Portable DNA Sequencing: Analysis Methods and Applications

Jared Simpson
Ontario Institute for Cancer Research
&
Department of Computer Science
University of Toronto

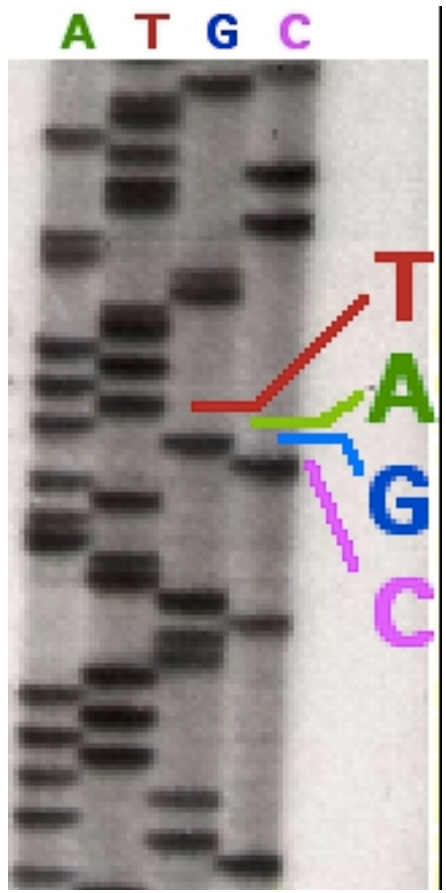
CSC2417
December 7th, 2016

Timeline of DNA Sequencing

1970s

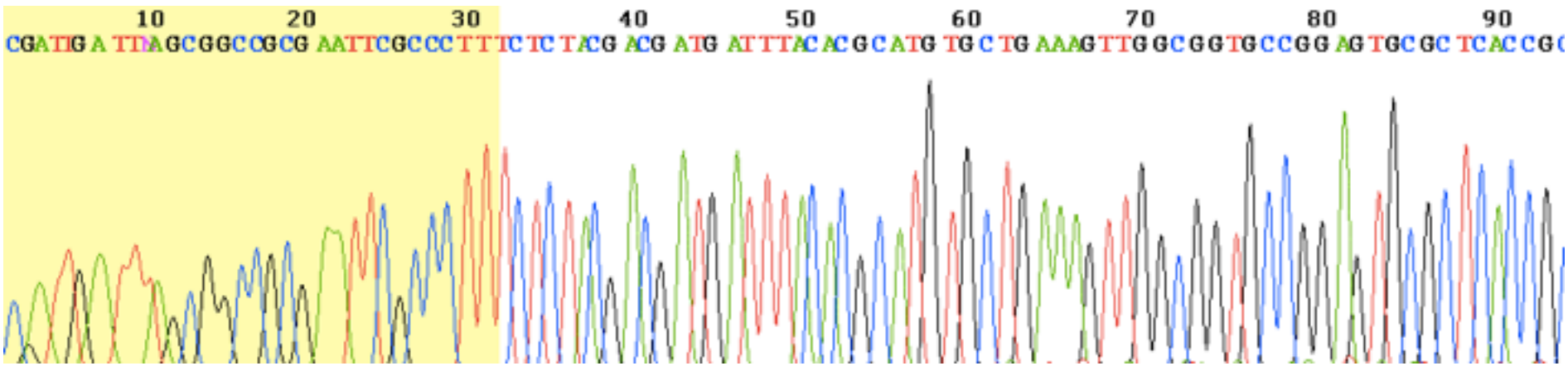
Sanger sequencing invented

Manual process; low throughput



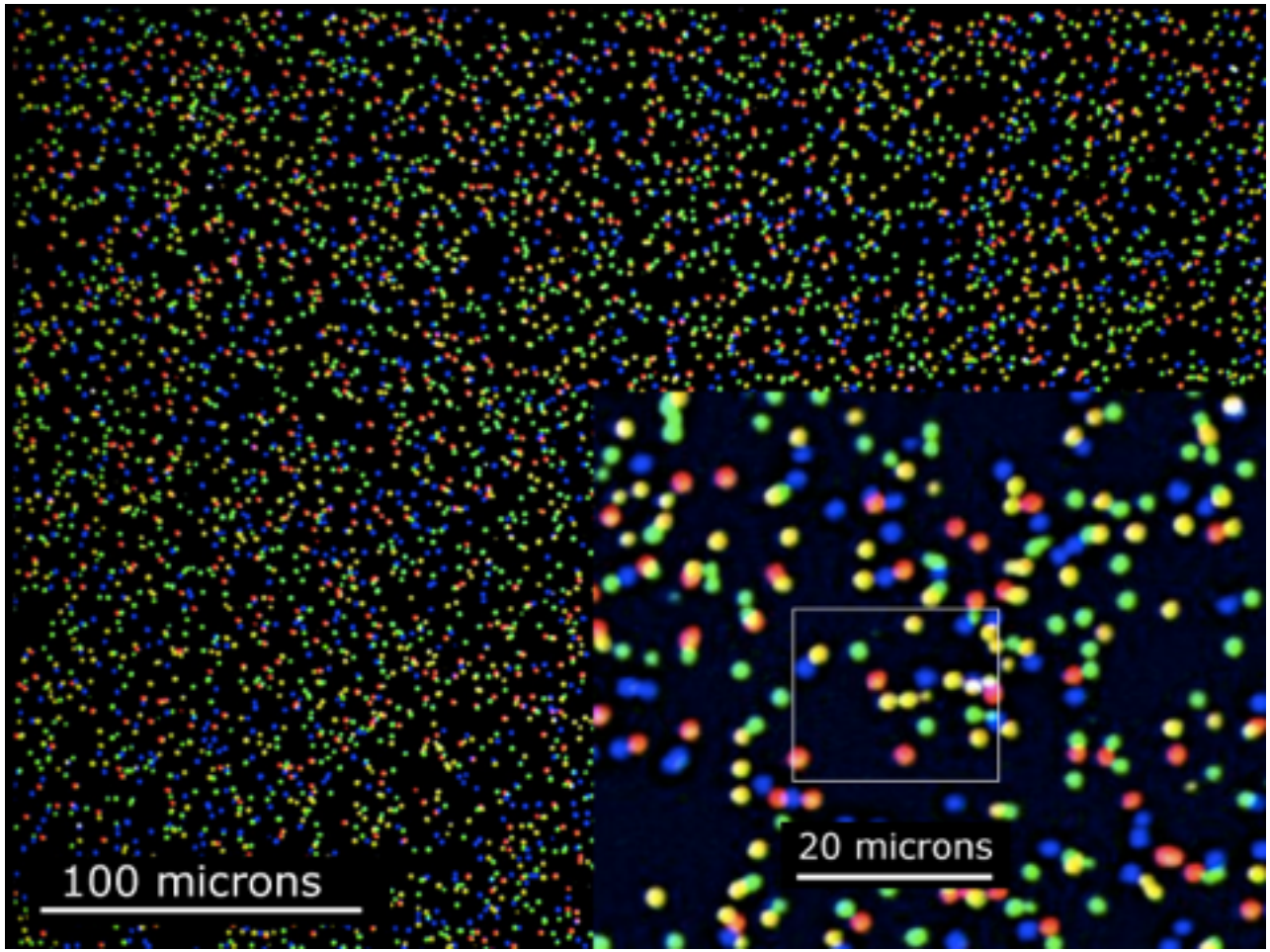
Timeline of DNA Sequencing

1980-1990s
Automation of Sanger sequencing
10-100 Megabases of data per run



Timeline of DNA Sequencing

2000s
Massively Parallel Sequencing
10-100s Gigabases per run



Timeline of DNA Sequencing

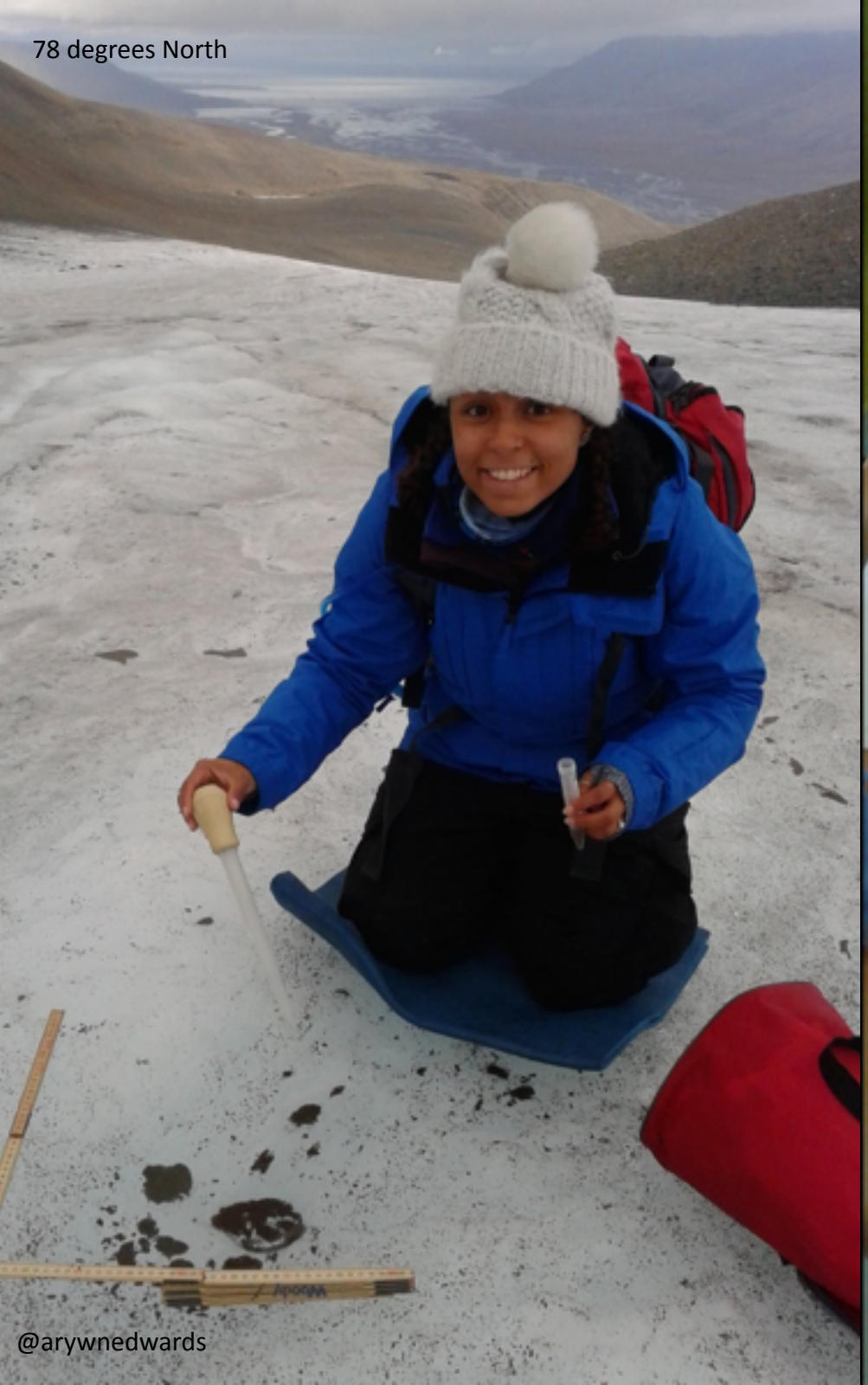
2010s

Centralisation of sequencing

18,000 human genomes/year

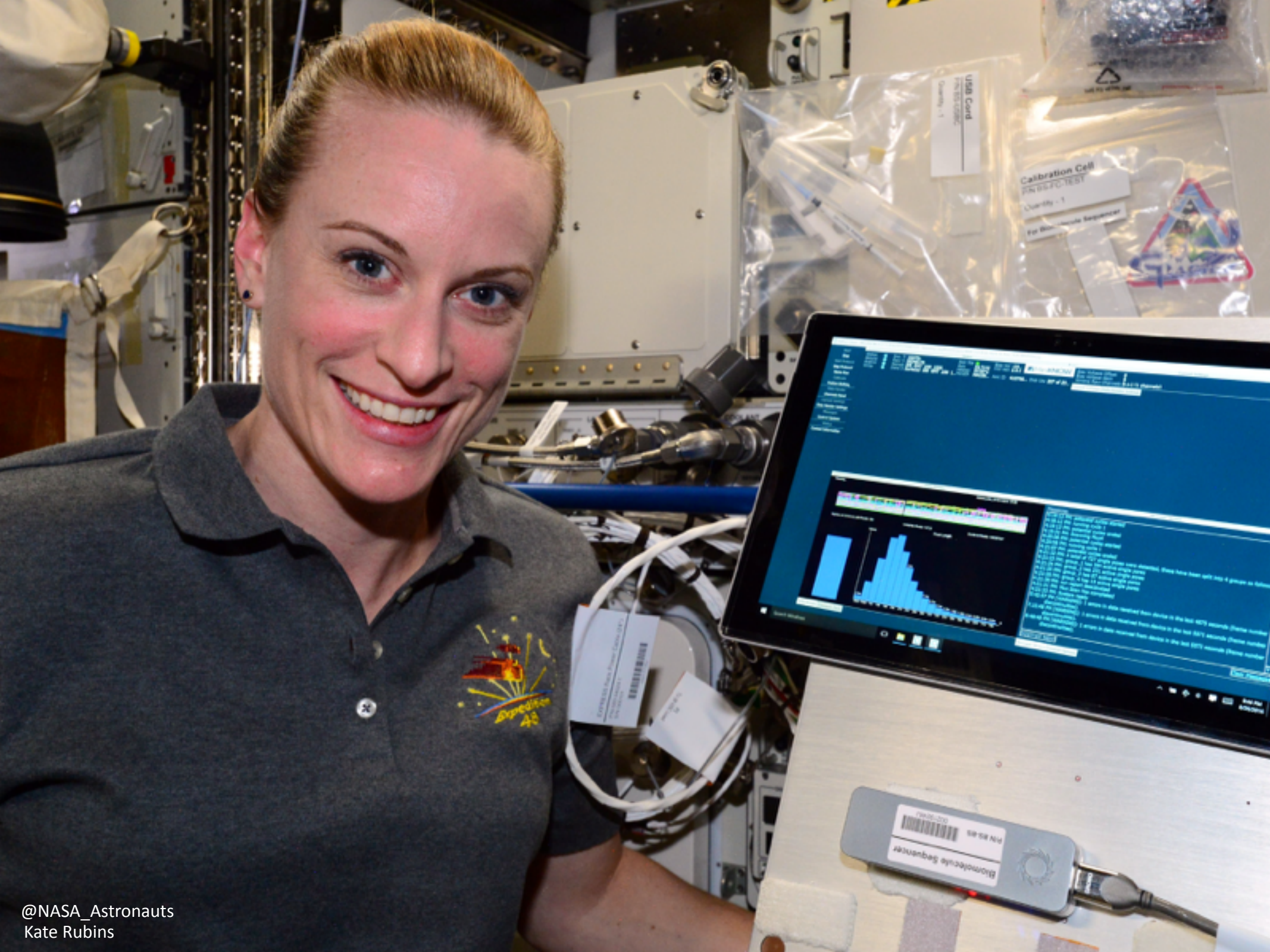
Miniature, Portable Sequencing

Joseph Bore running the MinION in Guinea, Quick et al Nature (2016)


78 degrees North
@arywnedwards


MinION


European astronaut centre
ESA CAVES
@explornaut



@NASA_Astronauts
Kate Rubins

Disclosure: ONT provides research funding to my lab

Why use genome sequencing during an outbreak?

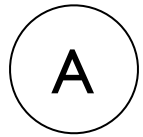
- Sequence viral genomes to calculate mutation rate
- Identify host adaptations
- Characterise response to therapy/immunization
- Relate cases to each other and monitor geographic spread

Why use genome sequencing during an outbreak?

- Sequence viral genomes to calculate mutation rate
- Identify host adaptations
- Characterise response to therapy/immunization
- **Relate cases to each other and monitor geographic spread**

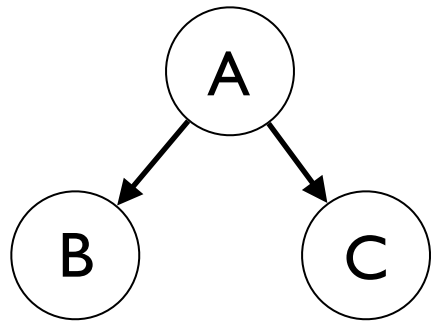
Ebola Surveillance

Ebola is passed through direct contact



Ebola Surveillance

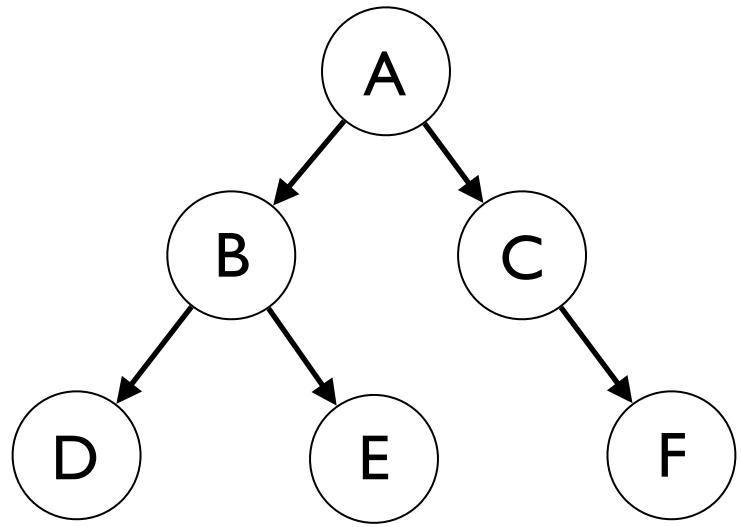
Ebola is passed through direct contact



Person A transmits the virus to B and C

Ebola Surveillance

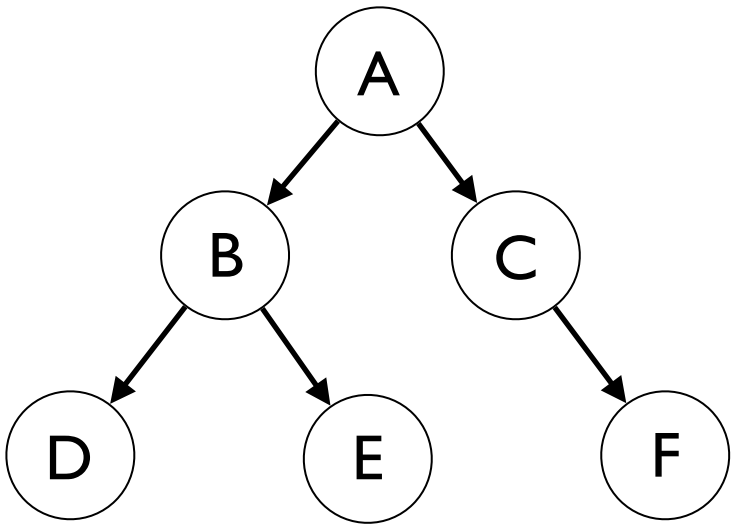
Ebola is passed through direct contact



Person A transmits the virus to B and C
Person B transmits the virus to D and E
Person C transmits the virus to F

