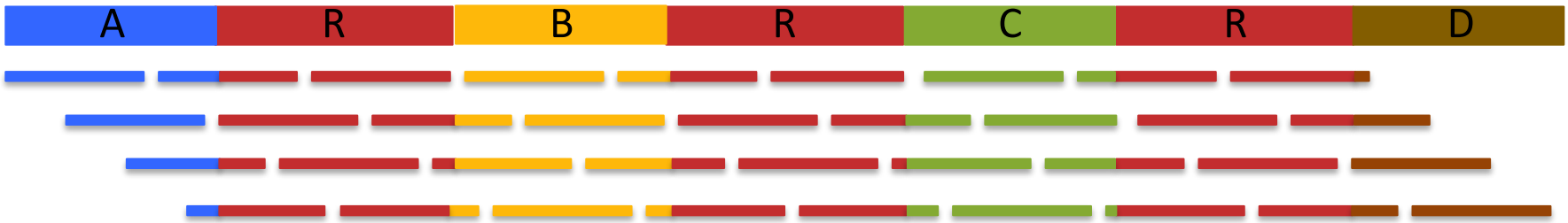
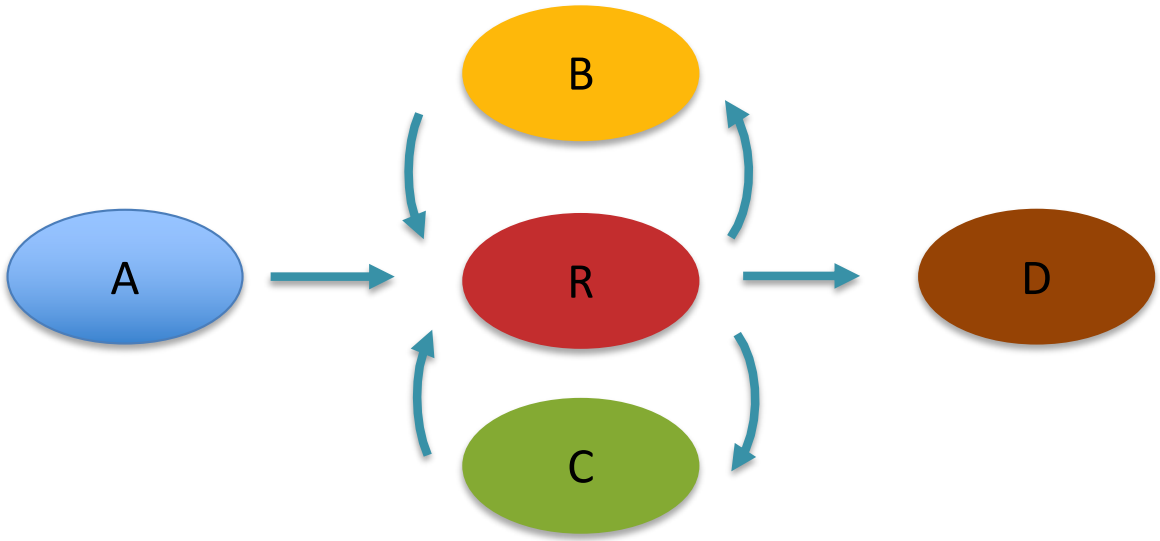
Nanopore Sequencing

Sam Kovaka
(Many slides by Michael Schatz)

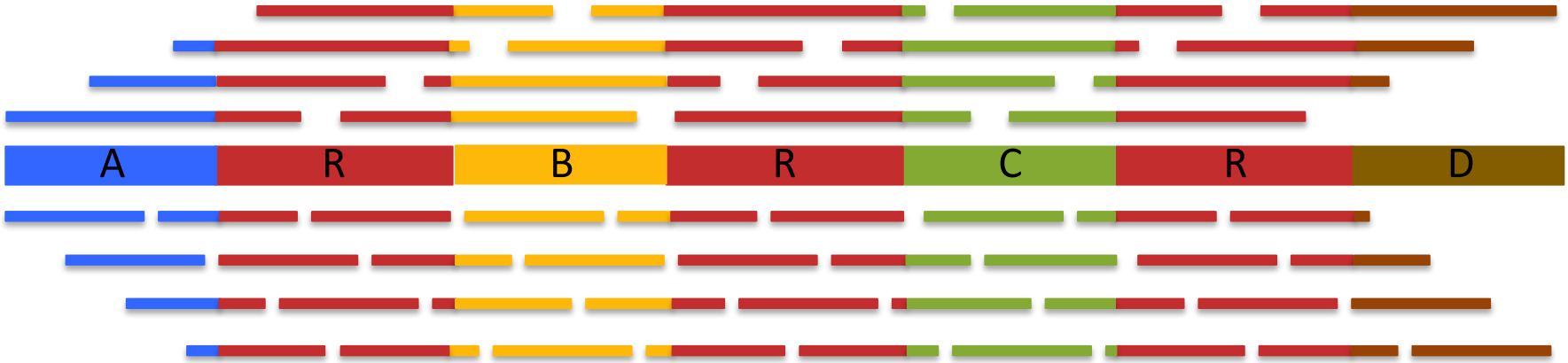
Mar 2, 2020
Lecture 11: Applied Comparative Genomics

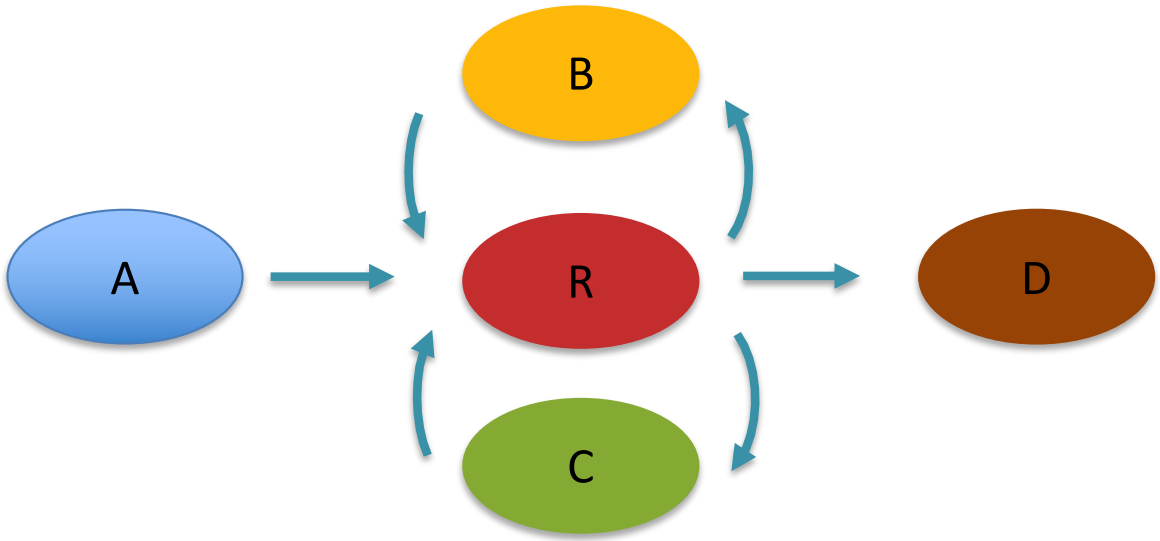
Assembly Complexity



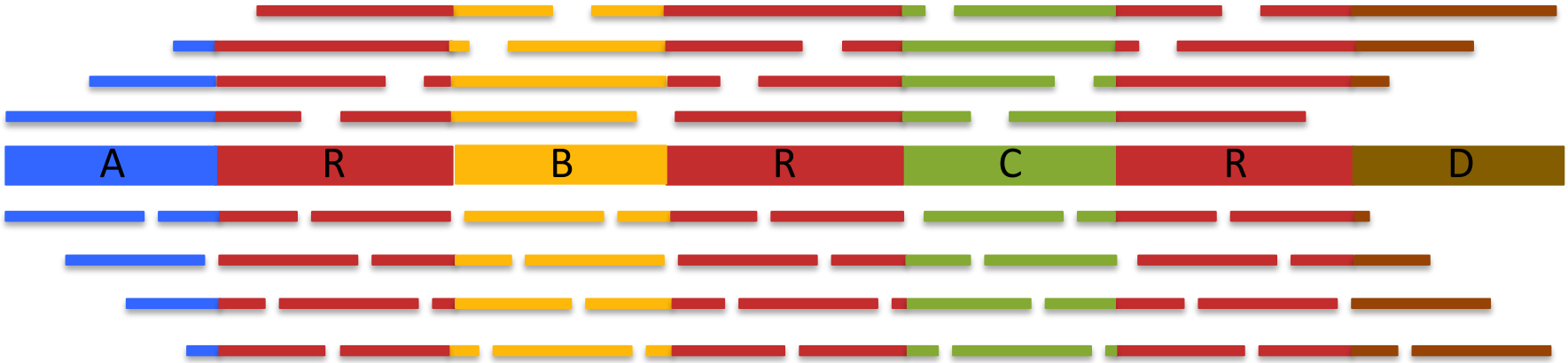


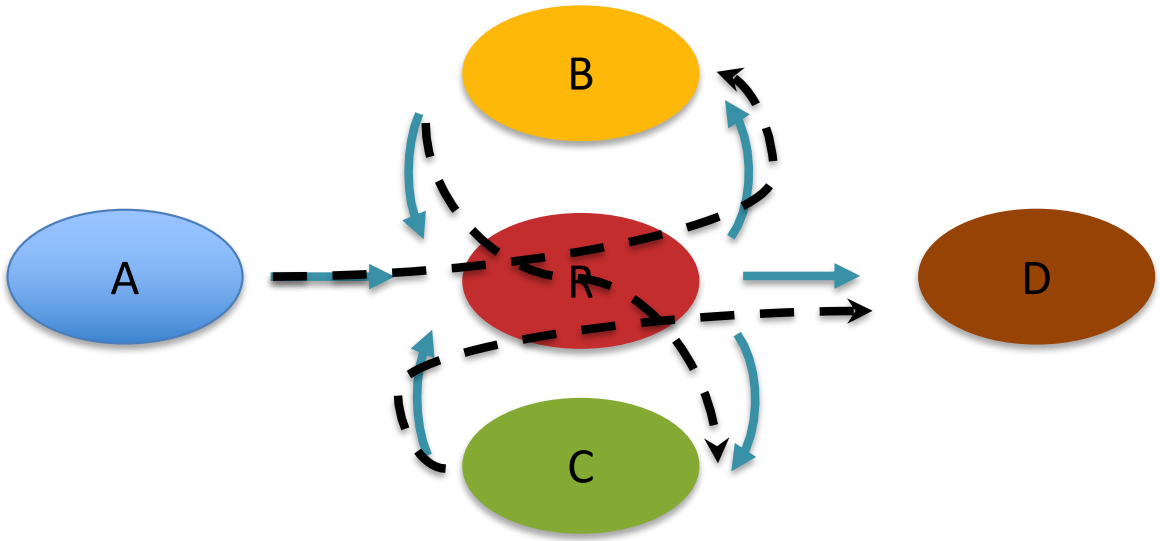
Assembly Complexity



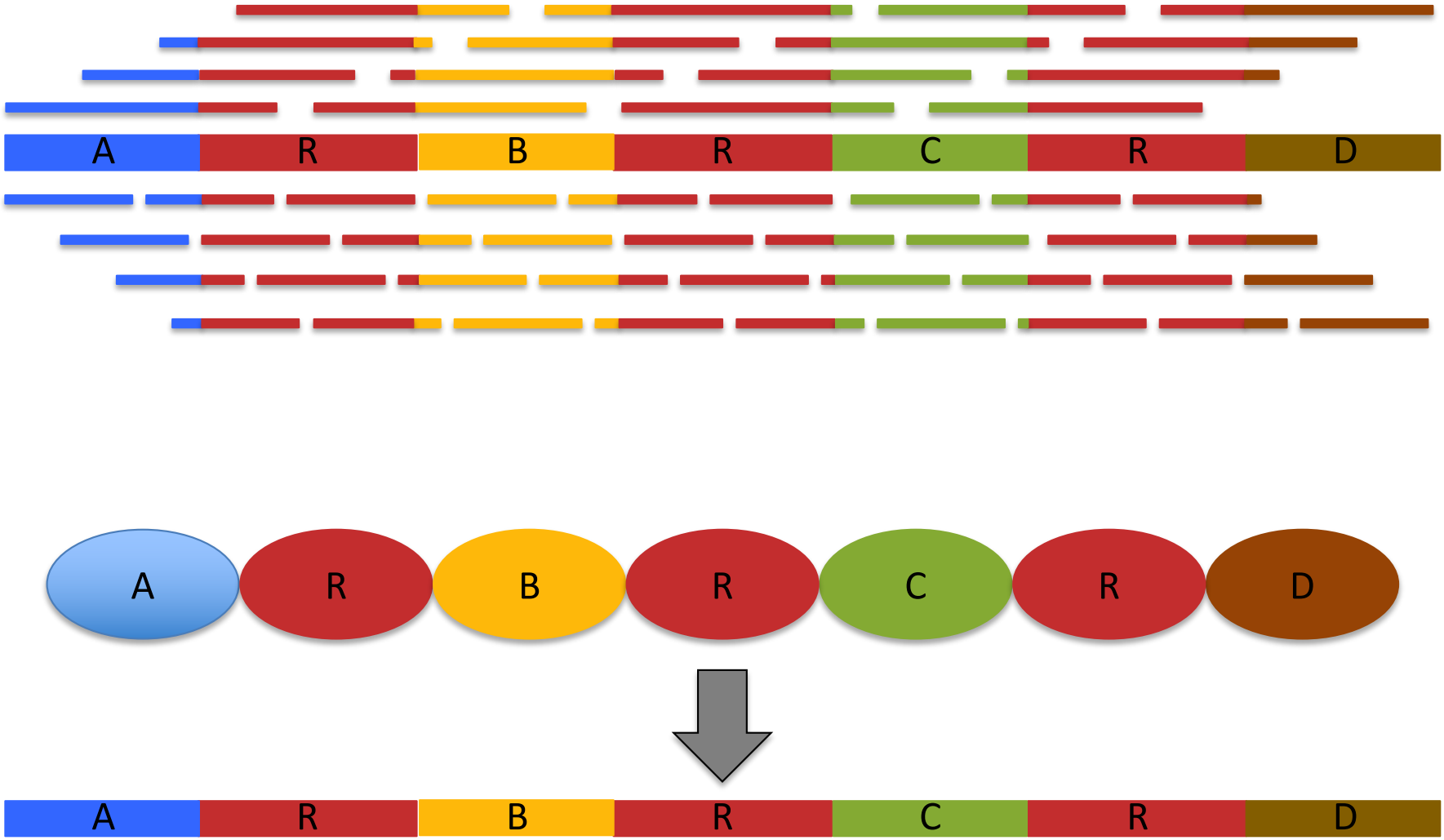


Assembly Complexity





Assembly Complexity

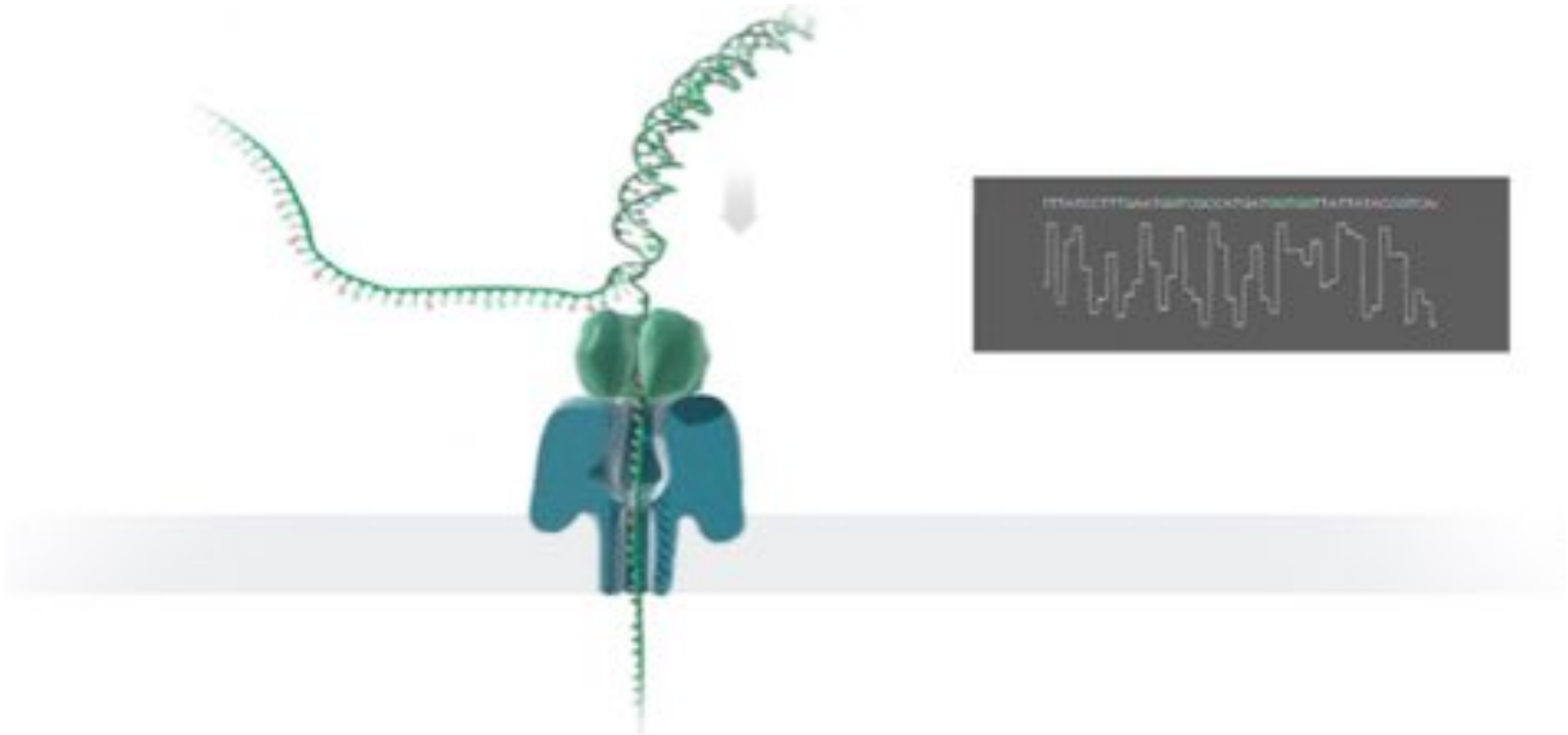


The advantages of SMRT sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology*. 14:405

Oxford Nanopore Technologies (ONT)

Nanopore Sequencing

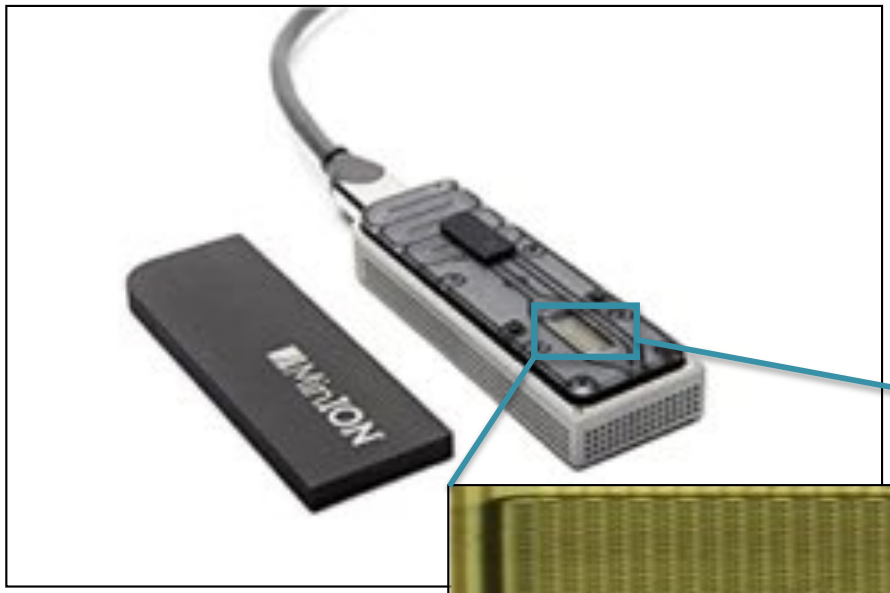
Sequences DNA/RNA by measuring changes in ionic current as nucleotide strand passes through a pore

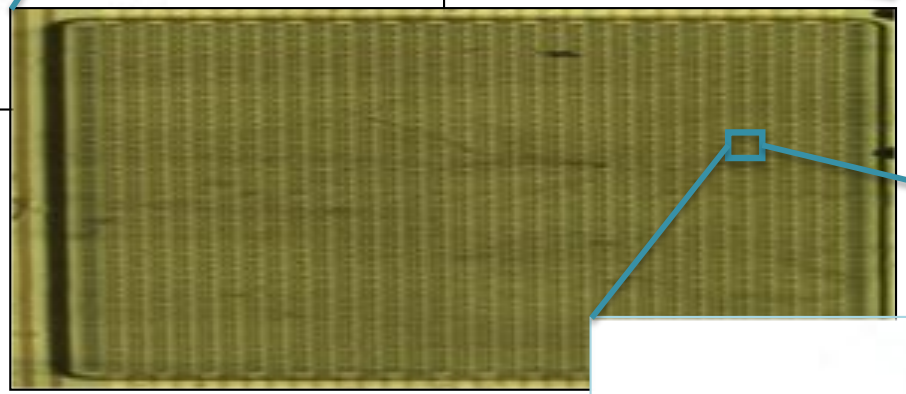


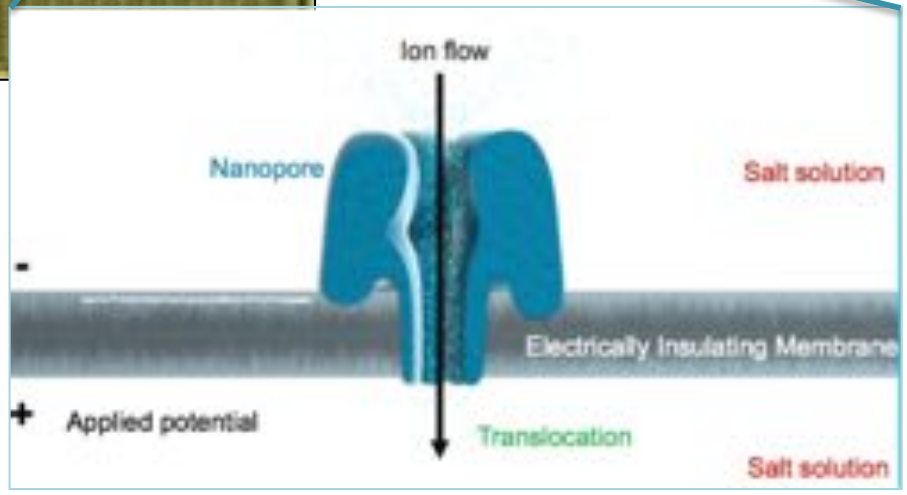
https://www.youtube.com/watch?v=CE4dW64x3Ts

nanoporetech.com/applications/dna-nanopore-sequencing

Oxford Nanopore MinION







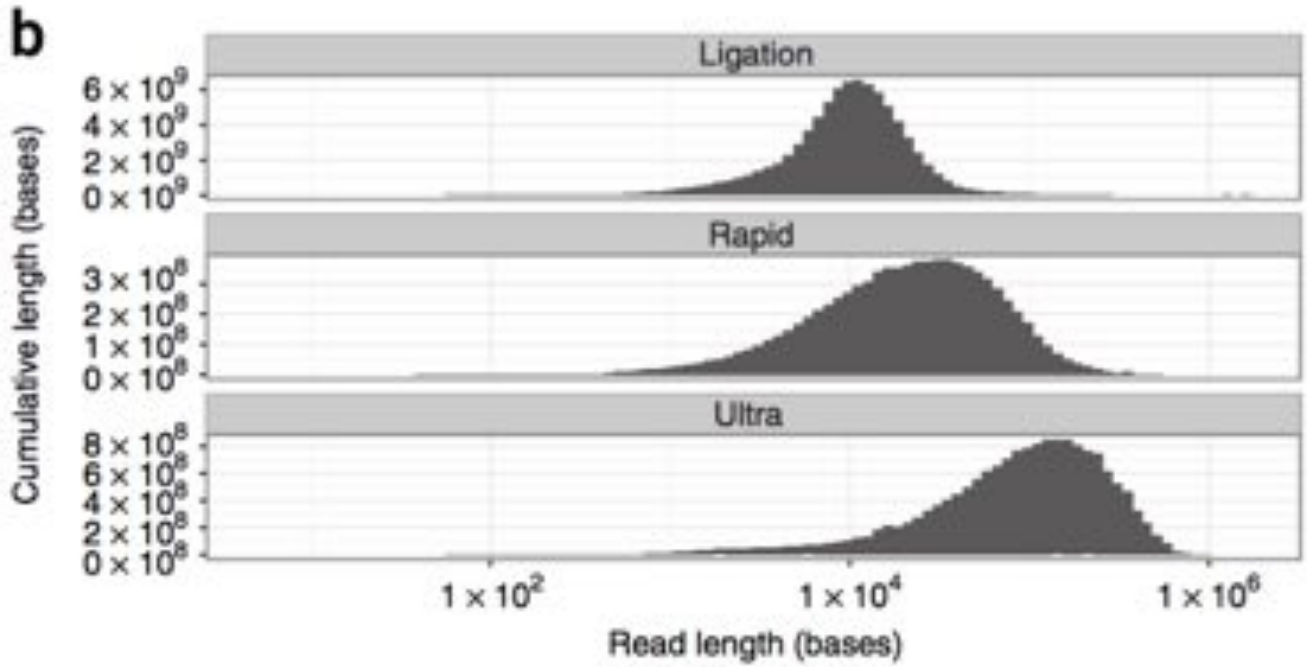
- Thumb drive sized sequencer powered over USB
- Contains 512 channels
- Four pores per channel, only one pore active at a time

- Early access began in 2014
- Officially released in 2015 (the same year I met Mike!)

