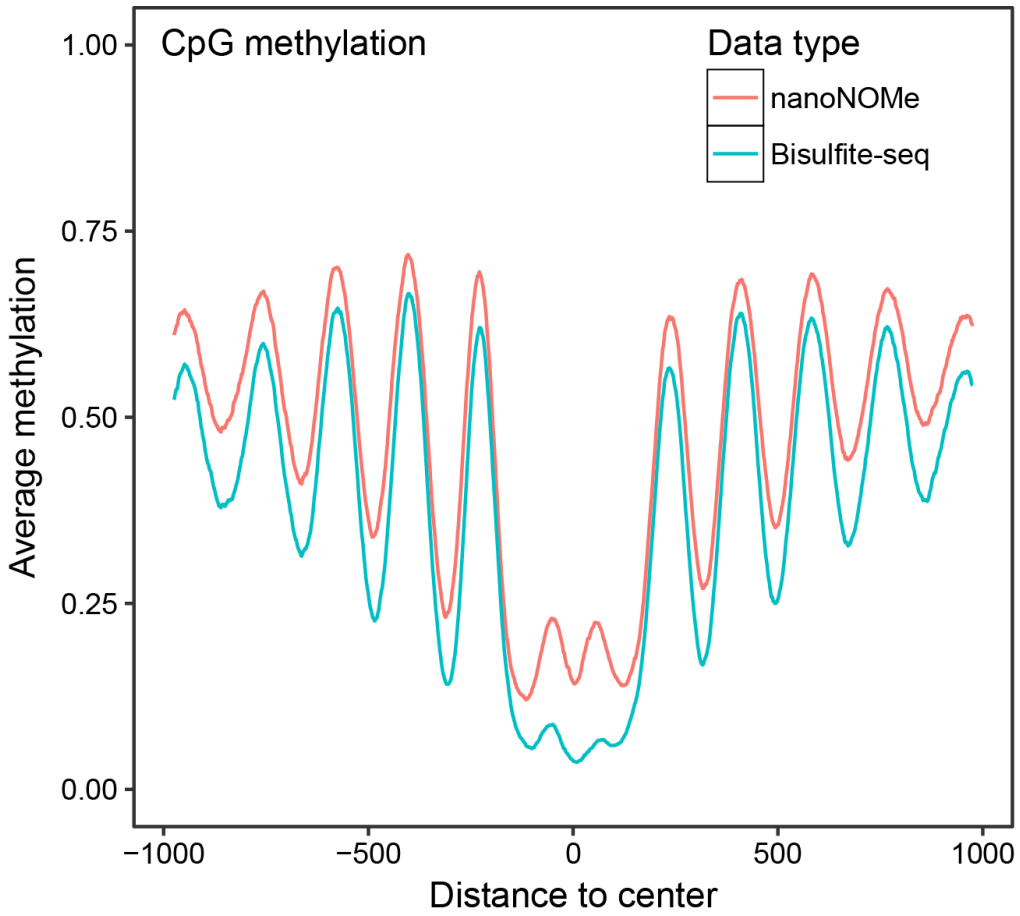


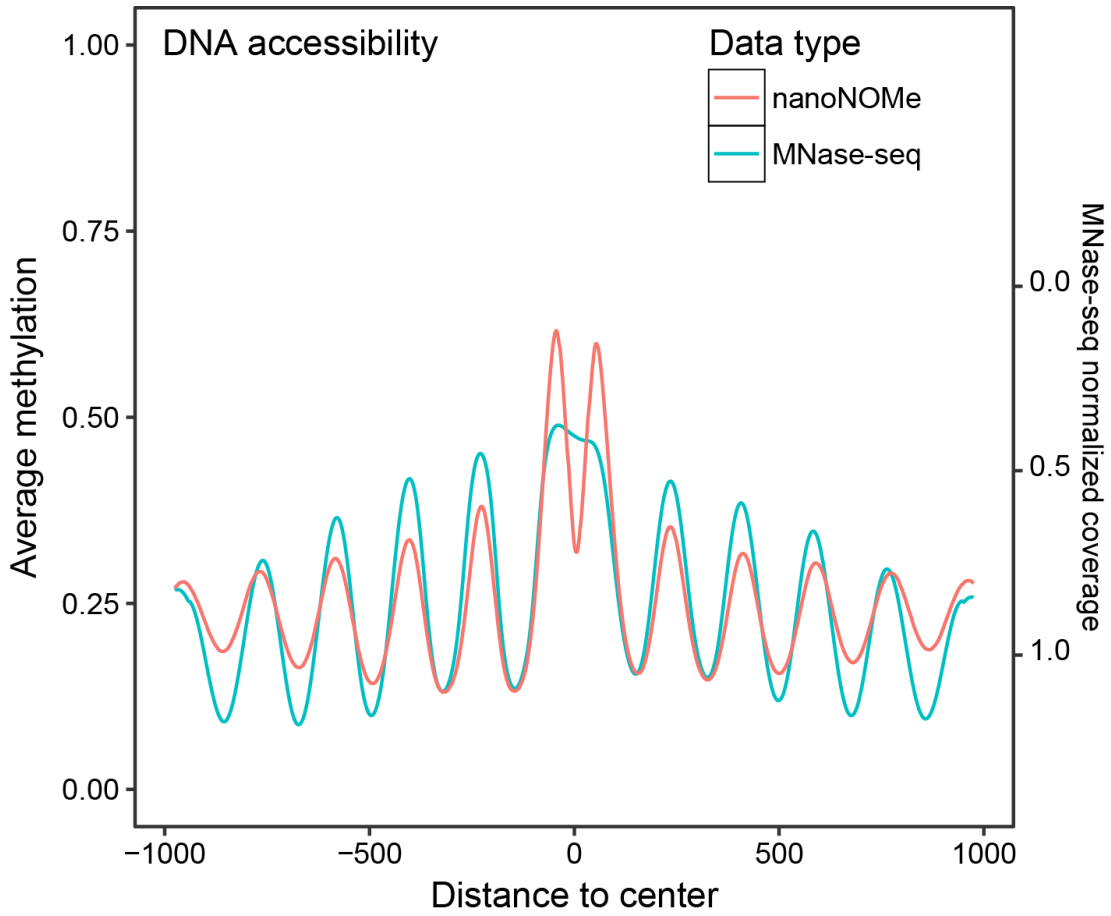


Lee et al. bioRxiv 2018

- 95% of CpGs as methylated in the 72% of all possible CpG k-mers
- 97% of GpCs as methylated in 89% of all possible GpC k-mers

NanoNOMe – DNAse Hypersensitive



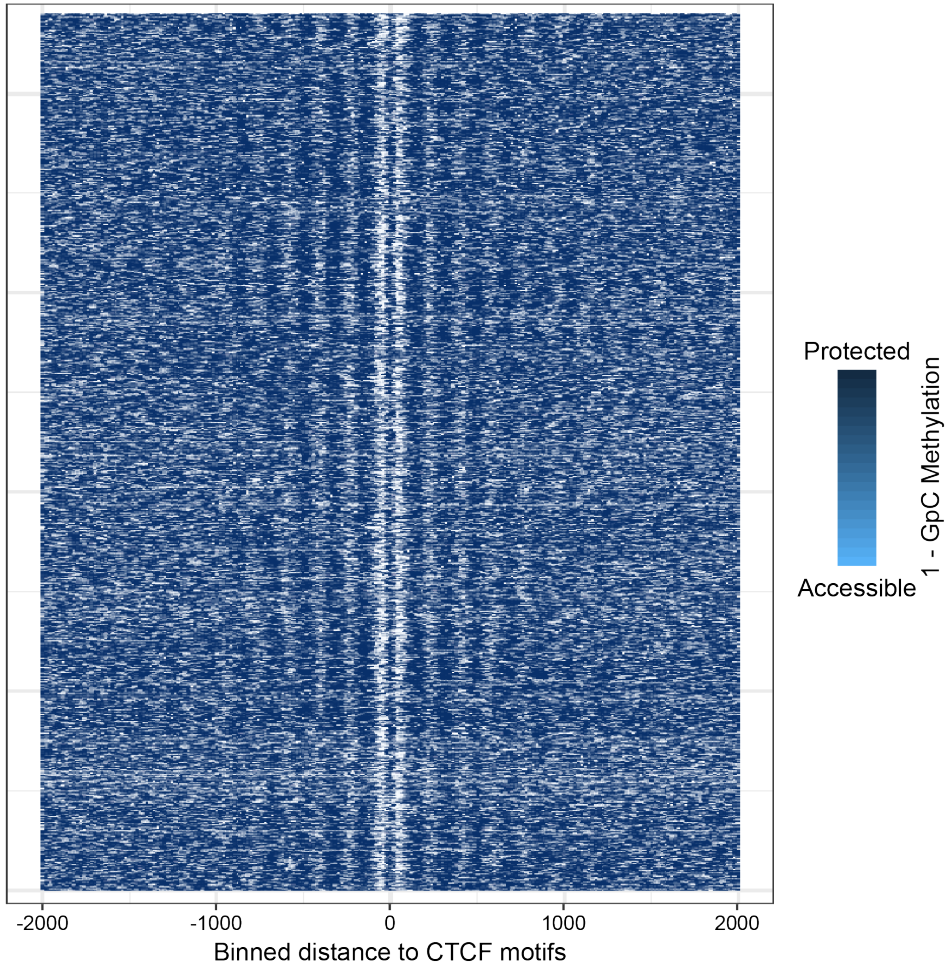


nanoNOMe signal near DNAse-seq peaks validates the methodology

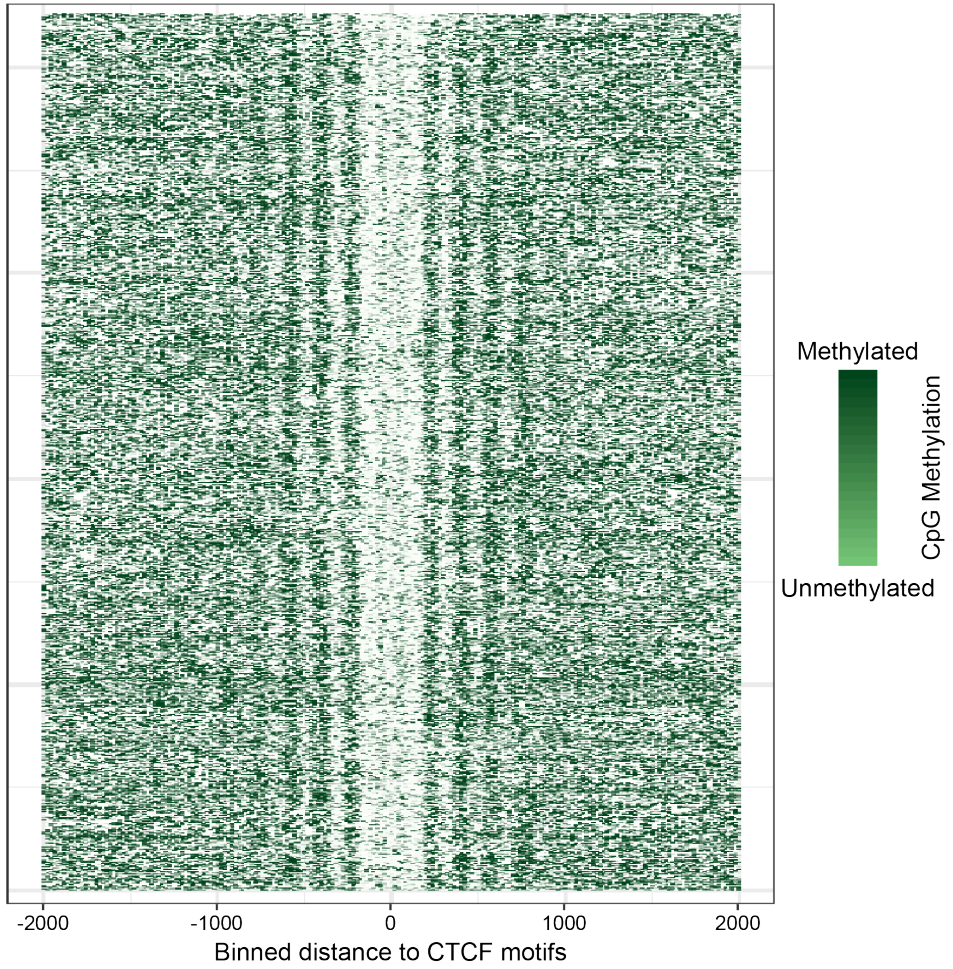
Lee et al. bioRxiv 2018

NanoNOMe: Aggregate CTCF binding sites

Chromatin Protection (1-GpC)



Endogenous Methylation (CpG)