Ebola Surveillance

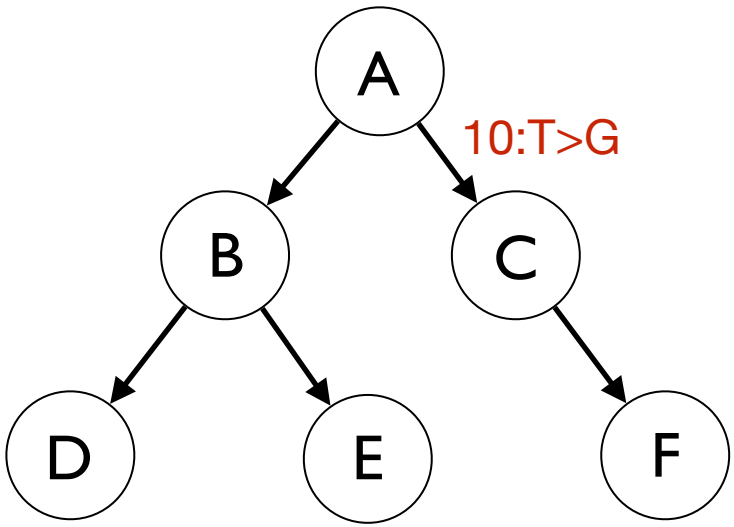
Ebola virus mutates at a rate of ~1.15 $\times 10^{-3}$ mutations/bp/year
These mutations allow us to track patterns of transmission




```
Ebola-A ...AGTAGCCTACGATACTACGATCGACTTA...
Ebola-B ...AGTAGCCTACGATATTACGATCGACTTA...
Ebola-C ...AGTAGCCGACGATACTACGATCGACTTA...
Ebola-D ...AGTTGCCTACGATATTACGATCGACTTA...
Ebola-E ...AGTAGCCTACGATATTACGATCGACATA...
Ebola-F ...AGTAGCCGACGATACTACGATGGACTTA...
```

Ebola Surveillance

Ebola virus mutates at a rate of ~1.15 $\times 10^{-3}$ mutations/bp/year
These mutations allow us to track patterns of transmission

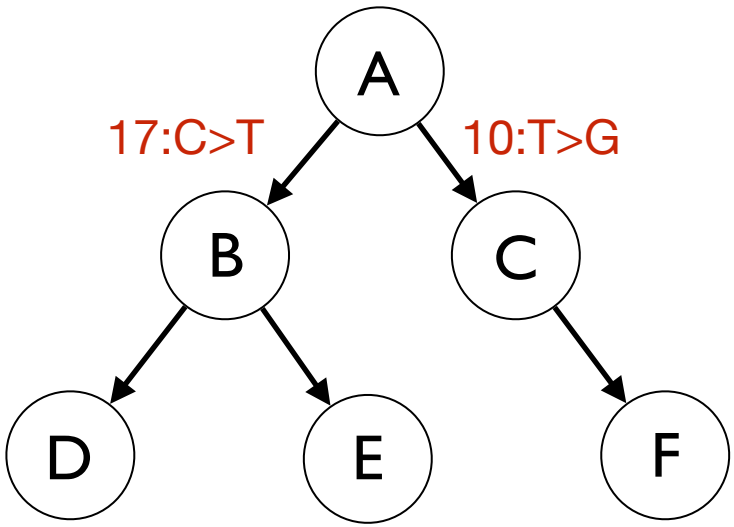



```
Ebola-A ...AGTAGCCTACGATACTACGATCGACTTA...
Ebola-B ...AGTAGCCTACGATATTACGATCGACTTA...
Ebola-C ...AGTAGCCGACGATACTACGATCGACTTA...
Ebola-D ...AGTTGCCTACGATATTACGATCGACTTA...
Ebola-E ...AGTAGCCTACGATATTACGATCGACATA...
Ebola-F ...AGTAGCCGACGATACTACGATGGACTTA...
```

T>G here indicates C/F lineage

Ebola Surveillance

Ebola virus mutates at a rate of ~1.15 $\times 10^{-3}$ mutations/bp/year
These mutations allow us to track patterns of transmission

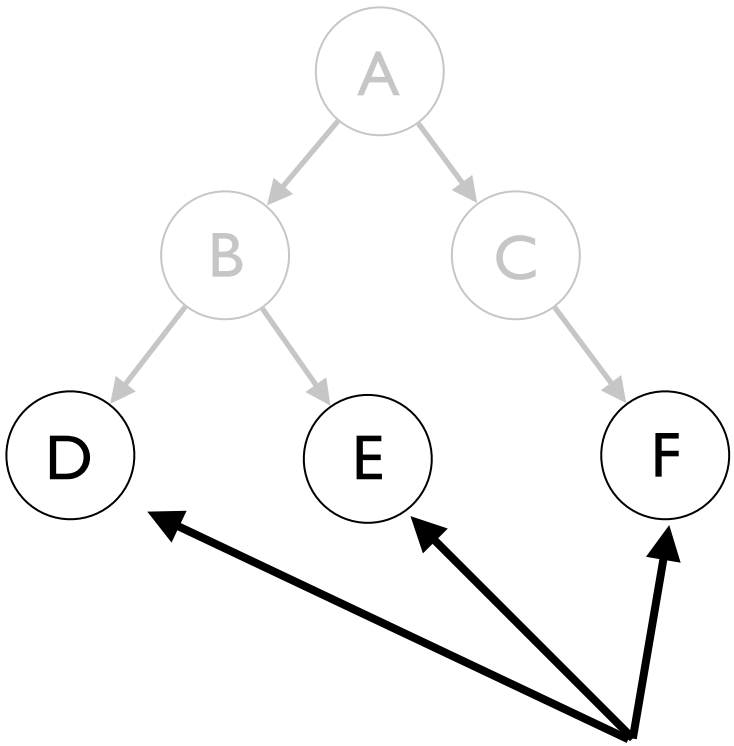



```
Ebola-A ...AGTAGCCTACGATACTACGATCGACTTA...
Ebola-B ...AGTAGCCTACGATATTACGATCGACTTA...
Ebola-C ...AGTAGCCGACGATACTACGATCGACTTA...
Ebola-D ...AGTTGCCTACGATATTACGATCGACTTA...
Ebola-E ...AGTAGCCTACGATATTACGATCGACATA...
Ebola-F ...AGTAGCCGACGATACTACGATGGACTTA...
```

C>T here indicates B/D/E lineage

Ebola Surveillance

We can't sequence every case in the outbreak but by sampling enough cases we can build a picture of how the virus is spreading (e.g. geographically)




```
Ebola-D ...AGTTGCCTACGATATTACGATCGACTTA...
Ebola-E ...AGTAGCCTACGATATTACGATCGACATA...
Ebola-F ...AGTAGCCGACGATACTACGATGGACTTA...
```

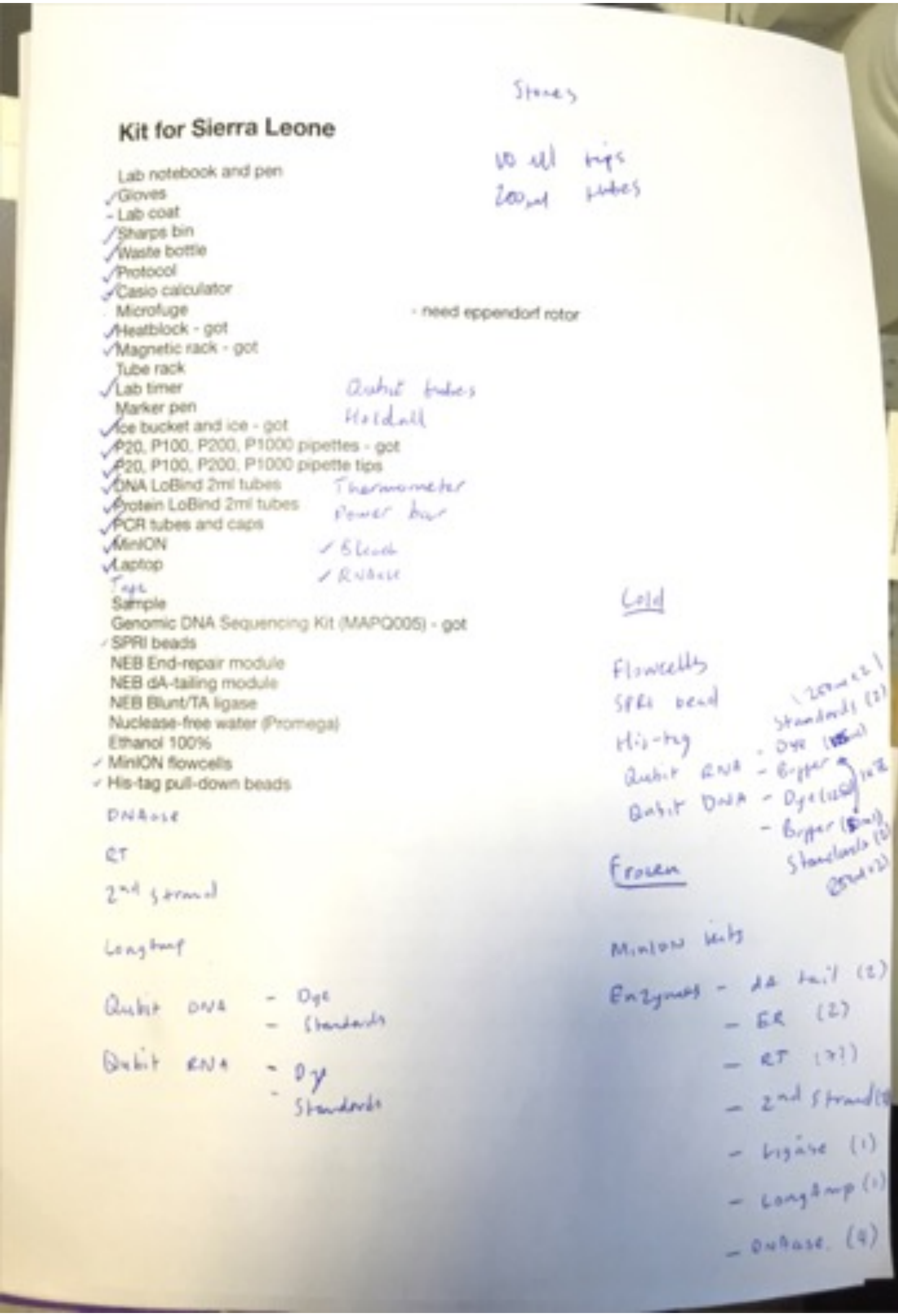
Sequence these cases; D and E share a mutation - possibly related?

Why use portable genome sequencing?

- Old process: take samples locally, send to lab in Europe/North America, sequence, return results to local epidemiologists
  - Slow, difficult to ship samples
- New process: use portable sequencers directly at the location they are needed

Image credit: Genome Research Limited


Heathrow
THE NORTH FACE



Kit for Sierra Leone

- Lab notebook and pen
- ✓ Gloves
- \- Lab coat
- ✓ Sharps bin
- ✓ Waste bottle
- ✓ Protocol
- ✓ Casio calculator
- Microfuge - need eppendorf rotor
- ✓ Heatblock - got
- ✓ Magnetic rack - got
- Tube rack
- ✓ Lab timer
- Marker pen
- ✓ Ice bucket and ice - got
- ✓ P20, P100, P200, P1000 pipettes - got
- ✓ P20, P100, P200, P1000 pipette tips
- ✓ DNA LoBind 2ml tubes
- ✓ Protein LoBind 2ml tubes
- ✓ PCR tubes and caps
- ✓ MinION
- ✓ Laptop
- Tape
- Sample
- Genomic DNA Sequencing Kit (MAPQ005) - got
- ✓ SPRI beads
- NEB End-repair module
- NEB dA-tailing module
- NEB Blunt/TA ligase
- Nuclease-free water (Promega)
- Ethanol 100%
- ✓ MinION flowcells
- ✓ His-tag pull-down beads

DNAase

RT

2nd strand

LongAmp

Qubit DNA - Dye - Standards

Qubit RNA - Dye - Standards

Stores

10 ul tips

200 ul tubes

Qubit tubes

Holdall

Thermometer

Power bar

✓ Bleach

✓ RNAase

Cold

Flowcells

SPRI bead

His-tag

Qubit RNA - Dye - Buffer - Standards (2)

Qubit DNA - Dye - Buffer - Standards (2)

Frozen

MinION kits

Enzymes - dA tail (2)

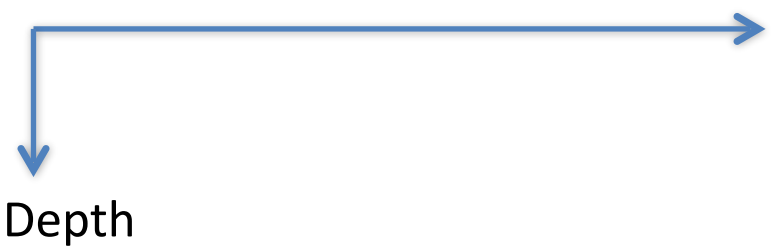
- ER (2)
- RT (?)
- 2nd strand
- ligase (1)
- LongAmp (1)
- DNAase (4)

Porton Down validation set, 89.1% coverage

Guinea 19 reactions v1, 98.1% coverage

Guinea 11 reactions v1, 95.9% coverage

Guinea 11 reactions v2, 98.4% coverage



Calculating a consensus sequence

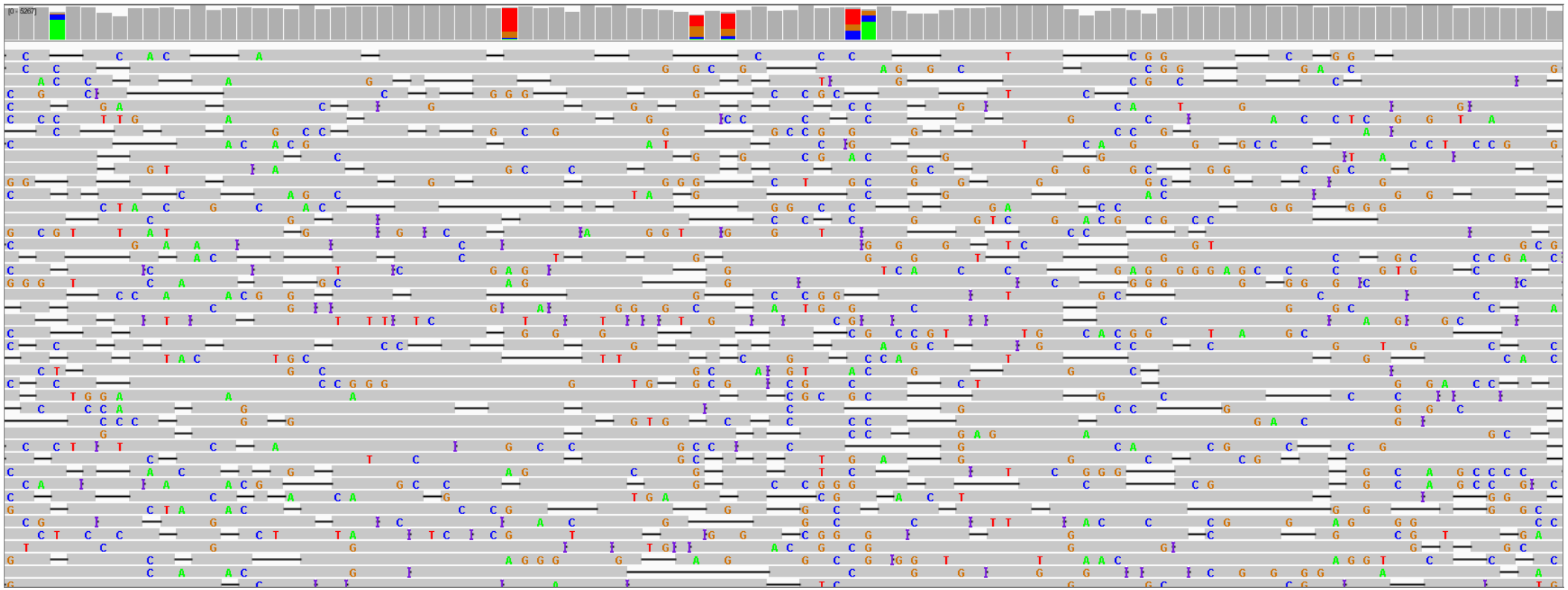
- Input: a set of nanopore reads ($r_1, r_2, \ldots, r_n$) from an Ebola genome (g)
- Output: the sequence of the Ebola genome, g

r_1 CAGATAGTCGGATGTTATGAACCAGATATATA
r_2 CAGACAGTCGGATGTTATGATCCAGATATGTA
r_3 CAGATAGTCGGATGTTATAATCCAGATATATA

---

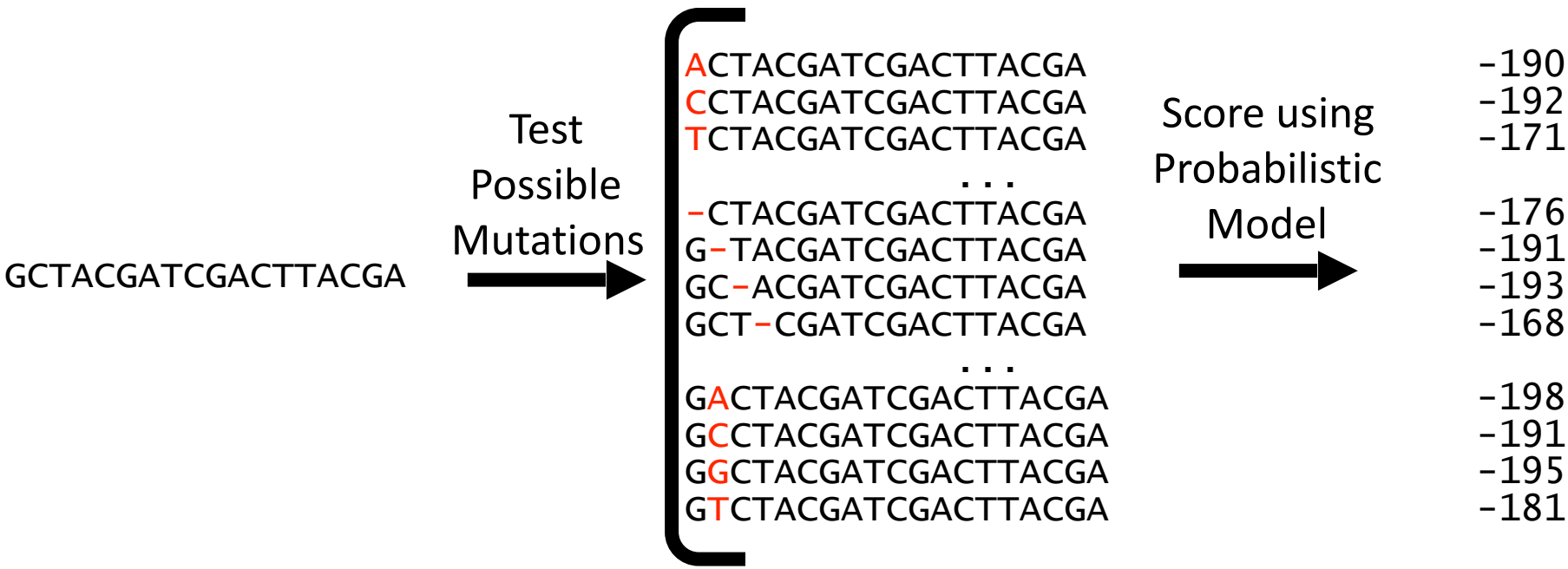
g CAGATAGTCGGATGTTATGATCCAGATATATA

Calculating a consensus sequence



Main challenge: the base-calling error rate is quite high (~10%)