Nanopore Read Lengths

A typical MinION run produces ~10Gbp worth of reads of with a mean read length of ~10Kbp

- "Ultra-long" runs produce many reads > 100Kbp, with the longest read ever observed ~2.3Mbp

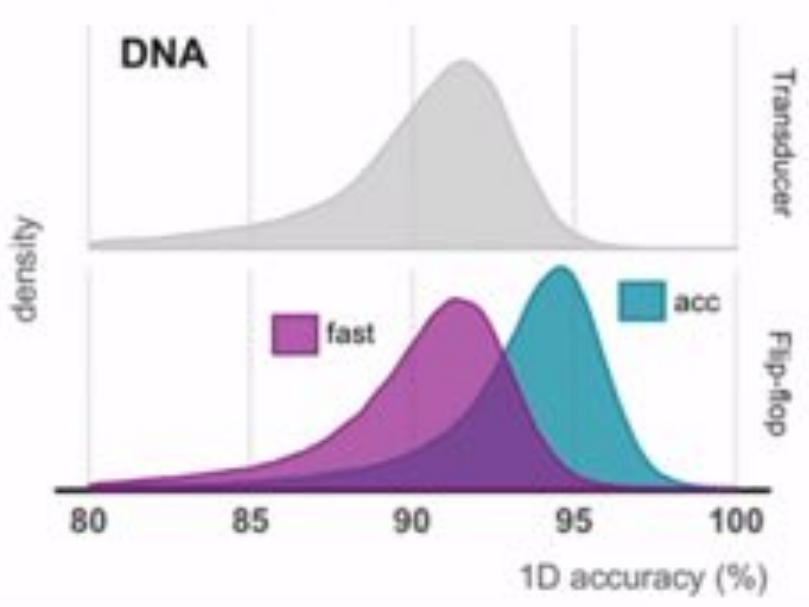


Nanopore sequencing and assembly of a human genome with ultra-long reads
Jain et al. (2018) Nature Biotechnology: https://www.nature.com/articles/nbt.4060

Nanopore Read Quality

ONT reads typically have a mean error rate of ~10%

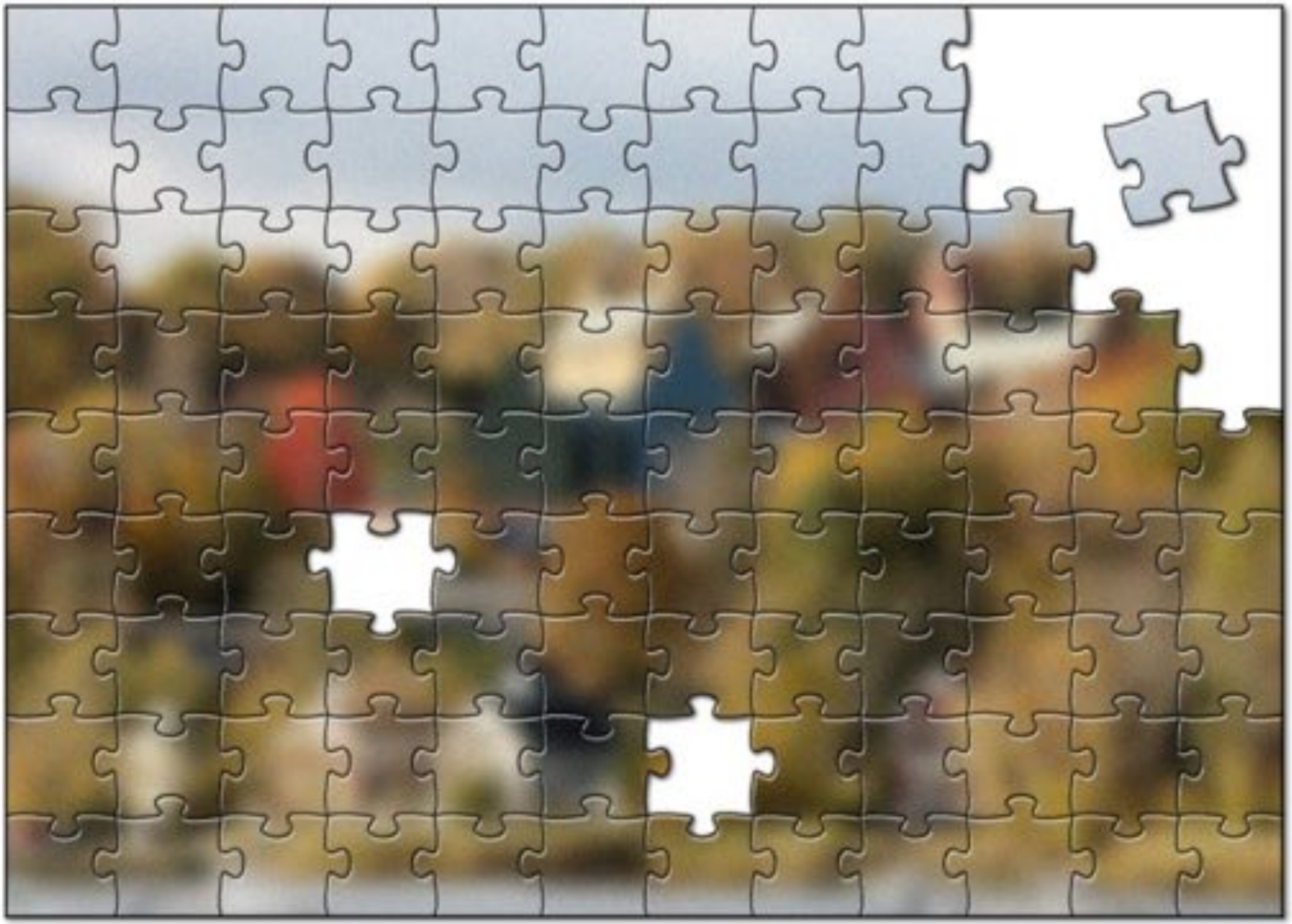
Predominantly indels



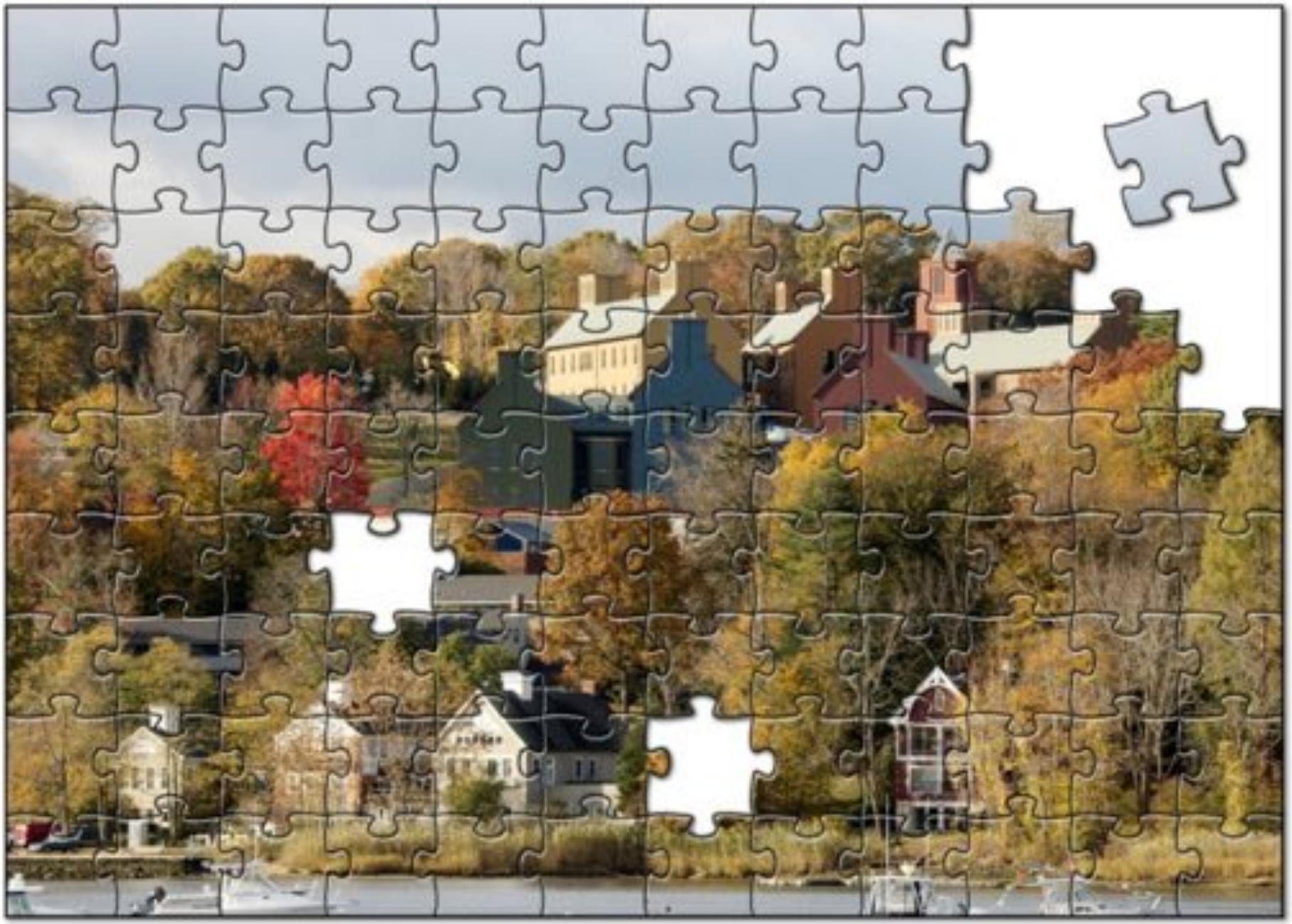
London Calling 2019 Keynote

```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
||||||||||||||||||||||||| ||||||| |||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAAGGCGGCTAGG
| |||||| |||||||||||||| |||| | |||||| |||||| ||||||
A-TATAAATCAGTTGATCCATTAAGAA-AGAAACGC-AAAGGC-GCTAGG

CAACCTTGAATGTAATCGCACTTGAAGAACAAGATTTTATTCCGCGCCCG
| |||||| |||| ||  ||||||||||||||||||||||||||||||||
C-ACCTTG-ATGT-AT--CACTTGAAGAACAAGATTTTATTCCGCGCCCG

TAACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA
| ||||||| |||||||||||||| || ||    |||||||||| |||||
T-ACGAATC-AGATTCTGAAAACA-ATGAT----ACCTCCAAAAGCACAA

-AGGAGGGGAAAGGGGGGAATATCT-ATAAAAGATTACAAATTAGA-TGA
 ||||||   ||     |||||||| || |||||||||||||| || |||
GAGGAGG---AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA

ACT-AATTCACAATA-AATAACACTTTTA-ACAGAATTGAT-GGAA-GTT
||| ||||||||| | |||||||||||| ||| ||||||| |||| |||
ACTAAATTCACAA-ATAATAACACTTTTAGACAAAATTGATGGGAAGGTT

TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA
|| ||||||||| ||||||| ||| |||| |||||| ||||| |||||||
TC-GAGAGATCC-AAACAAT-GGCGATCG-CTTTGACGTTACAAATCAAA

ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACTTATTCACAATTAG
||||||| |||||||||  |||||| ||||| ||||||||||||||||||
ATCCAGT-GAAAATATA--TTATGC-ATCCA-GAACTTATTCACAATTAG
```

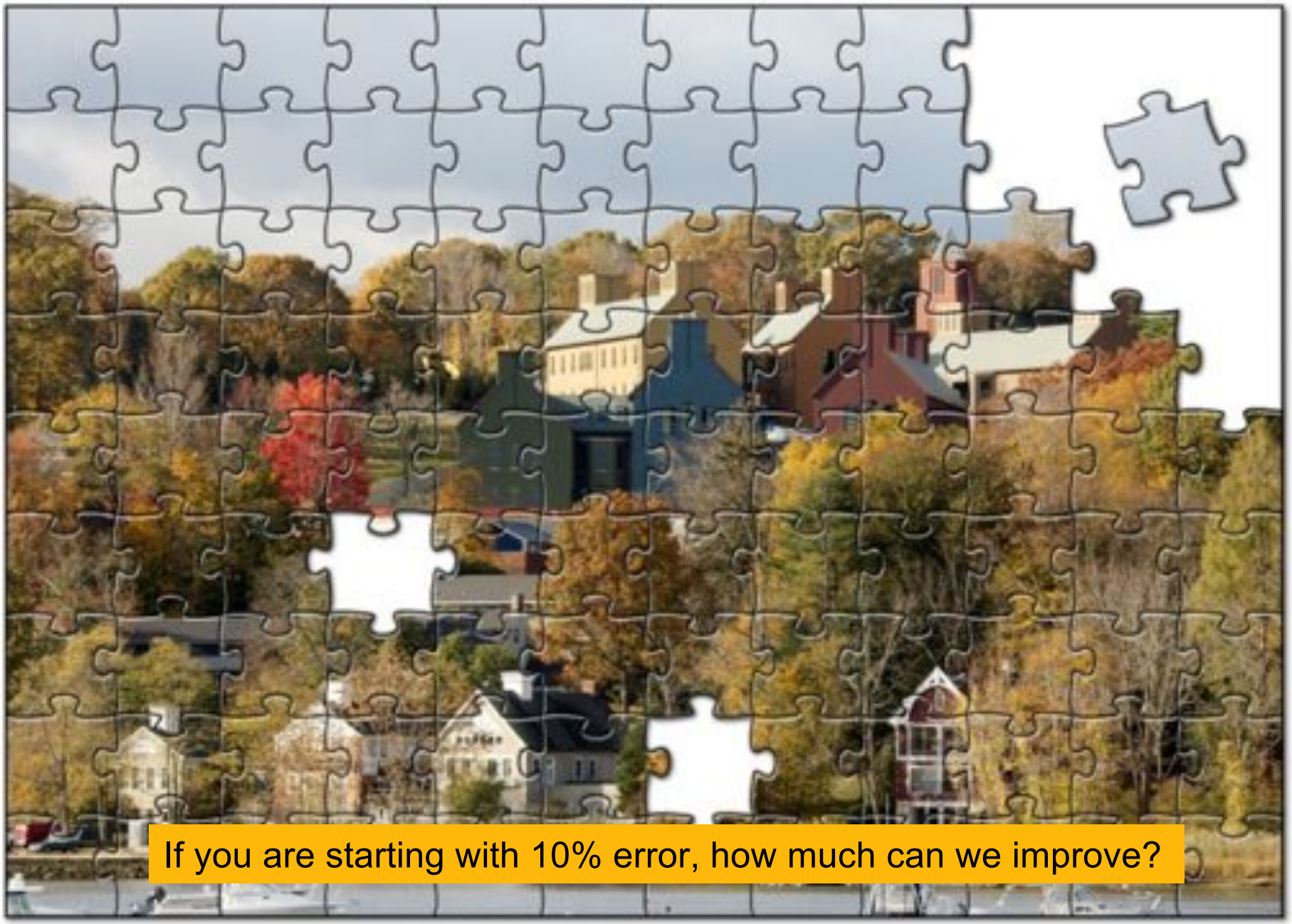
Single Molecule Sequences



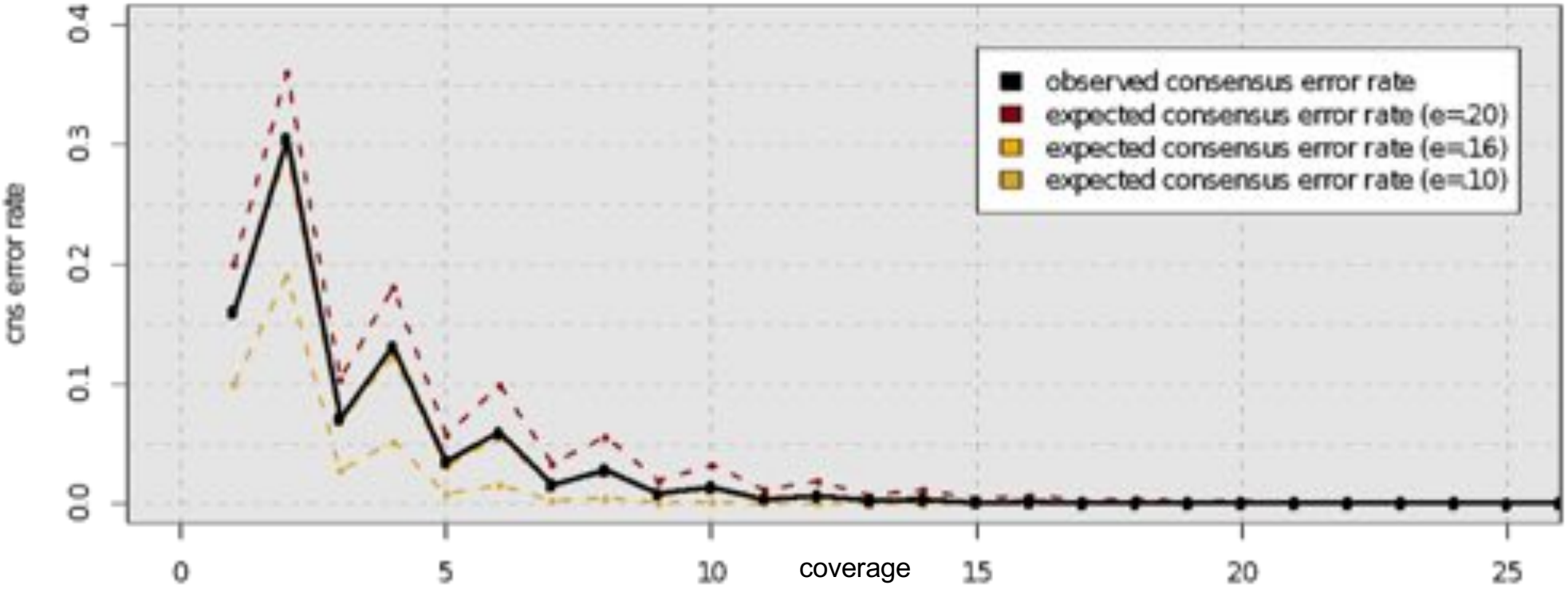
“Corrective Lens” for Sequencing



“Corrective Lens” for Sequencing



Consensus Accuracy and Coverage



Coverage can overcome *random* errors

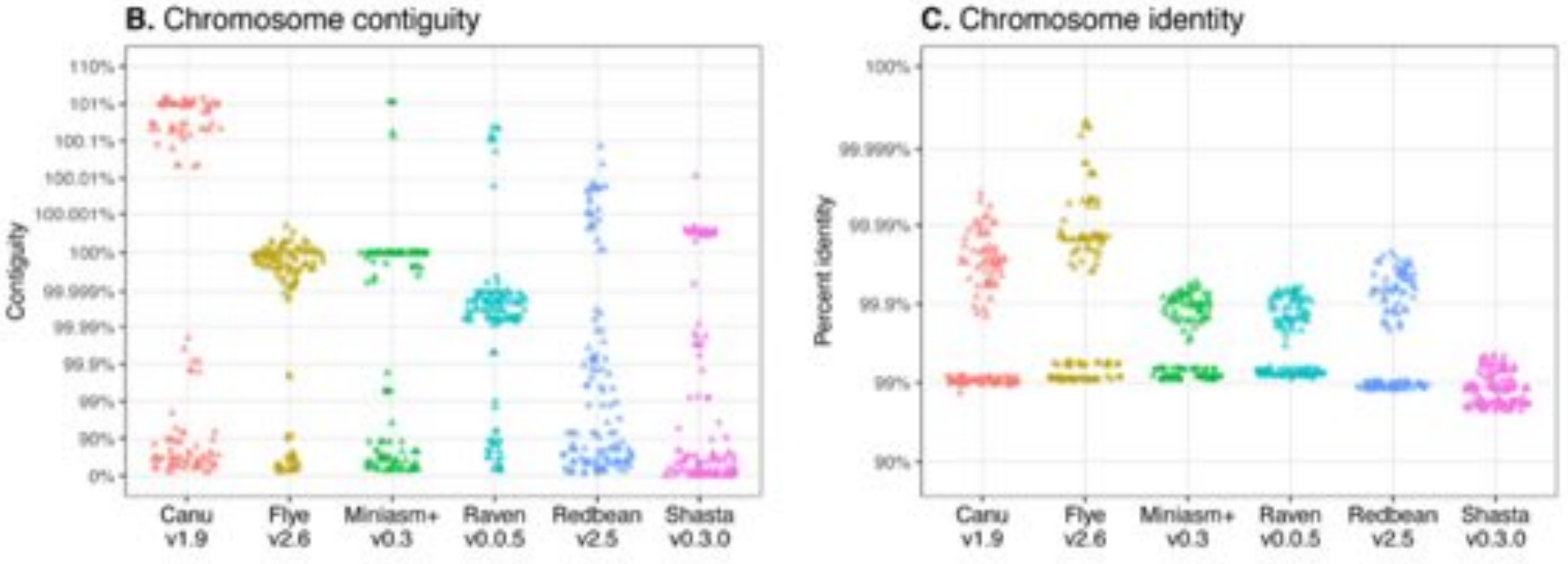
- Dashed: error model from binomial sampling; Solid: observed accuracy
- Unfortunately, ONT has some non-random errors (mainly homopolymers)

$$CNS\,Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

ONT Assembly Accuracy

Assemblies are quite contiguous, but percent identity maxes out ~99%

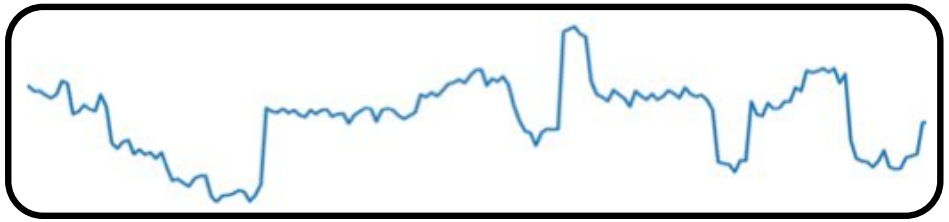
- Depends on organism - basecallers are mainly trained on human
- Polishing from other technologies necessary for reference-quality



Benchmarking of long-read assemblers for prokaryote whole genome sequencing
Wick and Holt (2019) F1000 Research: https://doi.org/10.12688/f1000research.21782.1

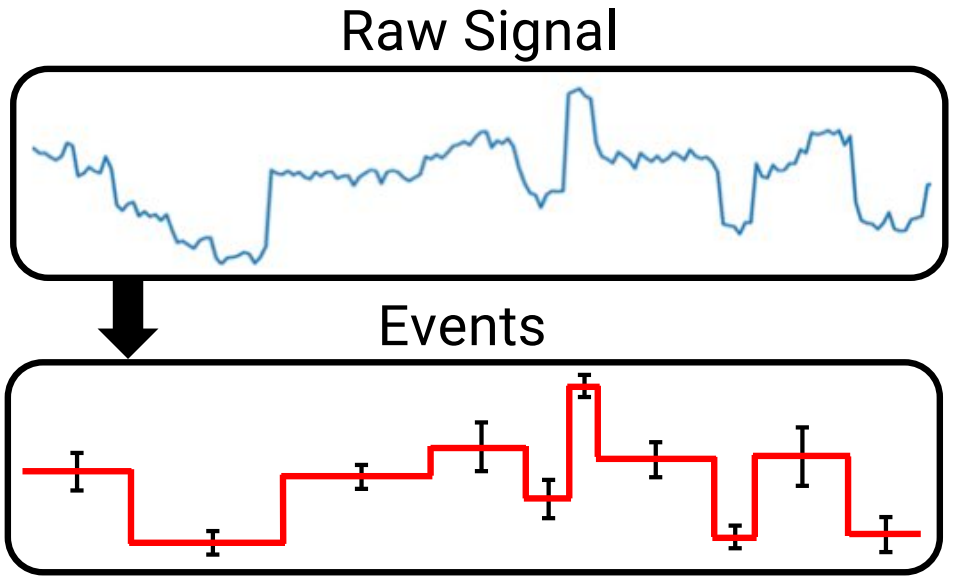
Nanopore Basecalling

Raw Signal



Translation of raw signal into basepairs

Nanopore Basecalling

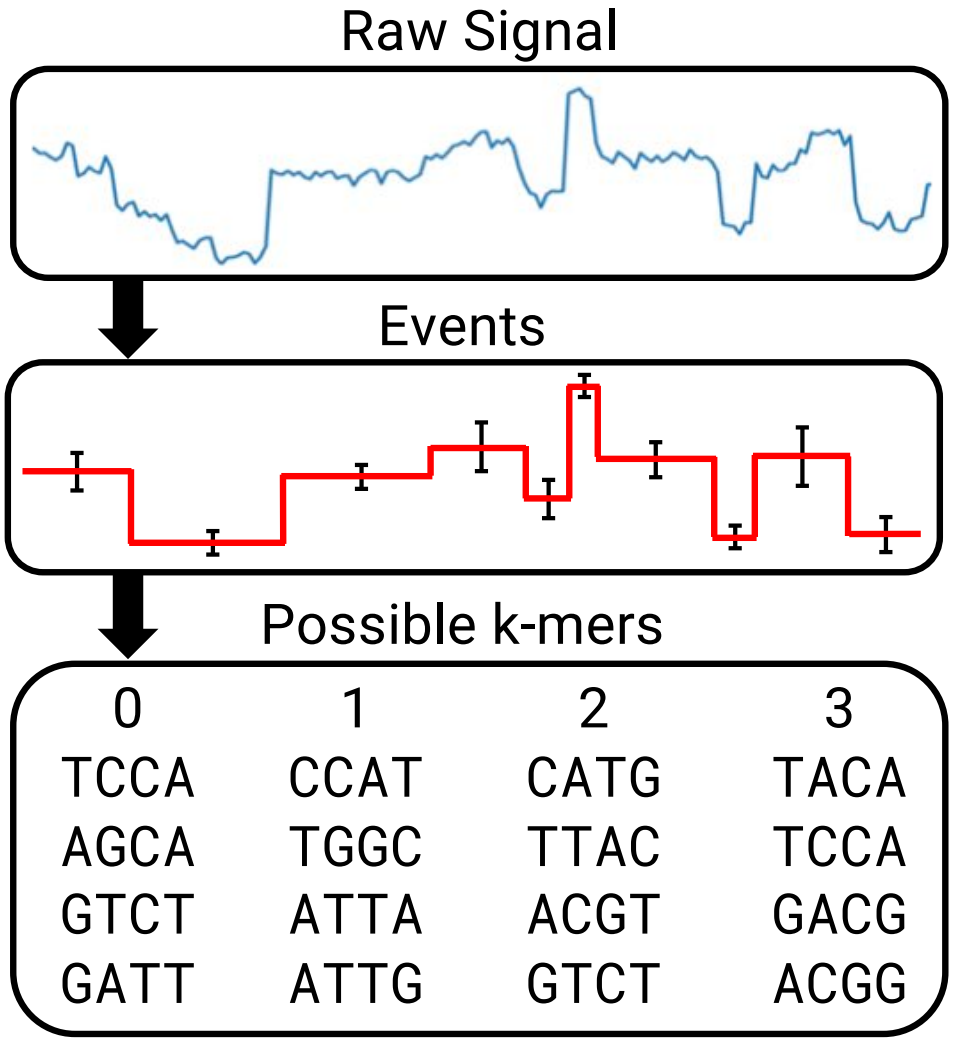


Translation of raw signal into basepairs

Early basecallers began by estimating k-mer boundaries using “events”, which were then input to an HMM

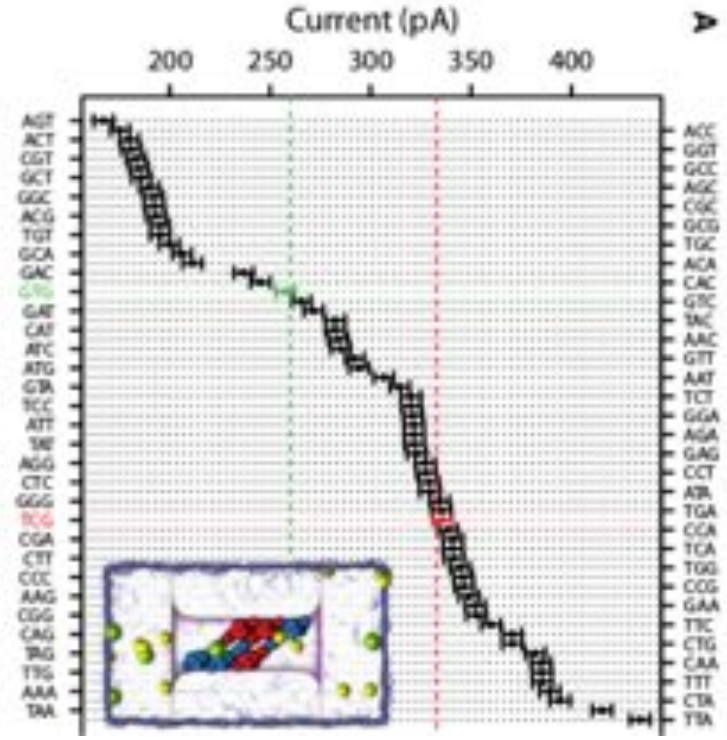
Modern basecalers use neural networks directly on raw signal