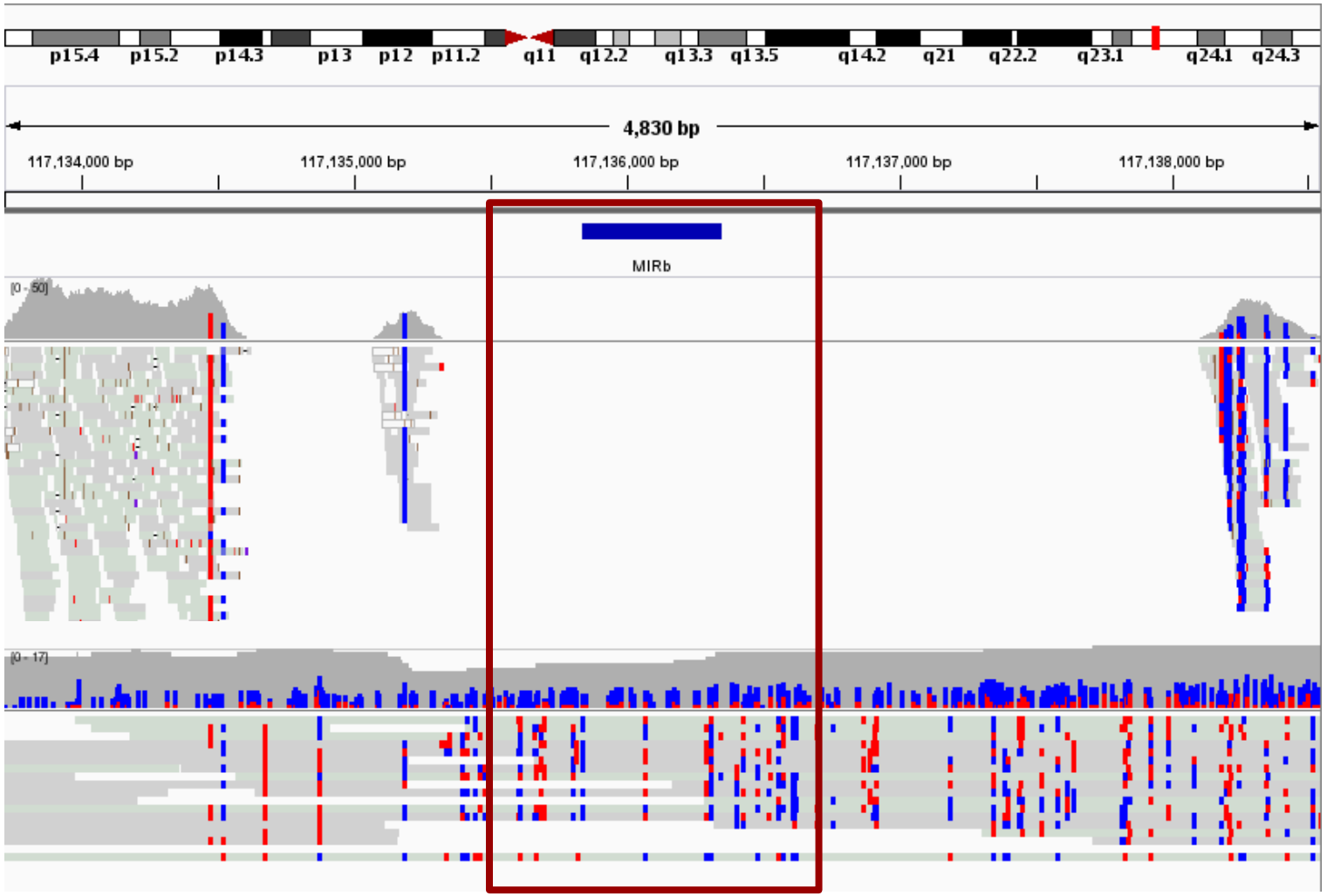


Lee et al. bioRxiv 2018

Methylation in Repetitive Regions

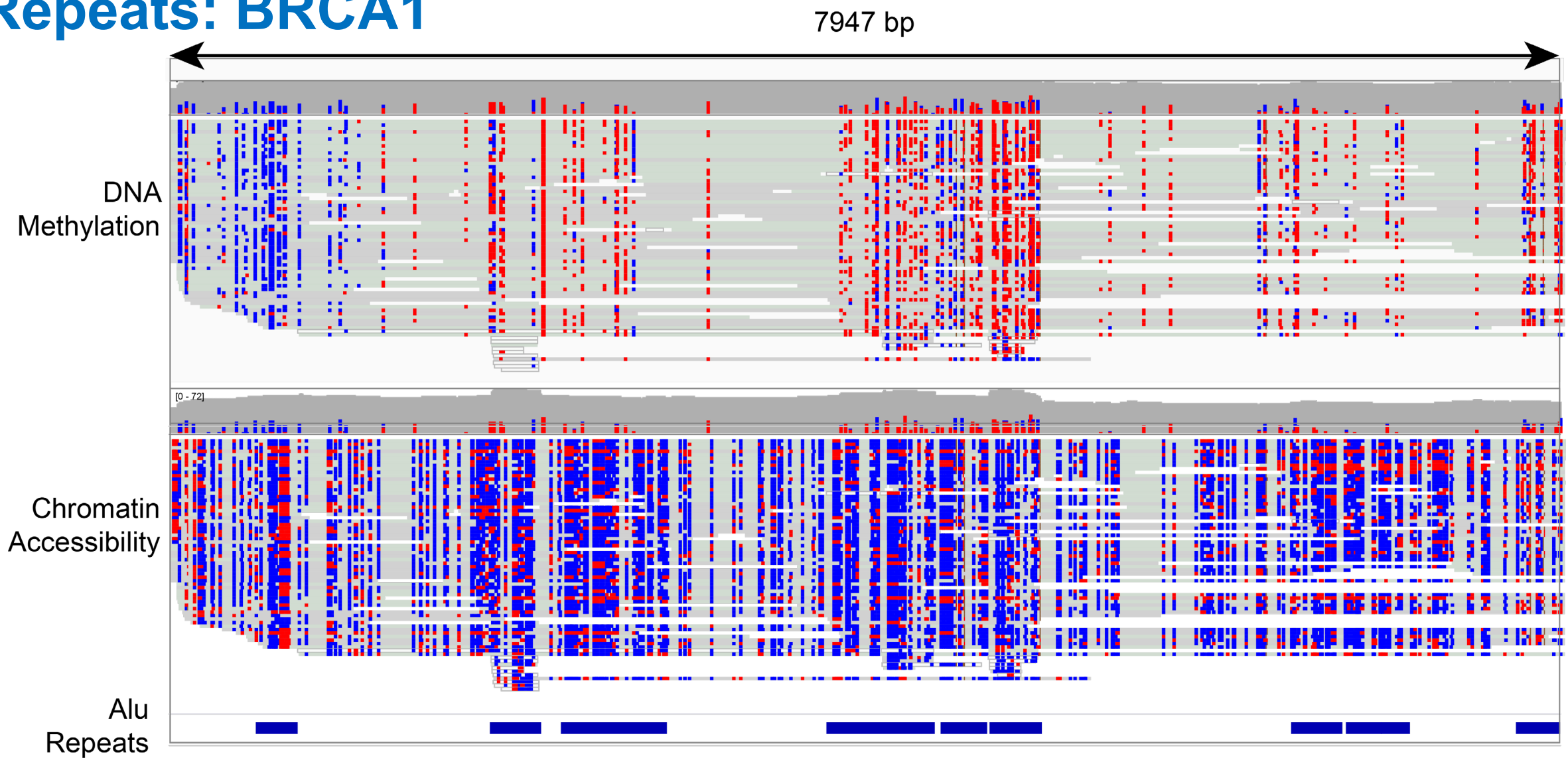
Bisulfite Sequencing (Illumina)

nanopolish methcall (Nanopore)



Regions unmappable by NGS are mappable with long reads

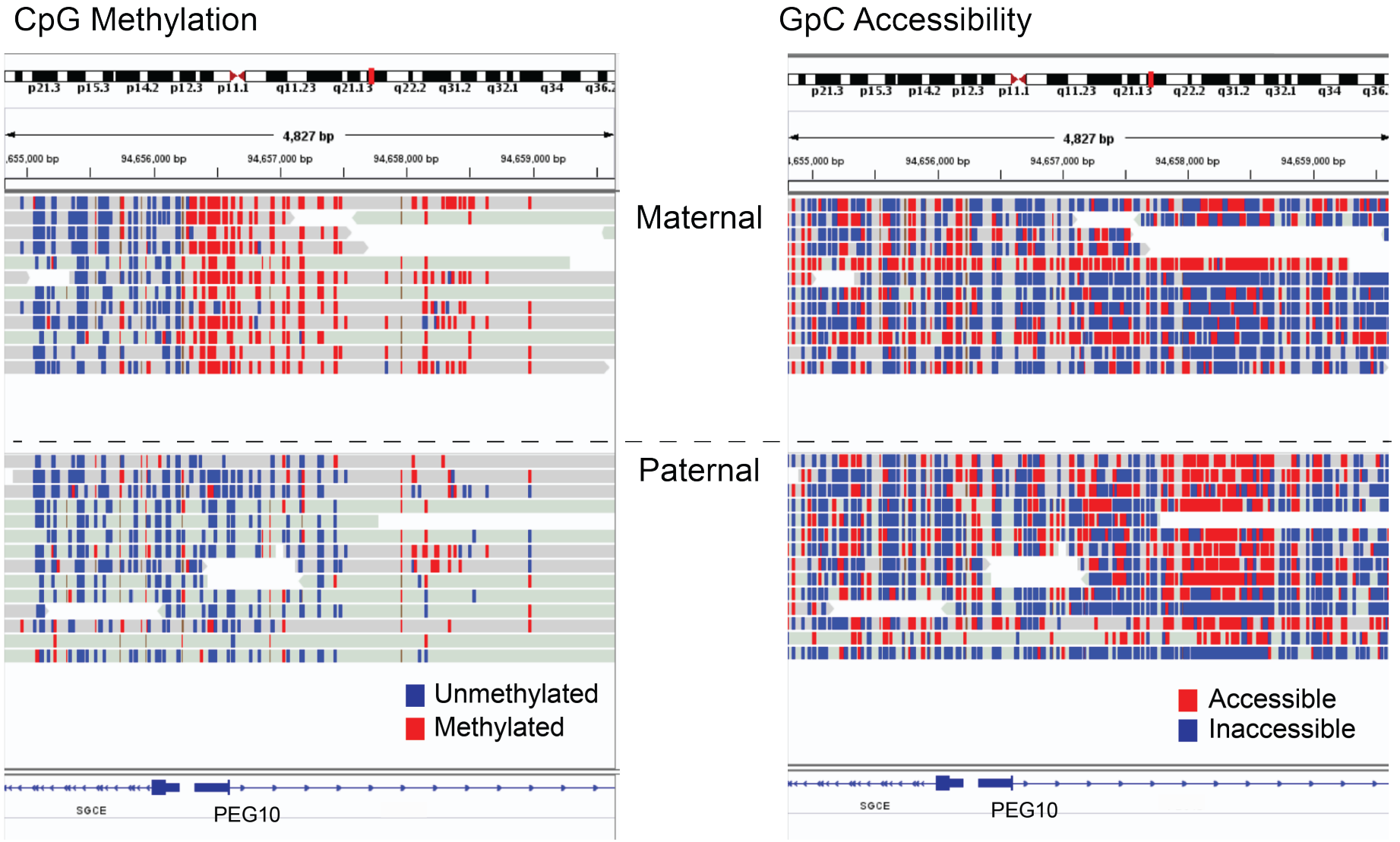
Repeats: BRCA1



Reference genome doesn't have many of these repeats properly – for BRCA1 region we aligned our reads against a custom GM12878 genome assembly (Jain et al)

Allele Specific Chromatin and Methylation

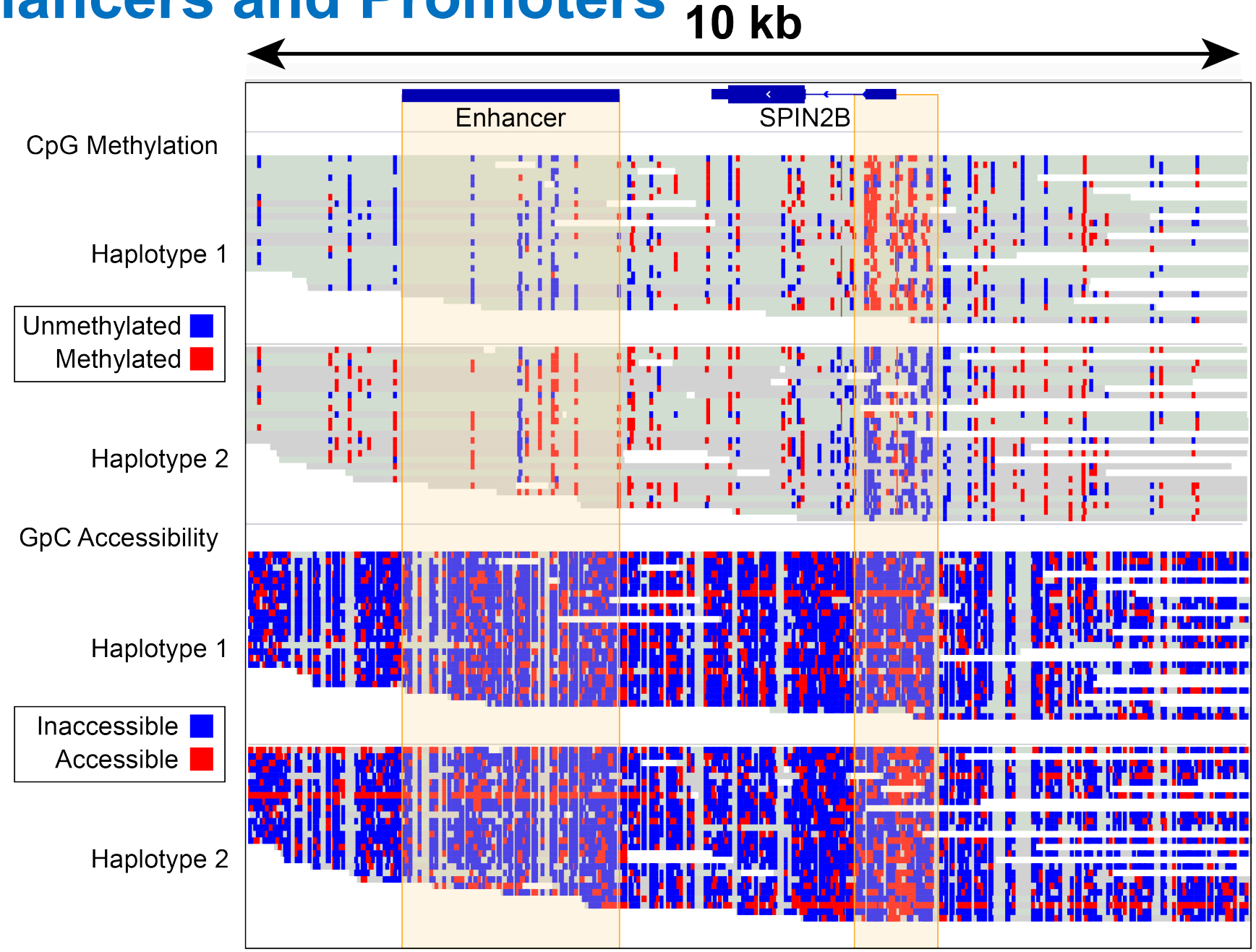
- Using long reads, we are likely to encounter a SNP
- This allows for phased methylation and chromatin data
- Near *PEG10* (imprinted gene):
  - Maternal copy is methylated and inaccessible
  - Paternal copy is unmethylated and accessible