Calculating a consensus sequence



Nanopore Sequencing

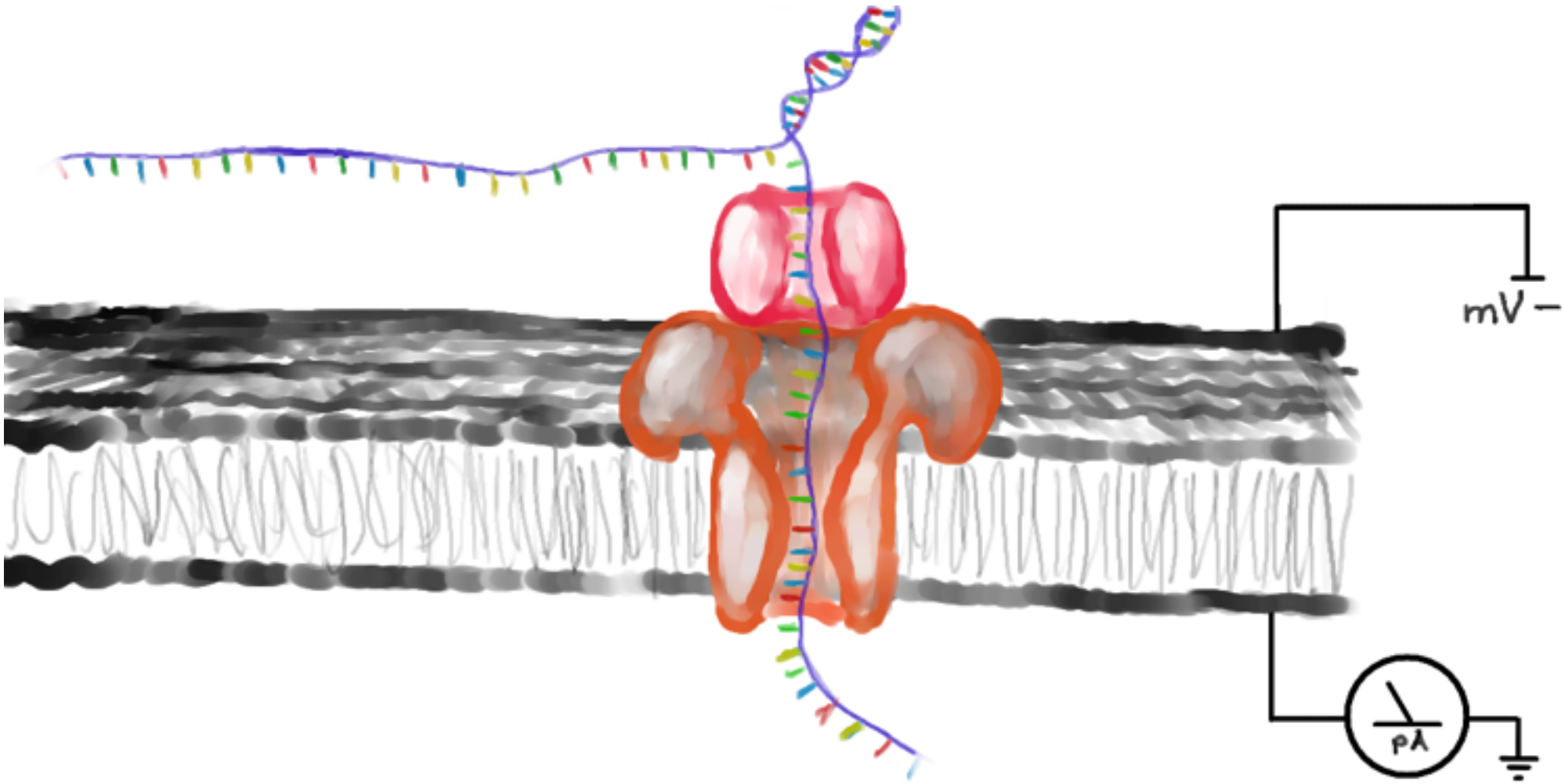
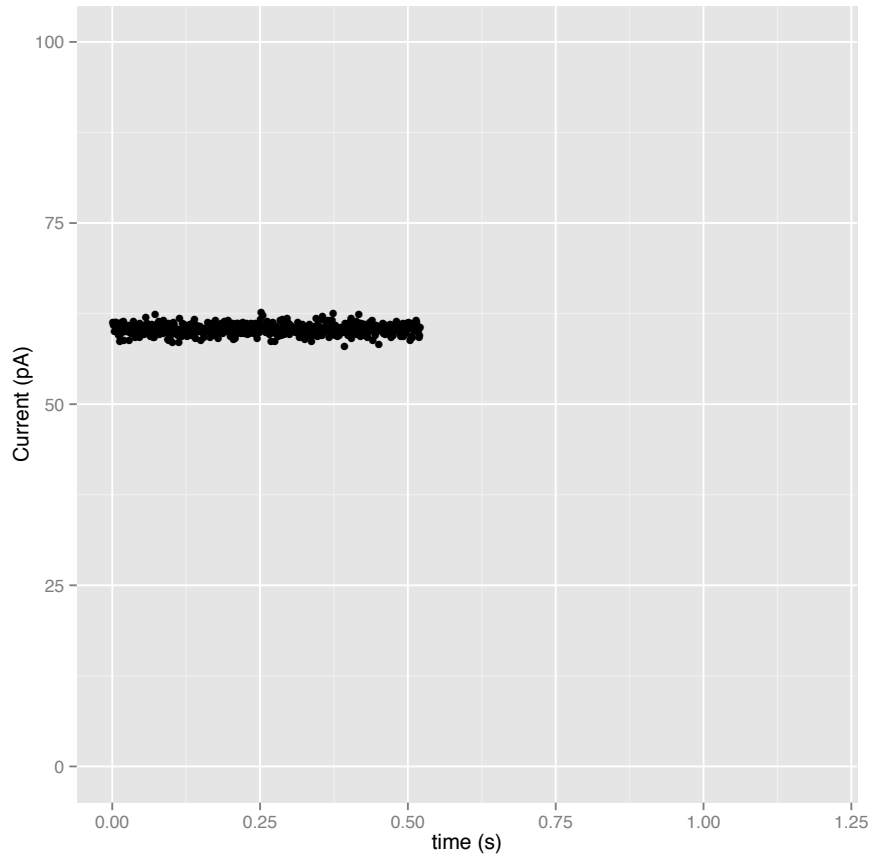
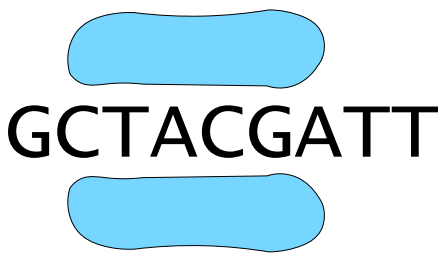