Nanopore Basecalling



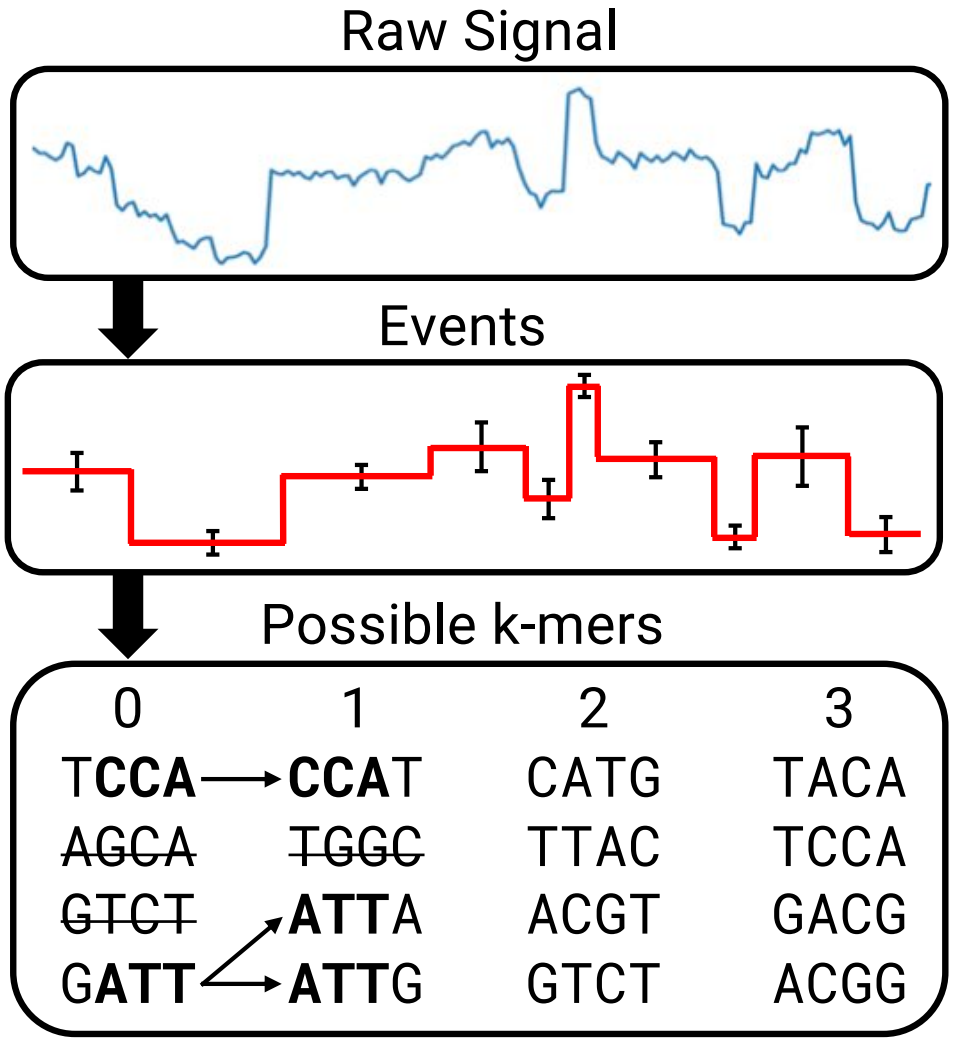
(Based on probability of event matches)

ONT releases k-mer models with expected current distribution of every k-mer

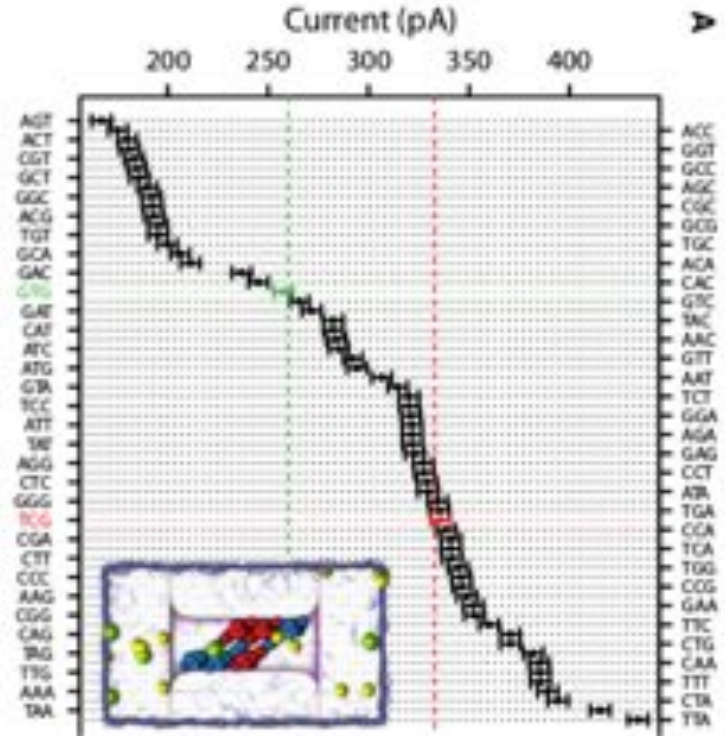


DNA Base-Calling from a Nanopore Using a Viterbi Algorithm
Timp et al. (2012) *Biophysical Journal*

Nanopore Basecalling

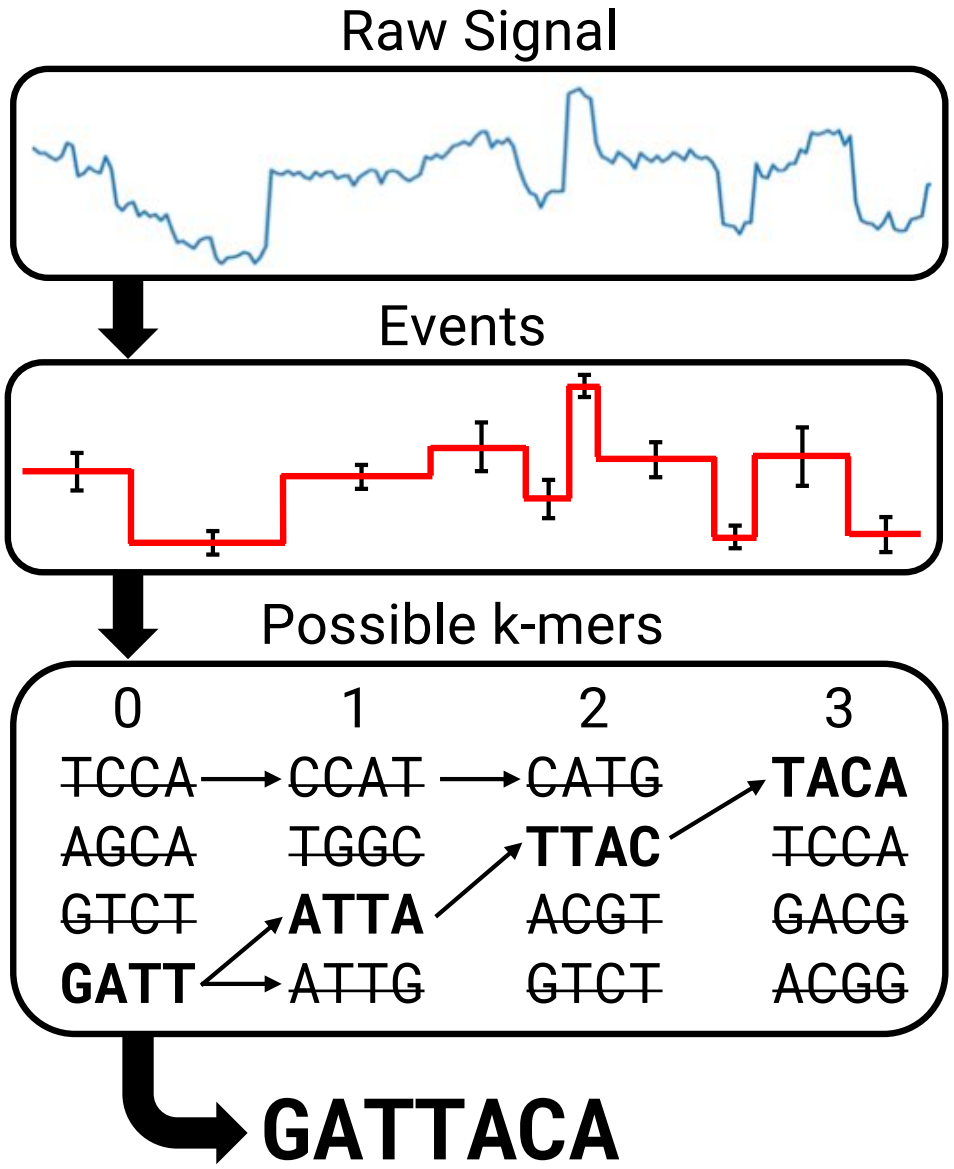


Certain k-mers can be eliminated based on possible transitions

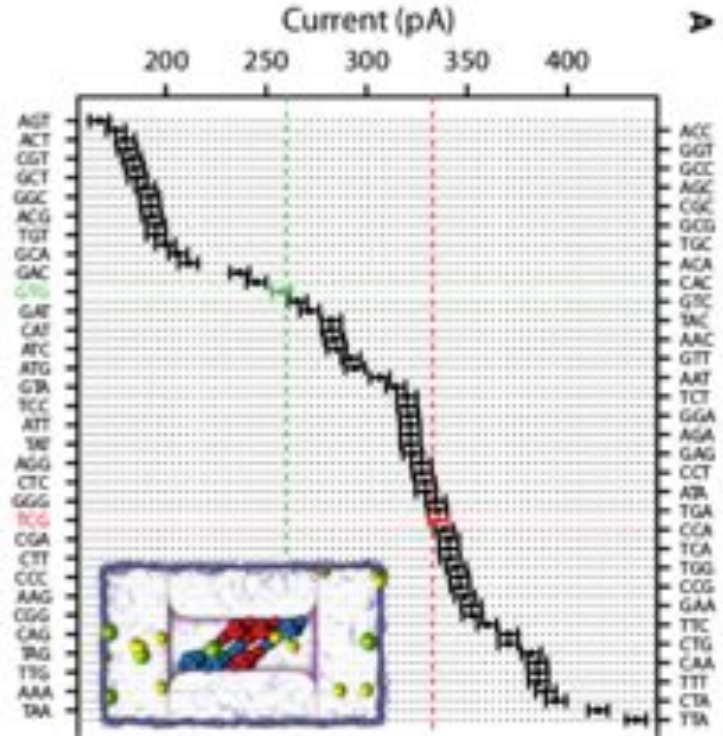


DNA Base-Calling from a Nanopore Using a Viterbi Algorithm
Timp et al. (2012) *Biophysical Journal*

Nanopore Basecalling



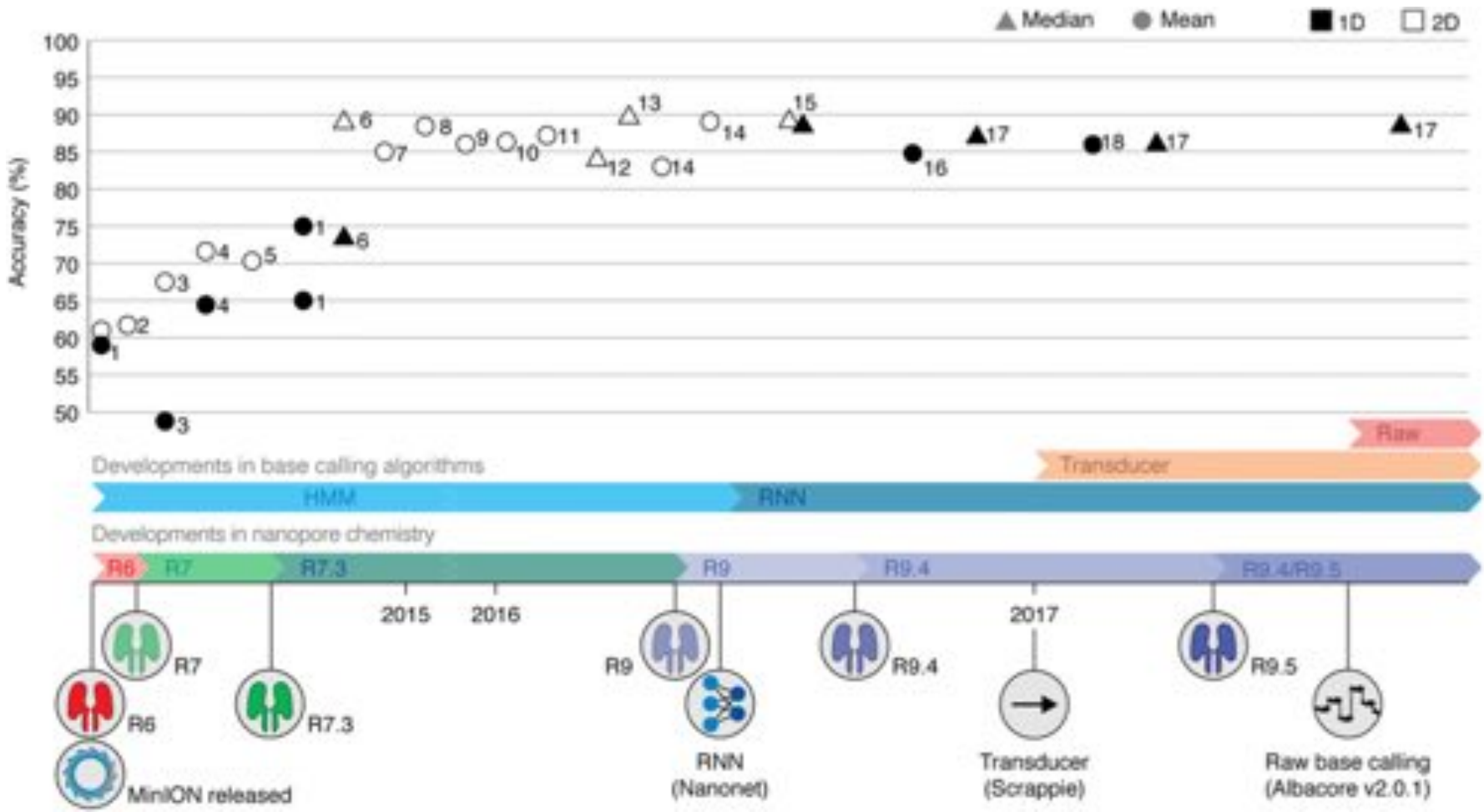
Final sequence determined by most probable k-mers



"DNA Base-Calling from a Nanopore Using a Viterbi Algorithm"
Timp et al. (2012) *Biophysical Journal*

Basecaller/Pore Timeline

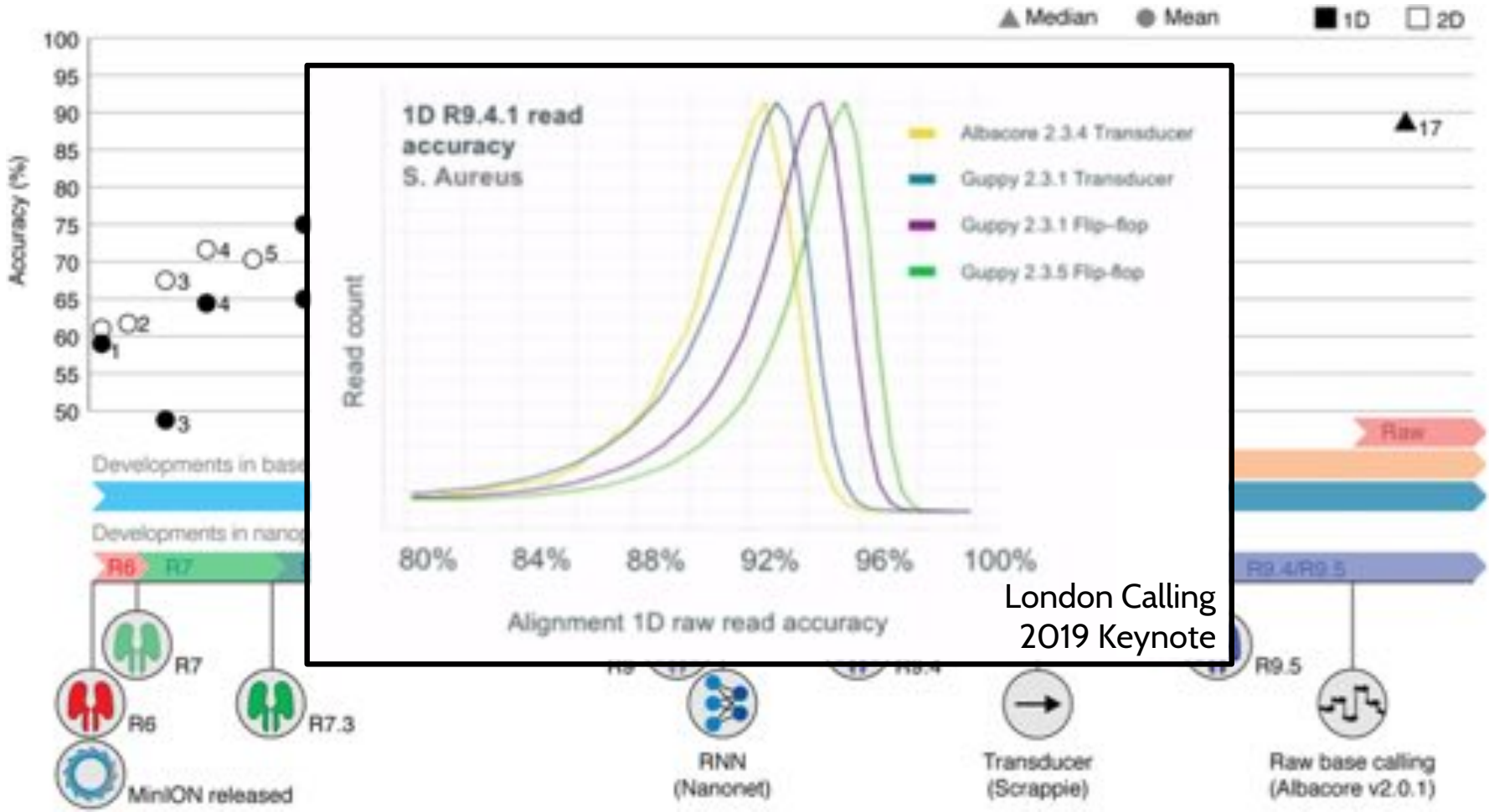
Development of both pore chemistry and basecalling algorithms is responsible for improvement in accuracy



From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy
Rang *et al* (2018) *Genome Biology*. https://doi.org/10.1186/s13059-018-1462-9

Basecaller/Pore Timeline

Development of both pore chemistry and basecalling algorithms is responsible for improvement in accuracy

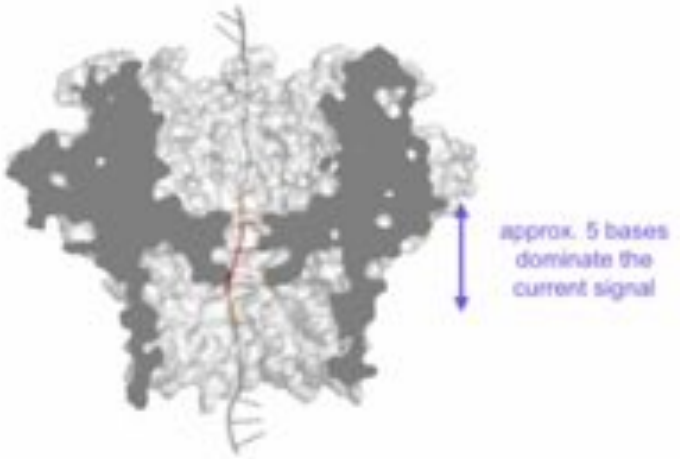


From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy
Rang *et al* (2018) *Genome Biology*. https://doi.org/10.1186/s13059-018-1462-9

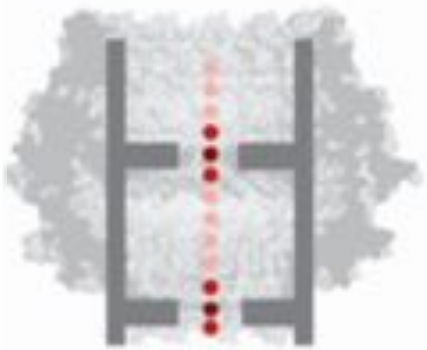
New Pore Chemistries

ONT recently released “R10” pore chemisity

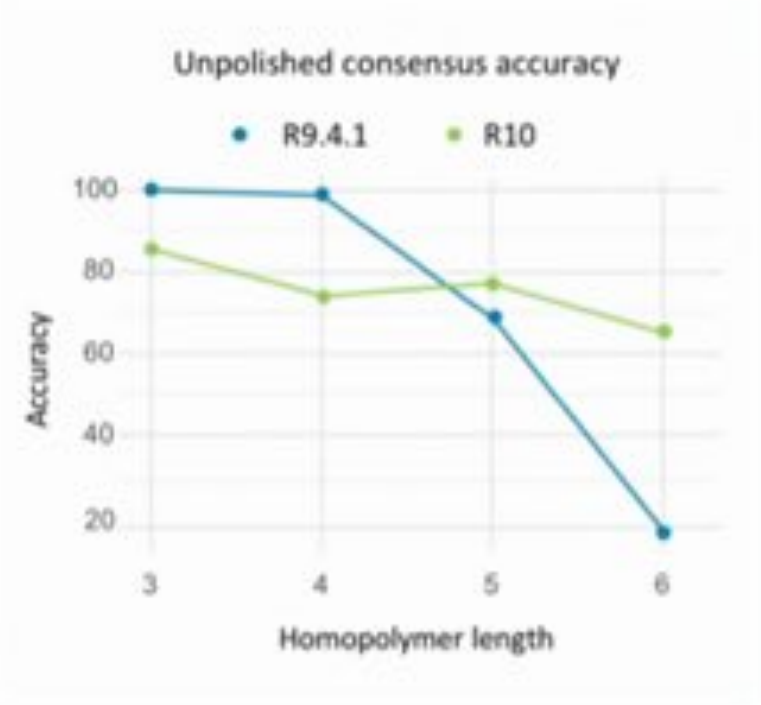
- Two bottlenecks where nucleotides affect current
- Spans longer homopolymers
- May still have non-random errors, but profile is different



Standard pore chemistry
“R9”



Multiple points of contribution
“R10”



From 2018 London Calling Keynote
https://vimeo.com/272526835

More Throughput

MinION

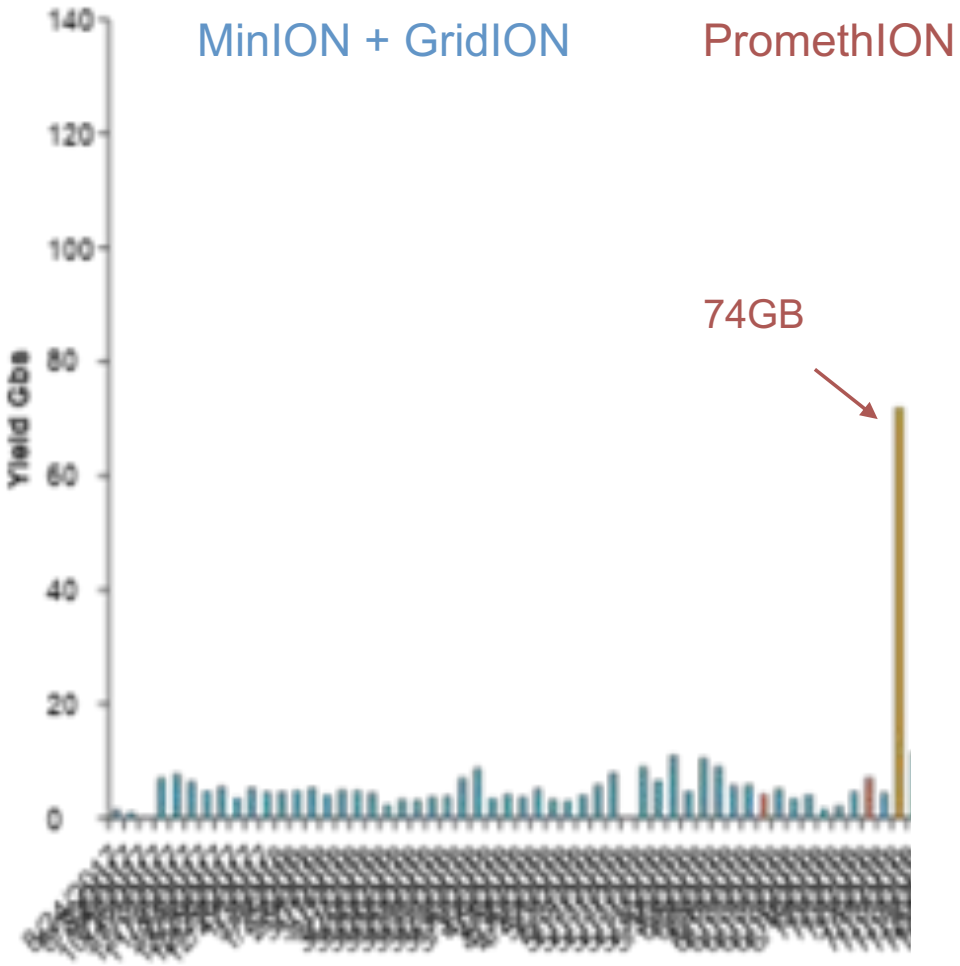
Quick Mobile Sequencing
$1k / instrument
5-8 GB / day