Lee et al. bioRxiv 2018

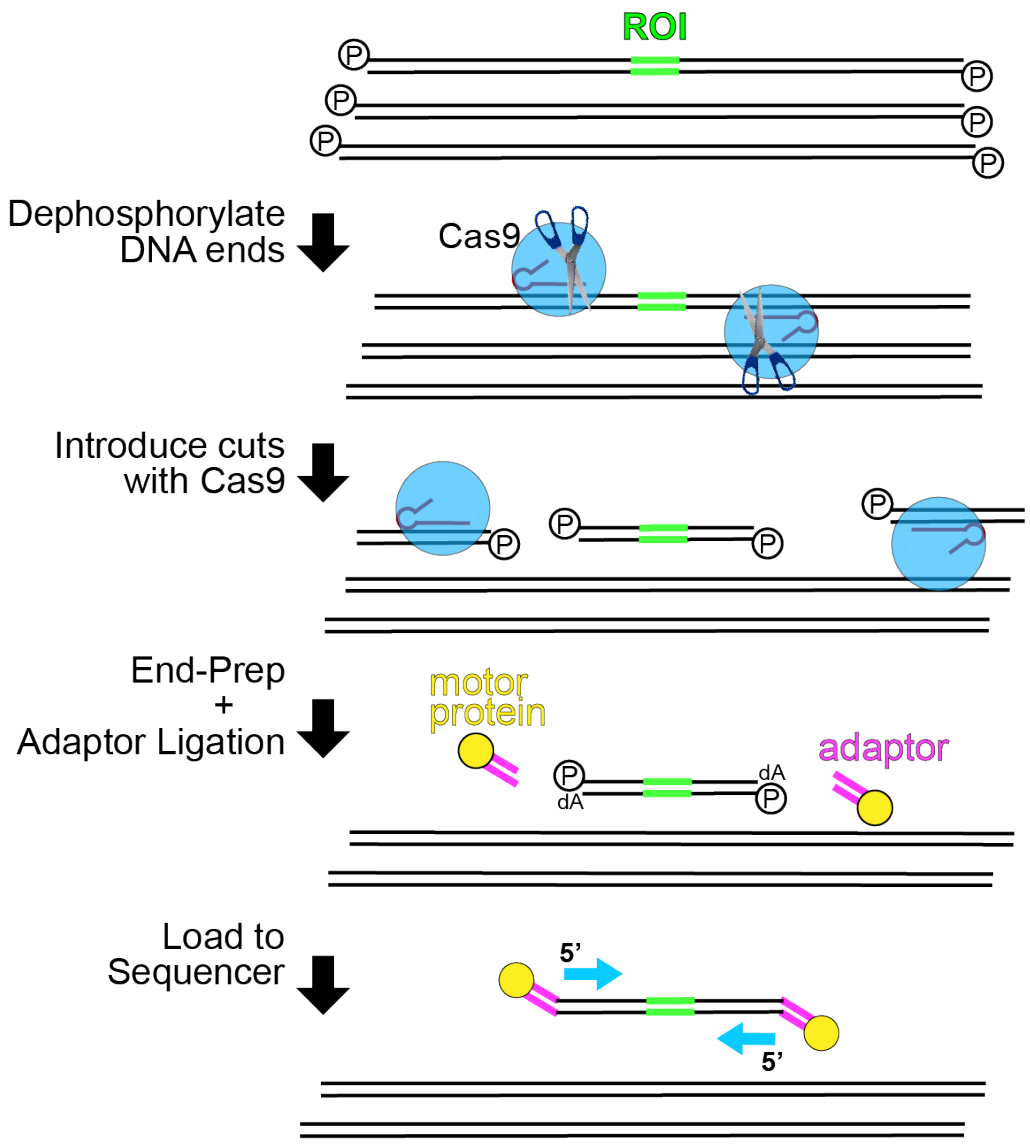
Coordinated Enhancers and Promoters

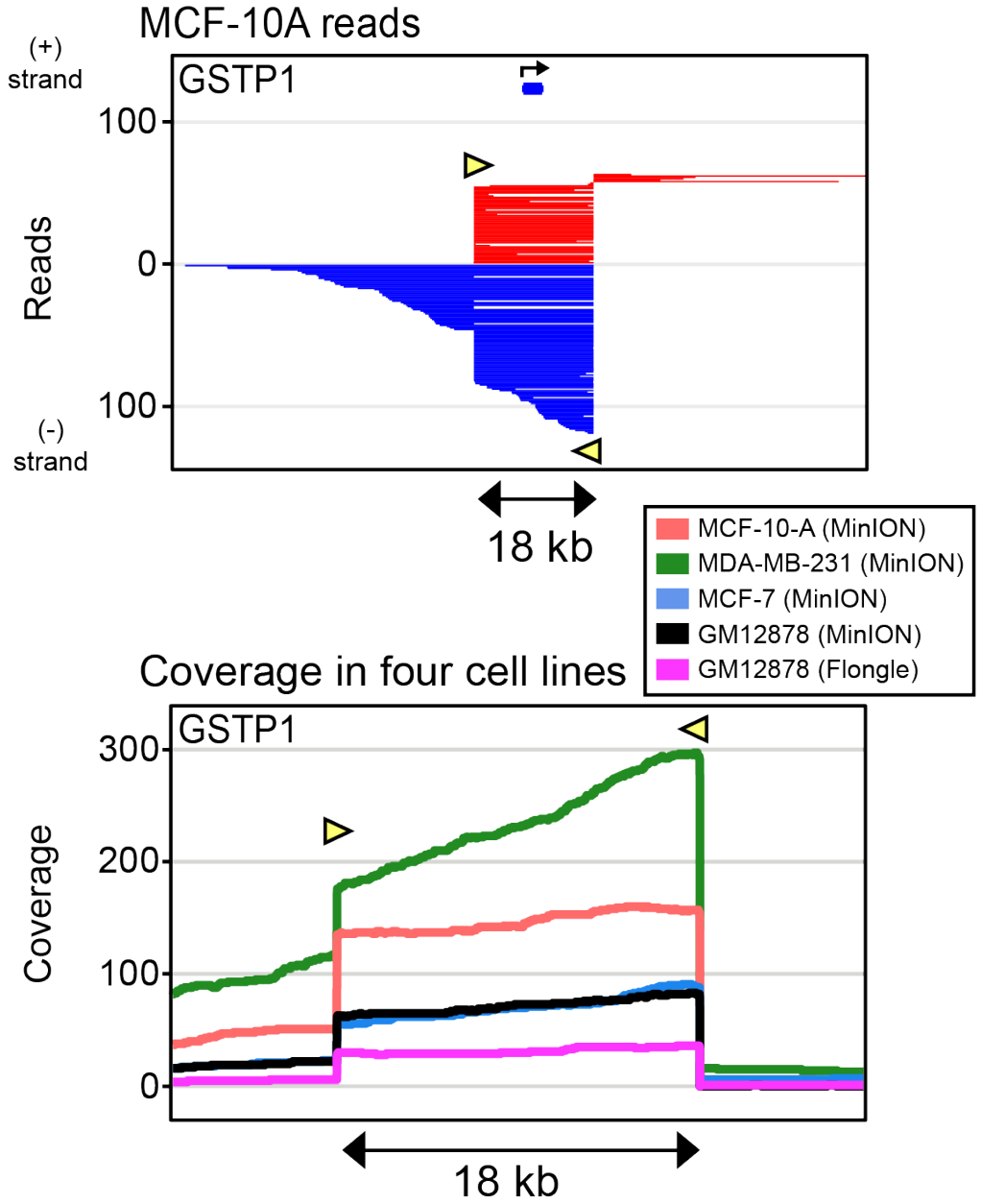
Using long reads, we can examine methylation and chromatin at some promoters and enhancers at the same time



Lee et al. bioRxiv 2018

Cas9 enrichment Method

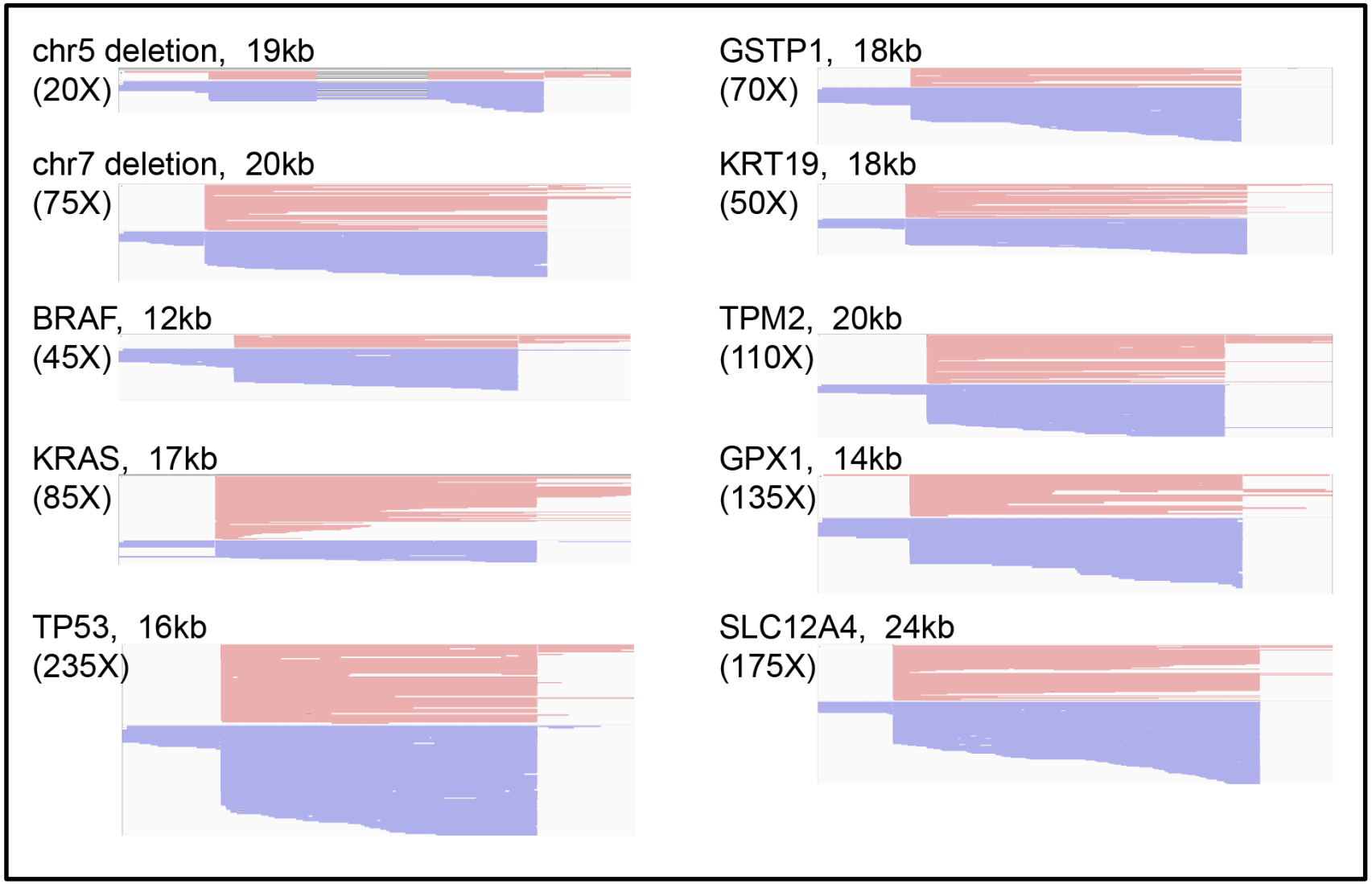




Gilpatrick et al bioRxiv (2019)