Illustration by David Eccles

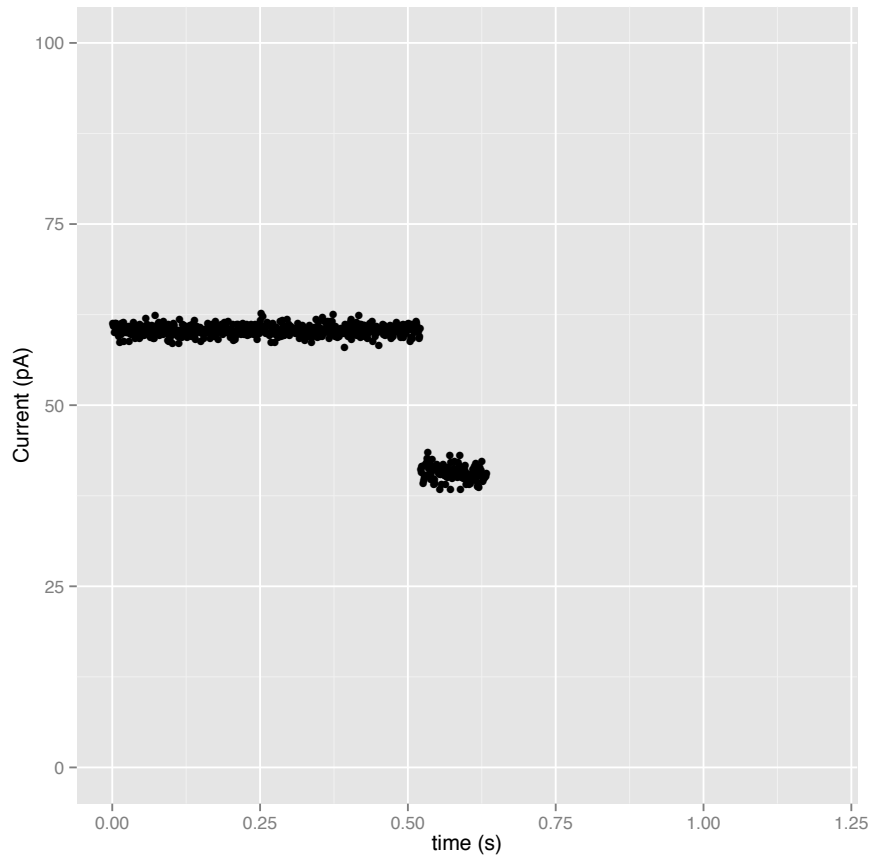
Nanopore Sequencing



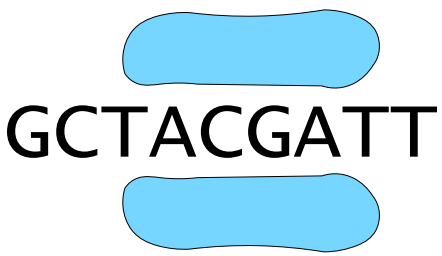
Nanopore Sequencing



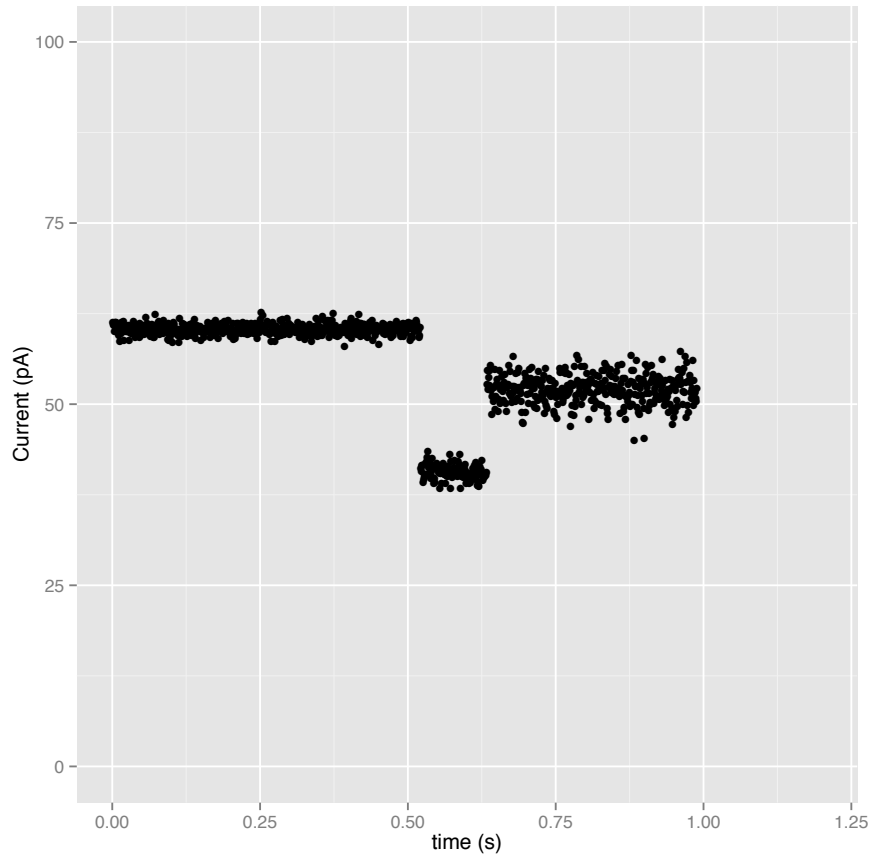
Sample Current



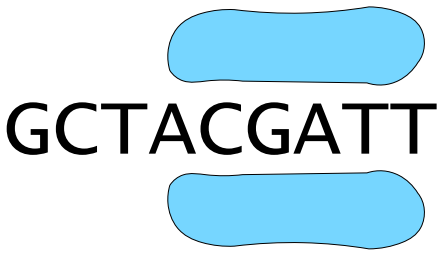
Nanopore Sequencing



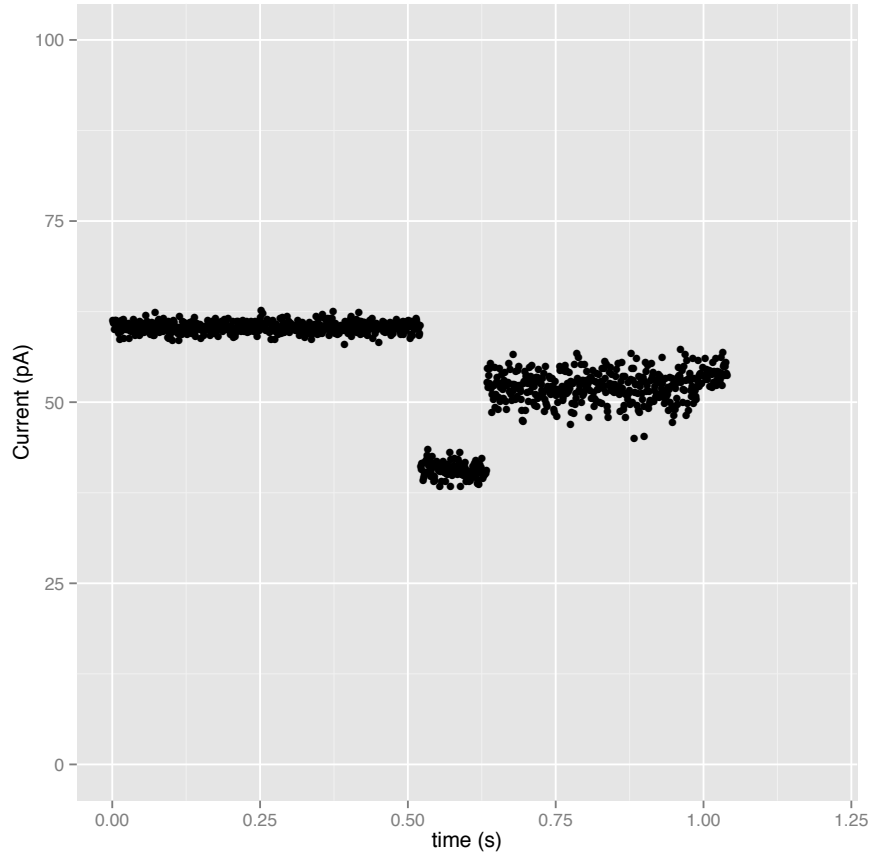
Sample Current



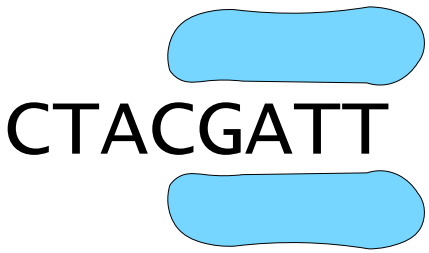
Nanopore Sequencing



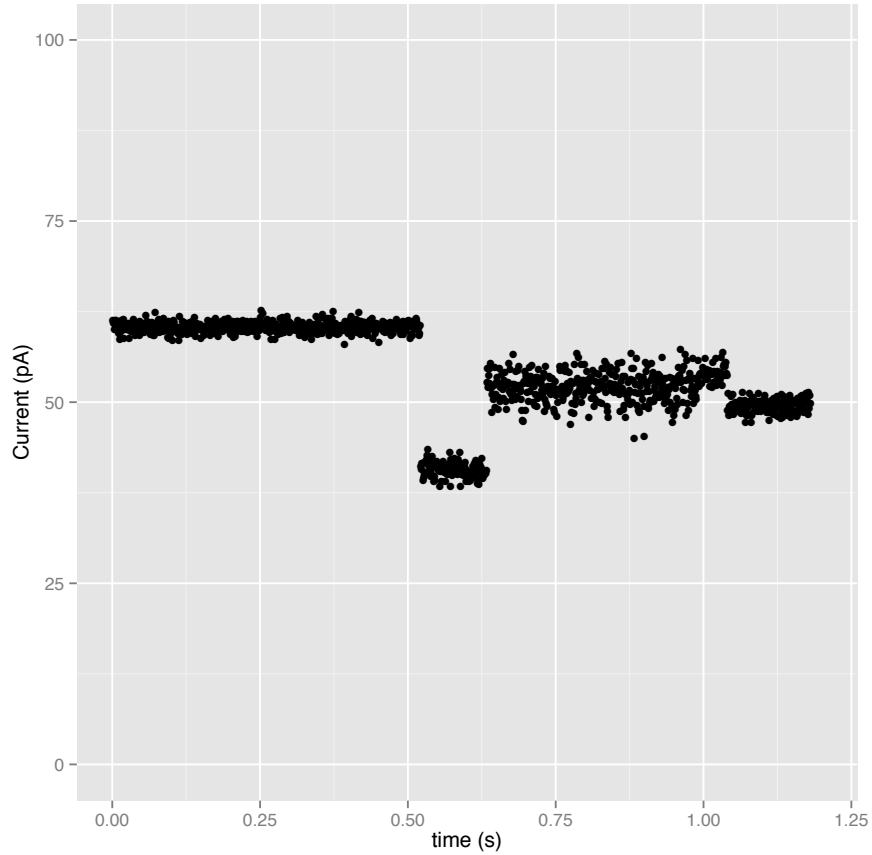
Sample Current



Nanopore Sequencing

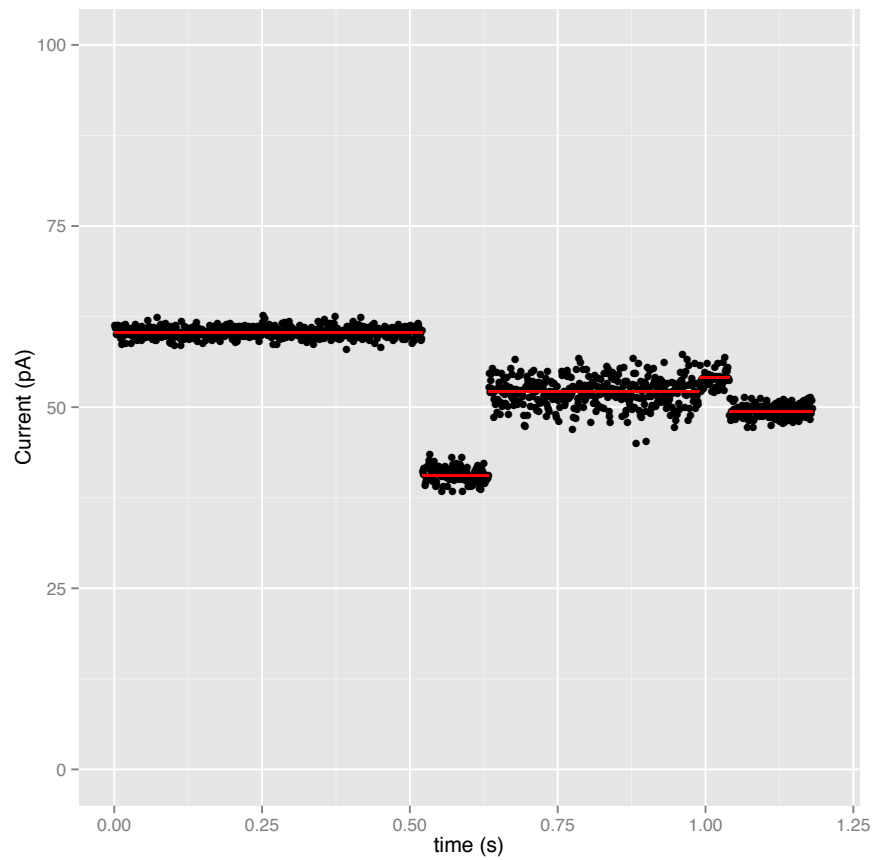


Sample Current

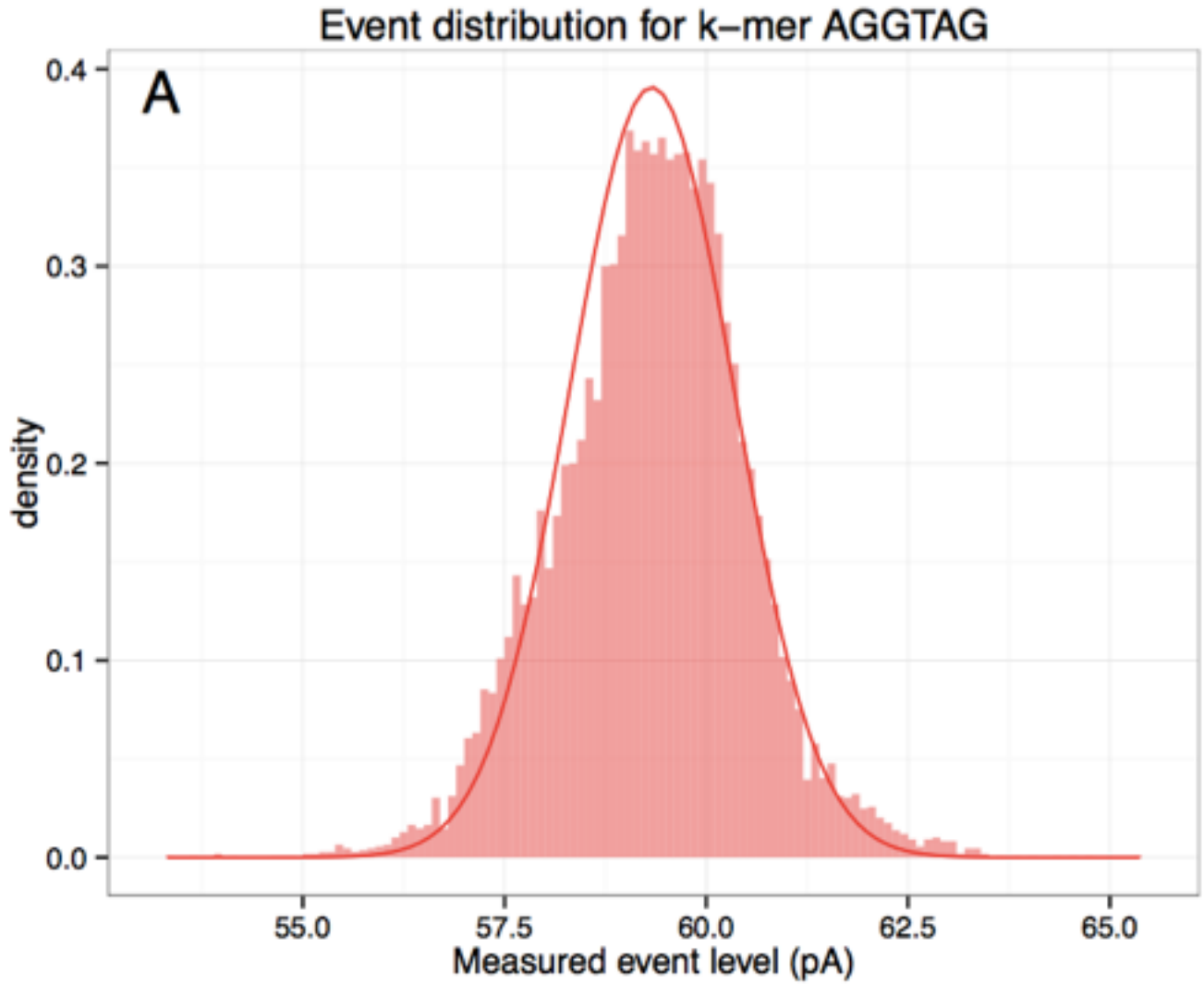


Event Detection

Event	mean current (pA)	current stdv	duration (s)
1	60.3	0.7	0.521
2	40.6	1.0	0.112
3	52.2	2.0	0.356
4	54.1	1.2	0.291
5	49.5	1.5	0.141



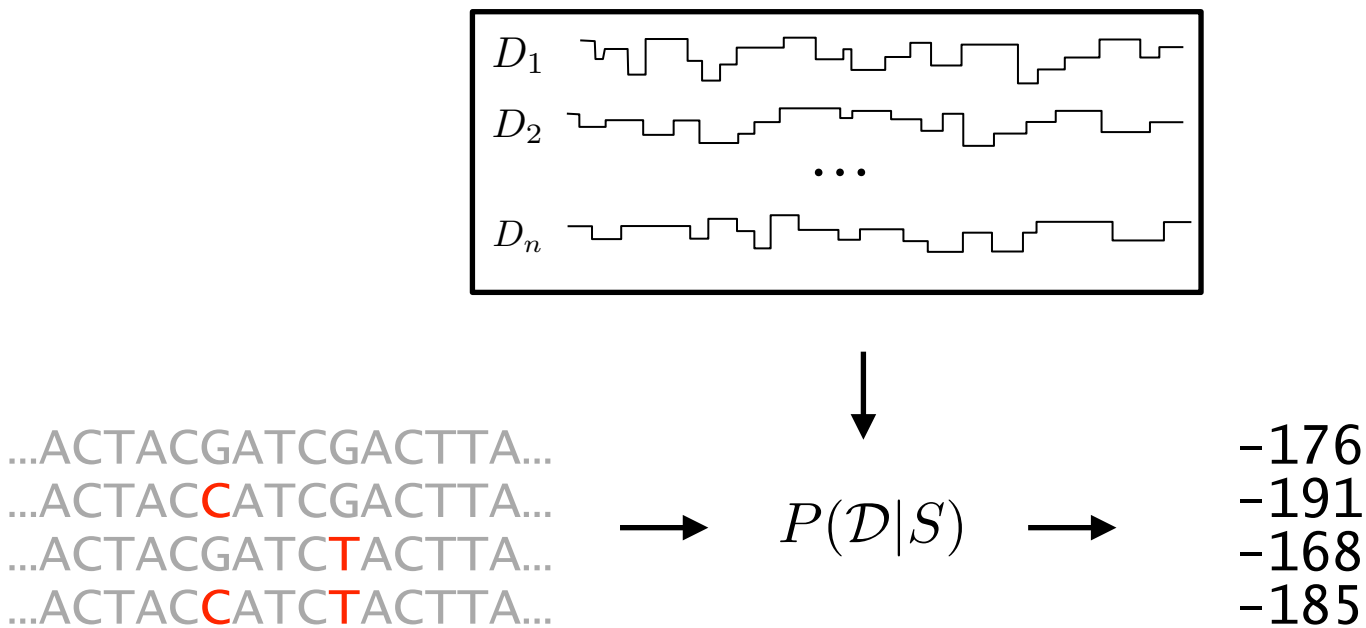
Event levels are k-mer dependent



Pore Models

6-mer	μ_k	σ_k
AAAAAA	54.2	1.2
AAAAAC	56.4	0.9
…	…	…
TTTTTG	61.7	1.1
TTTTTT	62.3	0.8

SNP Calling Pipeline

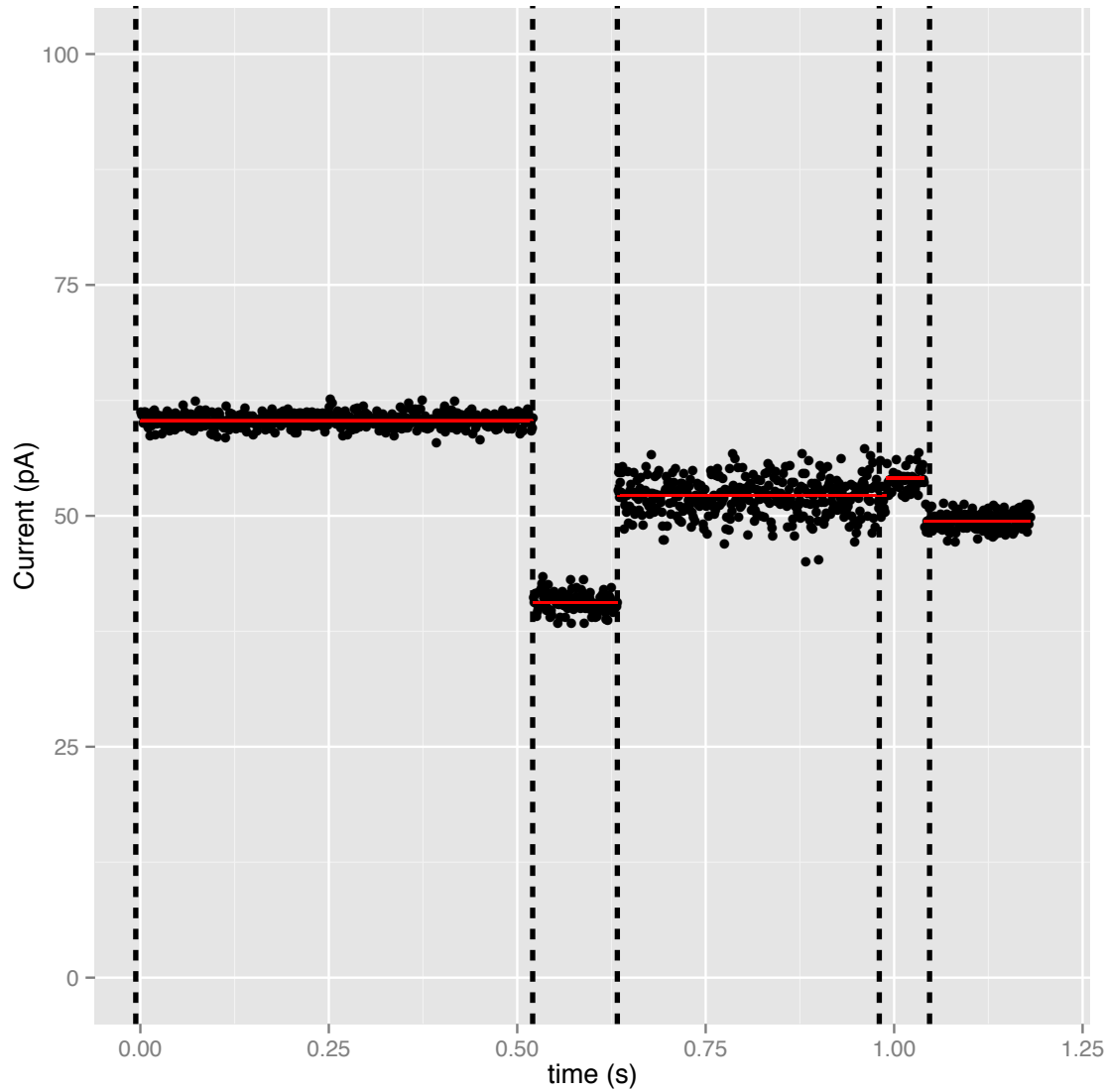


We need a function $P(\mathcal{D}|S)$ that calculates the probability of observing some nanopore data (D) given a proposed sequence (S).

A first model

Unrealistic assumption:

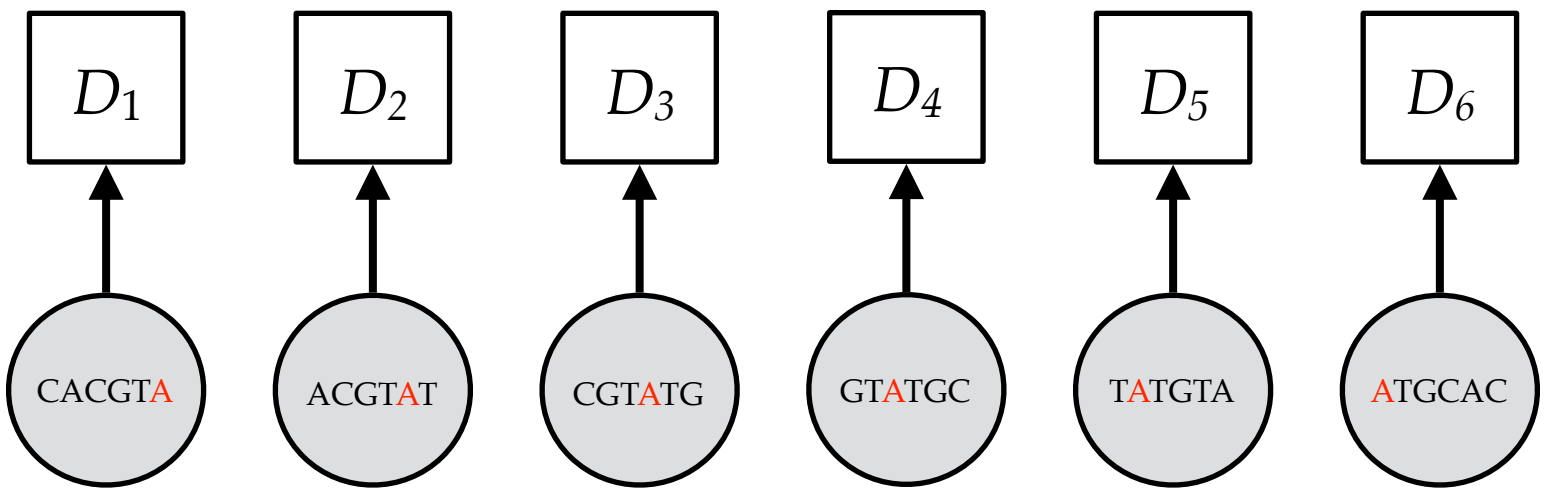
- every base generates exactly one event



A first model

Unrealistic assumption:

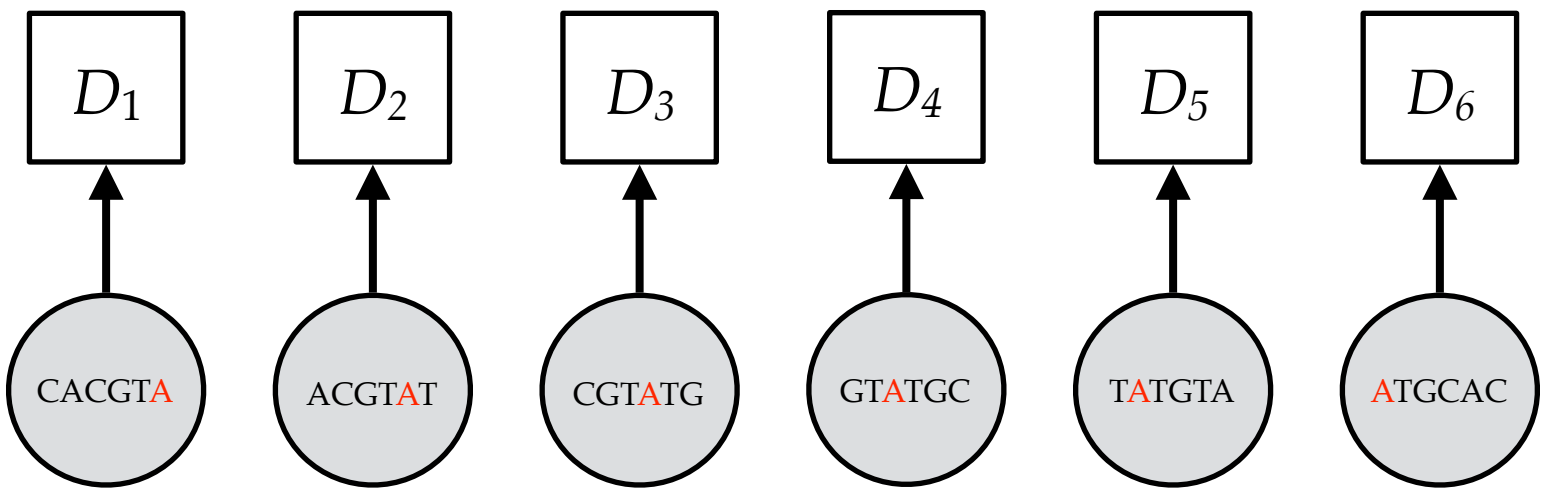
- every base generates exactly one event