PromethION

High Throughput Desktop Sequencer
$75k / instrument
>>1000GB / day

Nanopore Performance at CSHL

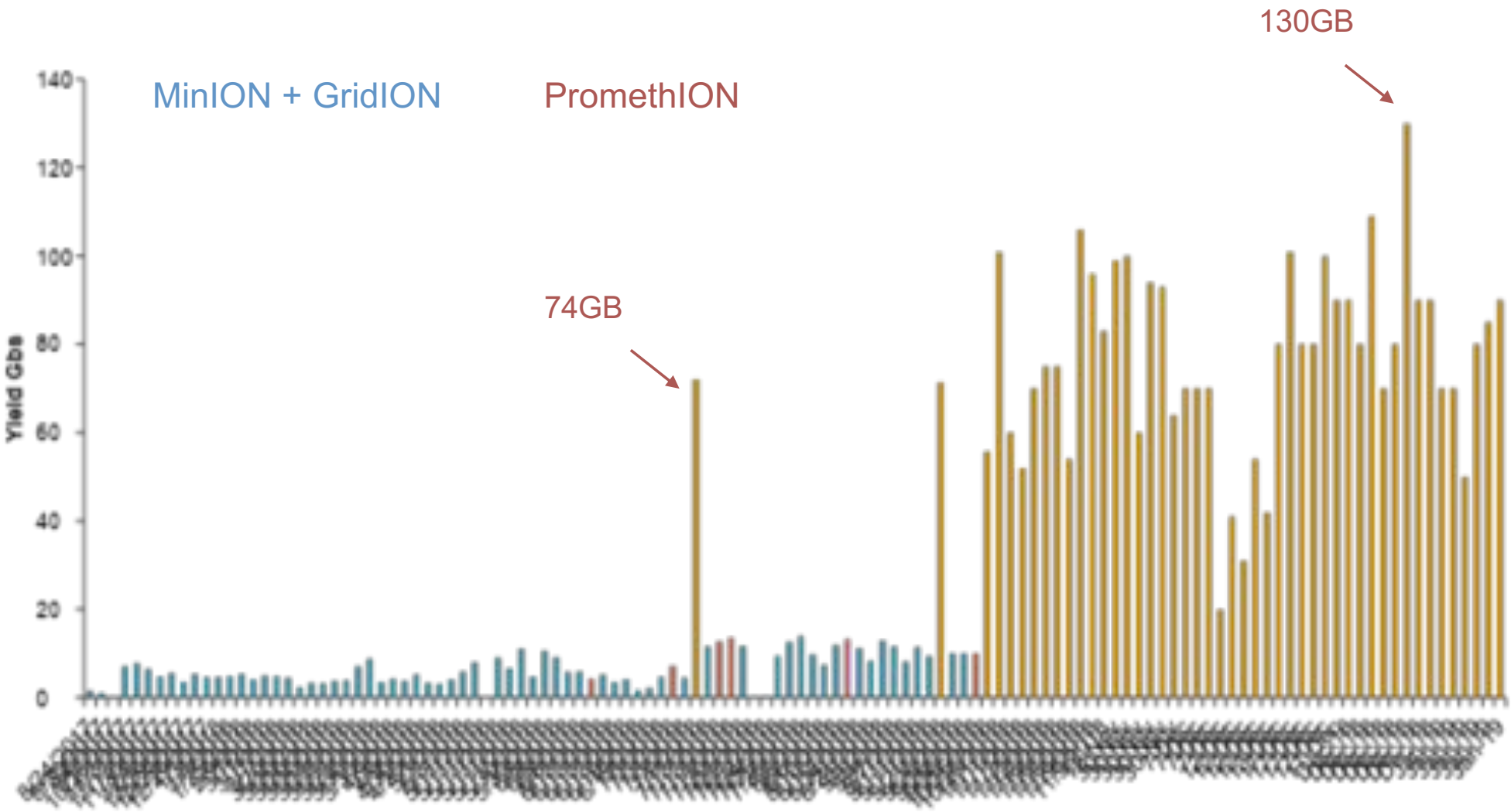
Sara Goodwin



Part of collaboration between JHU and CSHL to sequence 100 tomato genomes in 100 days

Nanopore Performance at CSHL

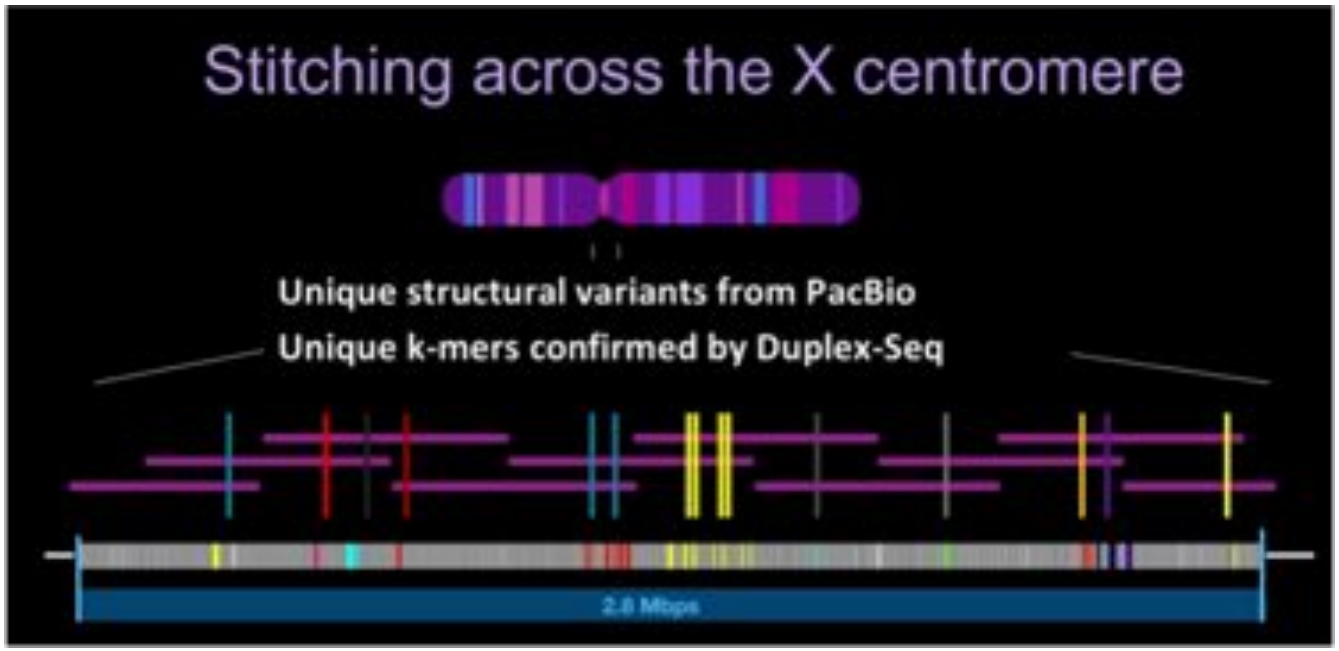
Sara Goodwin



Telomere-to-Telomere (T2T)

T2T consortium aims to finish the human genome

- Many gaps still exist, particularly around centromeres
- Uses ultra-long ONT reads, in addition to PacBio HiFi and other tech
- Finished chrX, close to finishing chr8, only 22 more to go



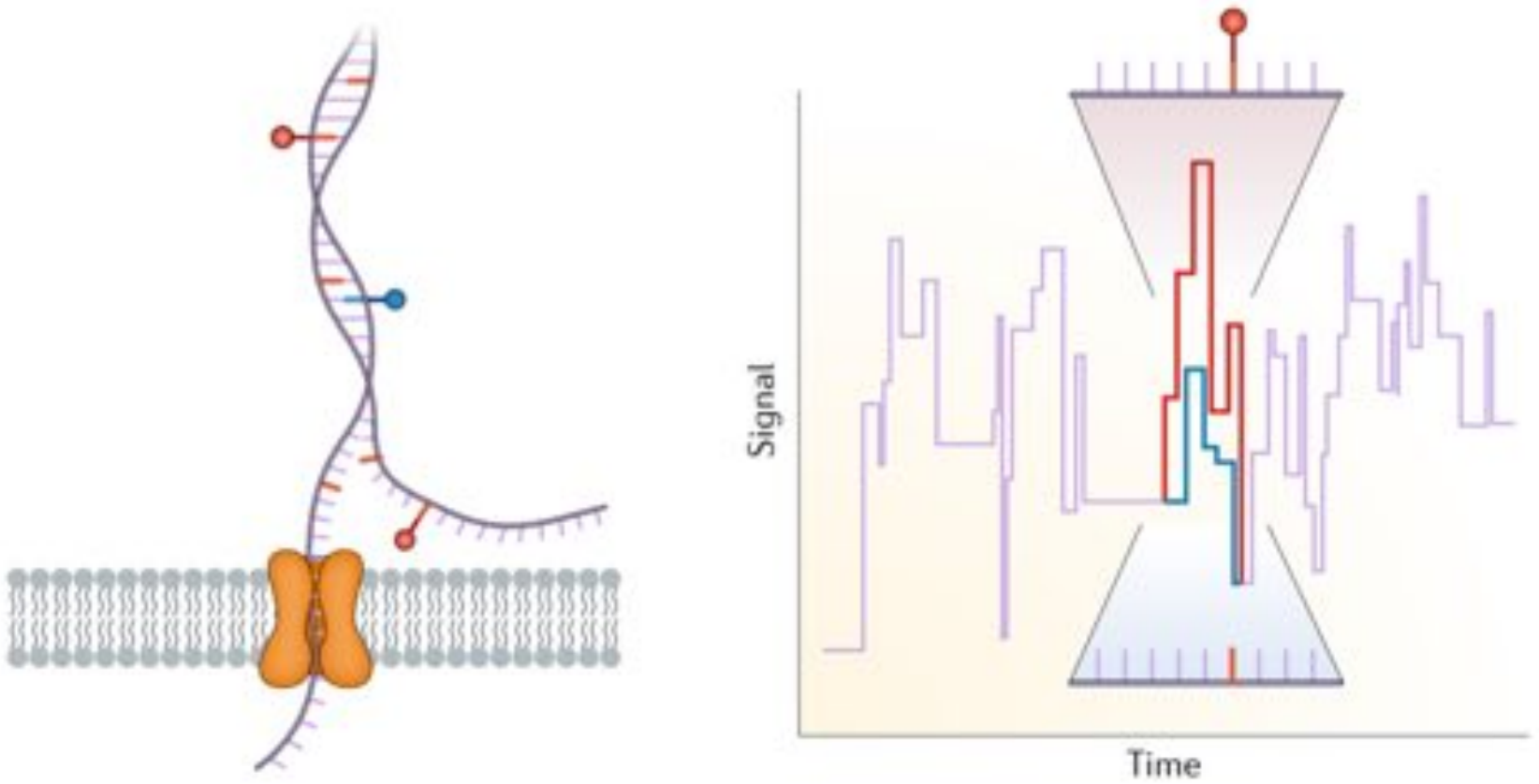
Karen Miga, London Calling 2019

Telomere-to-telomere assembly of a complete human X chromosome
Miga et al. (2019) *BioRxiv.* https://doi.org/10.1101/735928

DNA Modification Detection

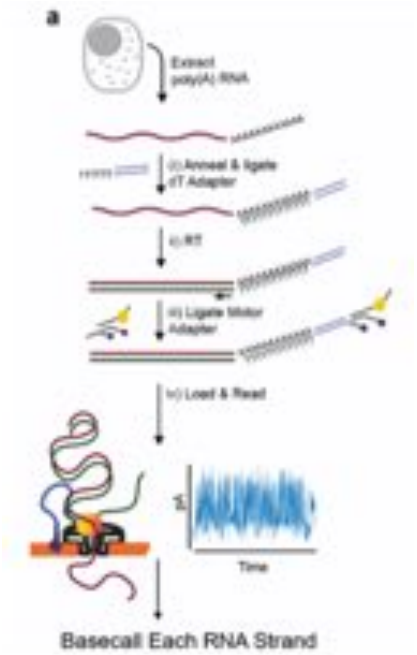
ONT can detect methylation from raw signal

- Or any other modification that changes ionic current



Piercing the dark matter: bioinformatics of long-range sequencing and mapping
Sedlazeck et al. (2018) *Nature Reviews Genetics.* 19:329

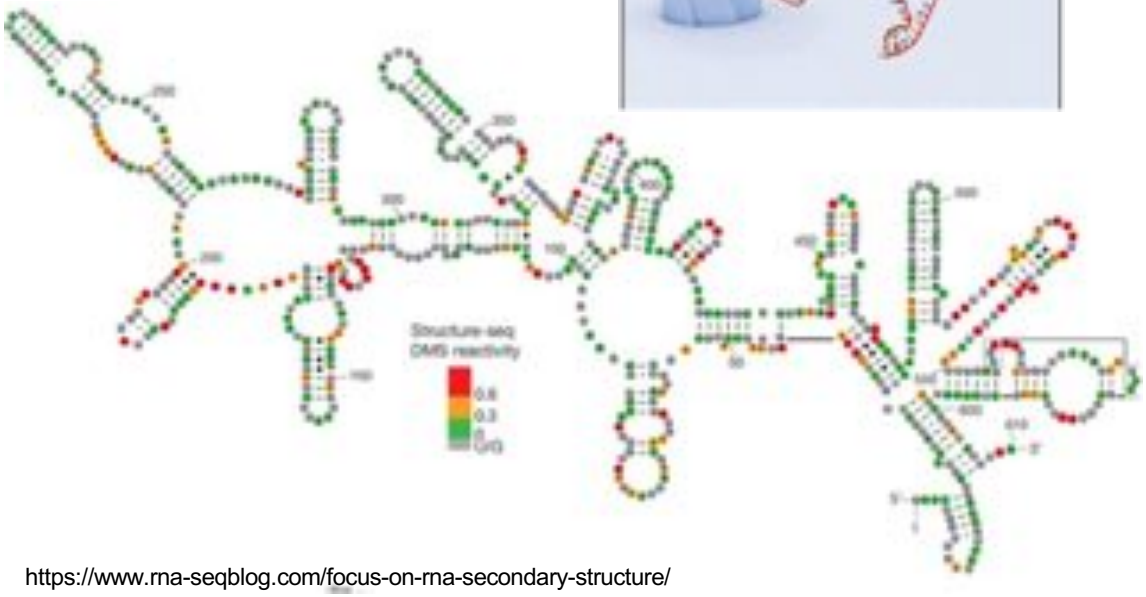
Nanopore Direct RNA-seq

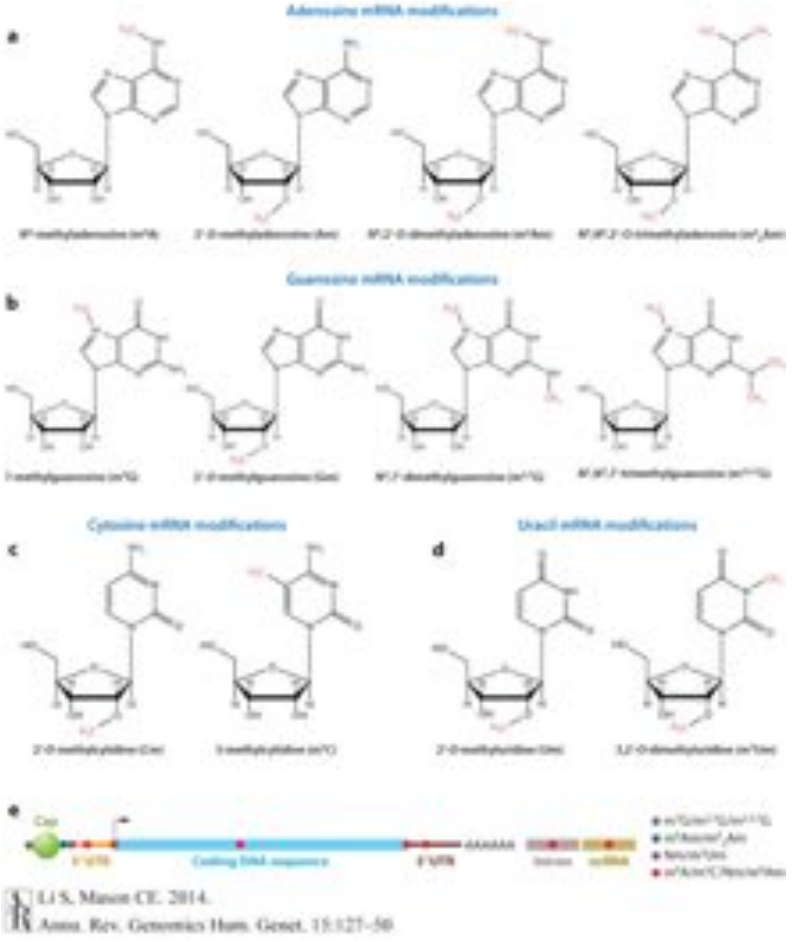


ONT can sequence RNA molecules directly

cDNA sequencing erases modifications and structure

Direct RNA-seq has potential to read both



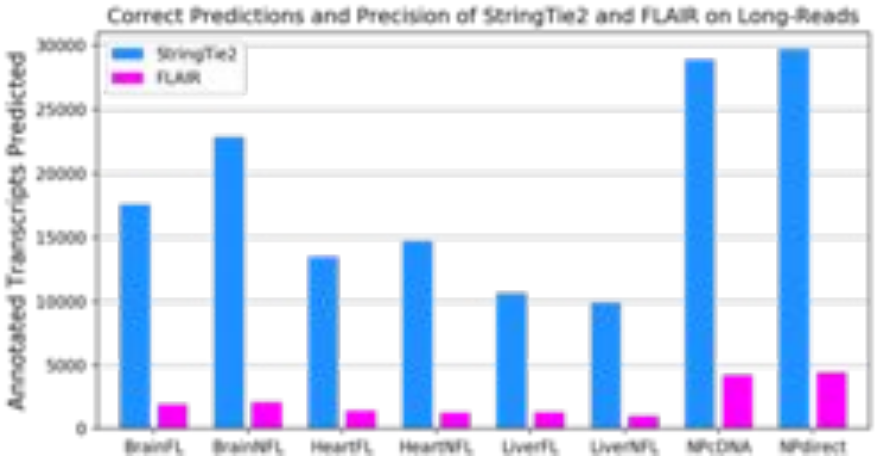


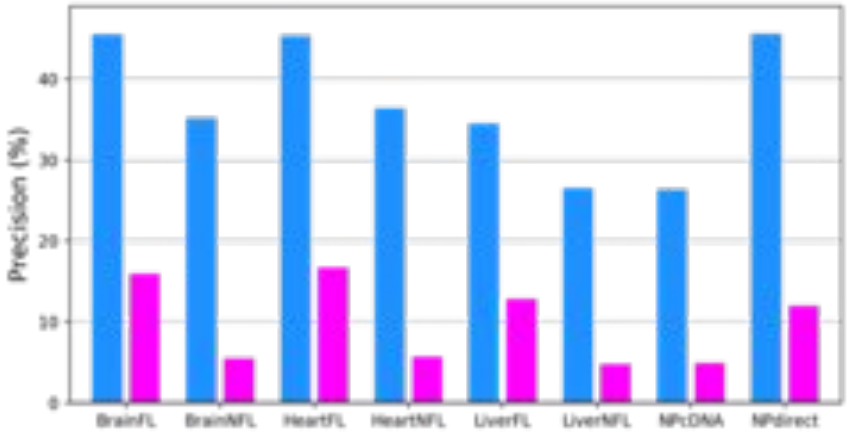
https://www.rna-seqblog.com/focus-on-rna-secondary-structure/

Long-Read RNA-seq Assembly

Both direct and cDNA ONT RNA-seq produce some transcript fragments

- I helped develop StringTie2, which applies short-read transcriptome assembly methods to long-reads





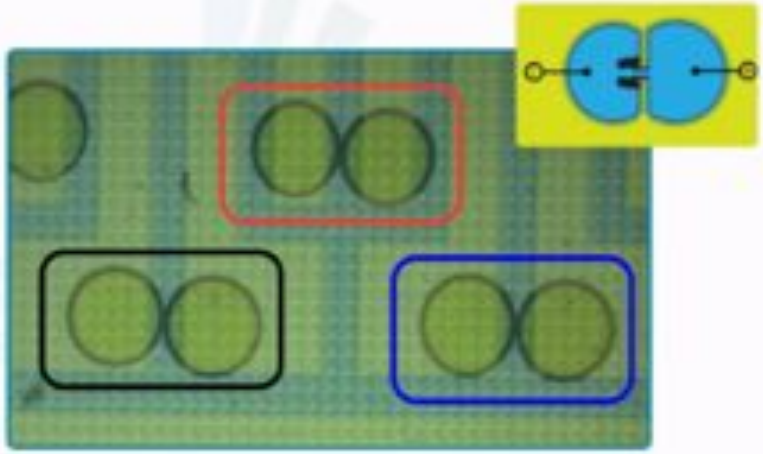
Transcriptome assembly from long-read RNA-seq alignments with StringTie2

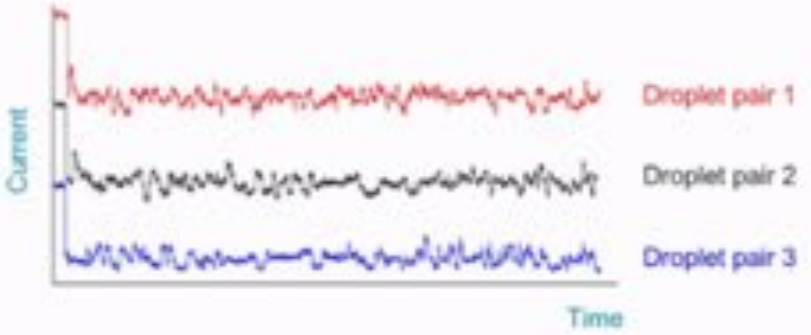
Sam Kovaka, Aleksey V. Zimin, Geo M. Pertea, Roham Razaghi, Steven L. Salzberg & Mihaela Pertea

Genome Biology **20**, Article number: 278 (2019) | Cite this article

VolTRAX - Library Prep (+ sequencing?)

- 3.6 kb DNA, standard ligation library preparation
- Sample + pores in one droplet
- Pores inserted, then library sequenced
- Droplet size ~ 10nL, could be 4.5 nL with current chip





Proof of concept array demonstrated

- No crosstalk - data taken directly from cartridge
- Wide range of experiments possible
- Will include MinKNOW control and feedback
- Data being collected for model training

NANOPORE

Extremely Portable Sequencing!

Kate Rubins sequencing DNA on the ISS

Ebola Surveillance

LETTER

doi:10.1038/nature16996

Real-time, portable genome sequencing for Ebola surveillance

Joshua Quick[1]*, Nicholas J. Loman[1]*, Sophie Duraffour[2,3]*, Jared T. Simpson[4,5]*, Ettore Severi[6]*, Lauren Cowley[7]*, Joseph Akoi Bore[2], Raymond Koundouno[2], Gytis Dudas[8], Amy Mikhail[7], Nobila Ouédraogo[9], Babak Afrough[2,10], Amadou Bah[2,11], Jonathan H. J. Baum[2,3], Beate Becker-Ziaja[2,3], Jan Peter Boettcher[2,12], Mar Cabeza-Cabrerizo[2,3], Álvaro Camino-Sánchez[7], Lisa L. Carter[2,13], Juliane Doerrbecker[2,3], Theresa Enkirch[2,14], Isabel García-Dorival[2,15], Nicole Hetzelt[2,12], Julia Hinzmann[2,12], Tobias Holm[2,3], Liana Eleni Kafetzopoulou[2,16], Michel Koropogui[2,17], Abigael Kosgey[2,18], Eeva Kuisma[2,10], Christopher H. Logue[2,10], Antonio Mazzarelli[2,19], Sarah Meisel[2,3], Marc Mertens[2,20], Janine Michel[2,12], Didier Ngabo[2,10], Katja Nitzsche[2,3], Elisa Pallasch[2,3], Livia Victoria Patrono[2,3], Jasmine Portmann[2,21], Johanna Gabriella Repits[2,22], Natasha Y. Rickett[2,15,23], Andreas Sachse[2,12], Katrin Singethan[2,24], Inês Vitoriano[2,10], Rahel L. Yemanaberhan[2,3], Elsa G. Zekeng[2,15,23], Trina Racine[25], Alexander Bello[25], Amadou Alpha Sall[26], Ousmane Faye[26], Oumar Faye[26], N'Faly Magassouba[27], Cecelia V. Williams[28,29], Victoria Amburgey[28,29], Linda Winona[28,29], Emily Davis[29,30], Jon Gerlach[29,30], Frank Washington[29,30], Vanessa Monteil[31], Marine Jourdain[31], Marion Bererd[31], Alimou Camara[31], Hermann Somlare[31], Abdoulaye Camara[31], Marianne Gerard[31], Guillaume Bado[31], Bernard Baillet[31], Déborah Delaune[32,33], Koumpingnin Yacouba Nebie[34], Abdoulaye Diarra[34], Yacouba Savane[34], Raymond Bernard Pallawo[34], Giovanna Jaramillo Gutierrez[35], Natacha Milhano[6,36], Isabelle Roger[34], Christopher J. Williams[6,37], Facinet Yattara[27], Kuiama Lewandowski[10], James Taylor[38], Phillip Rachwal[38], Daniel J. Turner[39], Georgios Pollakis[15,23], Julian A. Hiscox[15,23], David A. Matthews[40], Matthew K. O'Shea[41], Andrew McD. Johnston[41], Duncan Wilson[42], Emma Hutley[42], Erasmus Smit[43], Antonino Di Caro[2,19], Roman Wölfel[2,44], Kilian Stoecker[2,44], Erna Fleischmann[2,44], Martin Gabriel[2,3], Simon A. Weller[38], Lamine Koivogui[45], Boubacar Diallo[34], Sakoba Keita[17], Andrew Rambaut[8,46,47], Pierre Formenty[34], Stephan Günther[2,3] & Miles W. Carroll[2,10,48,49]

Ebola Surveillance

LETTER

doi:10.1038/nature16996

Real-time, portable genome sequencing for Ebola surveillance

Joshua Quick[1]*, Nicholas J. Loman[1]*, Sophie Duraffour[2,3]*, Jared T. Simpson[4,5]*, Ettore Se
Joseph Akoi Bore[2], Raymond Koundouno[2], Gytis Dudas[6], Amy Mikhail[7], Nobila Ouédraog
Amadou Bah[2,11], Jonathan H. J. Baum[2,3], Beate Becker-Ziaja[2,3], Jan Peter Boettcher[2,12], Mar
Álvaro Camino-Sánchez[2], Lisa L. Carter[2,13], Juliane Doerrbecker[2,3], Theresa Enkirch[2,14], Is
Nicole Hetzelt[2,12], Julia Hinzmann[2,12], Tobias Holm[2,3], Liana Eleni Kafetzopoulou[2,16], Miche
Eeva Kuisma[2,10], Christopher H. Logue[2,10], Antonio Mazzarelli[2,19], Sarah Meisel[2,3], Marc M
Didier Ngabo[2,10], Katja Nitzsche[2,3], Elisa Pallasch[2,3], Livia Victoria Patrono[2,3], Jasmine Port
Natasha Y. Rickett[2,15,23], Andreas Sachse[2,12], Katrin Singethan[2,24], Inês Vitoriano[2,10], Rahel
Elsa G. Zekeng[2,15,23], Trina Racine[25], Alexander Bello[25], Amadou Alpha Sall[26], Ousmane Fa
N'Faly Magassouba[27], Cecelia V. Williams[28,29], Victoria Amburgey[28,29], Linda Winona[28,29], Er
Frank Washington[29,30], Vanessa Monteil[31], Marine Jourdain[31], Marion Berard[31], Alimou Cam
Abdoulaye Camara[31], Marianne Gerard[31], Guillaume Bado[31], Bernard Baillet[31], Déborah Dela
Abdoulaye Diarra[34], Yacouba Savane[34], Raymond Bernard Pallawo[34], Giovanna Jaramillo Gu
Isabelle Roger[34], Christopher J. Williams[6,37], Facinet Yattara[27], Kuiama Lewandowski[10], Jam
Daniel J. Turner[39], Georgios Pollakis[15,23], Julian A. Hiscox[15,23], David A. Matthews[40], Matthe
Andrew McD. Johnston[41], Duncan Wilson[42], Emma Hutley[42], Erasmus Smit[43], Antonino Di
Kilian Stoecker[2,44], Erna Fleischmann[2,44], Martin Gabriel[2,3], Simon A. Weller[38], Lamine Koi
Sakoba Keita[17], Andrew Rambaut[8,46,47], Pierre Formenty[34], Stephan Günther[2,3] & Miles W. C

Figure 1 | Deployment of the portable genome surveillance system in Guinea. a, We were able to pack all instruments, reagents and disposable consumables within aircraft baggage. **b**, We initially established the genomic surveillance laboratory in Donka Hospital, Conakry, Guinea. **c**, Later we moved the laboratory to a dedicated sequencing laboratory in Coyah prefecture. **d**, Within this laboratory we separated the sequencing instruments (on the left) from the PCR bench (to the right). An uninterruptable power supply can be seen in the middle that provides power to the thermocycler. (Photographs taken by J.Q. and S.D.)