Using a panel of guideRNAs

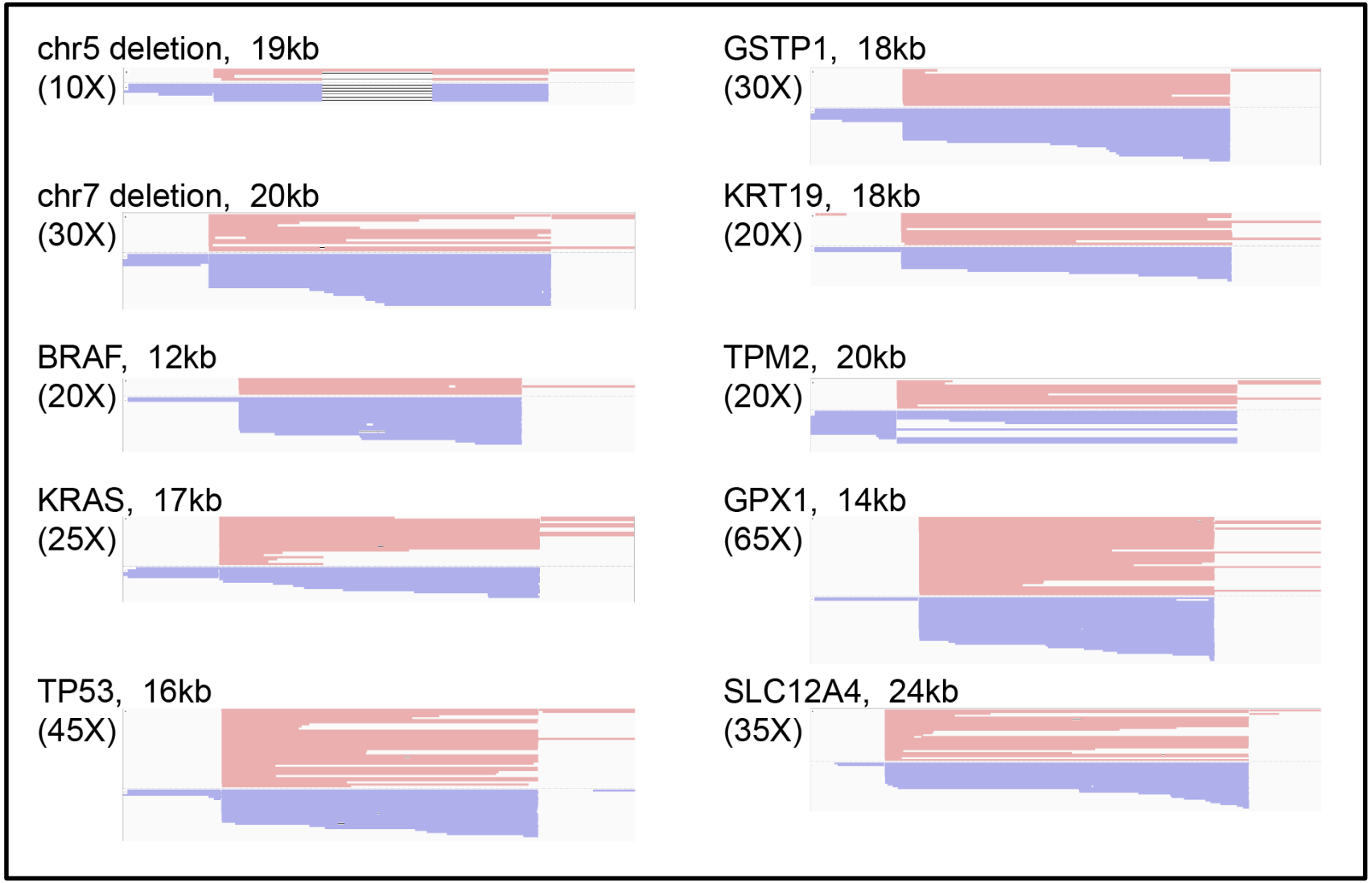
- Yield from
- 3ug GM12878 gDNA
- MinION Flow cell



Gilpatrick et al bioRxiv (2019)

Using a panel of guideRNAs

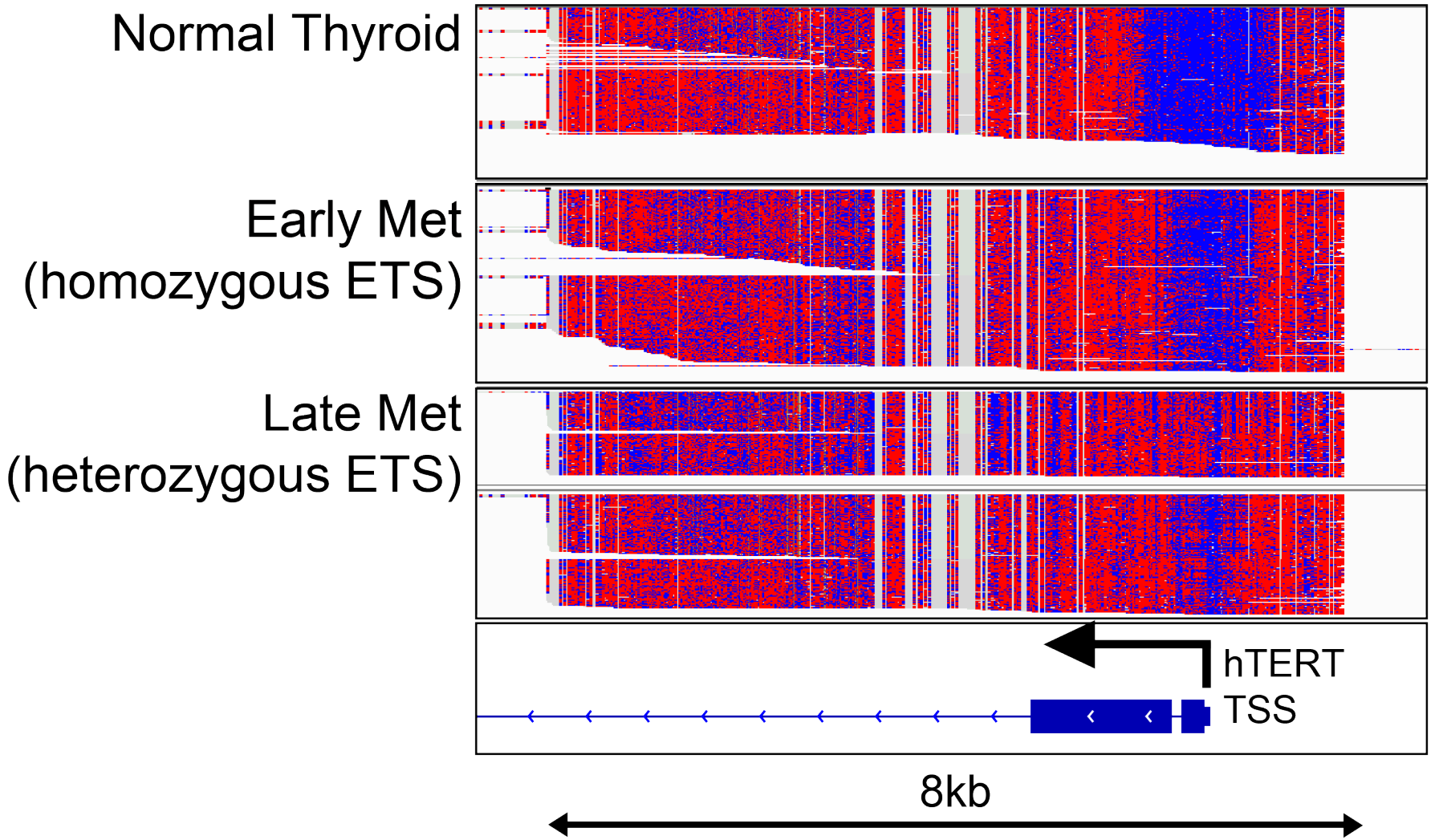
- Yield from
- 3ug GM12878 gDNA
- Flongle Flow cell



Gilpatrick et al bioRxiv (2019)

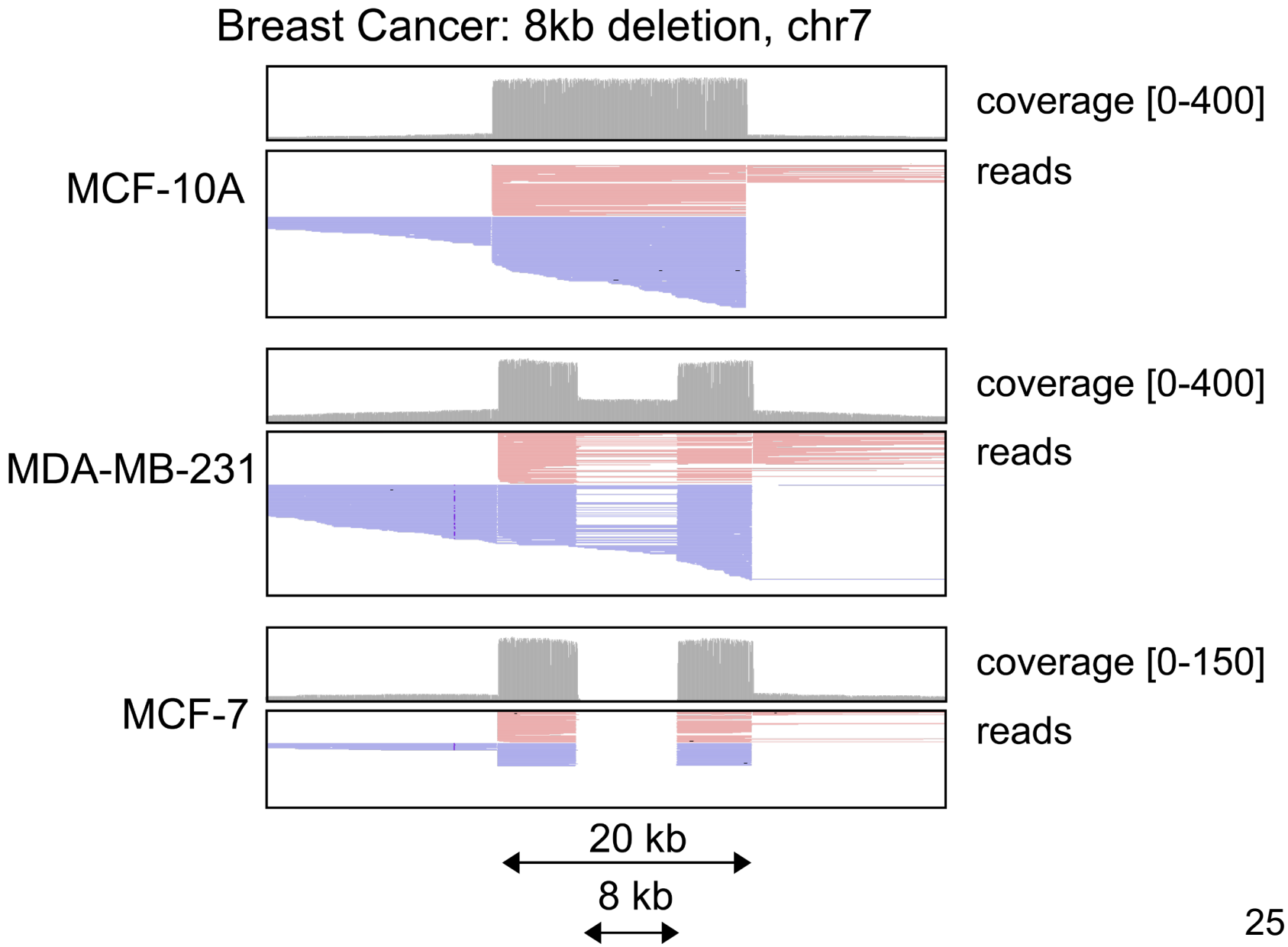
Enrichment of hTERT region

- We observe an “erosion” of the unmethylated (blue) CpG island in the promoter of hTERT in progressive cancer samples
- In the late metastasis a mutation in the ETS binding site of the promoter occurs in one of the alleles
- The mutant allele appears to have a more unmethylated island than the WT allele