A first model

Unrealistic assumption:

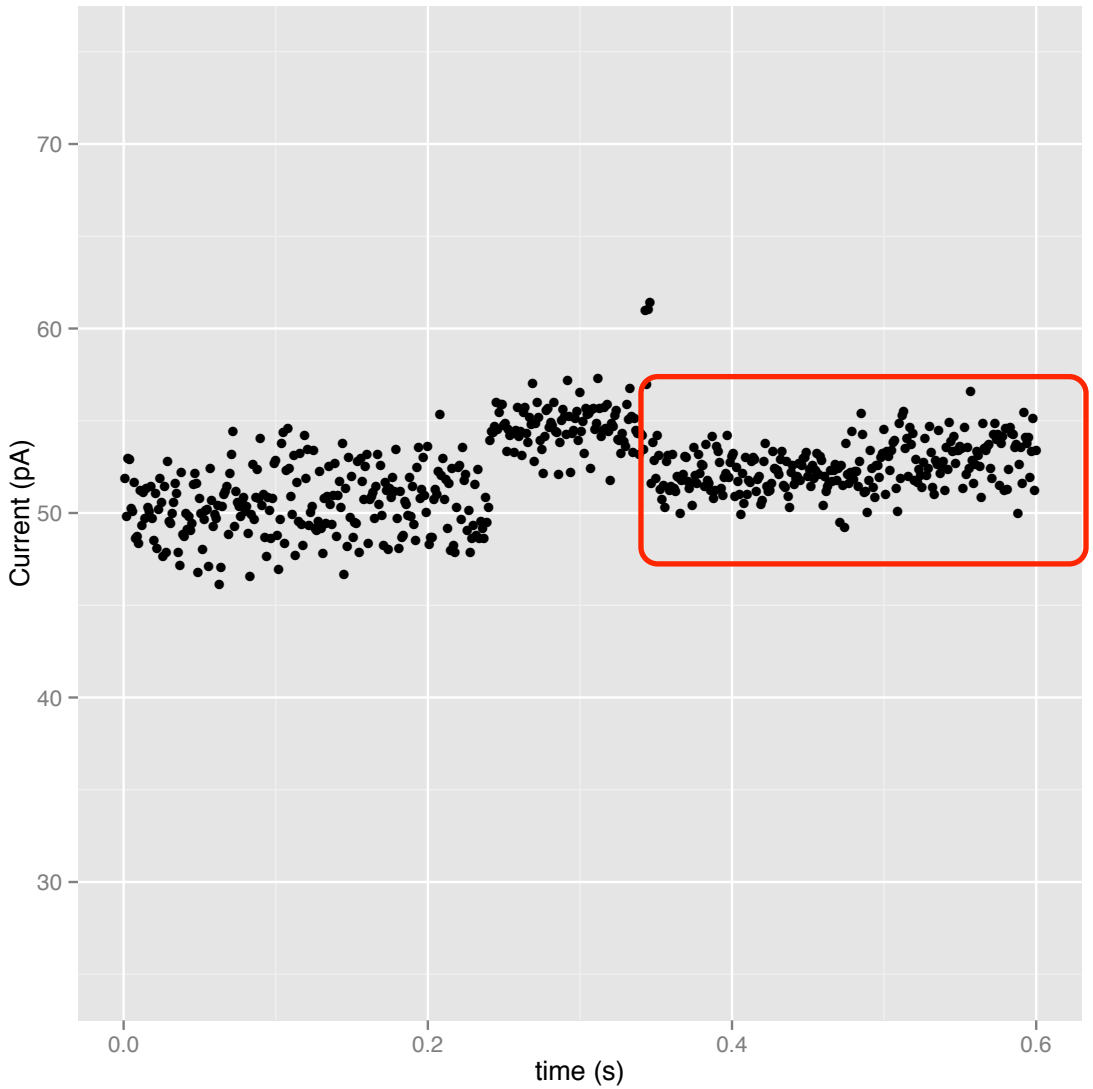
- every base generates exactly one event



$$P(\mathcal{D}|S_h) = \prod_{i=1}^{n} P(D_i|S_{h,i})$$

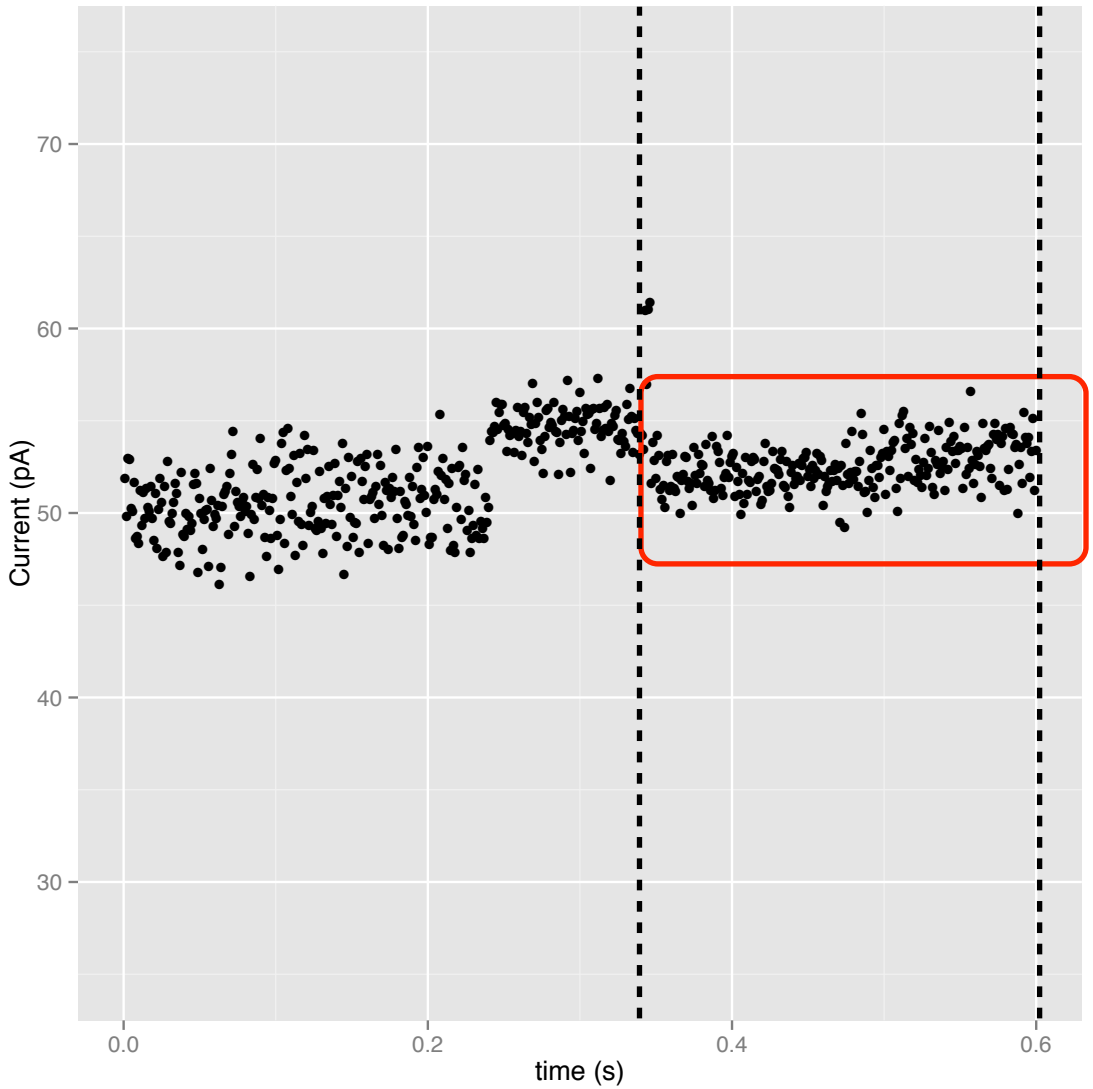
$$P(\mathcal{D}_i|S_{h,i}) = \mathcal{N}(\mu_{S_{h,i}}, \sigma^2_{S_{h,i}})$$

Complications - Segmentation Errors



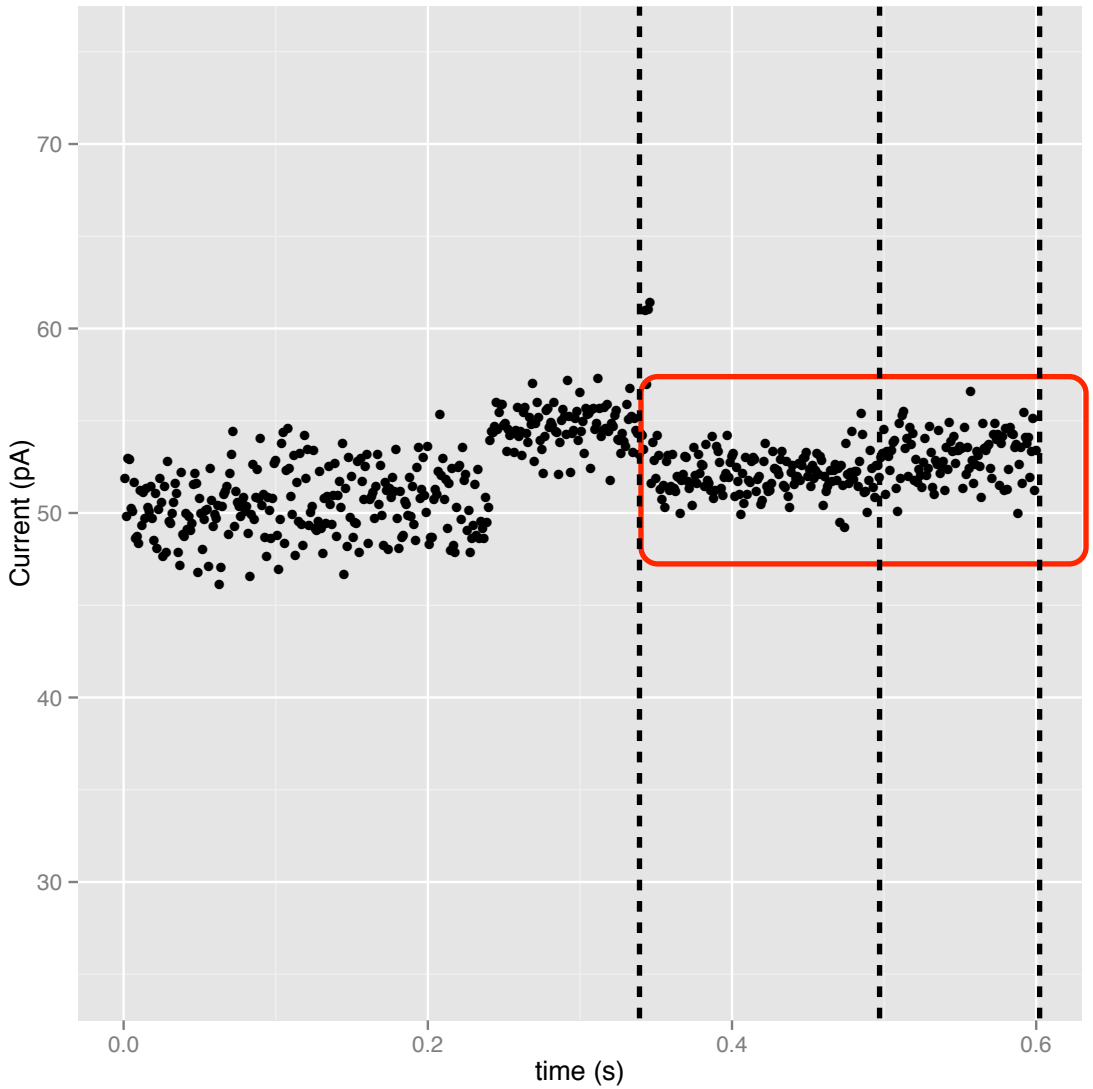
One event or two ?

Complications - Segmentation Errors



One event or two ?

Complications - Segmentation Errors

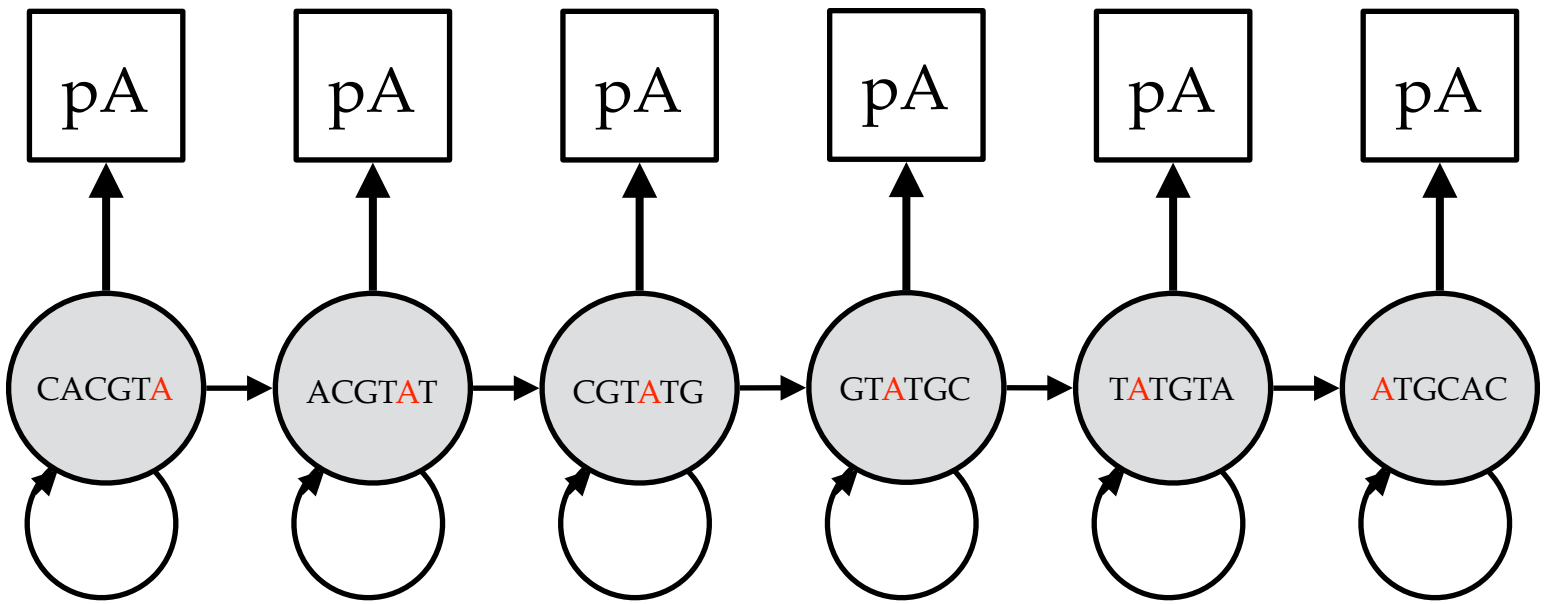


One event or two ?

Hidden Markov Model

Slightly more realistic assumption:

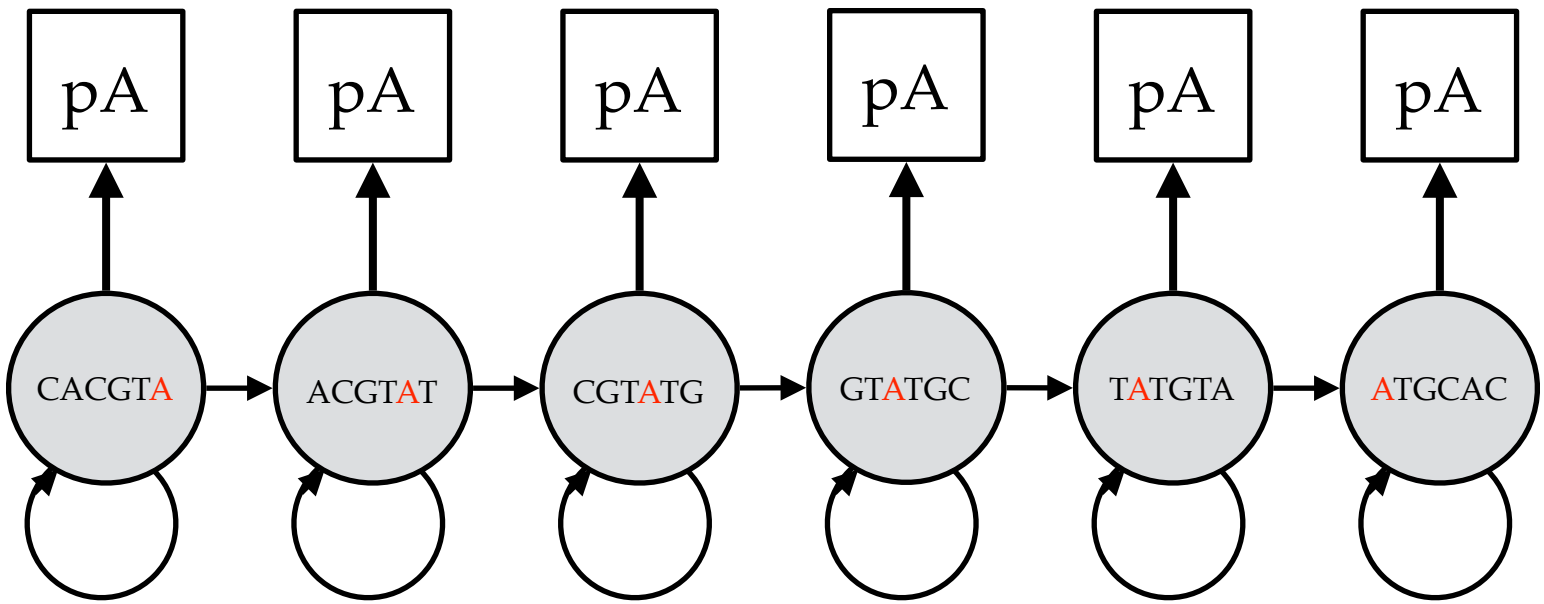
- every base generates at least one event



Hidden Markov Model

Slightly more realistic assumption:

- every base generates at least one event

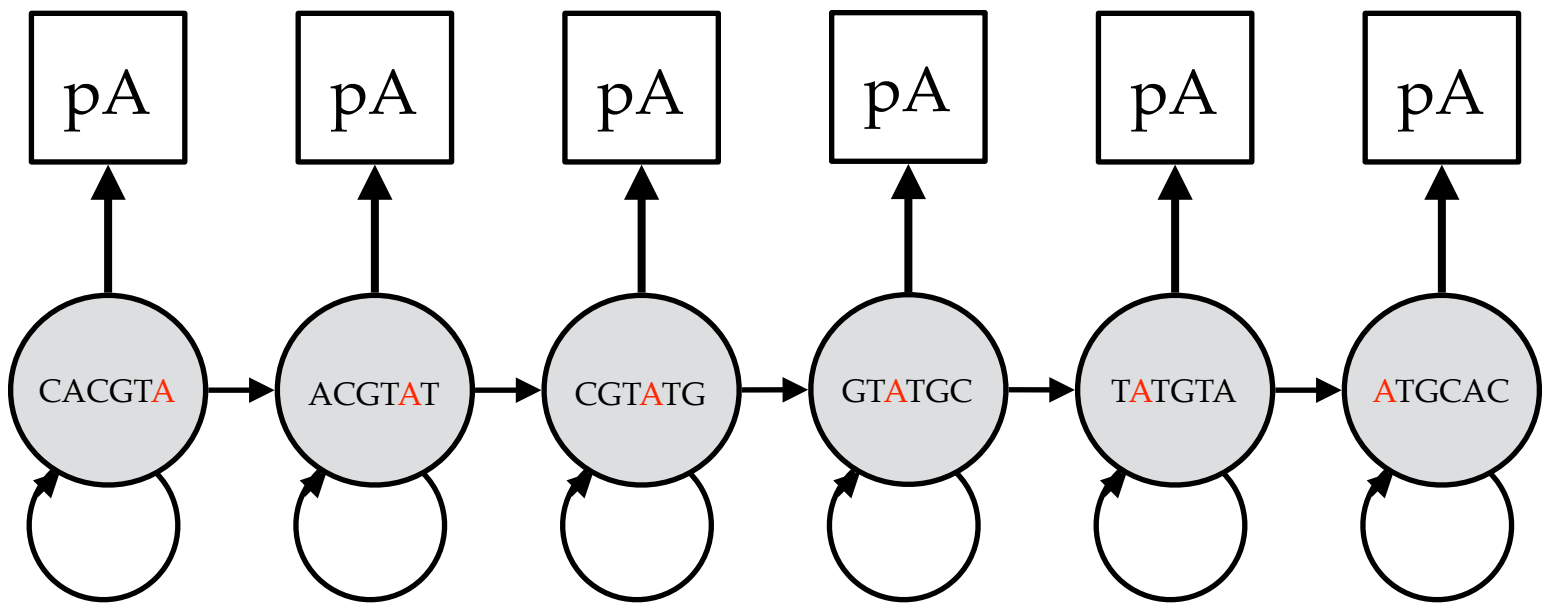


$$P(\pi, D|S_h) = \prod_{i=1}^{n} P(D_i|\pi_i, \mu_{S_{h,i}}, \sigma_{S_{h,i}}) P(\pi_i|\pi_{i-1}, S_h)$$

Hidden Markov Model

Slightly more realistic assumption:

- every base generates at least one event



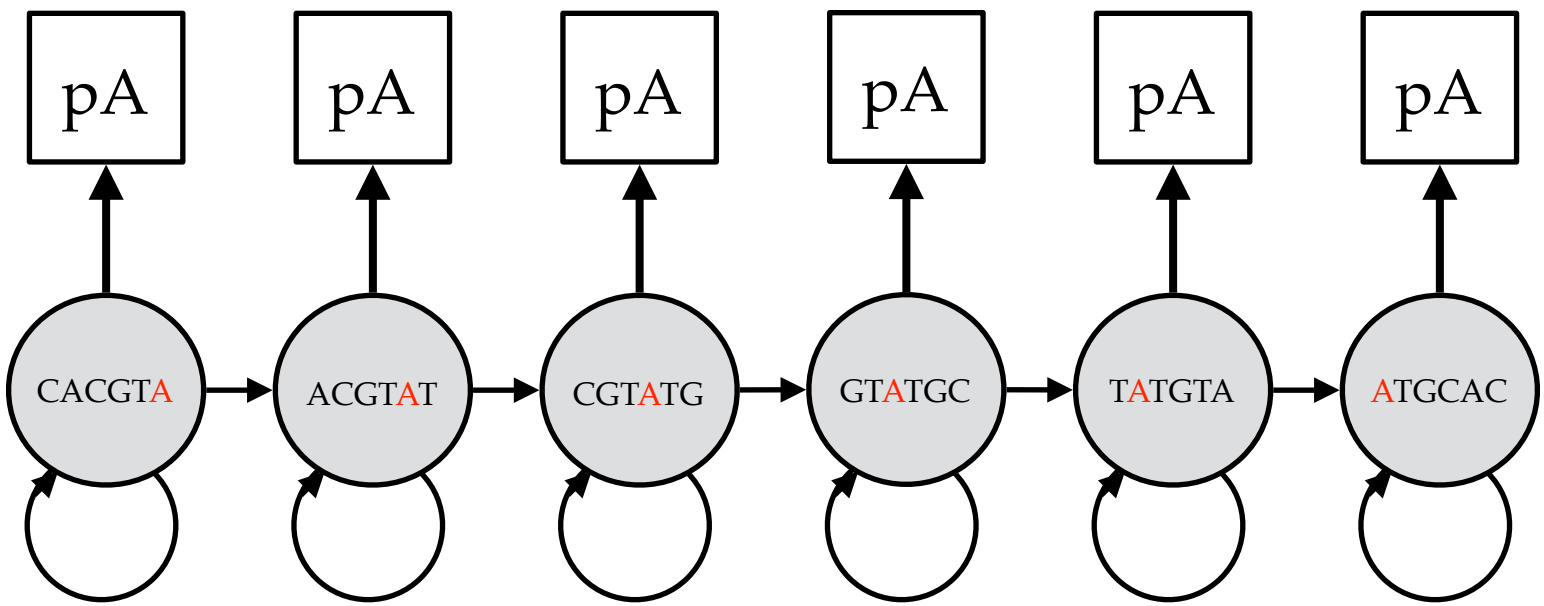
$$P(\pi, D|S_h) = \prod_{i=1}^{n} P(D_i|\pi_i, \mu_{S_{h,i}}, \sigma_{S_{h,i}}) P(\pi_i|\pi_{i-1}, S_h)$$

State path Emission function Transition function

Hidden Markov Model

Slightly more realistic assumption:

- every base generates at least one event



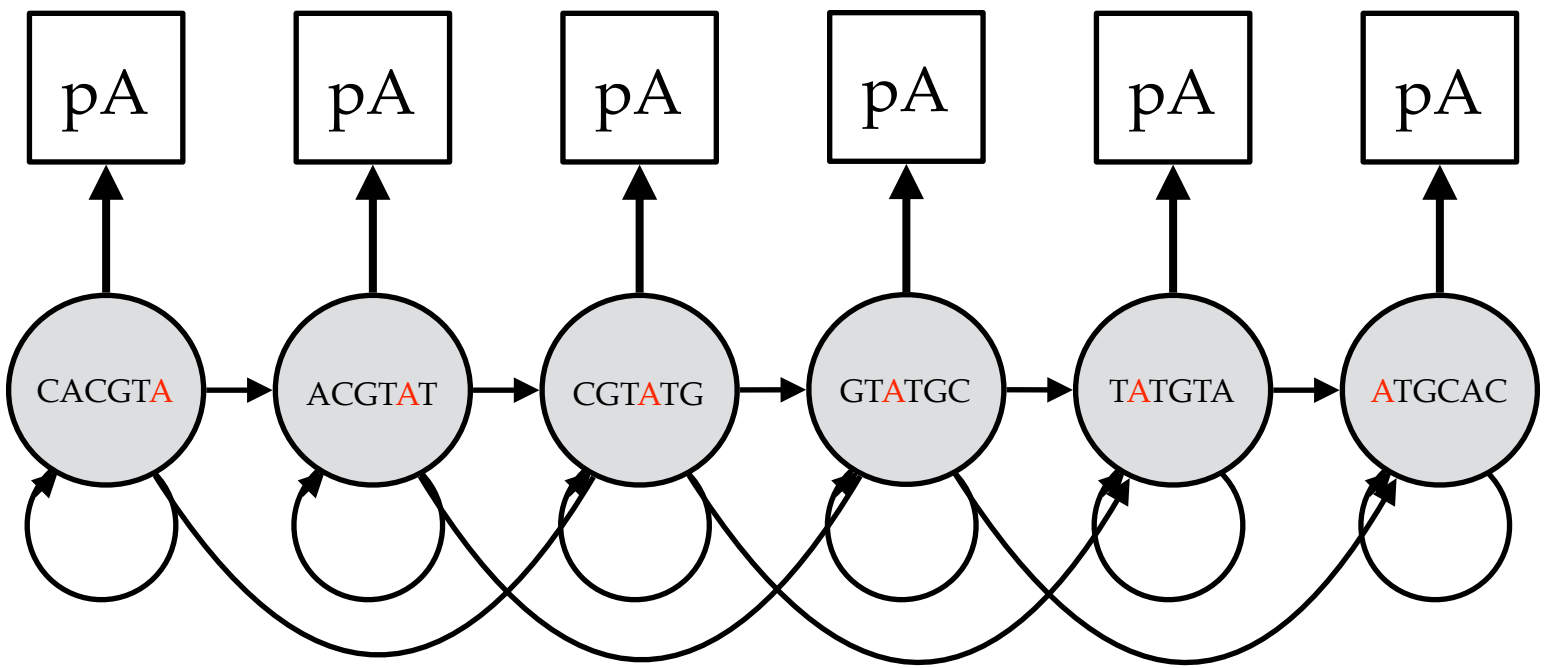
$$P(\pi, D|S_h) = \prod_{i=1}^{n} P(D_i|\pi_i, \mu_{S_{h,i}}, \sigma_{S_{h,i}}) P(\pi_i|\pi_{i-1}, S_h)$$

$$P(D|S_h) = \sum_{\pi} P(\pi, D|S_h)$$

Realistic model

Assumptions:

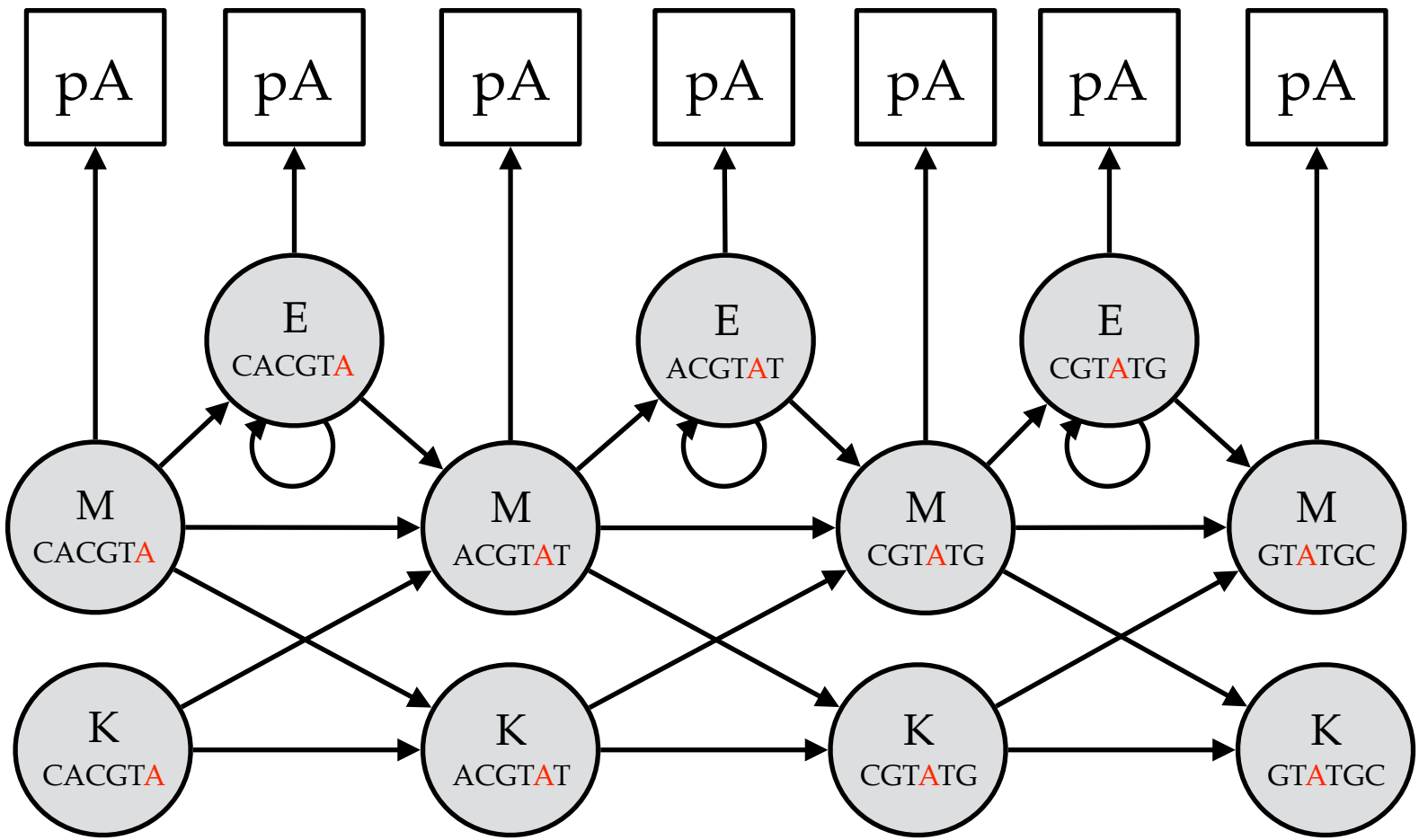
- ~~every base generates at least one event~~



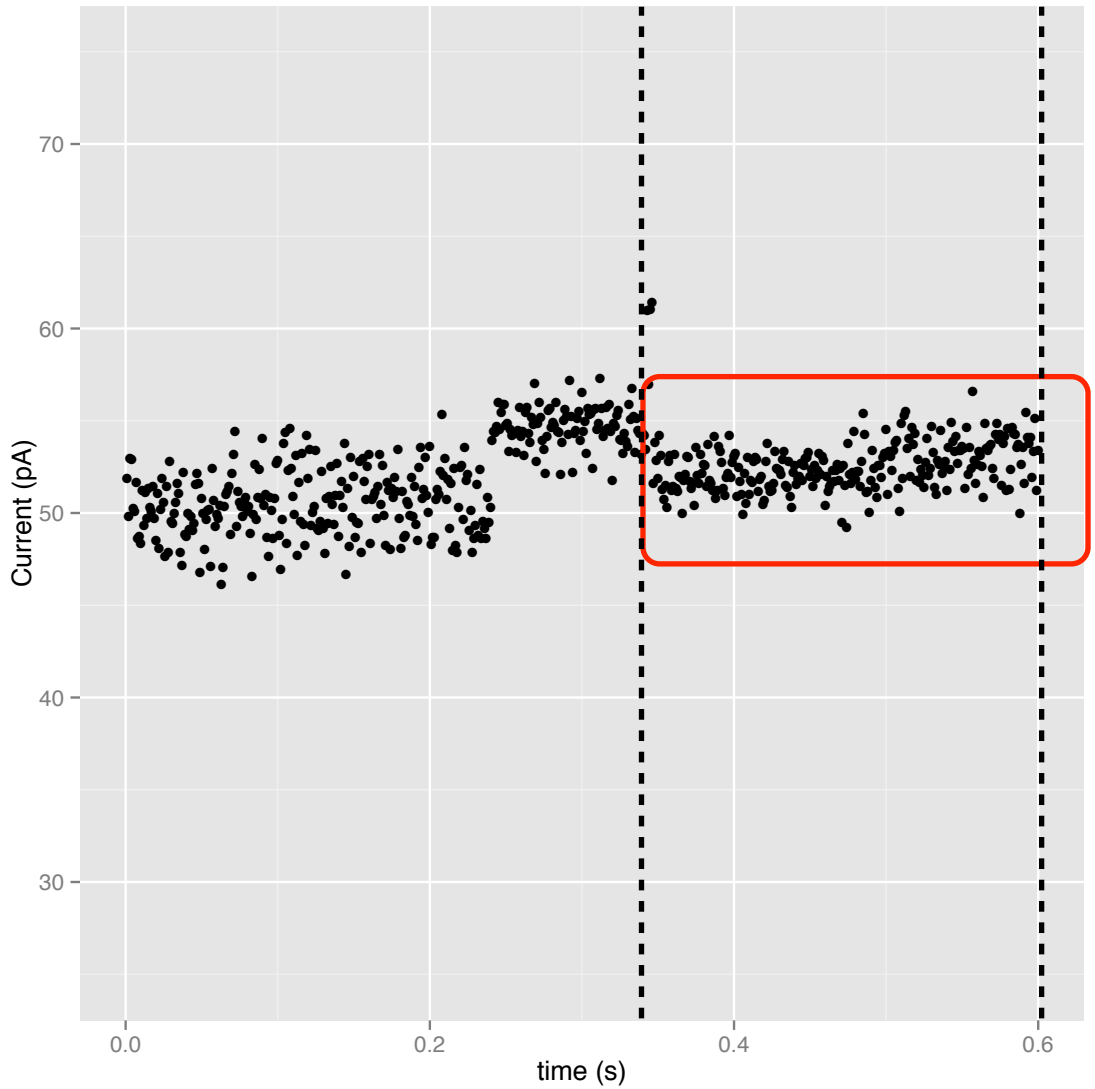
Realistic model

Assumptions:

- ~~every base generates at least one event~~



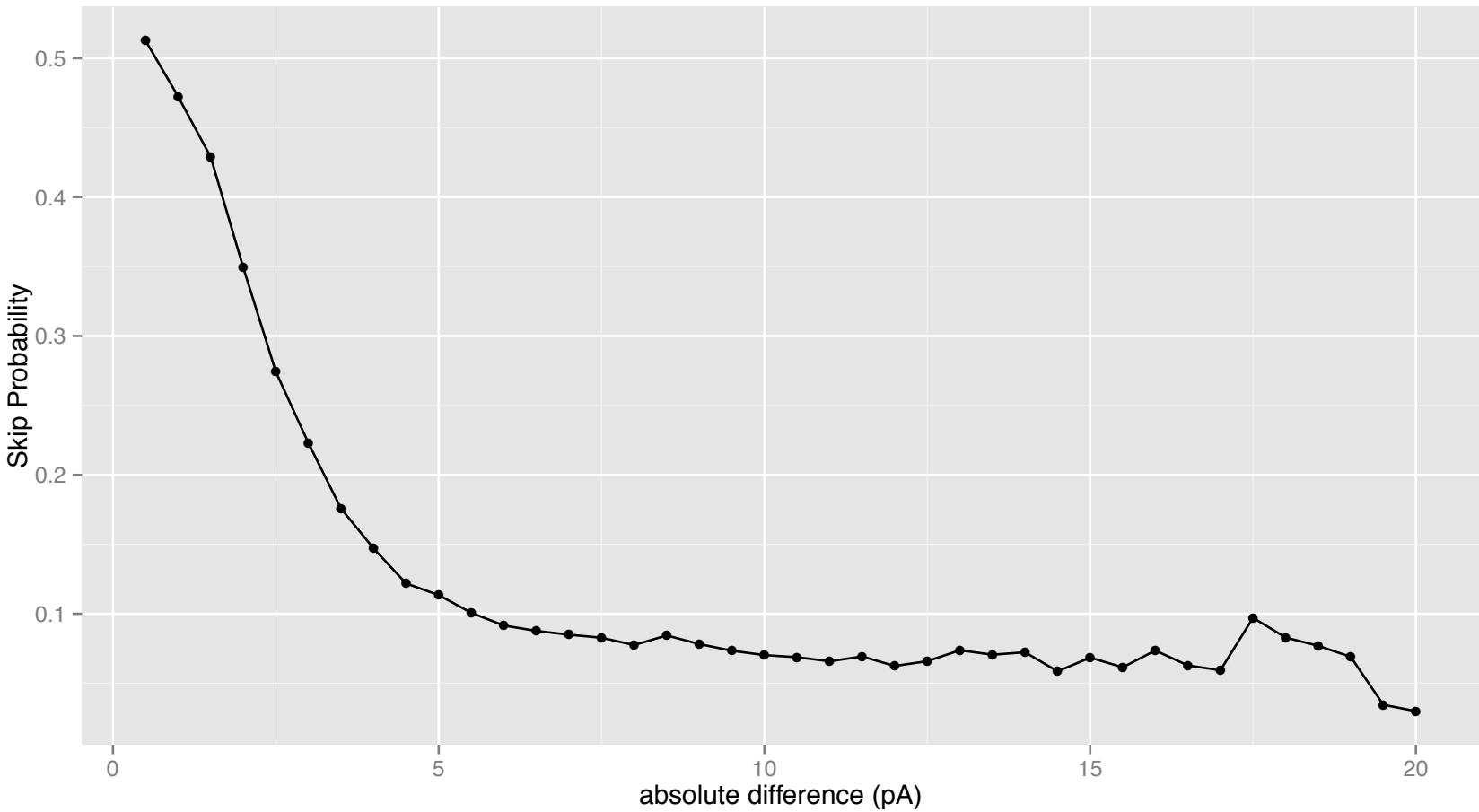
Transition Probabilities



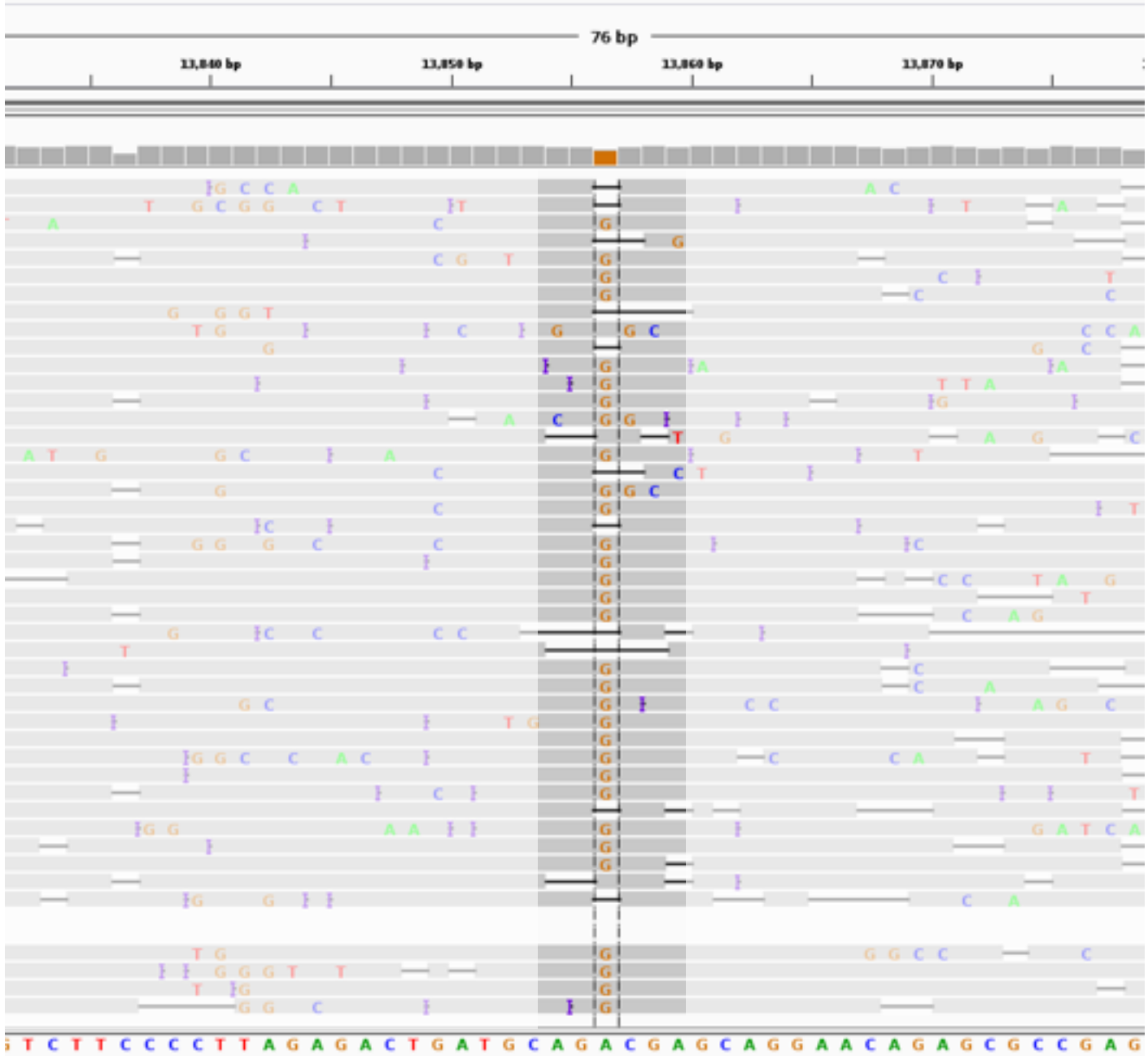
One event or two ?

Transition Probabilities

- Transition probabilities depend on how similar the event levels of adjacent *k*-mers are:

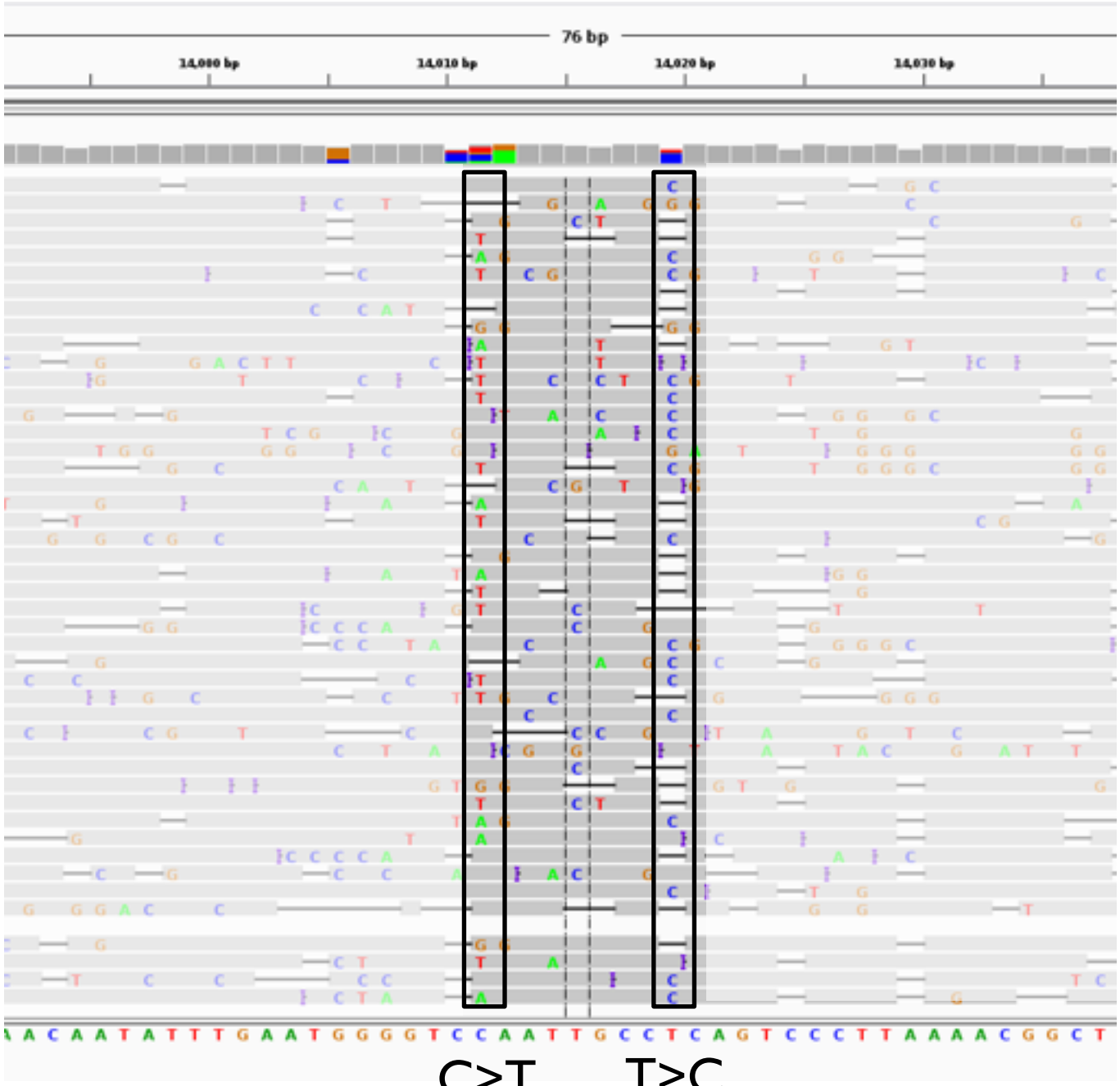


SNP Calling



A>G

SNP Calling



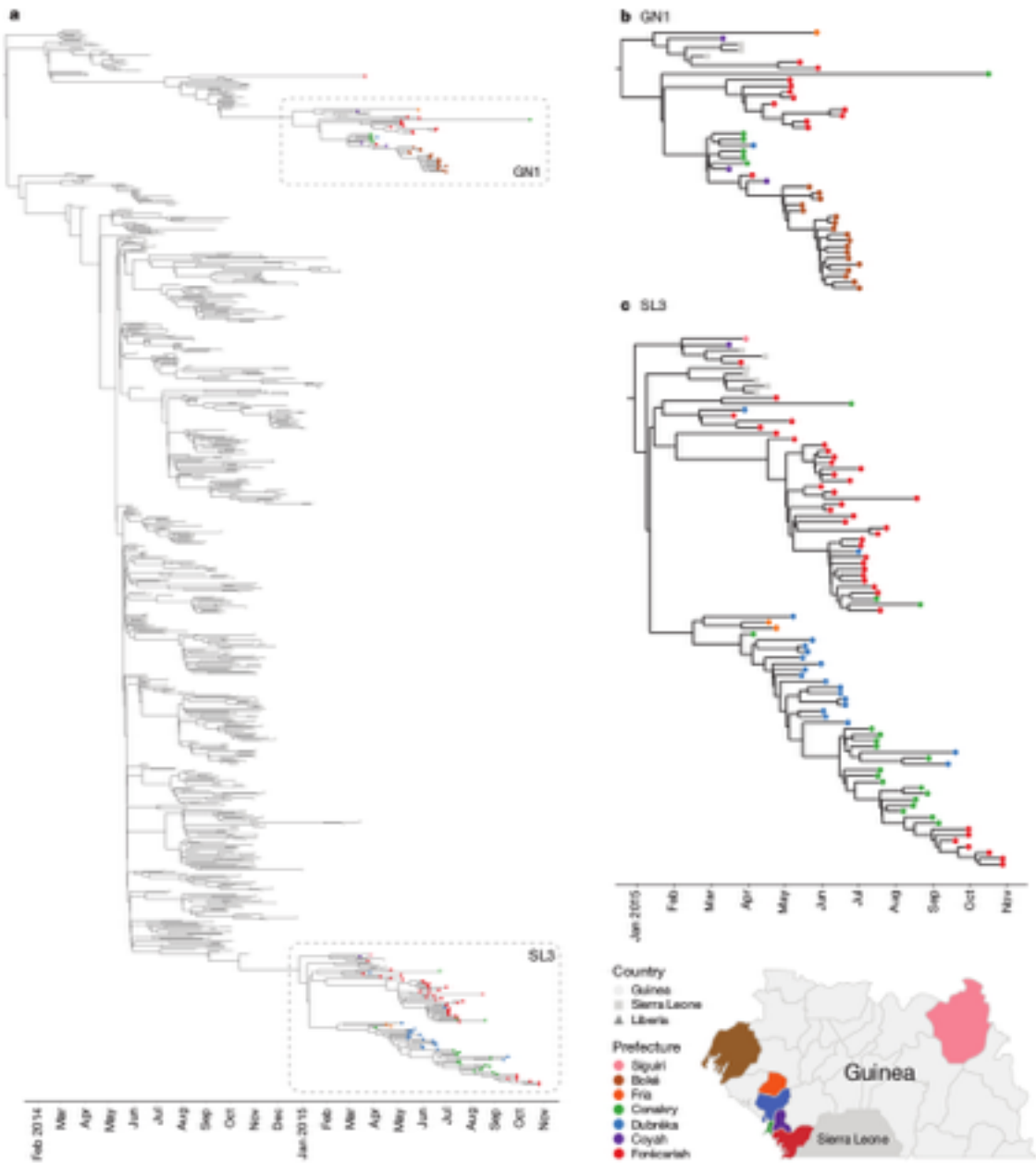
Ebola Genomes

We sequenced and reconstructed 142 Ebola genomes using the MinION and our hidden Markov model-based analysis software (nanopolish)

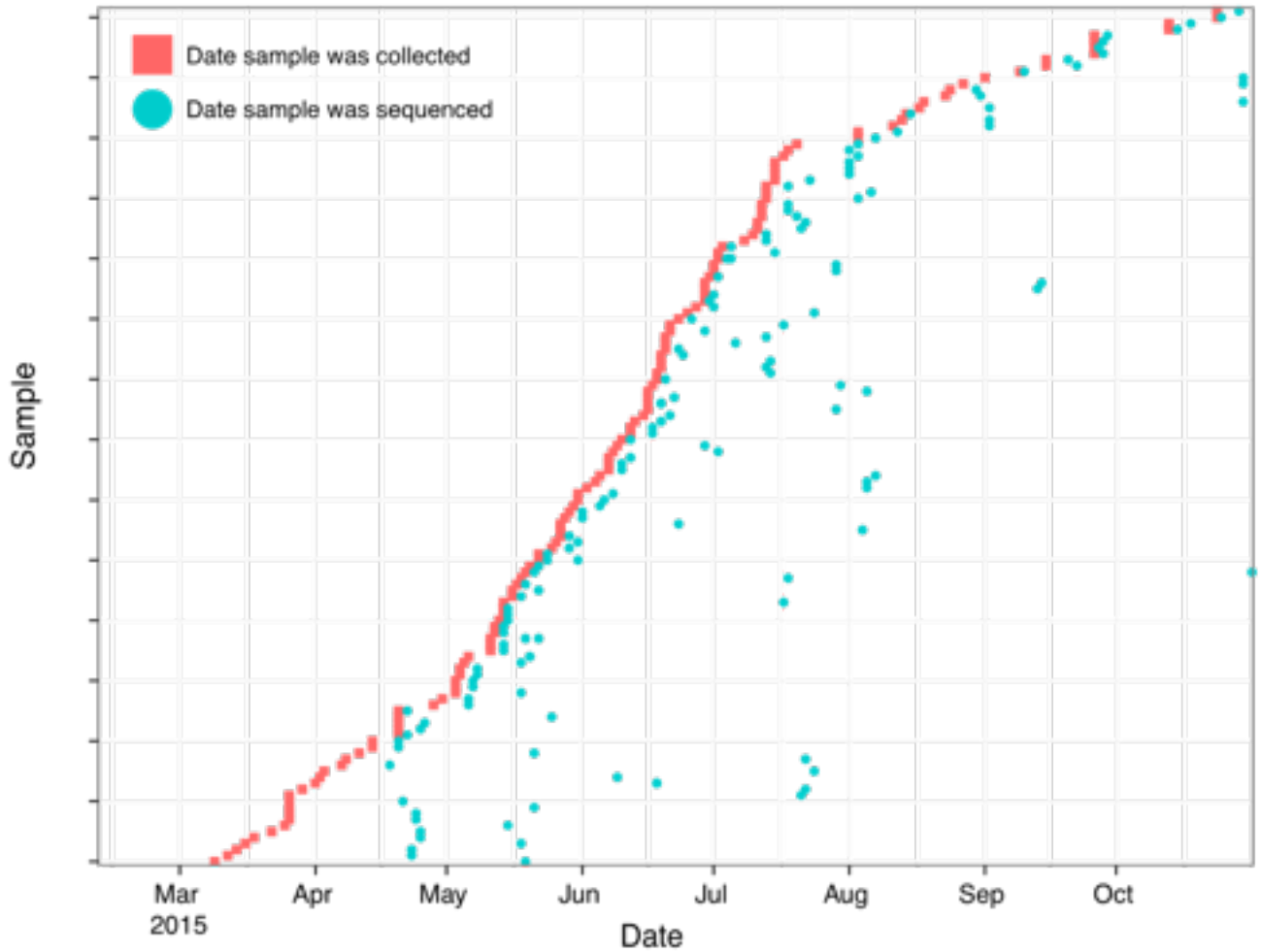
```
Ebola-1     ...AGTAGCCTACGATACTACGATCGACTTA...
Ebola-2     ...AGTAGCCTACGATATTACGATCGACTTA...
Ebola-3     ...AGTAGCCGACGATACTACGATCGACTTA...
Ebola-4     ...AGTTGCCTACGATATTACGATCGACTTA...
                          ...
Ebola-141   ...AGTAGCCTACGATACTACGATCGAGTTA...
Ebola-142   ...AGTAGGCTACGATACTACGATCGACTTA...
```