Ebola Surveillance

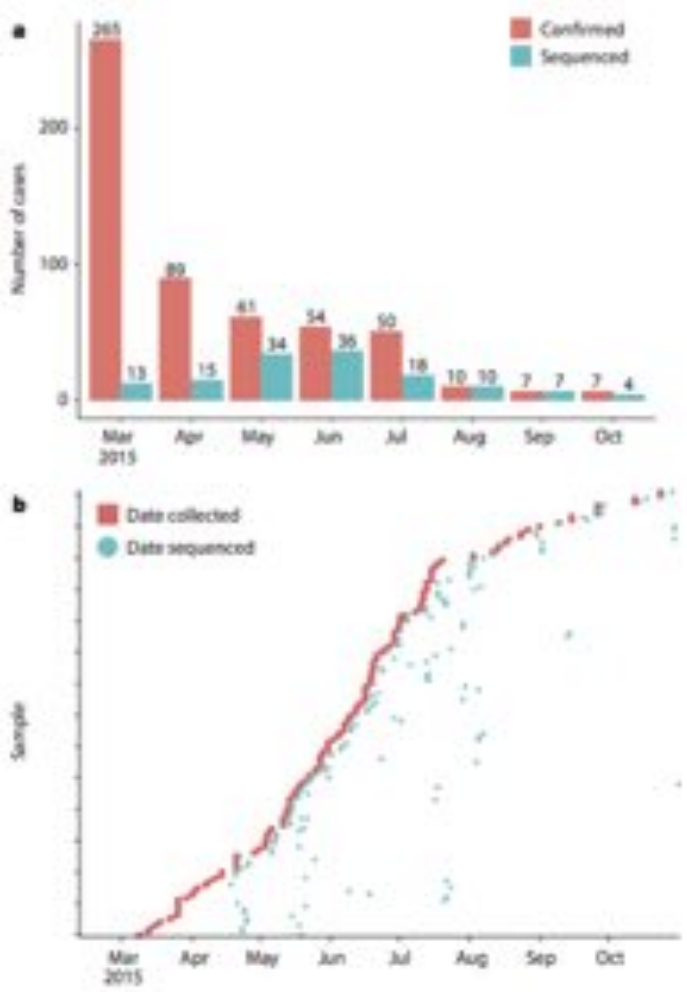


Figure 2 | Real-time genomics surveillance in context of the Guinea Ebola virus disease epidemic. a. Here we show the number of reported cases of Ebola virus disease in Guinea (red) in relation to the number of EBOV new patient samples ($n = 137$, in blue) generated during this study. **b.** For each of the 142 sequenced samples, we show the relationship between sample collection date (red) and the date of sequencing (blue). Twenty-eight samples were sequenced within three days of the sample being taken, and sixty-eight samples within a week. Larger gaps represent retrospective sequencing of cases to provide additional epidemiological context.

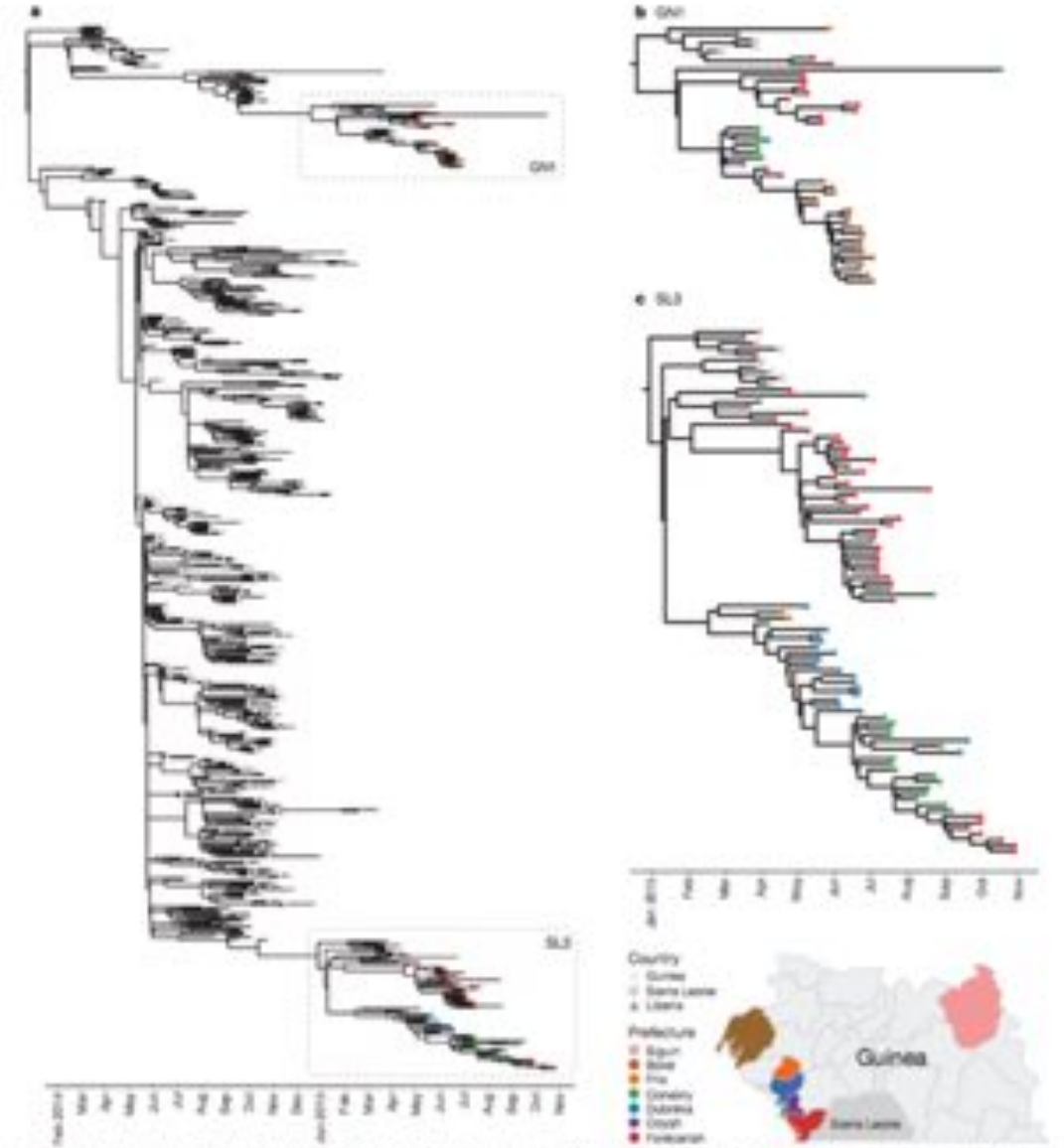


Figure 3 | Evolution of EBOV over the course of the Ebola virus disease epidemic. a. Time-scaled phylogeny of 603 published sequences with 125 high quality sequences from this study. The shape of nodes on the tree demonstrates country of origin. Our results show Guinean samples (coloured circles) belong to two previously identified lineages, GN1 and SL3. **b.** GN1 is deeply branching with early epidemic samples. **c.** SL3 is related to cases identified in Sierra Leone. Samples are frequently clustered by geography (indicated by colour of circle) and this provides information as to origins of new introductions, such as in the Boké epidemic in May 2015. Map figure adapted from SimpleMaps website (http://simplemaps.com/resources/svg-gin).

COVID-19 (😬) Surveillance

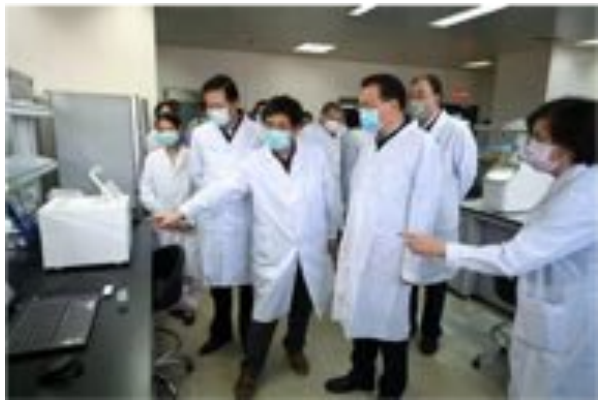
200 Oxford Nanopore sequencers have left UK for China, to support rapid, near-sample coronavirus sequencing for outbreak surveillance

ARTIC Network @NetworkArtic · Feb 5

First #nCoV2019 genome sequenced using the ARTIC @nanopore protocol and primer scheme by @Scalene is up on #GISAID. Sequenced by the Hangzhou CDC with assistance from @nanopore. artic.network/ncov-2019

Sequencing technology:	Nanopore MinION
Assembly method:	artic-ncov2019 and Genalice V1.1.0
Coverage:	> 2000x (average)
Originating lab:	Hangzhou Center for Disease Control and Prevention
Address:	568 Mingzh Road, Jianggan District 310021 Hangzhou China
Sample ID given by the sample provider:	BetaCoV/Hangzhou/HZCDC/0001/2020
Submitting lab:	Hangzhou Center for Disease Control and Prevention
Address:	568 Mingzh Road, Jianggan District 310021 Hangzhou China
Sample ID given by the submitting laboratory:	BetaCoV/Hangzhou/HZCDC/0001/2020
Authors:	Jun Li, Haoxiu Wang, Huai Yu, Linglong Miao, Kefan Yu, Zhou Sun, Qingxin Kong, Xin Qian, Shuchang Chen, Kuchu Wang

3 42 105



james hadfield
@hamesjadfield

When we started developing RAMPART the idea that you could analyse seq data in real time seemed rather aspirational. Today: a still-sequencing #SARSCoV2 @nanopore genome from Brazil already up on @nextstrain. Amazing to get to work with @jaquegj @CaddeProject @NetworkArtic et al

https://nanoporetech.com/about-us/news/novel-coronavirus-ncov-2019covid-19-information-and-updates

Less Throughput

Flongle

- An adapter for MinION for smaller tests or experiments
- Single-use, on-demand, cost efficient sequencing
- Suitable for quality checks, amplicons, smaller genomes, targeted regions, or those interested in diagnostics/other tests
- MinIT available to support IT/software needs

SmidgION (coming "soon")

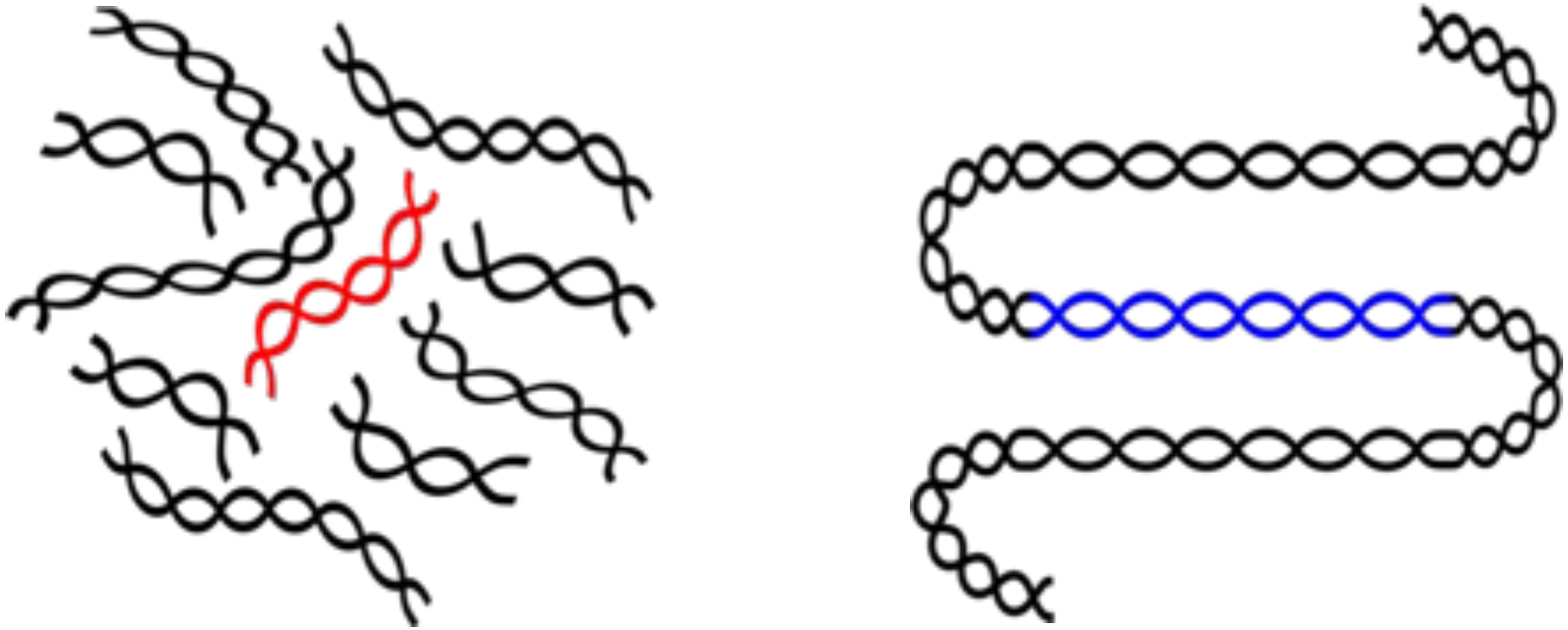
- Designed to be our smallest sequencing device so far
- Same nanopore sensing technology as MinION and PromethION
- Designed for use with a smartphone in any location

https://nanoporetech.com/products

Targeted Sequencing

Often you're only interested in certain sequences

- Can be challenging to reach sufficient coverage with low yield
- For example: **pathogen DNA** enrichment or **targeting genes**

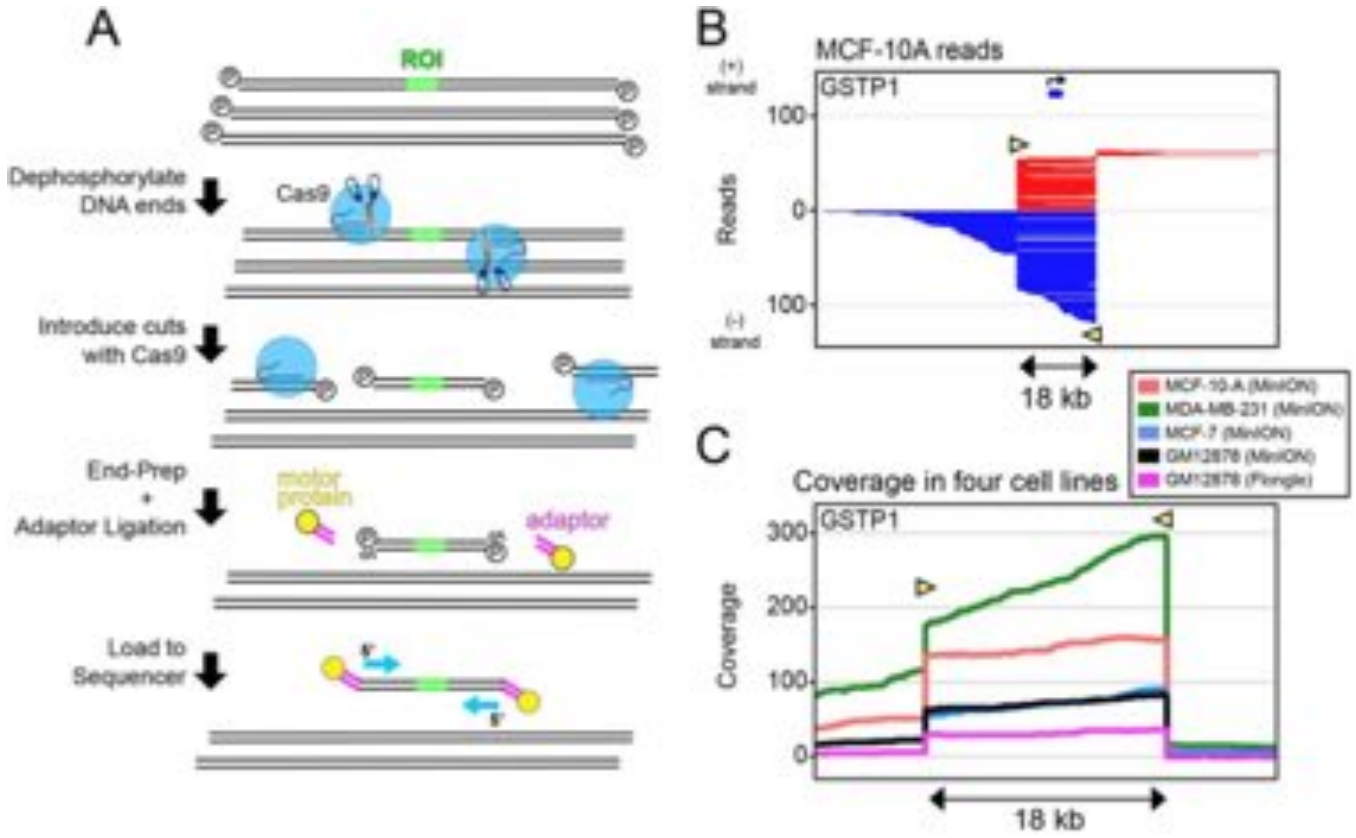


PCR doesn't work work well for ONT sequencing

- Limits read length and erases epigenetc modifications

CIRSPR/Cas9 Enrichment

Uses Cas9 to cut DNA around target region, then binds adapters to phosphorylated ends

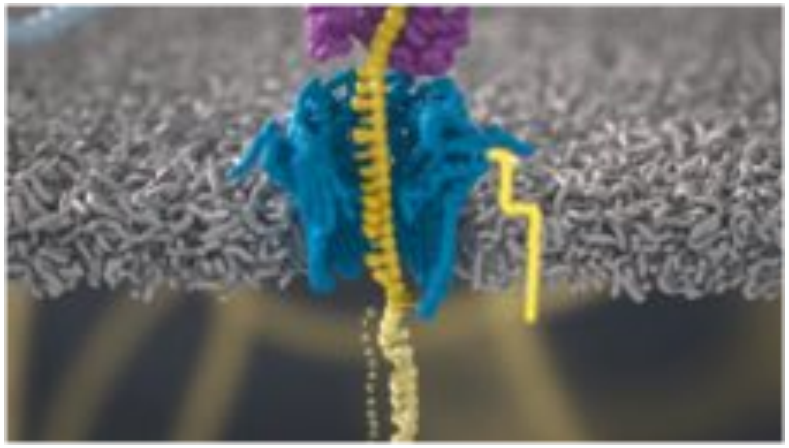


Targeted Nanopore Sequencing with Cas9 for studies of methylation, structural variants, and mutations

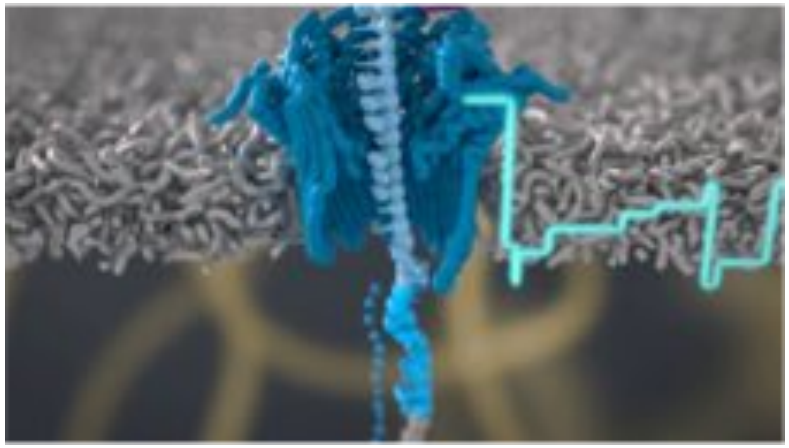
Timothy Gilpatrick, Isac Lee, James E. Graham, Etienne Raimondeau, Rebecca Bowen, Andrew Heron, Fritz J Sedlazeck, Winston Timp

ReadUntil Sequencing

ONT devices can selectively eject reads in real-time



Nanopore Sequencing



Nanopore Sequencing with ReadUntil

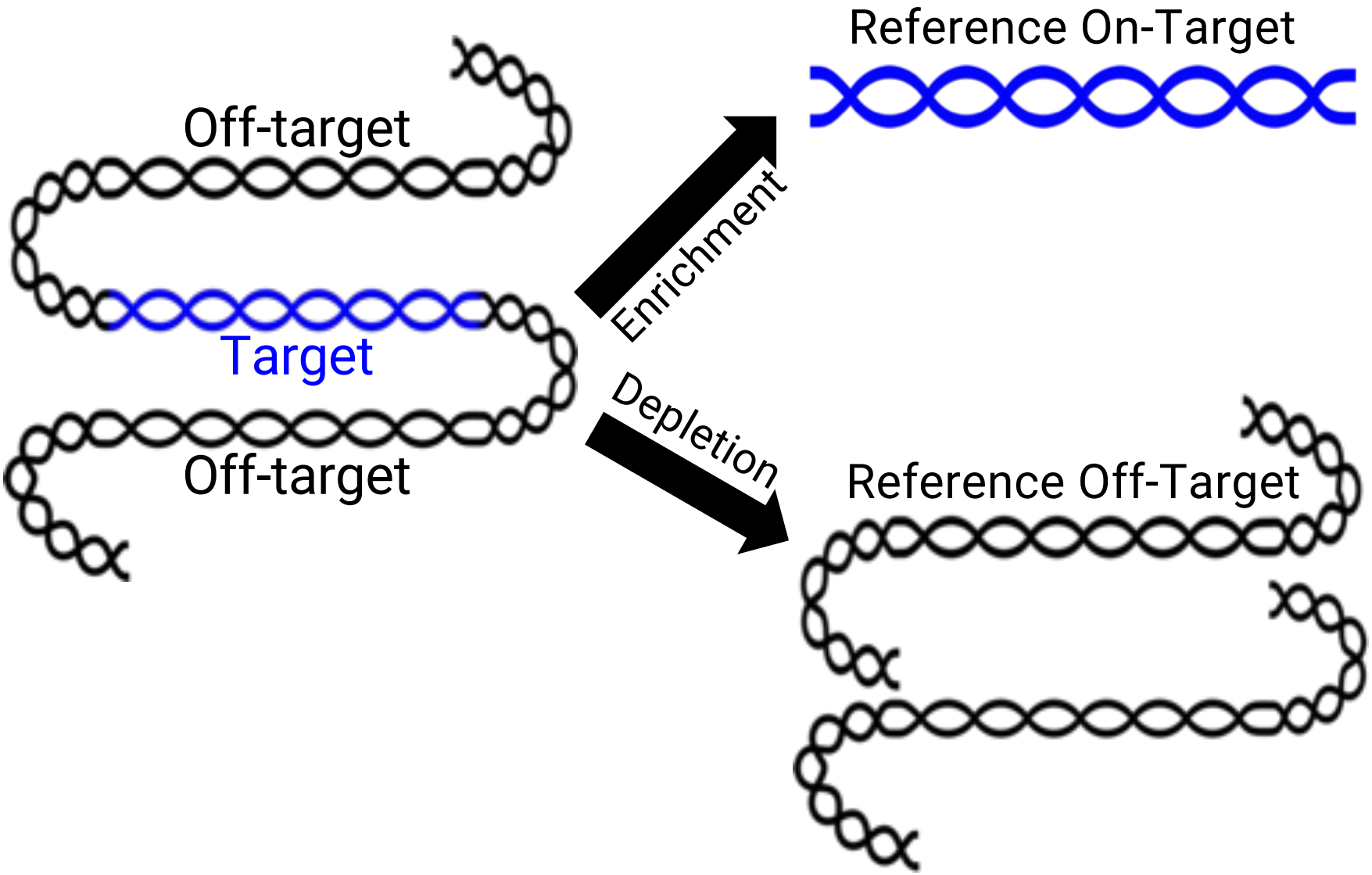
Enables targeted sequencing without addition sample prep

- Requires rapid real-time read identification

MinION has up to 512 active channels, each reading 450 bp/sec

- Actual number of active channels is variable

ReadUntil Sequencing



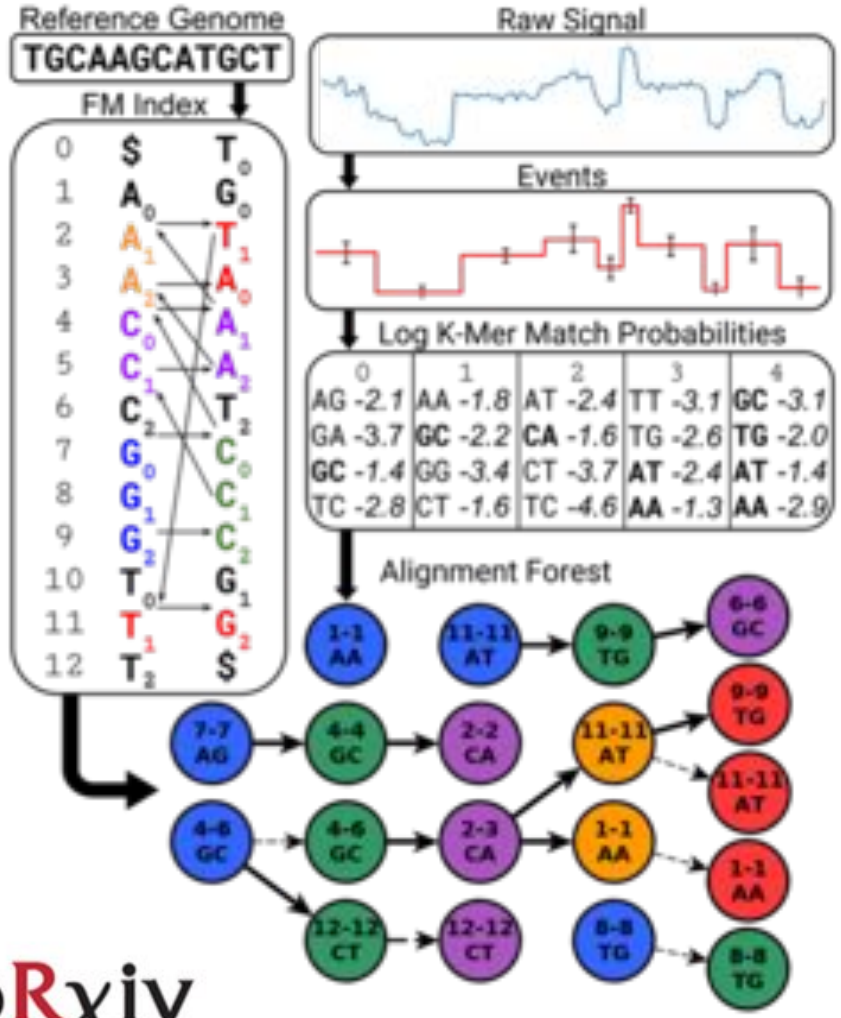
Utility for Nanopore Current ALignment to Large Expanses of DNA

UNCALLED

A novel streaming algorithm which maps raw nanopore signal as it is being sequenced

- Can map reads from all active MinION channels to reference tens of megabases in size
- Uses the mapping results to make ReadUntil decisions

github.com/skovaka/UNCALLED



bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

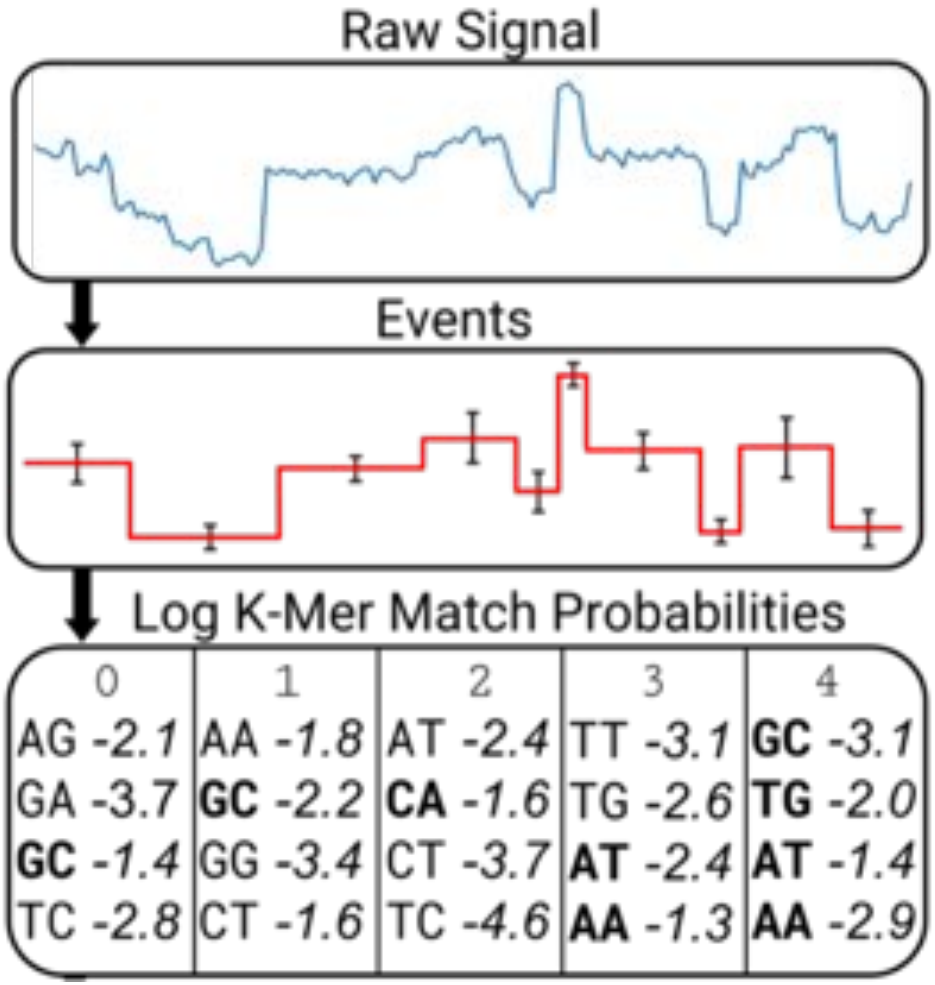
Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED

Sam Kovaka, Yunfan Fan, Bohan Ni, Winston Timp, Michael C Schatz

doi: https://doi.org/10.1101/2020.02.03.931923

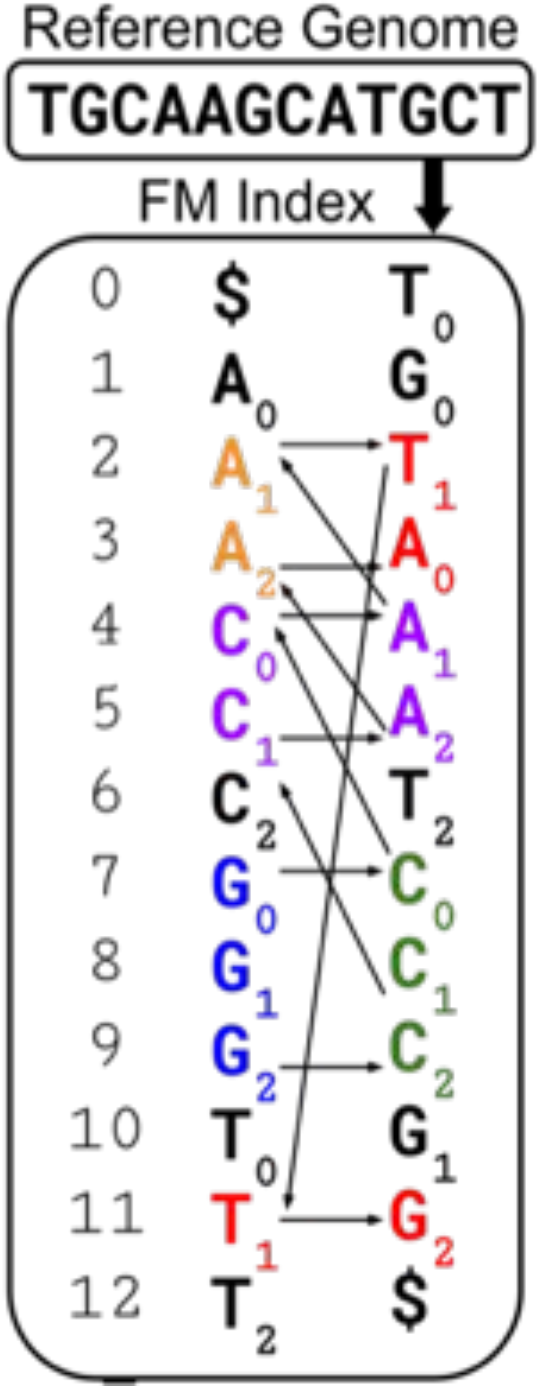
UNCALLED Signal Processing

- Stretches of similar signal are collapsed into **events**
  - Averages out noise and reduces amount of signal to process
  - Ideally each event represents a single k-mer, but many errors occur (50% stays, 1% skips)
- Probability of events matching each k-mer is then computed
  - Expected current for each k-mer modeled by normal distribution
  - ONT releases 6-mer models (I use 5-mers)

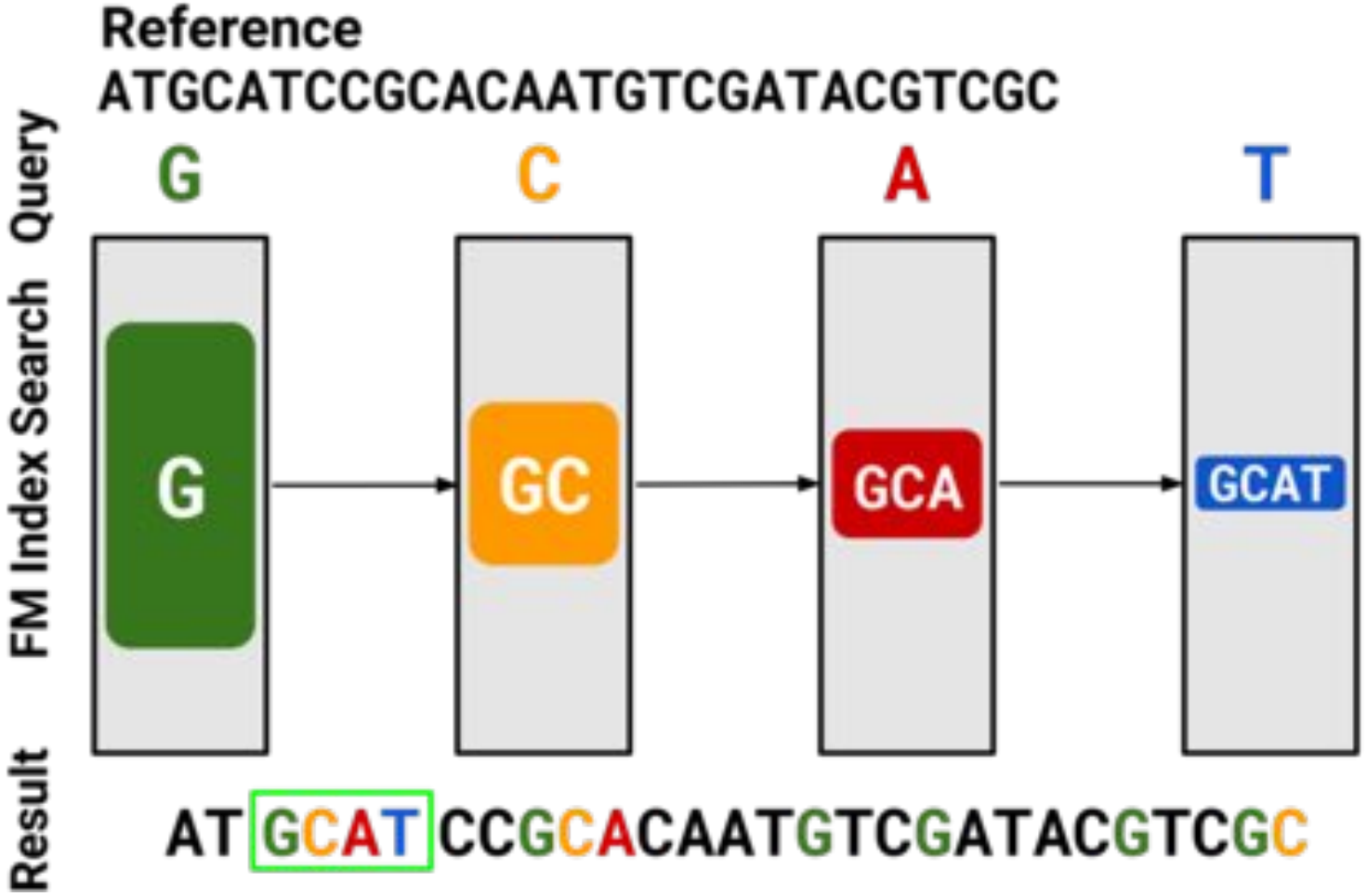


FM Index

- Used by many aligners such as BWA, Bowtie, and HISAT
- UNCALLED uses BWA's FM index
  - Interchangeable - started with my own implementation

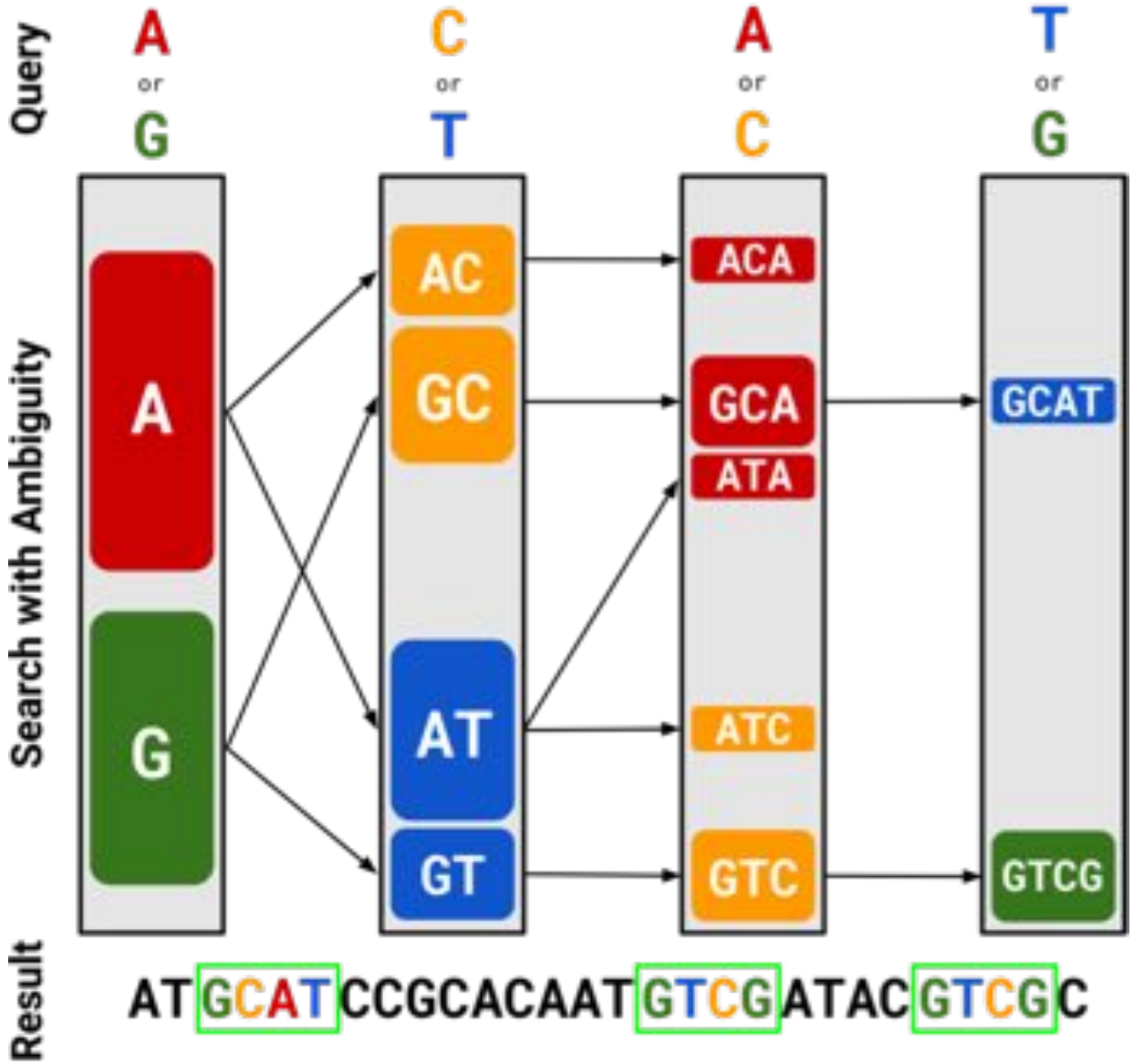


FM Index Search

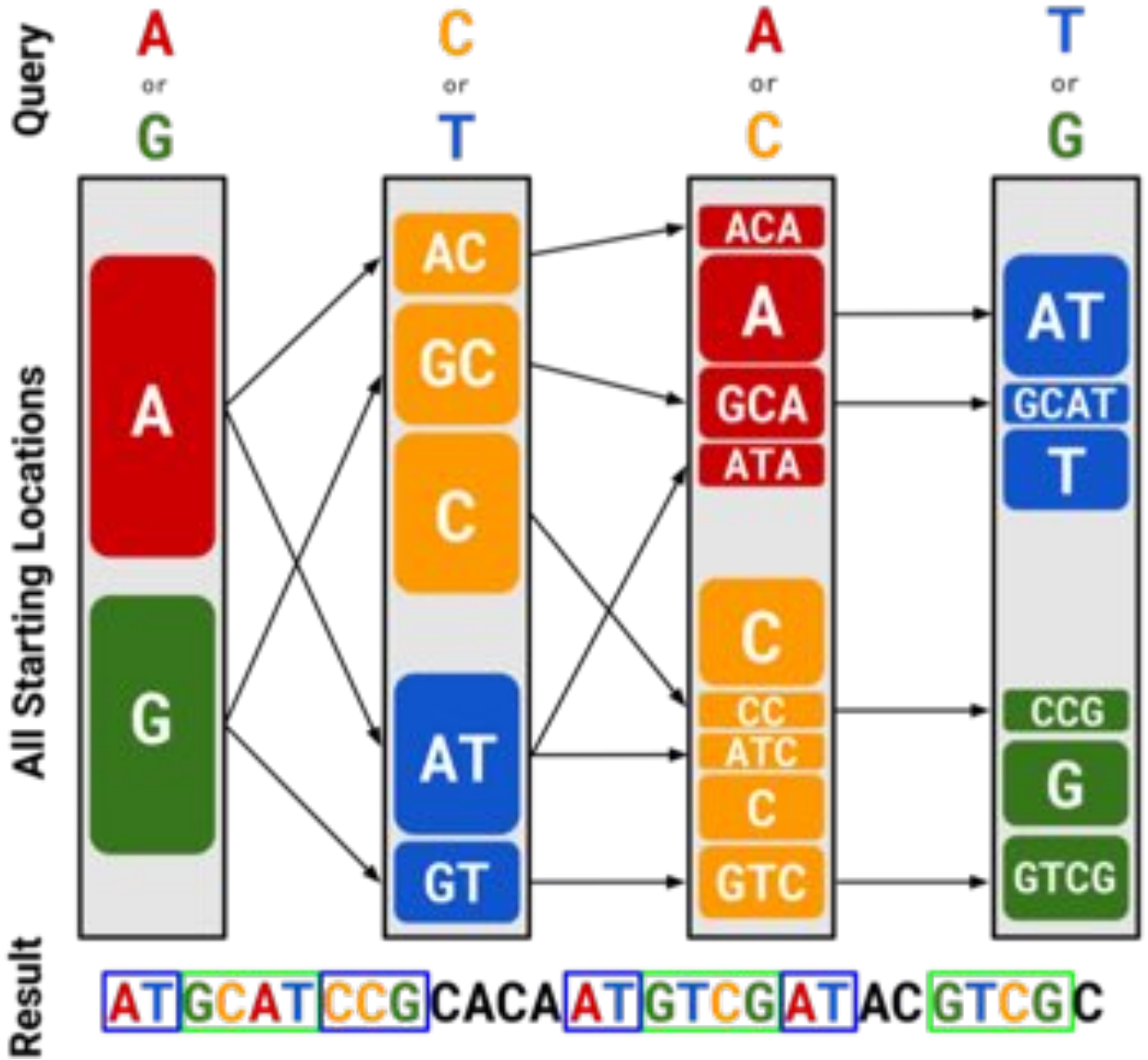


Size of range = number of possible alignments

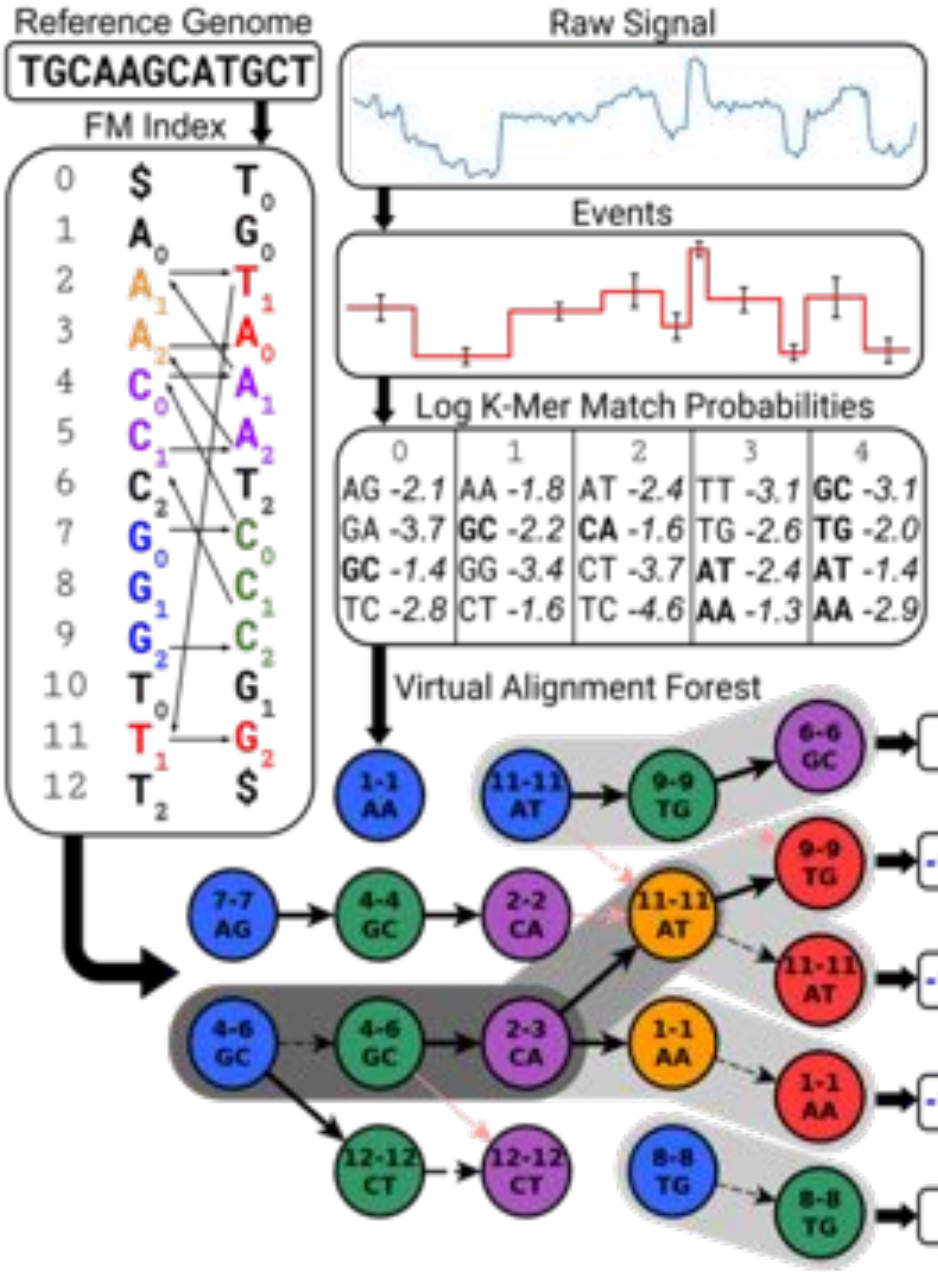
FM Index Search w/ Ambiguity



FM Index Search w/ Ambiguity



UNCALLED Algorithm



Conceptually all possible paths through the FM index form a forest of trees

Traversing trees is cache-inefficient. Instead every path is stored in a fixed-length buffer

- Stores cumulative log probabilities/event types
- Whole buffer must be copied when a path splits

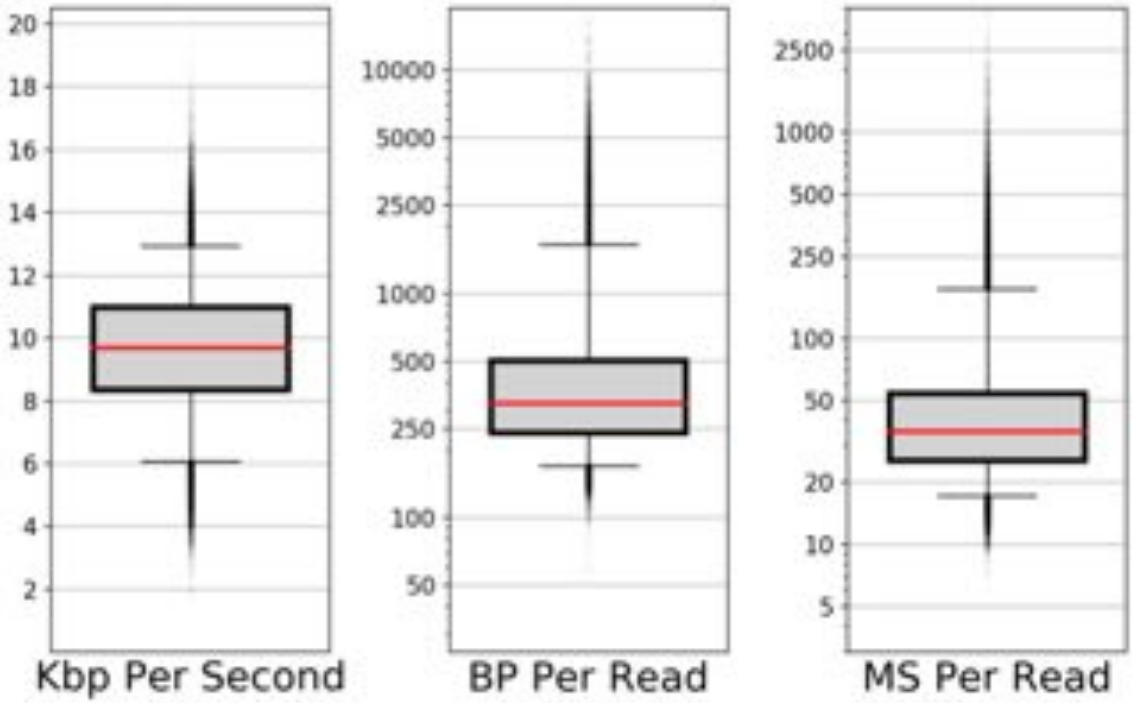
Once a buffer is full:

- Report a seed alignment
- Erase oldest event, making room for next
- Buffers keep rolling across read until no possible extension exists

Mapping Reads to *E. coli*

Mapped 100K E. coli reads to the E. coli reference genome

- Running on a single core of a 3.0 GHz Intel Xeon Gold 6136
- Estimated accuracy using minimap2 as ground truth

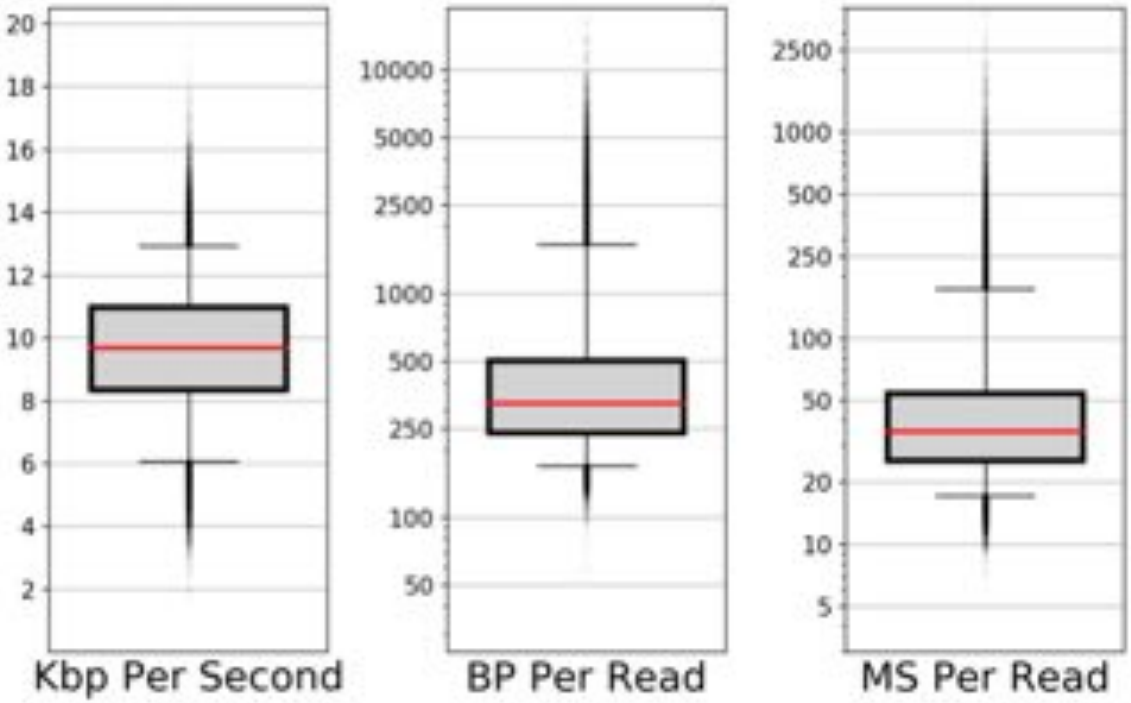


Mapping Reads to *E. coli*

Mapped 100K E. coli reads to the E. coli reference genome

- Running on a single core of a 3.0 GHz Intel Xeon Gold 6136
- Estimated accuracy using minimap2 as ground truth

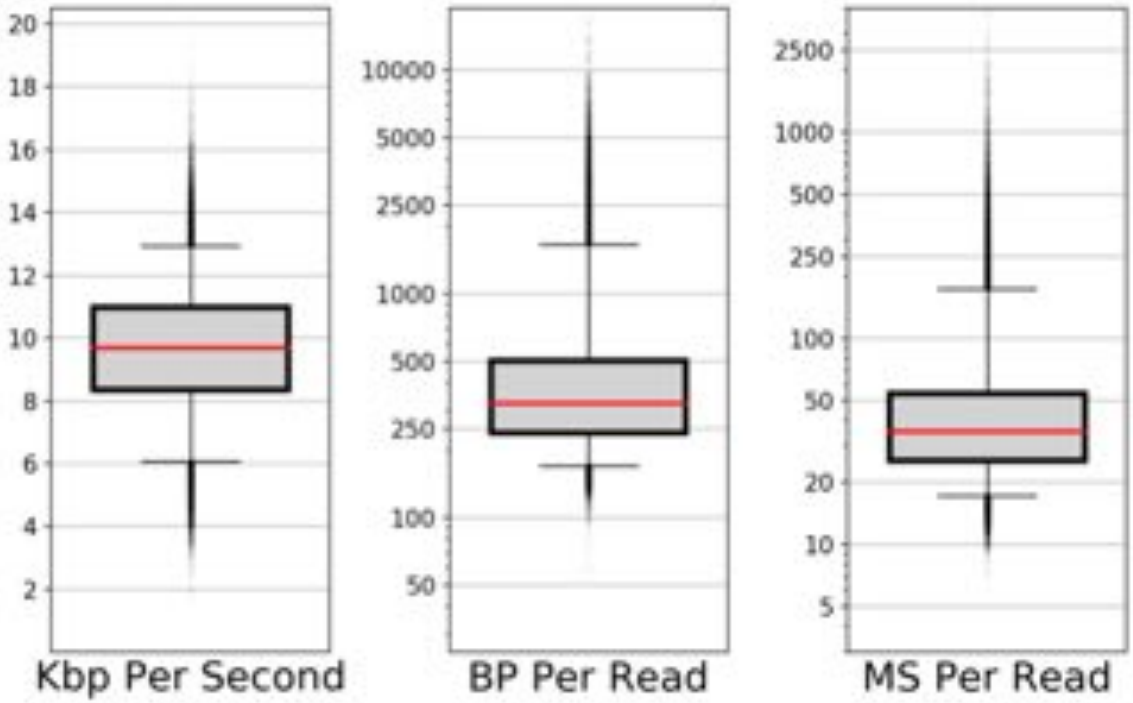
	P	N
T	85.70%	7.99%
F	1.03%	5.28%



Mapping Reads to *E. coli*

Mapped 100K E. coli reads to the E. coli reference genome

- Running on a single core of a 3.0 GHz Intel Xeon Gold 6136
- Estimated accuracy using minimap2 as ground truth



	P	N
T	85.70%	7.99%
F	1.03%	5.28%

Mean Guppy Q Scores:

	P	N
T	11.16	4.71
F	6.09	7.26

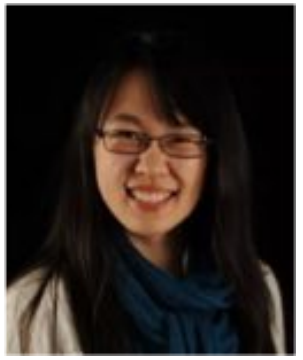
75.4% of “FPs” were not aligned by Minimap2

92.4% of FPs that Minimap2 aligned are explained by repeats

UNCALLED on Zymo

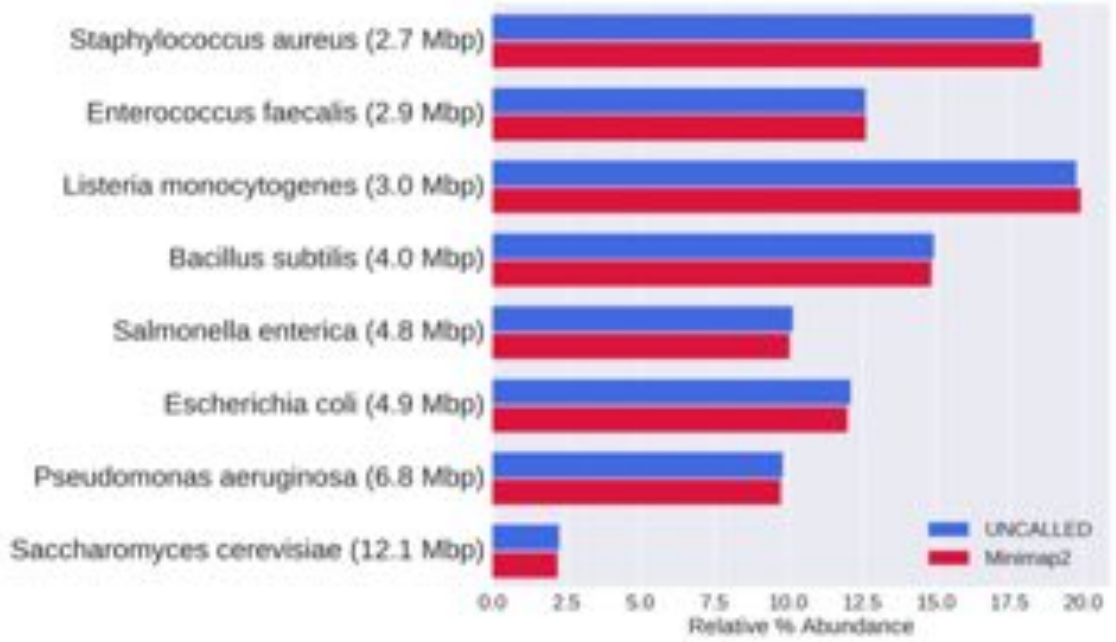
Tested UNCALLED on the Zymo high molecular weight mock microbial community standard

- Contains seven bacteria and one fungus
- Mapped reads to the full 41Mbp reference at a rate of ~6,000 bp/thread/sec
- Match minimap2 alignments with 96% accuracy



Sequencing by Yunfan Fan

Read lengths:
Median: 12.2Kbp
Mean: 15.9Kbp
N50: 24.7Kbp



Bacterial Depletion on Zymo

- Mapping to a reference of the all bacteria
- Ejecting reads if they map within the first 4,500bp (10 sec), enriching fungal sequence
- Sequenced same sample on flowcell of similar quality directly next to ReadUntil run as control
- Running with 48 threads on a 3.0 GHz Intel Xeon Gold 6136 (probably overkill)

Bacterial Depletion on Zymo

- Mapping to a reference of the all bacteria
- Ejecting reads if they map within the first 4,500bp (10 sec), enriching fungal sequence
- Sequenced same sample on flowcell of similar quality directly next to ReadUntil run as control
- Running with 48 threads on a 3.0 GHz Intel Xeon Gold 6136 (probably overkill)

Read decision accuracy:

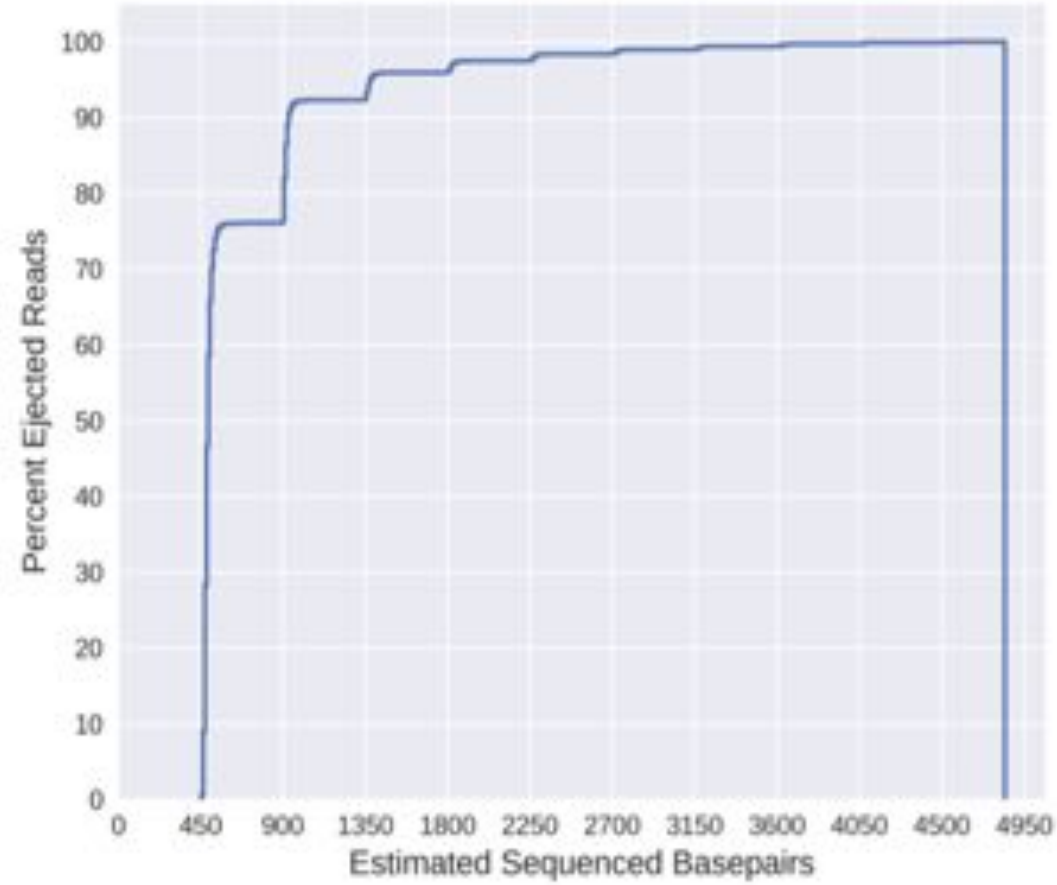
Correctly eject:	3,467,492	96.74%
Correctly kept:	49,282	99.78%

Bacterial Depletion on Zymo

- Mapping to a reference of the all bacteria
- Ejecting reads if they map within the first 4,500bp (10 sec), enriching fungal sequence
- Sequenced same sample on flowcell of similar quality directly next to ReadUntil run as control
- Running with 48 threads on a 3.0 GHz Intel Xeon Gold 6136 (probably overkill)

Read decision accuracy:

Correctly eject:	3,467,492	96.74%
Correctly kept:	49,282	99.78%

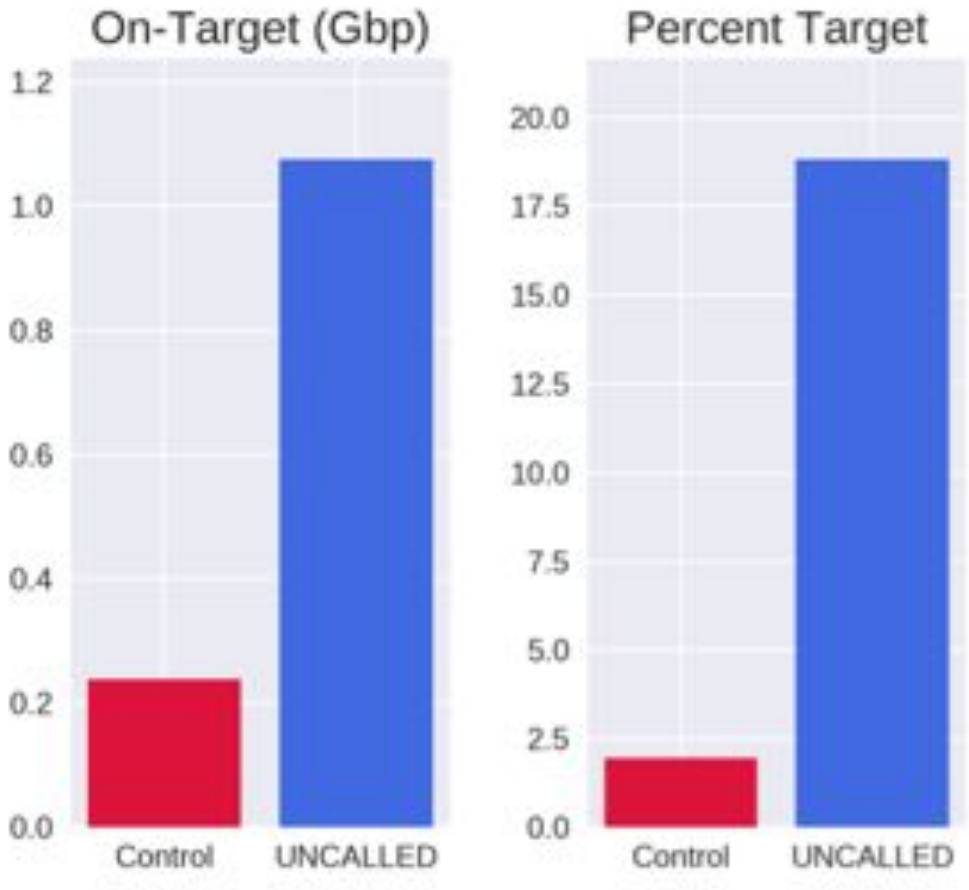


Bacterial Depletion on Zymo

- Mapping to a reference of the all bacteria
- Ejecting reads if they map within the first 4,500bp (10 sec), enriching fungal sequence
- Sequenced same sample on flowcell of similar quality directly next to ReadUntil run as control
- Running with 48 threads on a 3.0 GHz Intel Xeon Gold 6136 (probably overkill)

Read decision accuracy:

Correctly eject:	3,467,492	96.74%
Correctly kept:	49,282	99.78%

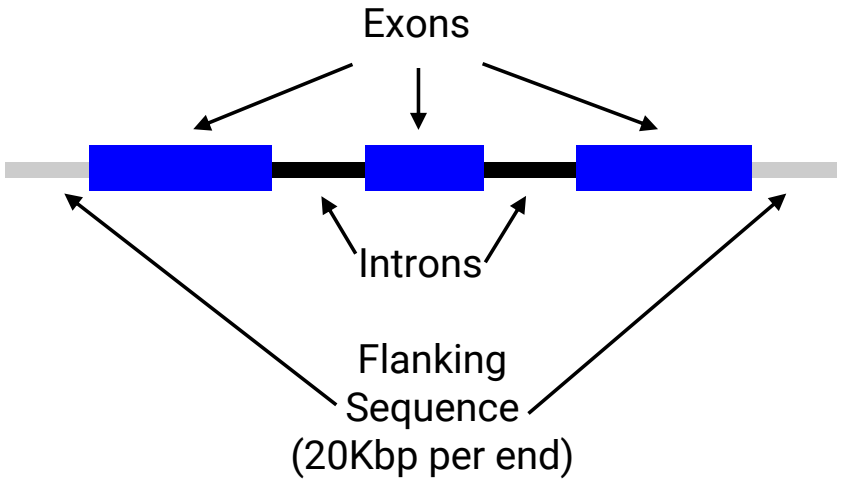


Fold change: **4.50** **9.63**

Human Gene Enrichment

Enriching GM12878 for genes associated with hereditary cancer

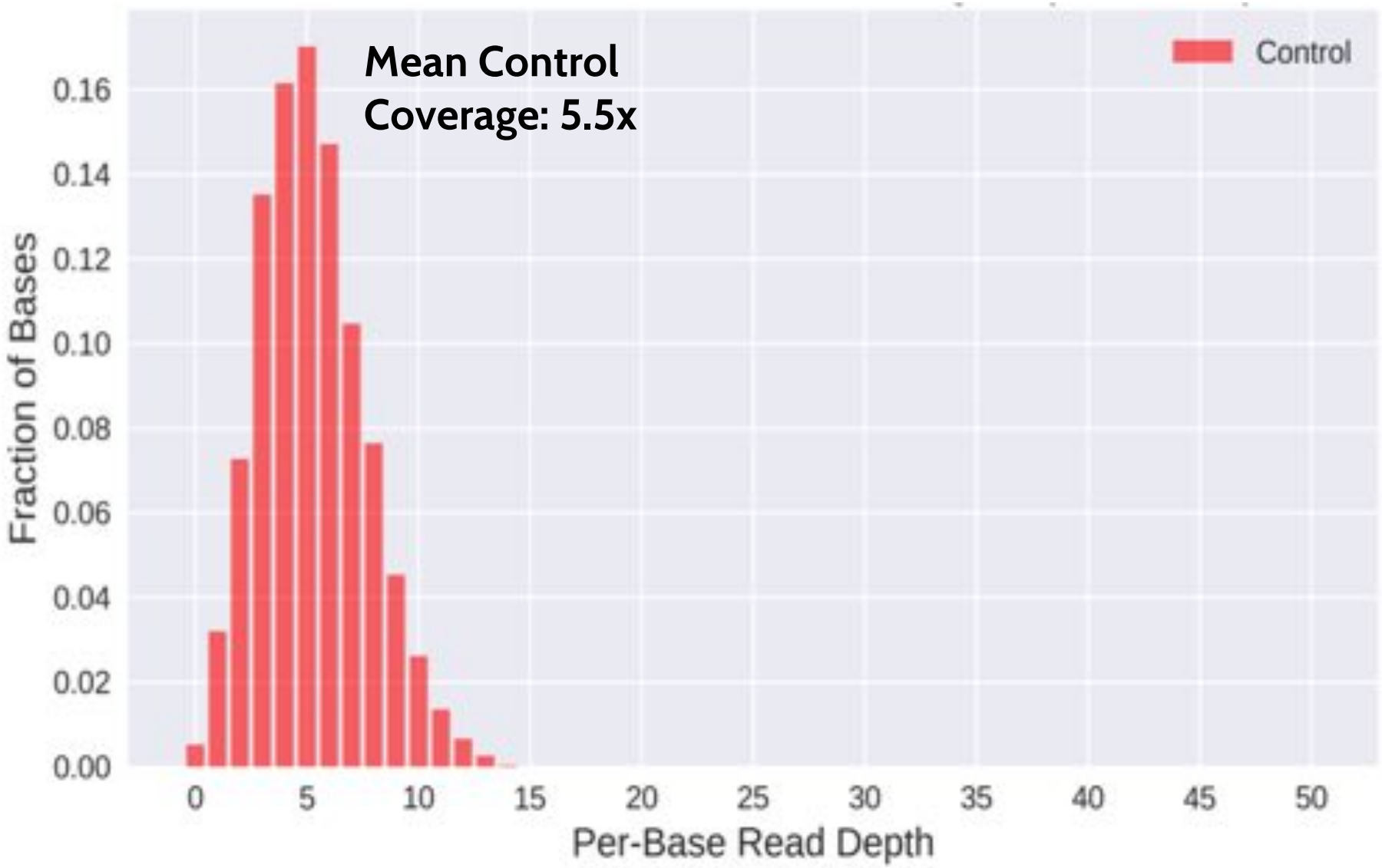
- Could provide a low-cost and portable method to identify, SNVs, SVs, and epigenetic modifications
- Mapping to a 18.6 Mbp reference, ejecting reads that don't map in 3 sec, ran for 72 hours
- Also masked low-complexity and repetitive regions, and performed two nuclease flushes on each run



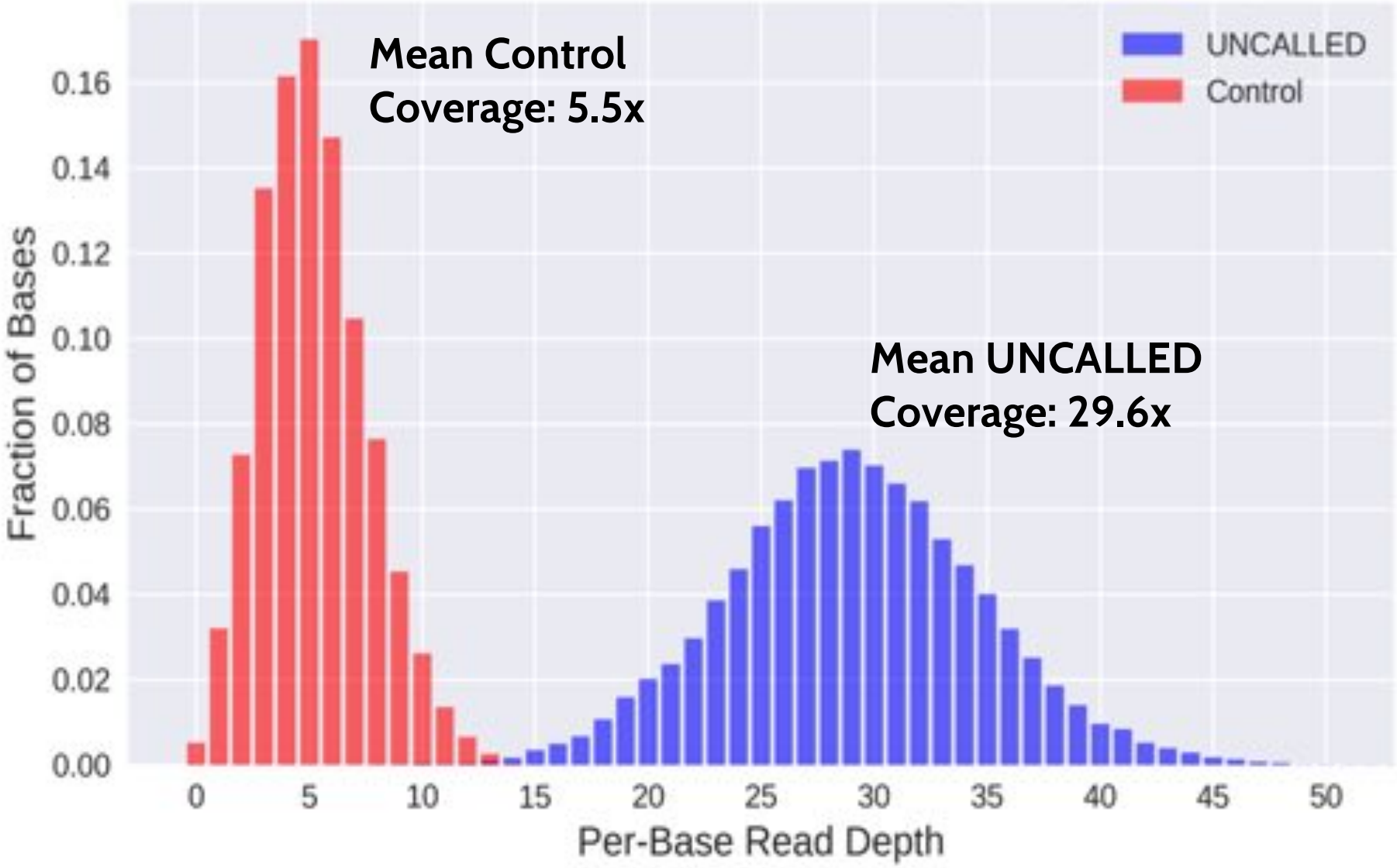
All 148 genes from every Invitae cancer panel:

ABRAXAS1	CDKN1B	EXT1	HOXB13	NOP10	RECQL	SDHC	XRCC2
AIP	CDKN1C	EZH2	HRAS	NTHL1	RECQL4	SDHB	
AKT1	CEBPA	FANCA	KIF1B	PALB2	REST	SDHD	
ALK	CDKN2A	FANCB	KIT	PALLD	RET	SLX4	
AP2S1	CEP57	FANCC	LZTR1	PARN	RINT1	SMAD4	
APC	CHEK2	FANCD2	MC1R	PDGFRA	RNF43	SMARCB1	
ATM	CFTR	FANCE	MAX	PHOX2B	RPL11	SMARCA4	
ATR	CPA1	FANCF	MEN1	PIK3CA	RPL15	SPINK1	
AXIN2	CTNNA1	FANCG	MET	PMS2	RPL26	SMARCE1	
BAP1	CTC1	FANCI	MITF	POLD1	RPL35A	STK11	
BARD1	DIS3L2	FANCL	MLH1	POT1	RPL5	SUFU	
BMPR1A	CTRC	FANCM	MLH3	POLE	RPS10	TERT	
BLM	CTR9	FLCN	MRE11	PRKAR1A	RPS19	TERC	
BRCA1	DICER1	FH	MSH2	PRSS1	RPS20	TINF2	
BRCA2	DKC1	GALNT12	MSH3	PTCH1	RPS24	TMEM127	
CASR	EGLN1	GATA1	MSH6	PTCH2	RPS26	TP53	
BUB1B	EGFR	GEN1	NBN	PTEN	RPS7	TSC1	
BRIP1	EPCAM	GATA2	MUTYH	RAD50	RTEL1	TSC2	
CDC73	ENG	GNA11	NF1	RAD51C	RUNX1	VHL	
CDK4	ERCC4	GPC3	NF2	RAD51D	SDHA	WRN	
CDH1	EXT2	GREM1	NHP2	RB1	SDHAF2	WT1	

Hereditary Cancer Gene Enrichment



Hereditary Cancer Gene Enrichment



Hereditary Cancer Gene Analysis

Small variant calling with Clair

	Dataset	Precision	Recall	F1
SNPs	Control (5x)	41.9%	39.4%	0.406
	UNCALLED (29.6x)	92.8%	97.6%	0.951
	ONT WGS (51.1x)	93.2%	98.5%	0.958
Indels	Control (5x)	37.6%	23.4%	0.288
	UNCALLED (29.6x)	80.4%	73.1%	0.766
	ONT WGS (51.1x)	79.9%	72.7%	0.761

Hereditary Cancer Gene Analysis

Small variant calling with Clair

	Dataset	Precision	Recall	F1
SNPs	Control (5x)	41.9%	39.4%	0.406
	UNCALLED (29.6x)	92.8%	97.6%	0.951
	ONT WGS (51.1x)	93.2%	98.5%	0.958
Indels	Control (5x)	37.6%	23.4%	0.288
	UNCALLED (29.6x)	80.4%	73.1%	0.766
	ONT WGS (51.1x)	79.9%	72.7%	0.761

Structural variant calling with Sniffles

- 50 SVs detected from UNCALLED reads
- 100% concordance with 50x coverage ONT WGS and 30x PacBio HiFi reads
- More than double the number of SVs detected with 50x coverage whole-genome Illumina sequencing

Hereditary Cancer Gene Analysis

Small variant calling with Clair

	Dataset	Precision	Recall	F1
SNPs	Control (5x)	41.9%	39.4%	0.406
	UNCALLED (29.6x)	92.8%	97.6%	0.951
	ONT WGS (51.1x)	93.2%	98.5%	0.958
Indels	Control (5x)	37.6%	23.4%	0.288
	UNCALLED (29.6x)	80.4%	73.1%	0.766
	ONT WGS (51.1x)	79.9%	72.7%	0.761

Structural variant calling with Sniffles

- 50 SVs detected from UNCALLED reads
- 100% concordance with 50x coverage ONT WGS and 30x PacBio HiFi reads
- More than double the number of SVs detected with 50x coverage whole-genome Illumina sequencing

Methylation calling with Nanopolish

.96 Pearson correlation coefficient with 50x ONT WGS sequencing

Found evidence of X-inactivation

Structural Variant Breakdown

56 concordant SVs overall

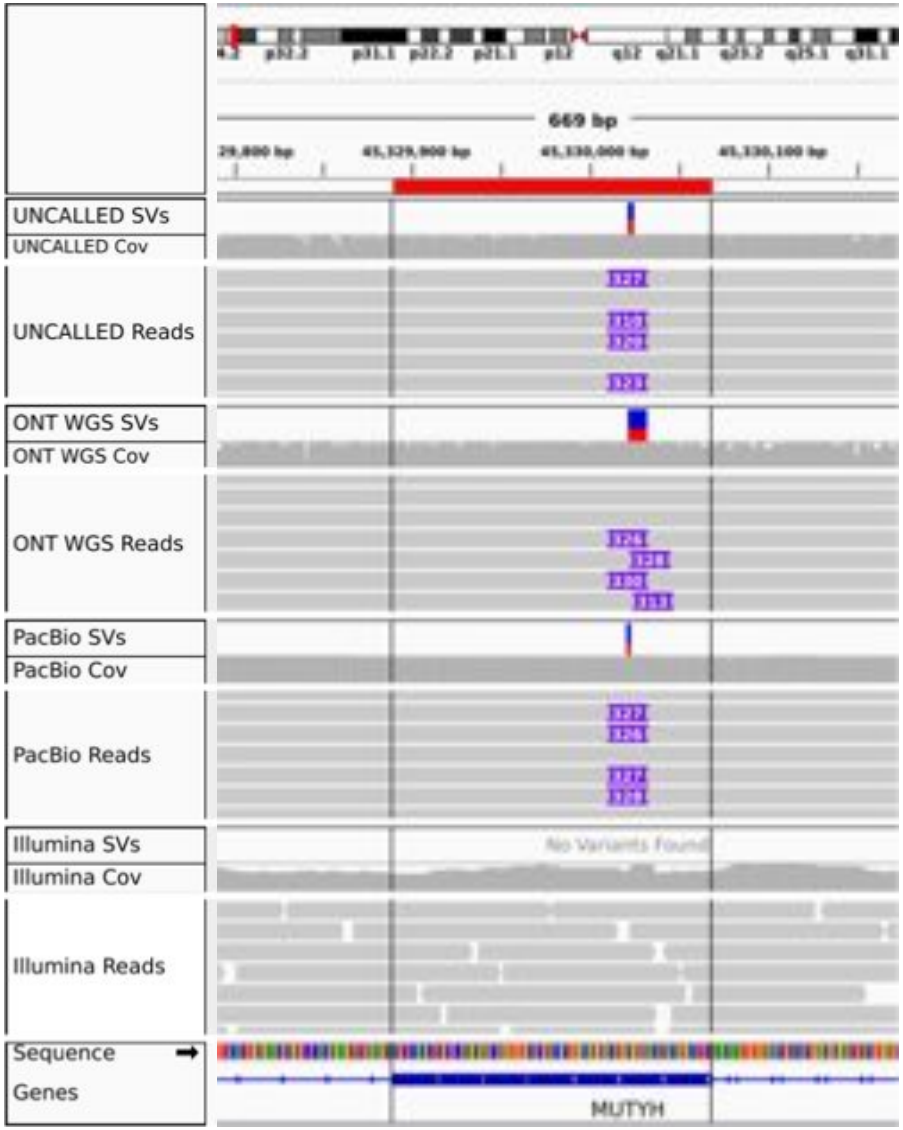
- 39 insertions, 17 deletions

Classified insertions and deletions by repeat alignment and overlap

- 56% of insertions and 41% of deletions are simple repeats or low-complexity
- 16% of insertions and 29% of deletions overlap or align to Alu elements

One SV was found in an exon

- Heterozygous Alu insertion in MUTYH
- Mutations in this gene promote colorectal and breast cancers



Class Project

- UNCALLED started as a project for this class!
- How we split the work between the three of us:
  - Collecting/parsing raw nanopore signal data
  - Signal processing/k-mer matching
  - FM Index construction/basic search algorithm
- All of us brainstormed how the algorithm should work
- We did not have a functional aligner in the end
  - Created a signal-based FM index (later turned out to be unnecessary)
  - Figured out how to compute event/k-mer match probabilities (but messed up signal normalization)
  - Could produce seed alignments based on a very simple algorithm (but had no way to filter the many many false positives)
- Despite the incompleteness it was a successful project!

Welcome to Applied Comparative Genomics

https://github.com/schatzlab/appliedgenomics2020

Questions?

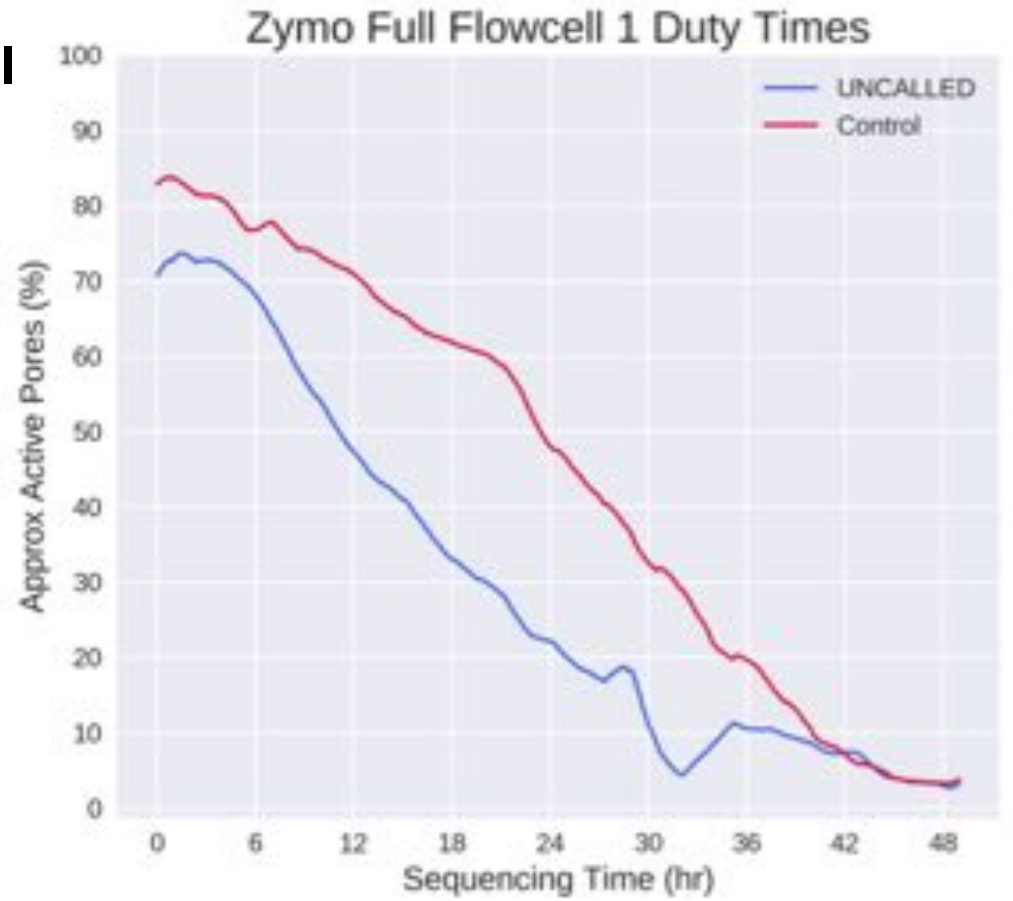
Reduced UNCALLED Yield

UNCALLED has ~50% the yield of control

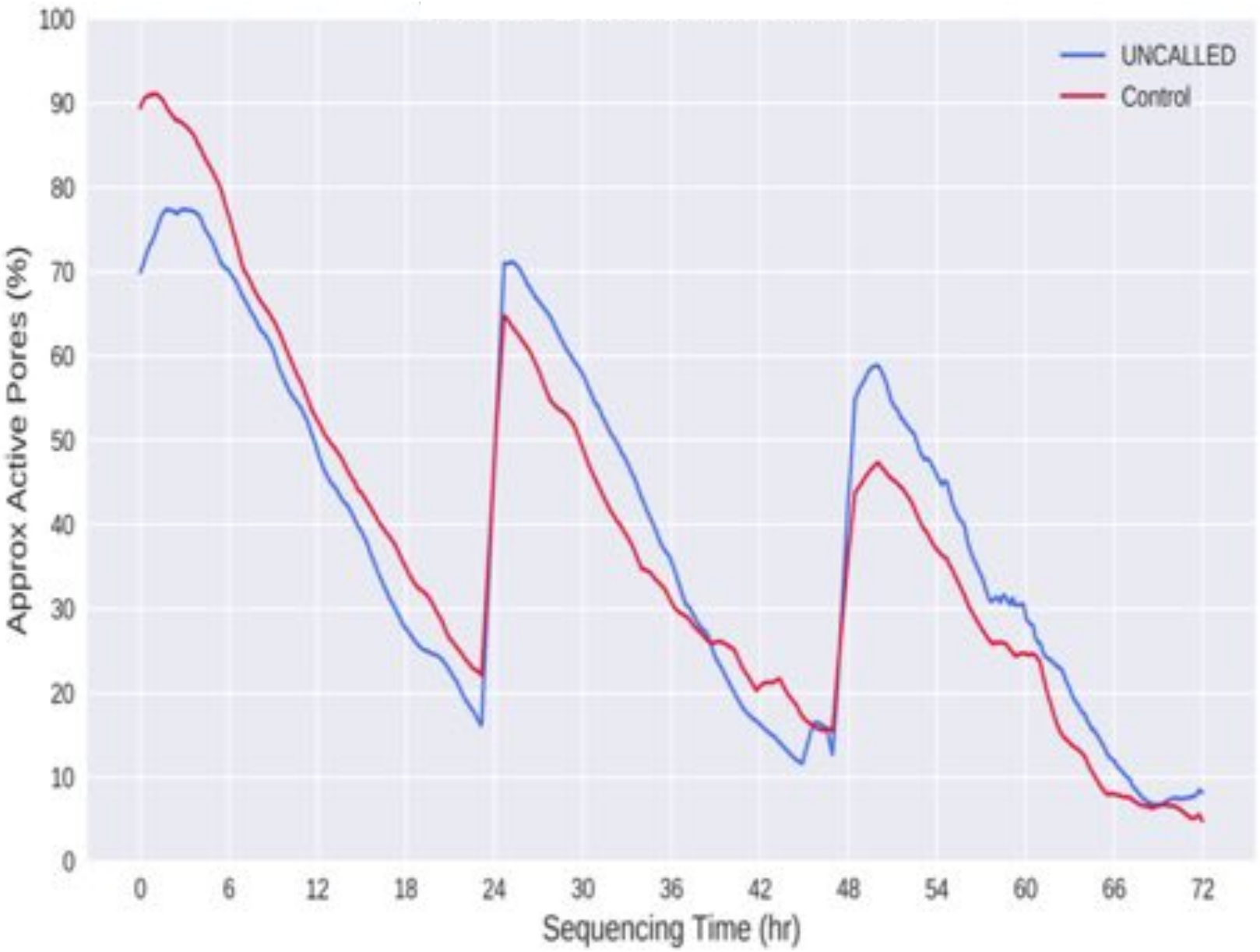
- Fewer pores are active throughout
- Partially explained by ejections causing more gaps between reads, but not entirely
- Pores *do not* seem to be “dying” faster

Likely due to ejections causing more pore blockages

- Maybe single stranded DNA self binding, or more reads mean more chances to clog



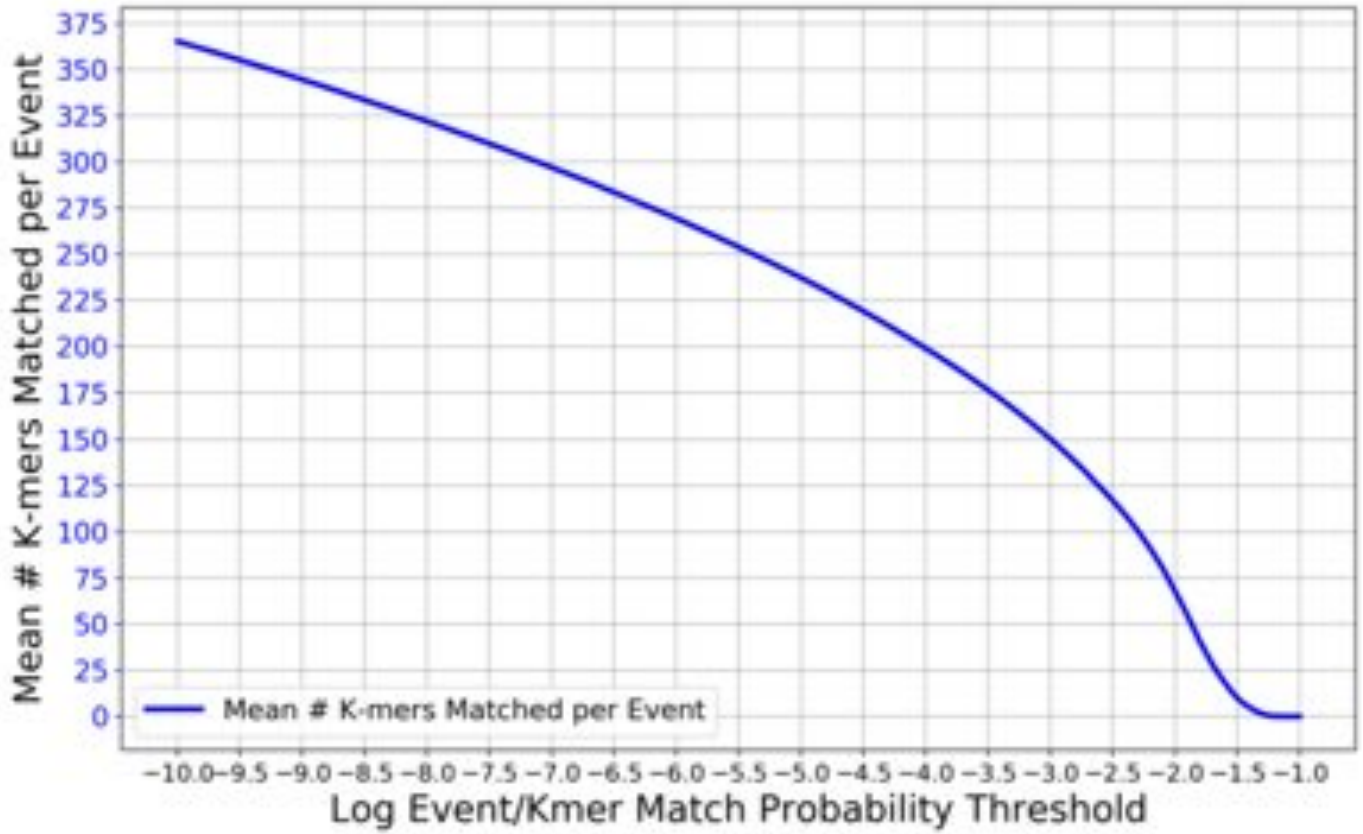
Nuclease Flush Results



For each event, we consider all k-mers which match with probability above some threshold

Lower threshold means

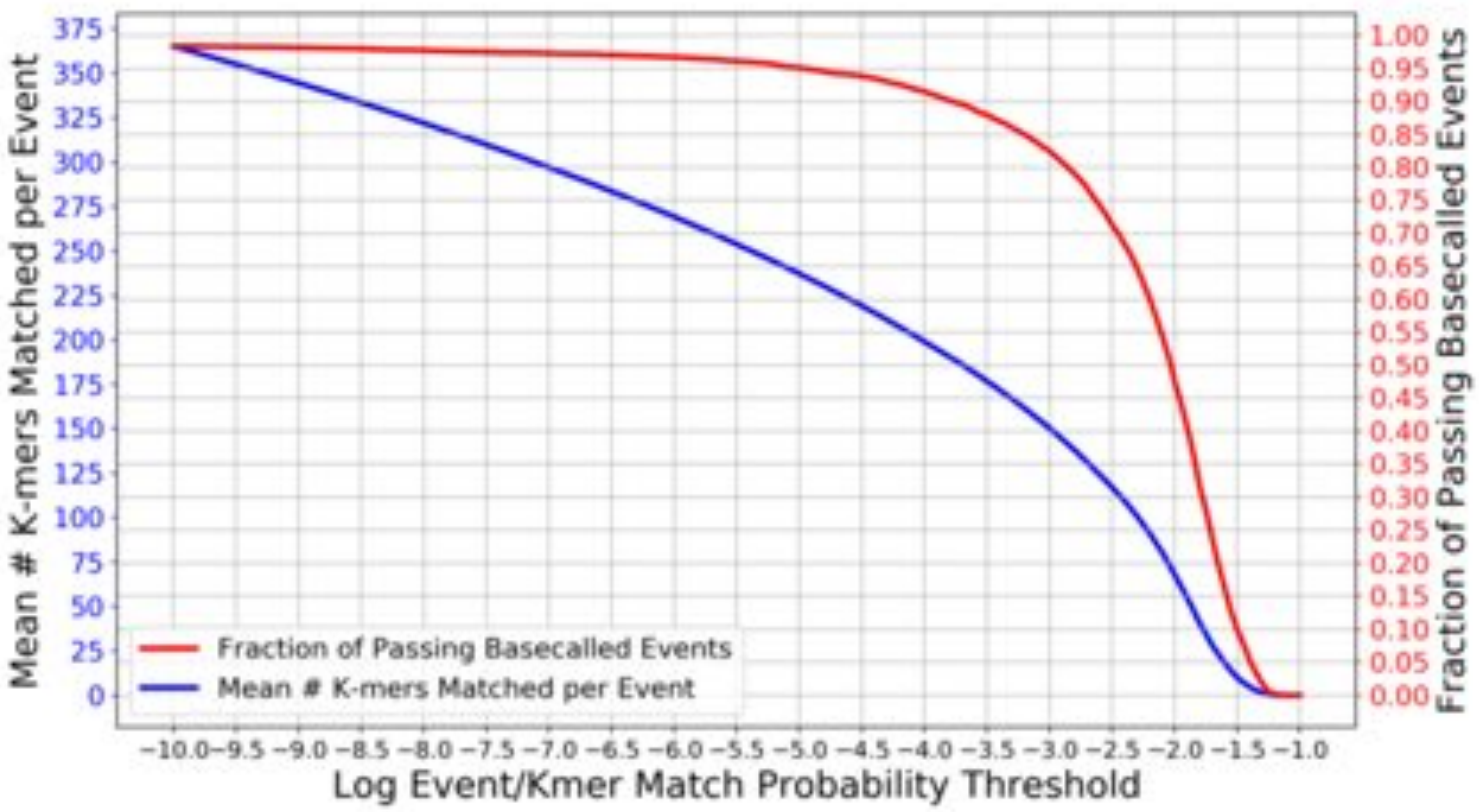
- Takes more time to process the event



For each event, we consider all k-mers which match with probability above some threshold

Lower threshold means

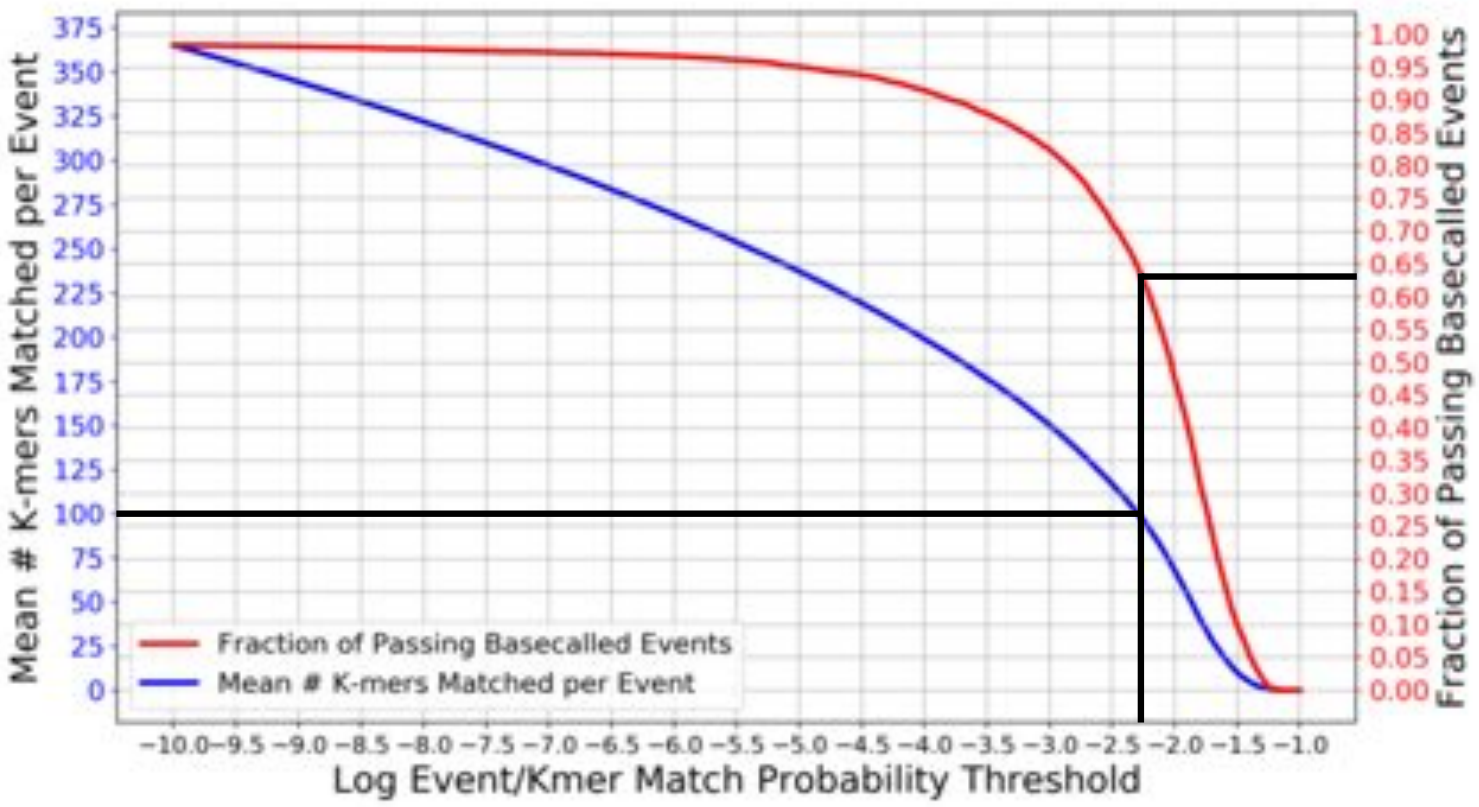
- Takes more time to process the event
- Higher chance of finding the right k-mer



For each event, we consider all k-mers which match with probability above some threshold

Lower threshold means

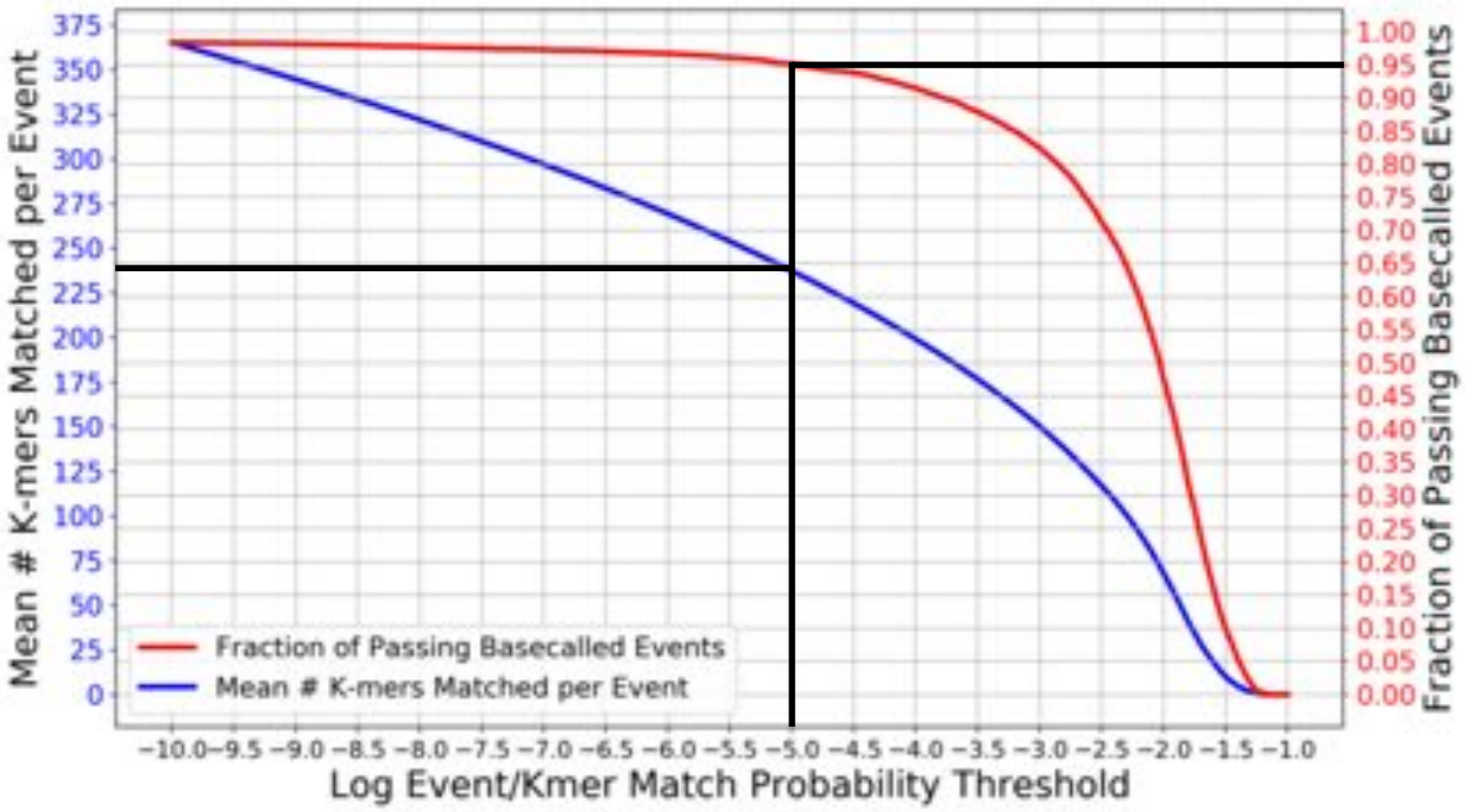
- Takes more time to process the event
- Higher chance of finding the right k-mer



For each event, we consider all k-mers which match with probability above some threshold

Lower threshold means

- Takes more time to process the event
- Higher chance of seeing correct k-mer



Smaller FM index ranges represent fewer genomic locations

- Fewer possible distinct sequences
- Less potential for branching

Threshold gets lower (considers more k-mers) as FM index range gets smaller

- Must change with reference size/repeat content