Structural Variants in Cancer

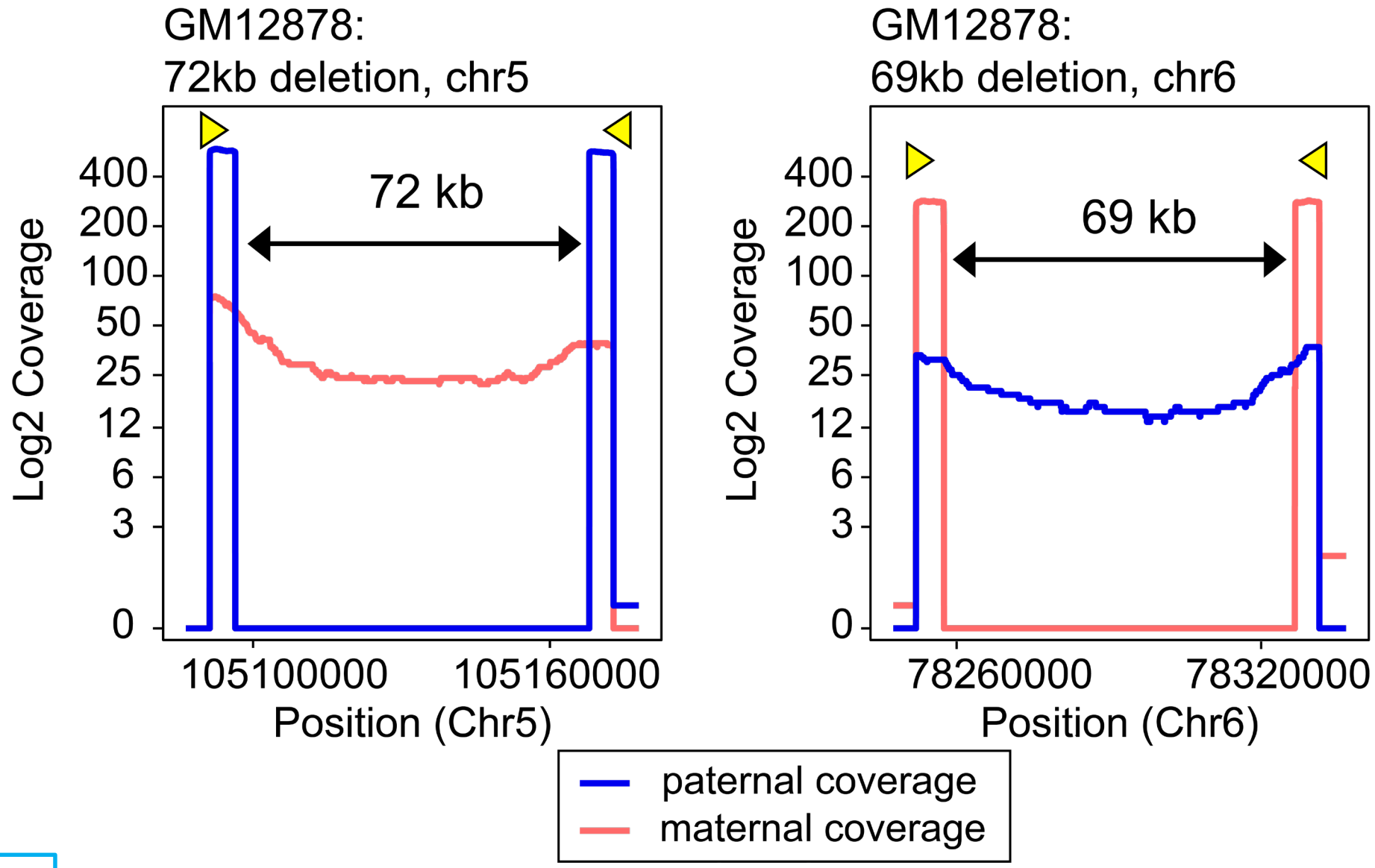
- Structural variants (SV), large insertions, deletions or translocations in the genome, are hard to detect with short-read sequencing
- Nanopore sequencing can map them well, and with targeted sequencing we can observe these



Gilpatrick et al bioRxiv (2019)

Structural Variation Detection

- Deleted allele has >200X coverage, but intact allele has 1/10 as much coverage
- Bias likely due to length of reads input into enrichment



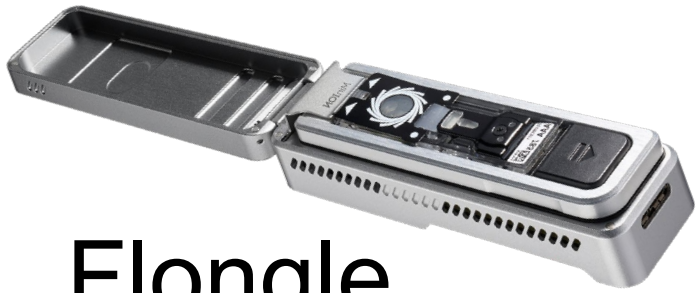
Gilpatrick et al bioRxiv (2019)

Single Nucleotide Variants

176 known SNVs exist in in span of 140kb in GM12878

MinION

Avg Cov : 100X



Flongle

Avg Cov : 30X

MinION		TP	Sensitivity	FP	PPV
default variant calls	SAMTOOLS	170	0.97	12	0.93
	NANOPOLISH	169	0.96	17	0.91
dual-strand filter	SAMTOOLS	142	0.81	3	0.98
	NANOPOLISH	156	0.89	1	0.99

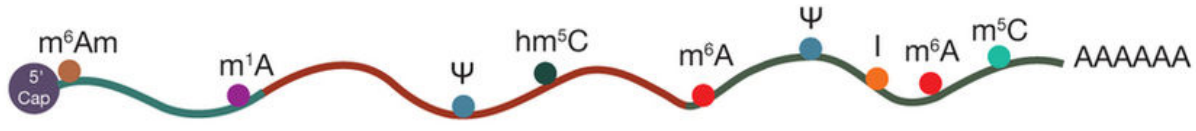
Flongle		TP	Sensitivity	FP	PPV
default variant calls	SAMTOOLS	138	0.78	10	0.93
	NANOPOLISH	160	0.91	20	0.89
dual-strand filter	SAMTOOLS	61	0.35	0	1.00
	NANOPOLISH	100	0.57	1	0.99

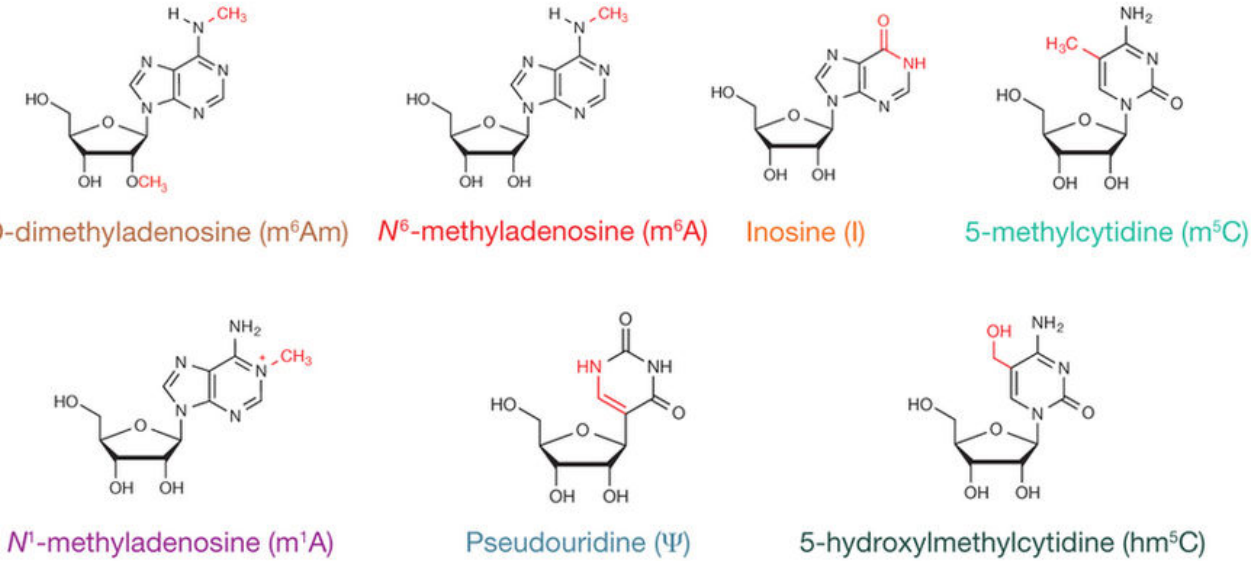
Gilpatrick et al bioRxiv (2019)

Questions in RNA:

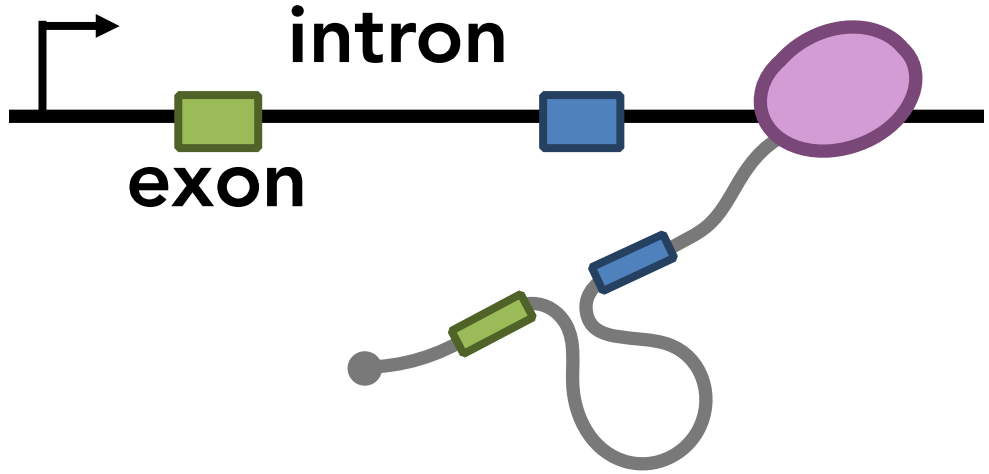
Pol II

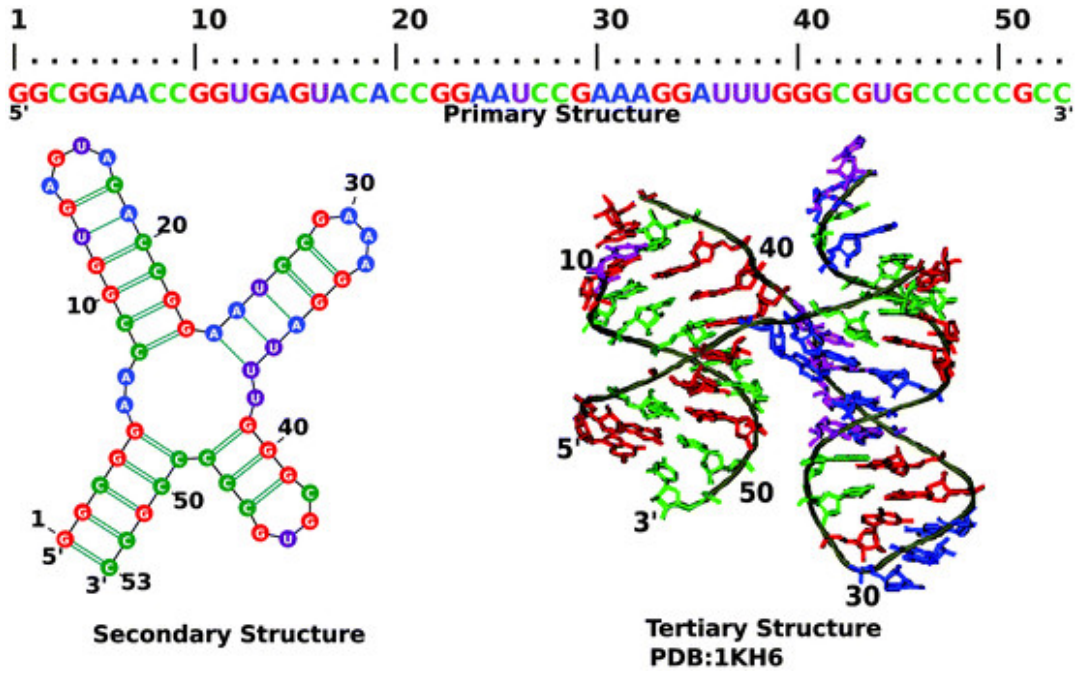
Li, Xiong, Yi, Nature Methods (2017)



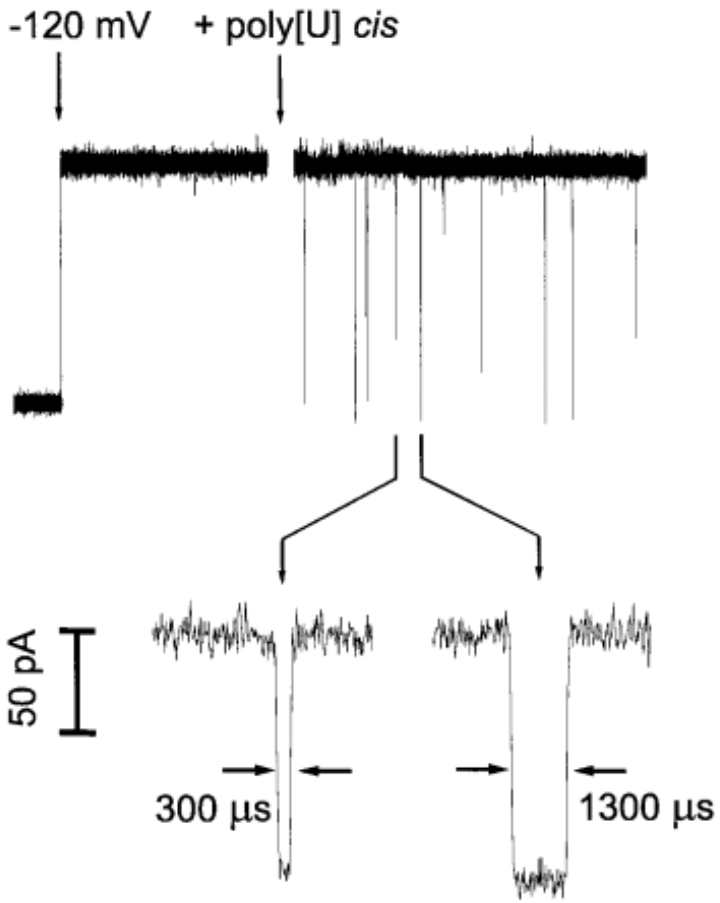


- RNA modifications
- RNA dynamics
- RNA structure

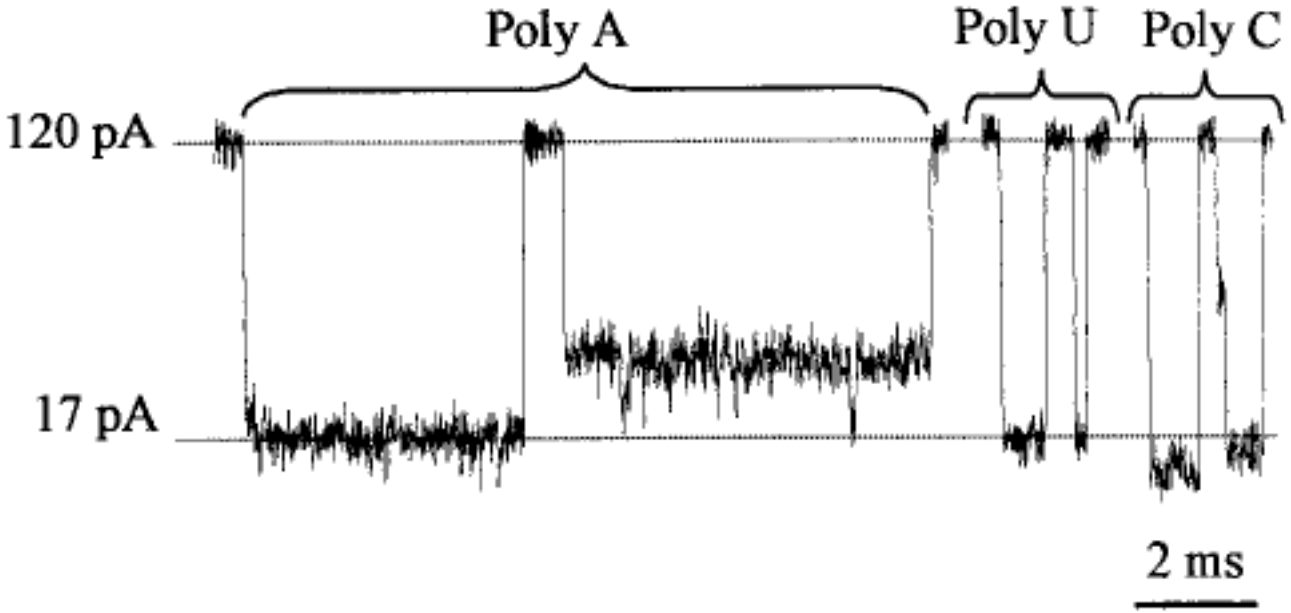




Earliest nanopore experiments analyzed RNA



Kasianowicz, Brandin, Branton, & Deamer
PNAS 1996



Akeson, Branton, Brandin & Deamer
Biophysical J. 1999

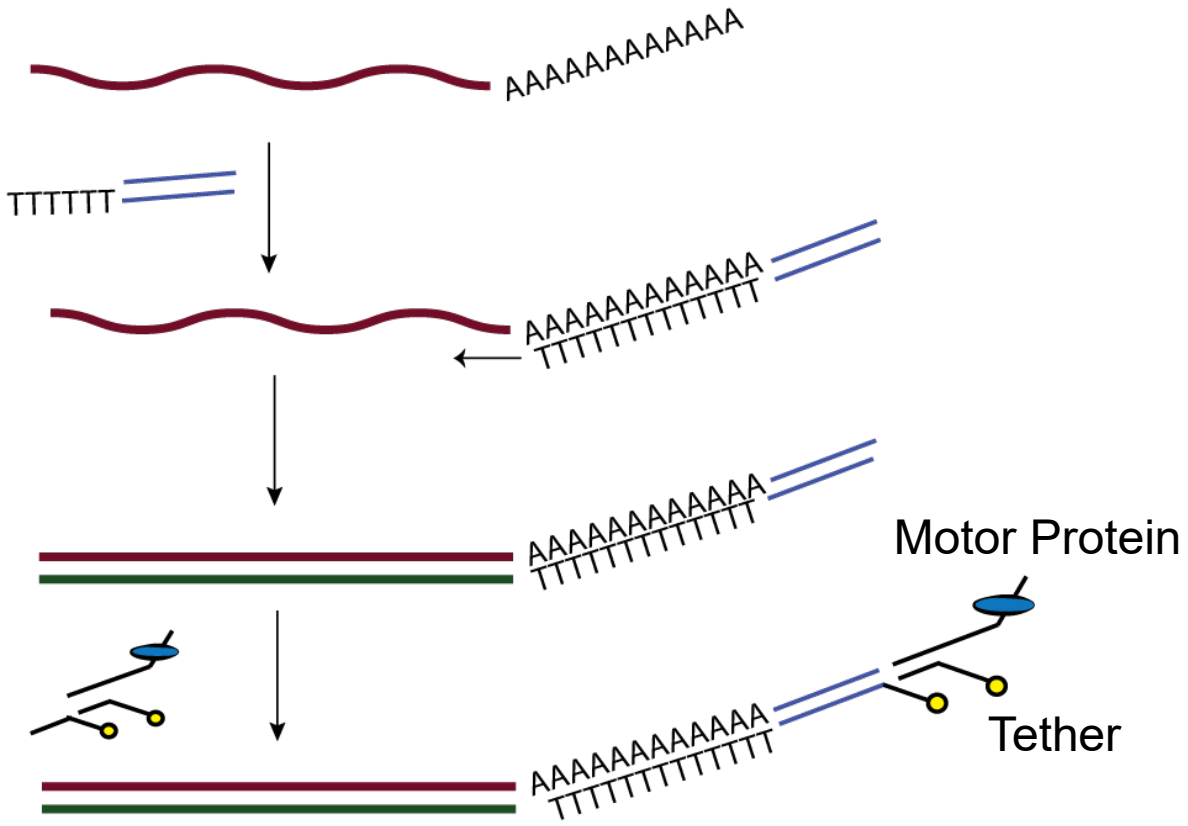
Direct RNA Sequencing

PolyA+ RNA captured

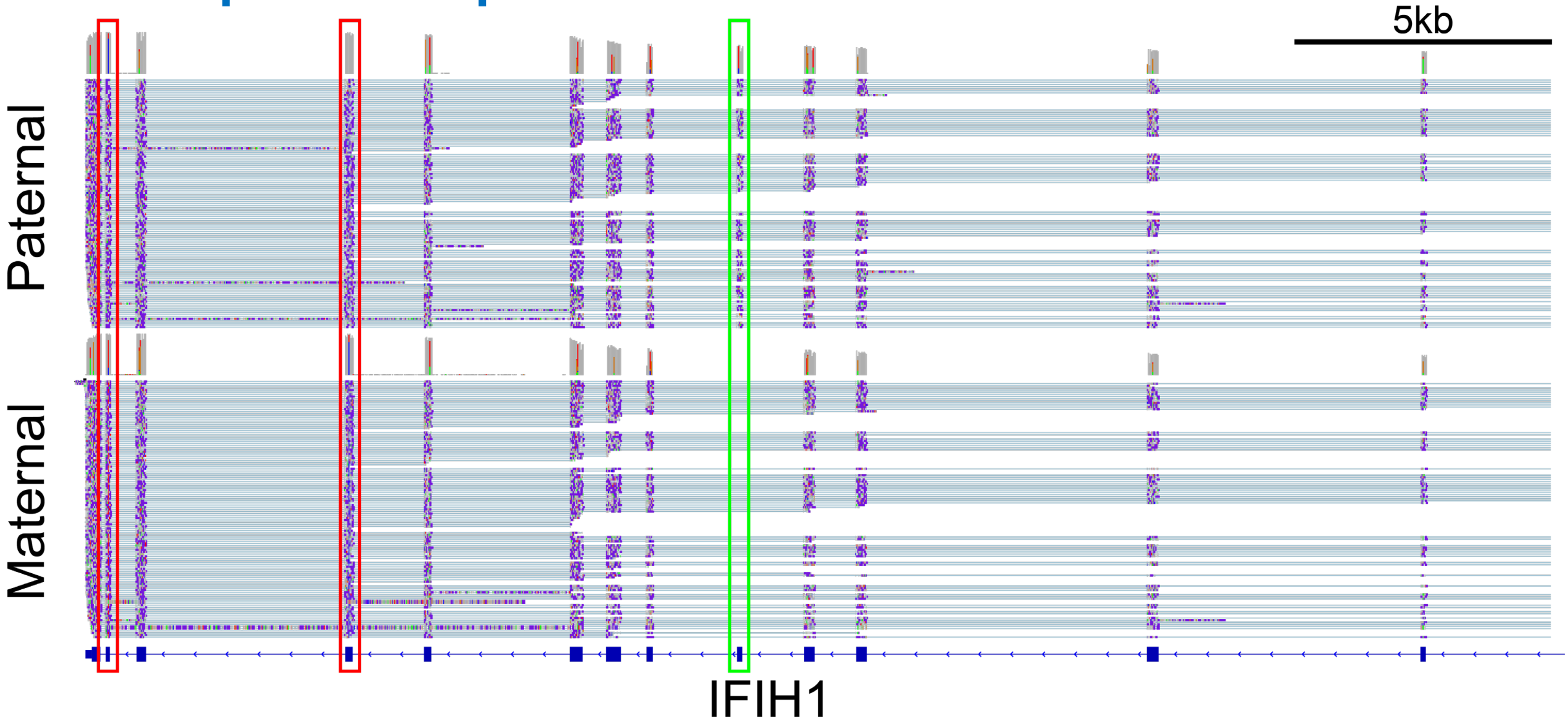
Splint poly-T adapter ligation

Reverse transcription (optional)

Sequencing adapter ligation

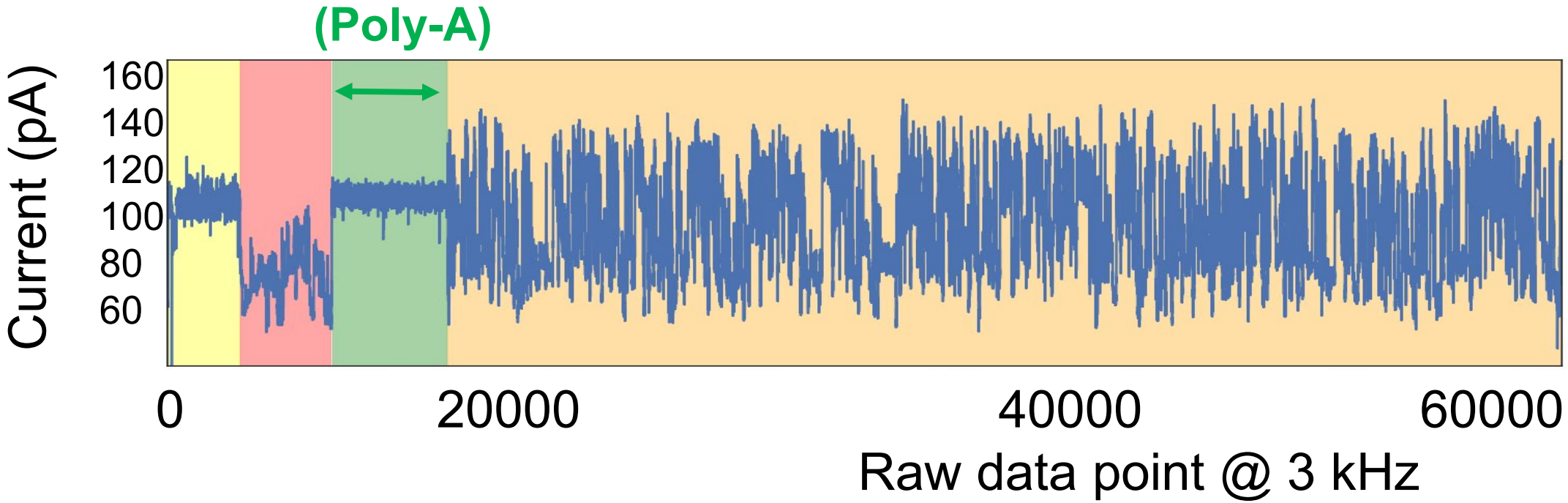


Allele Specific Expression



Long reads allow identification of allele specific expression, even when SNPs are far from the exon

Ionic current dwell time can be used to estimate poly-A tail lengths



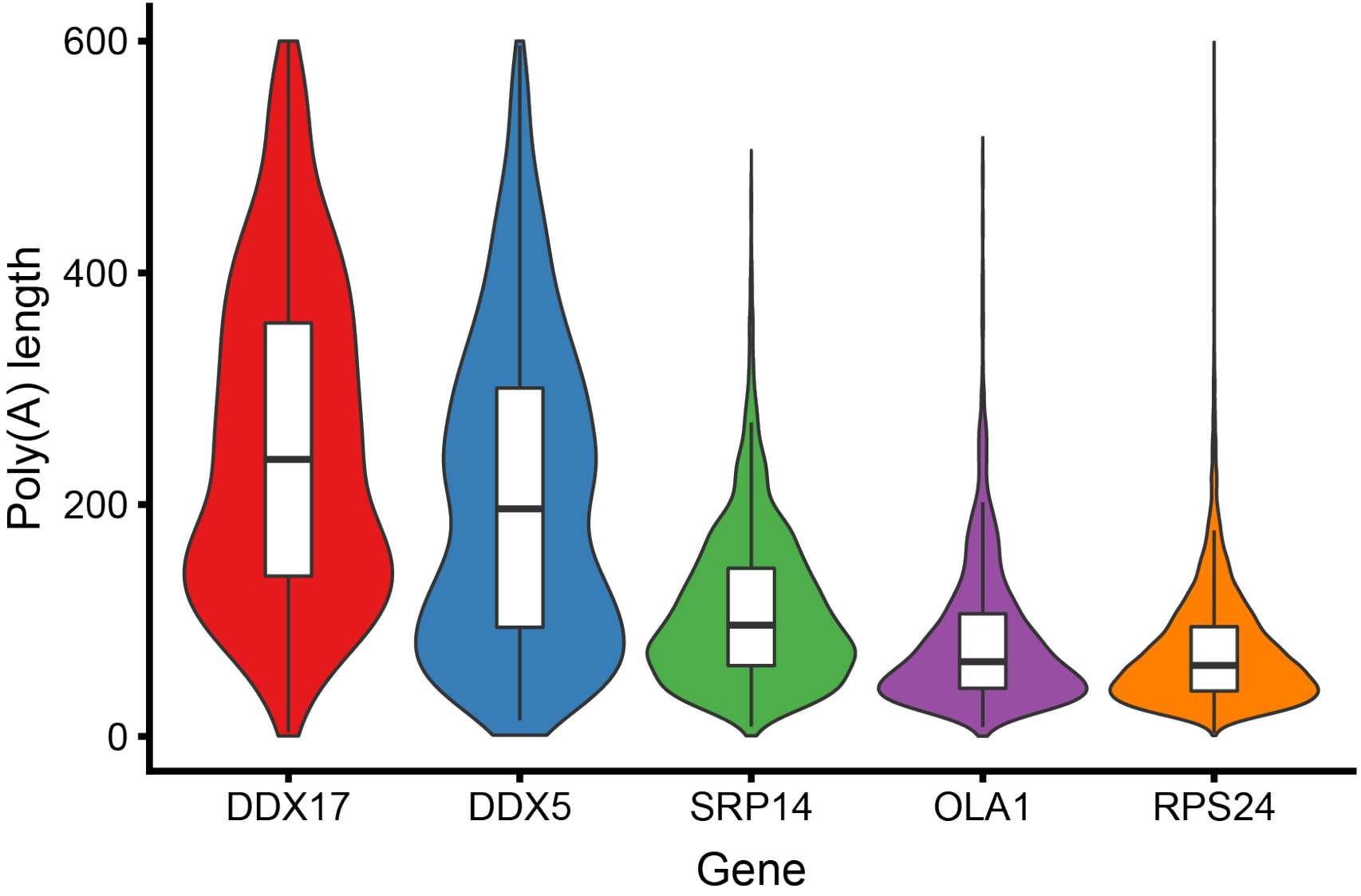
Predicting poly-A sequence length becomes tractable when consistent structural regions of dRNA reads can be identified and separated

Workman et al bioRxiv (2018)

PolyA estimator:
https://github.com/jts/nanopolish **polyA**

Poly (A) tails of genes

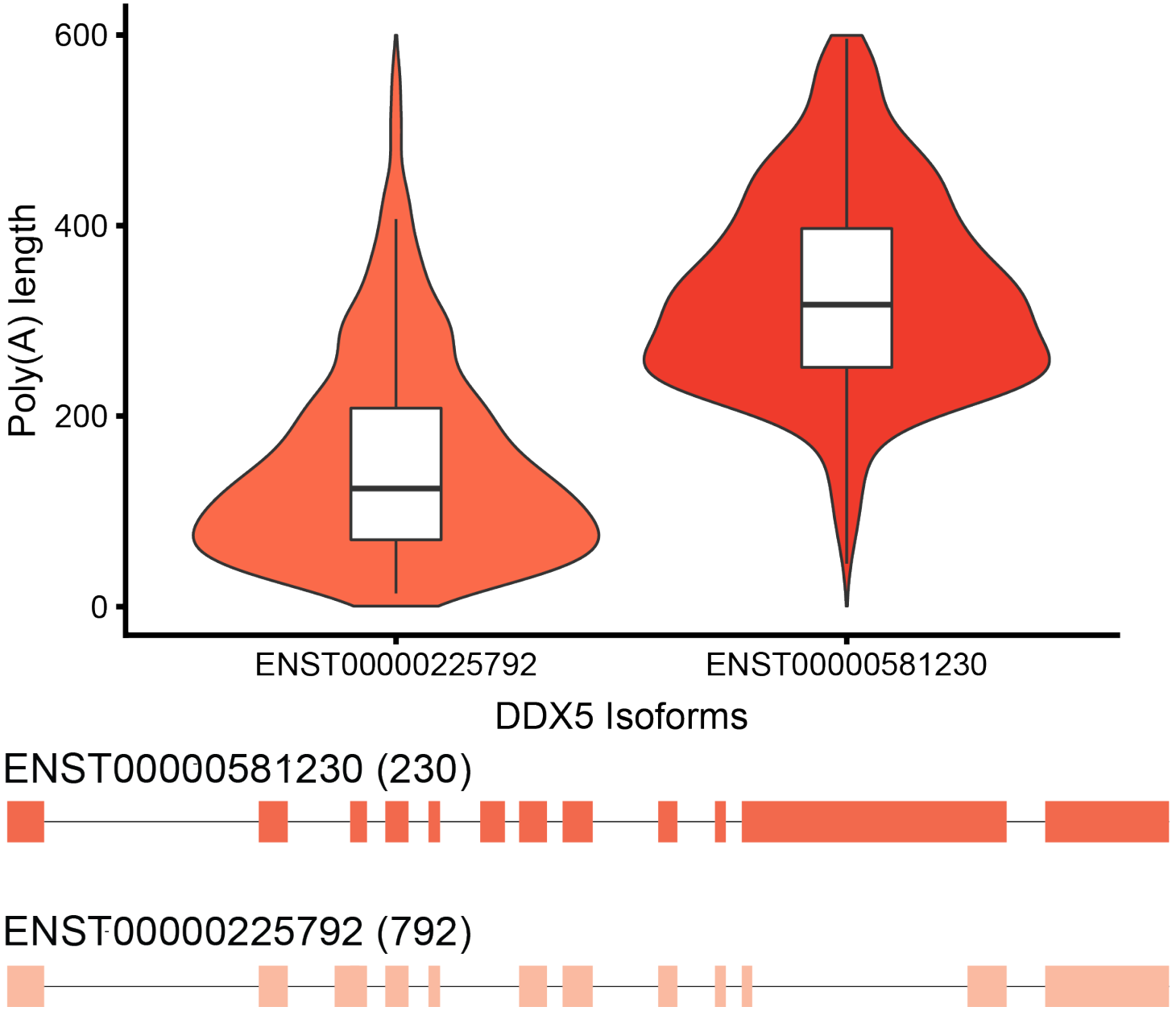
- Poly (A) tails of different genes with >500 reads
- Of those we plotted the longest 2, shortest 2 and median poly(A) length genes
- Some of the genes have interesting distributions, with surprisingly long poly(A) tails.



Workman et al bioRxiv (2018)

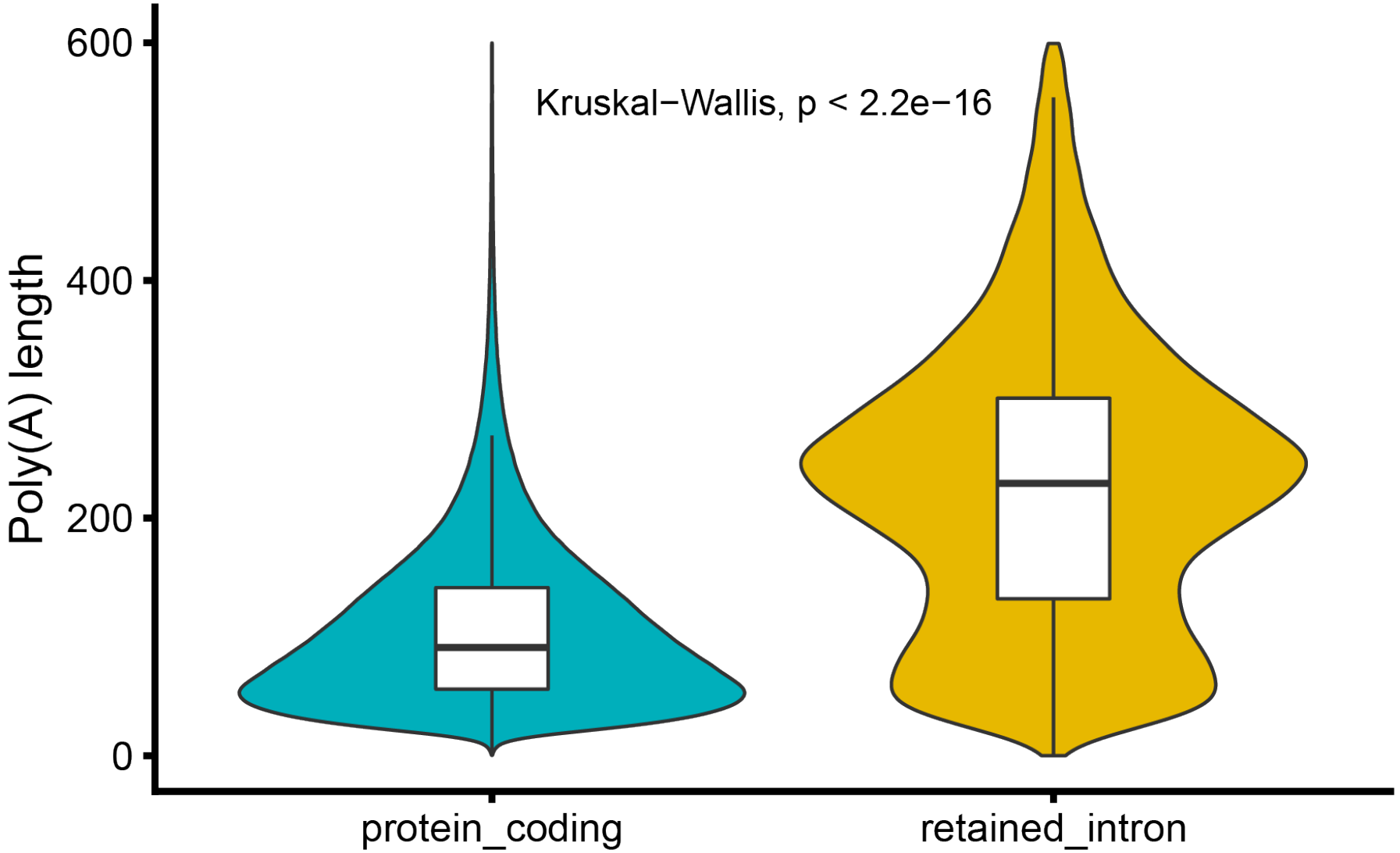
Poly (A) tails of isoforms

- Exploring further into different isoforms of *DDX5*
- We identified different isoforms had different poly(A) lengths (>25X coverage per isoform)
- Isoforms with retained introns had longer poly(A) tails



Workman et al bioRxiv (2018)

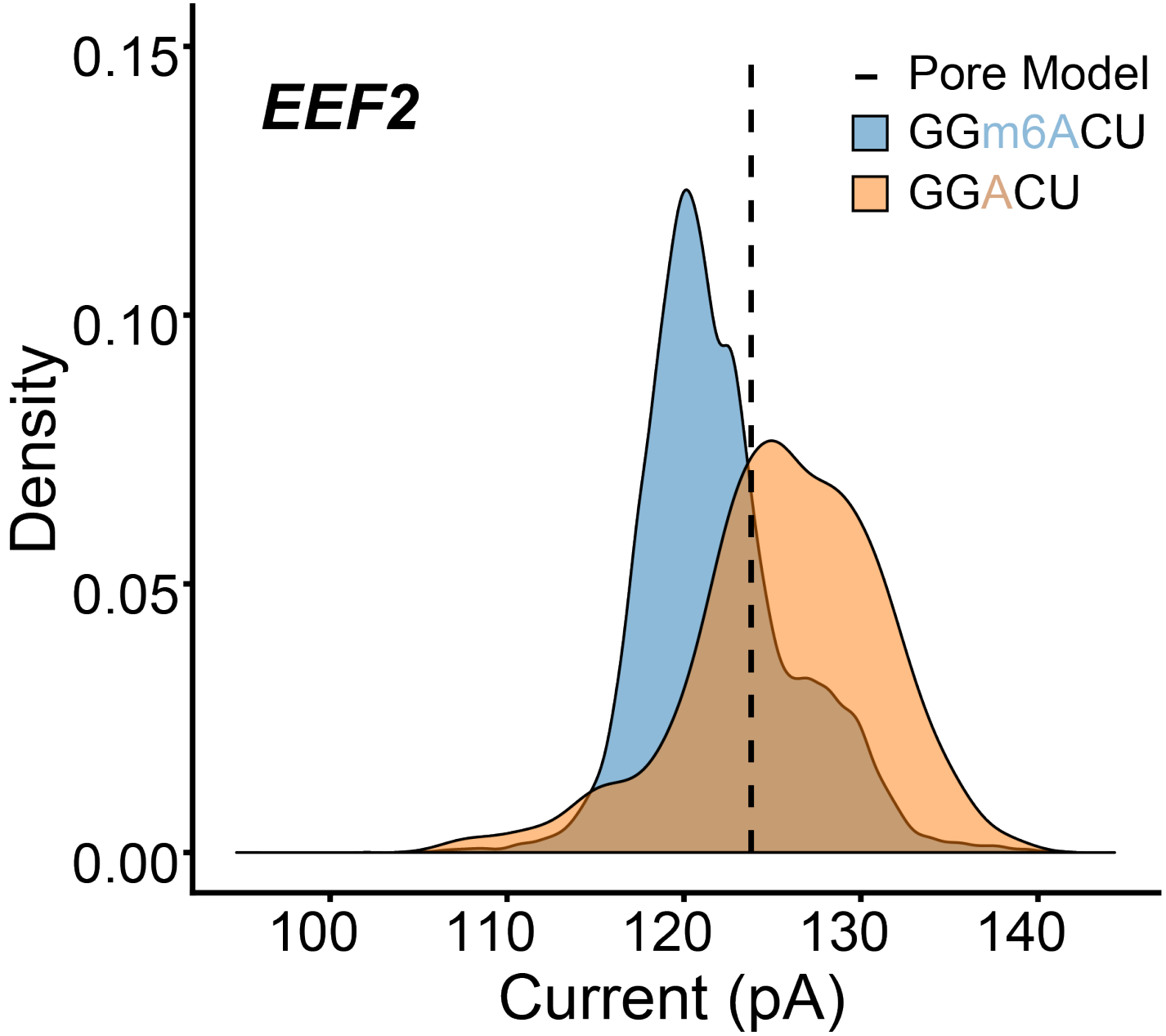
Poly(A) tails of transcript classes



Exploring this trend: transcripts with retained introns have longer poly(A) tails than spliced transcripts

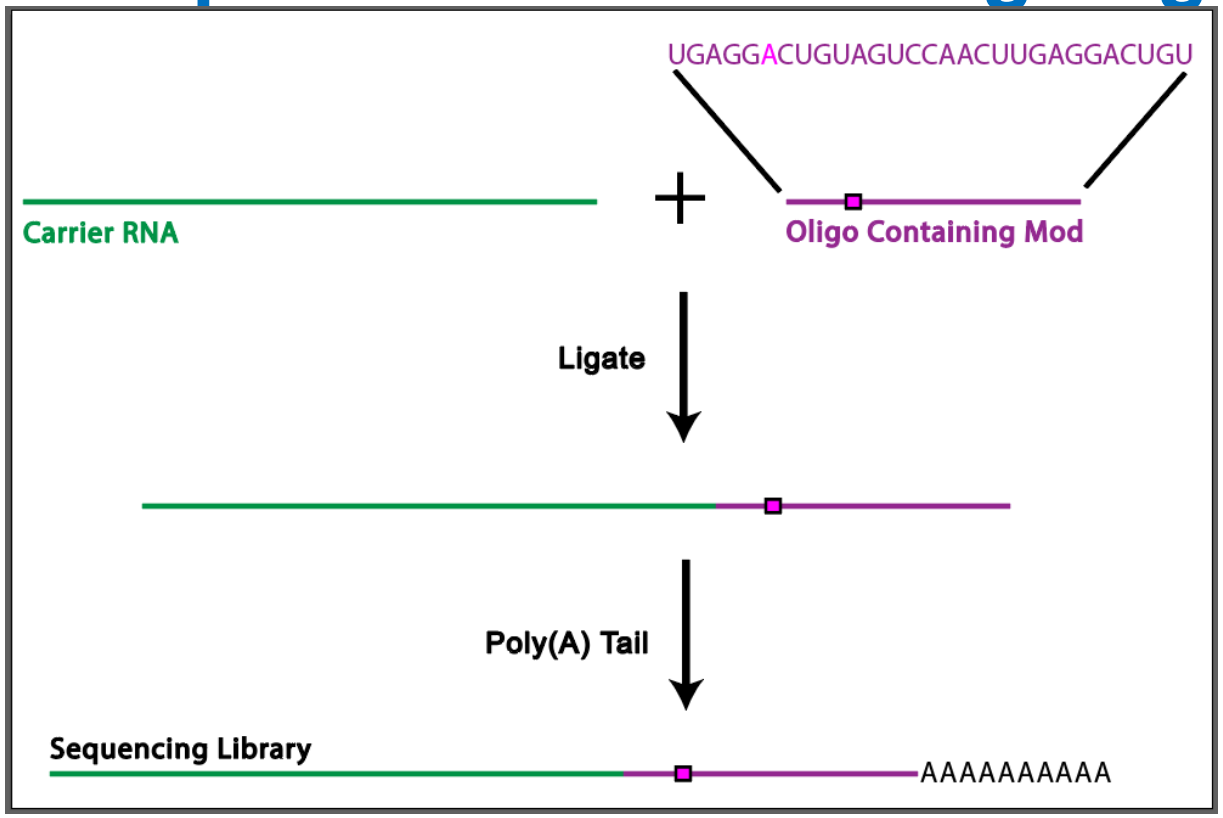
Exploring the dRNA for m6A

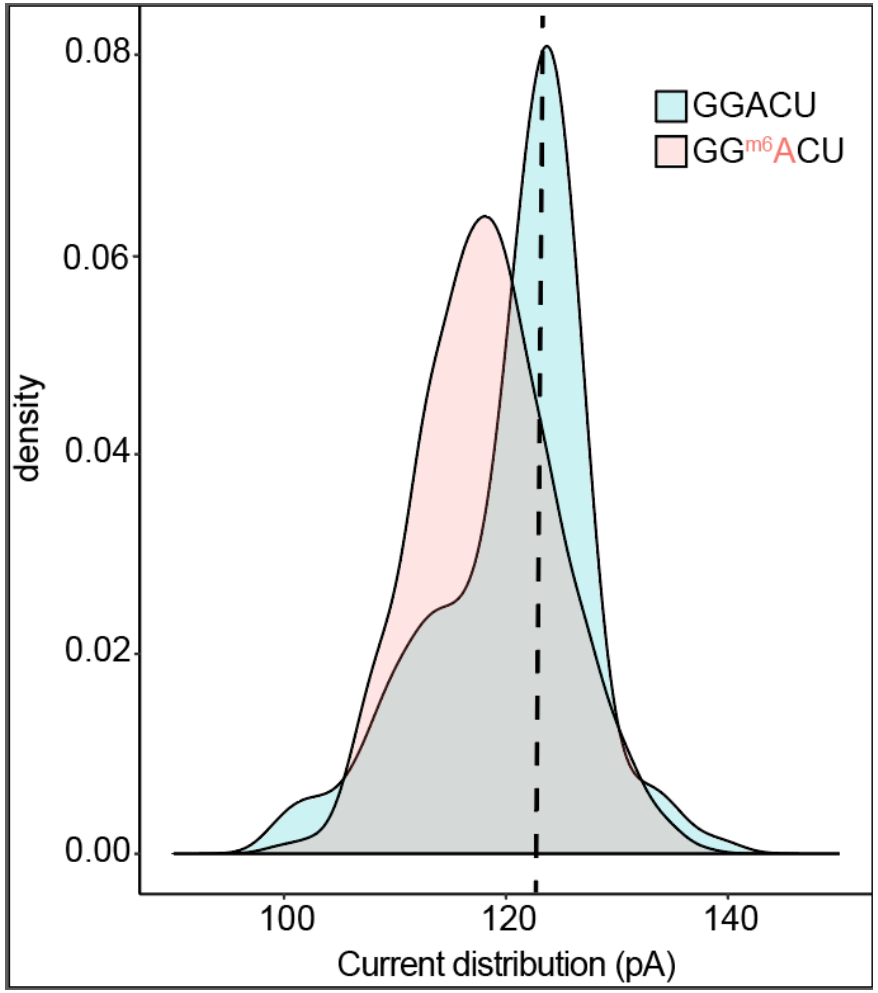
- Eukaryotic elongation factor 2 has a METTL3 motif GGACU (m6A writer) in the mRNA sequence
- Has been shown to have m6A via IP-seq methods (Meyer et al Cell 2012)
- Compared dRNA data with IVT'd dRNA signal



Workman et al bioRxiv (2018)

Training RNA basecaller to recognize modified sites requires truth sets: oligo ligation



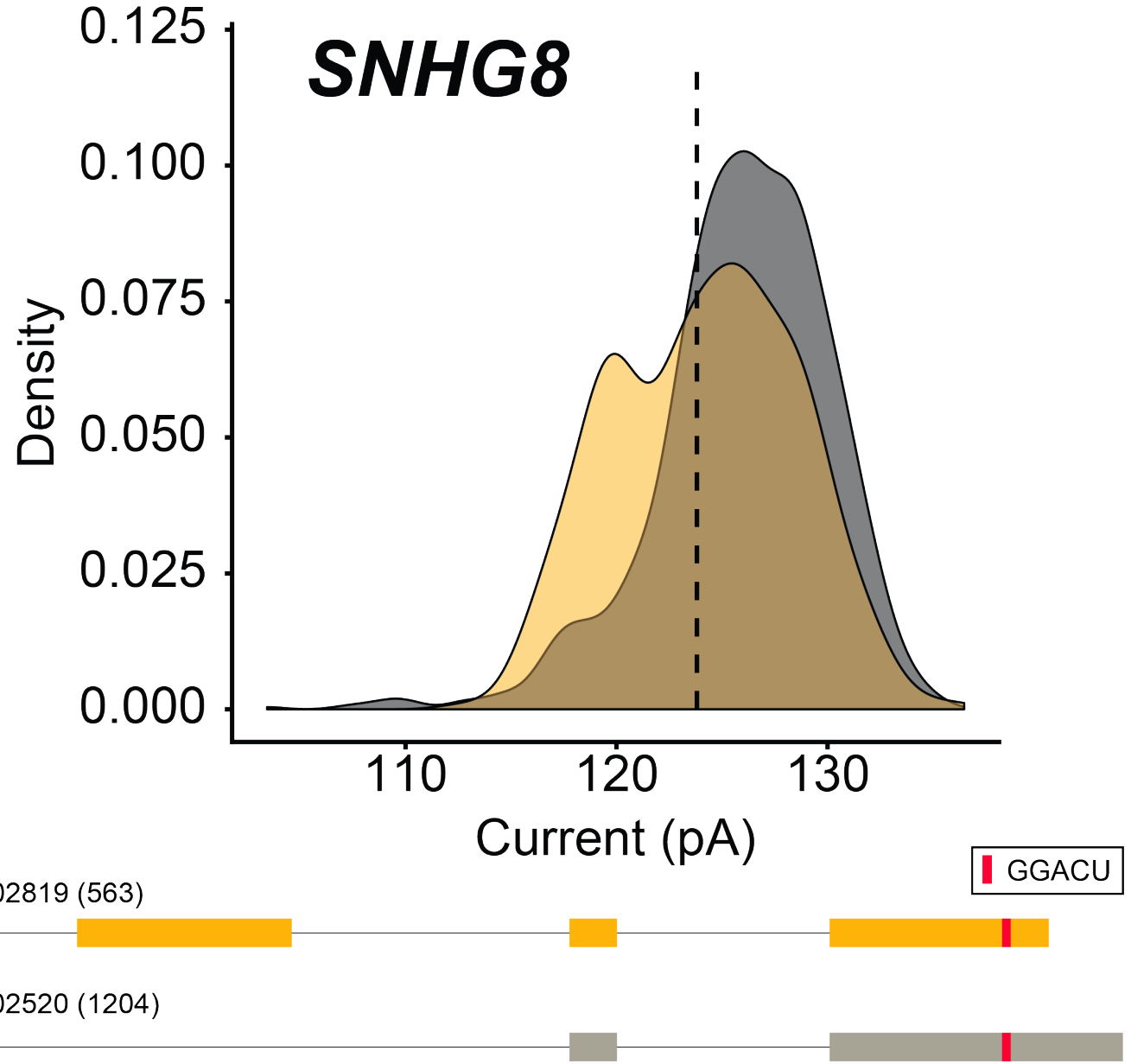


Short oligo (cheaper) containing METTL3 motif GGACU (m6A writer) ligated to handle (to achieve ~100b desired for seq)

Workman et al bioRxiv (2018)

Isoform Specific m6A modifications: *SNHG8*

- Examining isoform dependence of modification signal: METTL3 motif in *SNHG8* isoforms
- Different % of transcripts are modified dependent on isoform



Workman et al bioRxiv (2018)

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
- Targeted sequencing with Cas9 allows for long reads in targeted regions, sidestepping issues of cost.
- Direct RNA sequencing suggests we can measure isoforms, poly (A) tail lengths and even RNA modifications

Acknowledgments

National Human Genome Research Institute
1R01HG010538
1R01HG009190

OICR
Ontario Institute for Cancer Research

- Jared Simpson
- Paul Tang
- P.C. Zuzarte
- Michael Molnar

Nanopore RNA Consortia

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

JOHNS HOPKINS
BIOMEDICAL ENGINEERING

- Jawara Allen
- Brittany Avin
- Sheridan Cavalier
- Yunfan Fan
- Ariel Gershman
- Timothy Gilpatrick
- Isac Lee
- Brittany Pielstick
- Roham Razaghi
- Norah Sadowski
- Rachael Workman