Next computational challenge: Arrange these genomes into a phylogenetic tree

Ebola Phylogeny



Rapid sequencing



MRC Medical Research Council

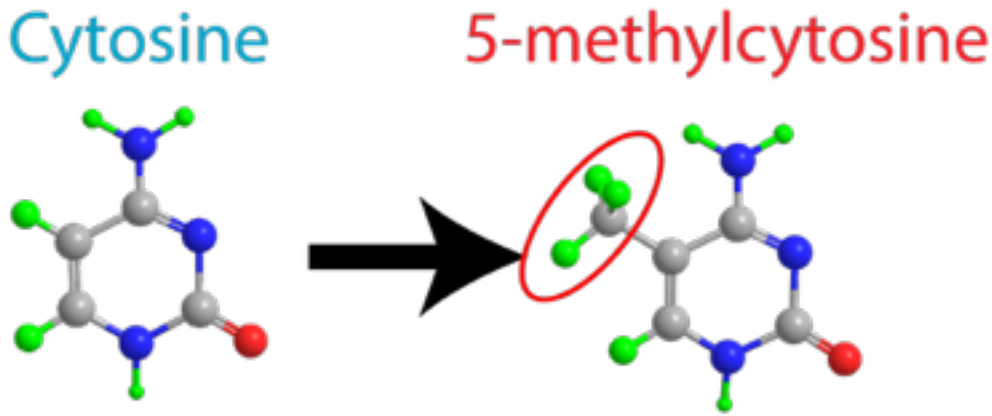
zibraproject.org

ZIBRA

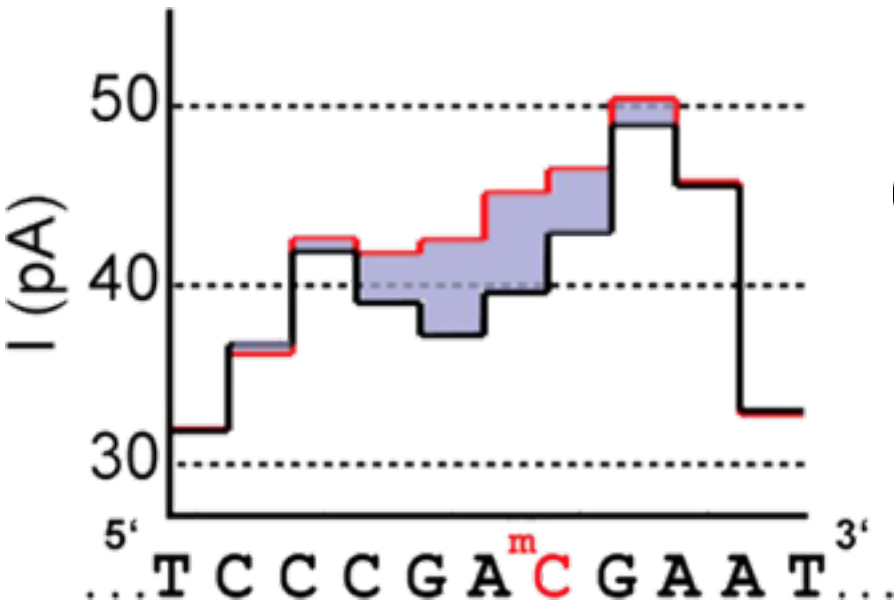
ZIKA IN BRAZIL REAL TIME ANALYSIS


CIÊNCIA NA ESTRADA

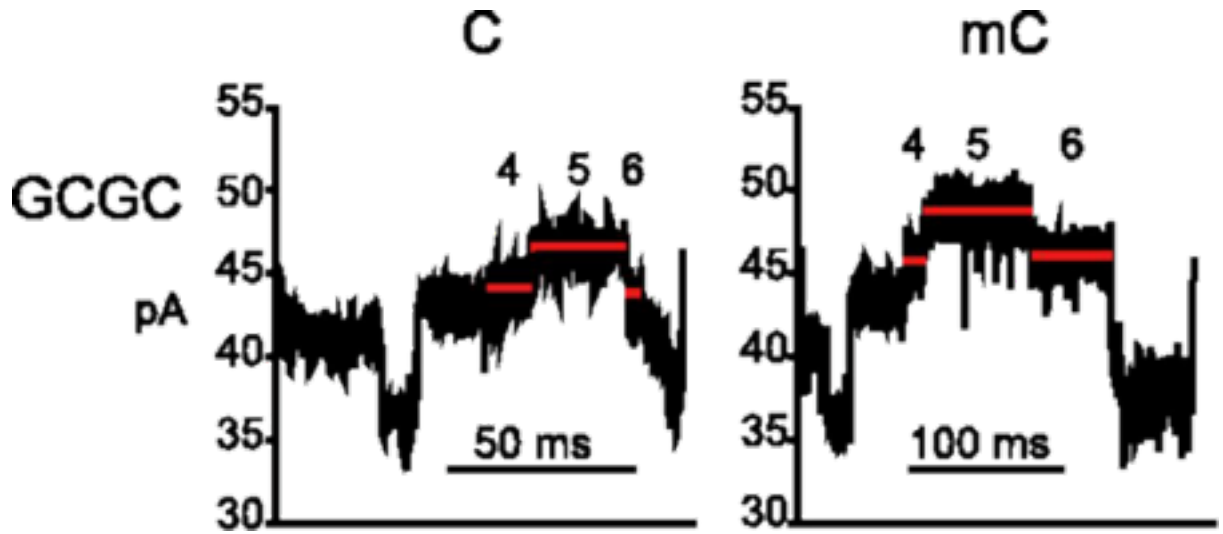
Nanopore Methylation



Laszlo, et al. *PNAS* (2013)



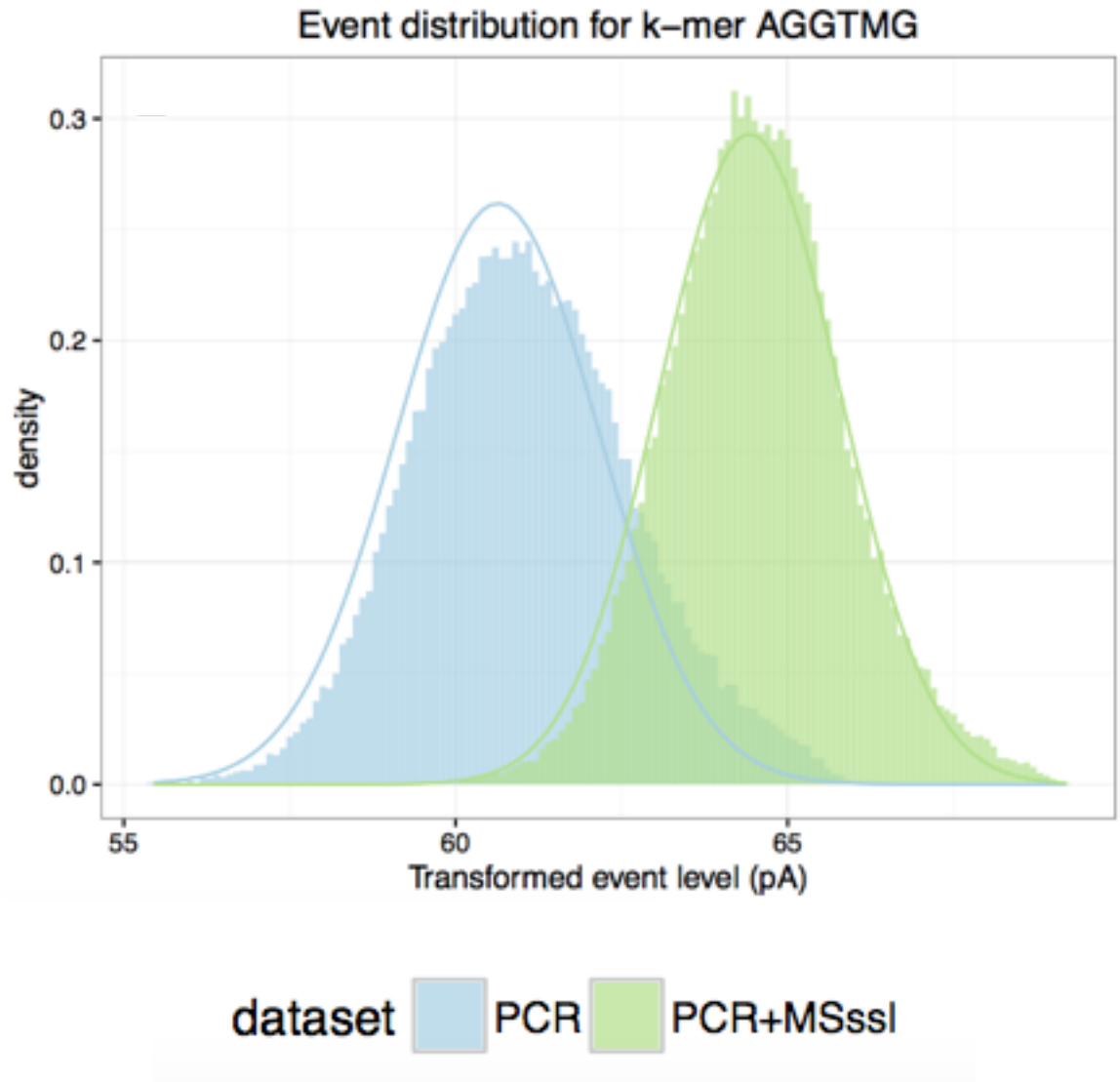
Schreiber, et al. *PNAS.* (2013)



Slide courtesy of Winston Timp

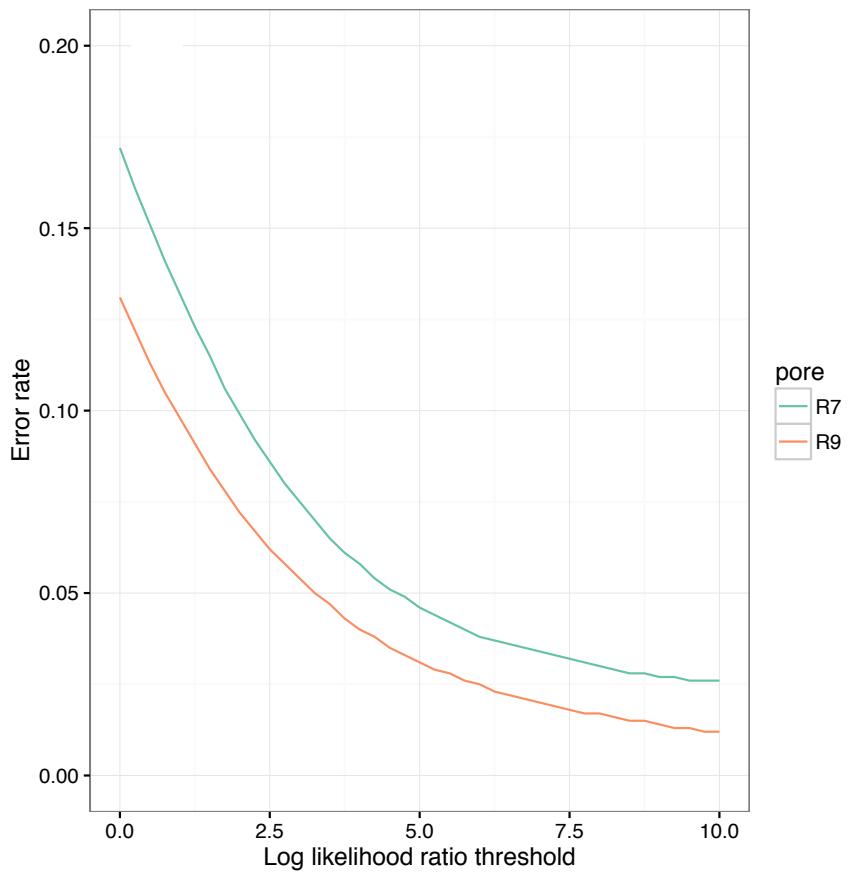
Training Methylation Models

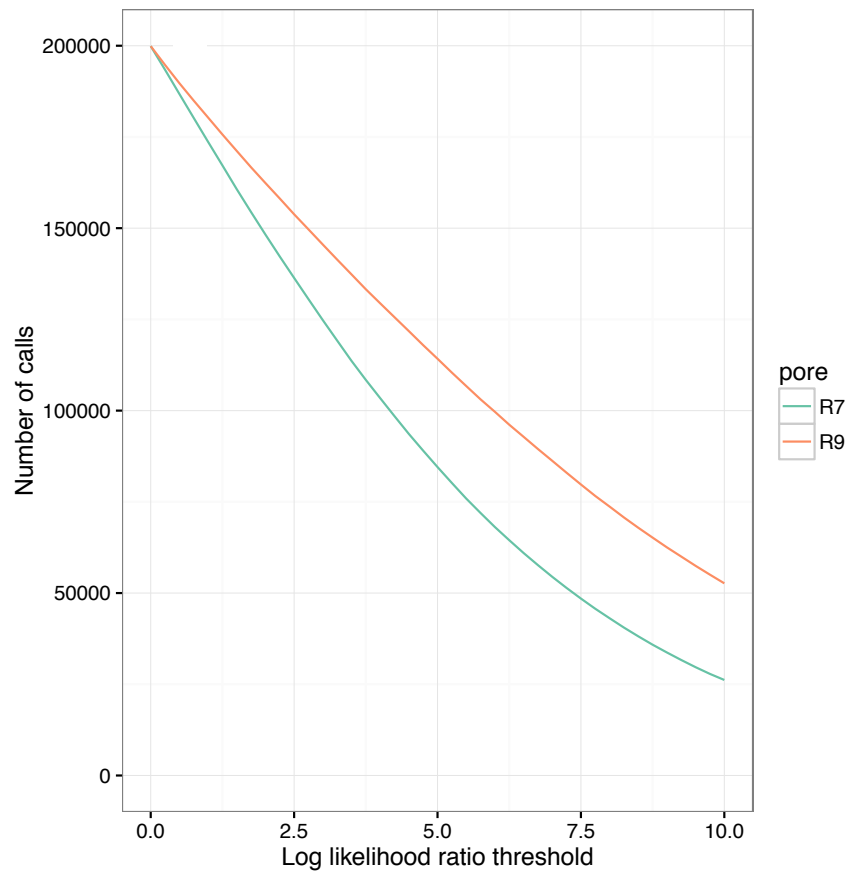
Learn emissions for *k*-mers over expanded alphabet using synthetically methylated DNA (w/ M.SssI)



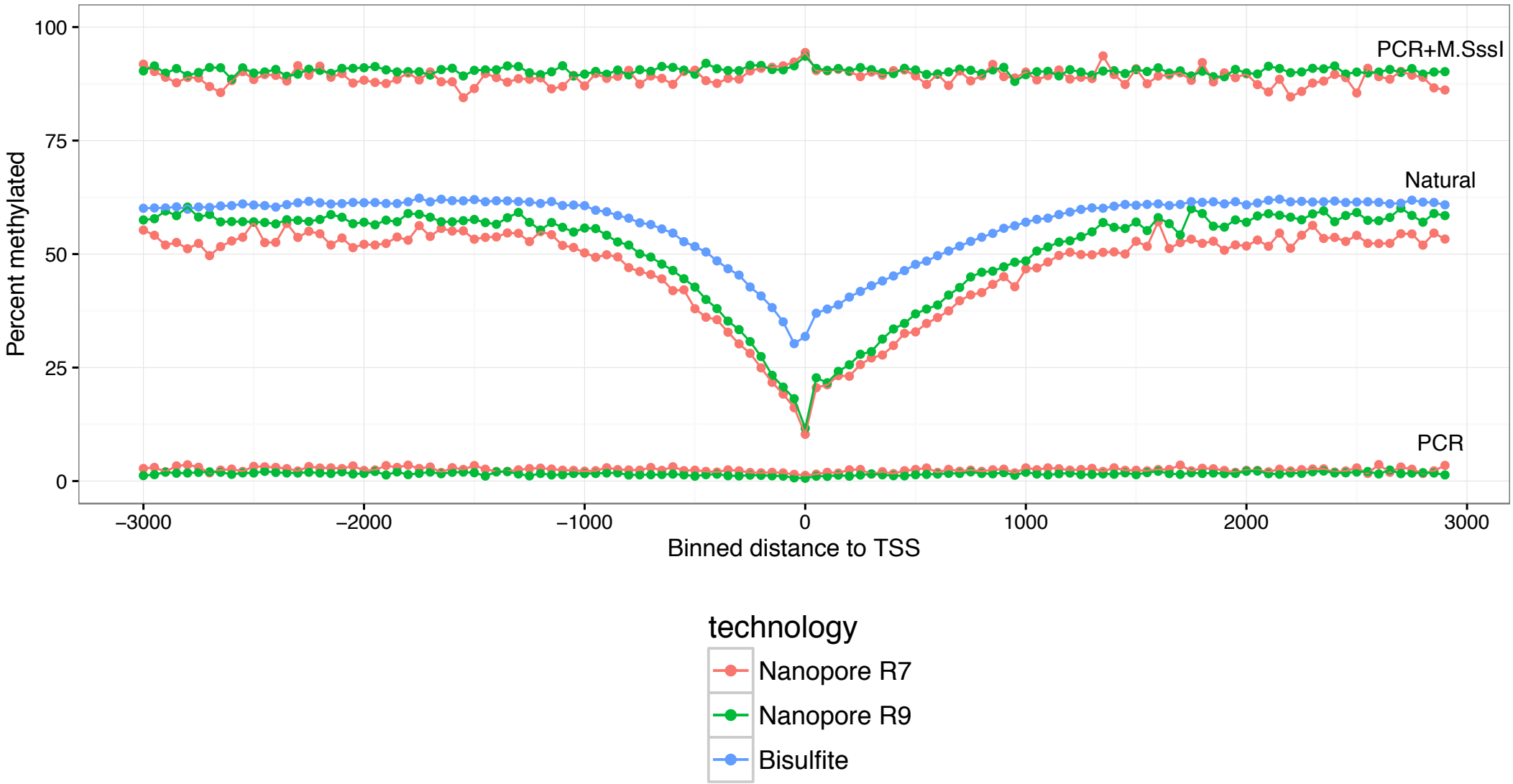
Evaluating Accuracy

- We test our model by calling methylation at randomly sampled sites from the human positive/negative controls
- Accuracy: 82% (R7) 87% (R9)





Low CpG methylation near TSS



Consensus accuracy improvement

Version	Coverage	Percent Identity	# Mismatches	# Indels
R7-SQK005	29X	99.47%	1,363	22,702
R7-SQK006	30X	99.81%	239	8,659
R9-SQK007	30X	99.89%	693	4,467
R9-SQK007	88X	99.95%	325	2,263

* nanopolish after CA/canu assembly

Summary/Outlook

- Overview of portable sequencing & applications
- Data is challenging to work with but similar to classic machine learning problems
- Human genome sequencing now possible (four human genomes were announced last week)

Jonathan Dursi

Matei David

Phil Zuzarte

Nick Loman

Josh Quick

Winston Timp

Rachael Workman

Funding: