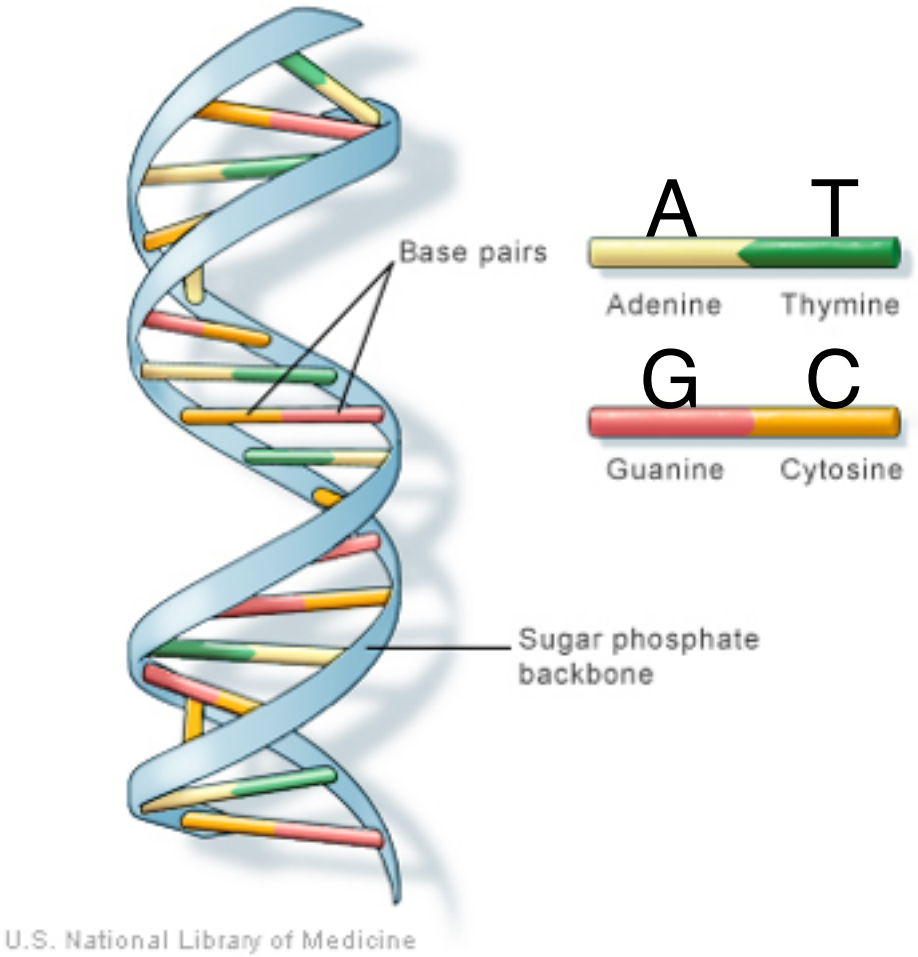
Portable DNA Sequencing: Analysis and Applications

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

Computing for Medicine
October 16th, 2018

DNA: The genome's molecule



Picture: http://ghr.nlm.nih.gov/handbook/basics/dna

Deoxyribonucleic acid

"Rungs" of DNA double-helix are base pairs. Pair combines two complementary

Complementary pairings: A-T, C-G

Single base also called a "nucleotide"

The Central Dogma

Short version:

DNA -> RNA -> Protein

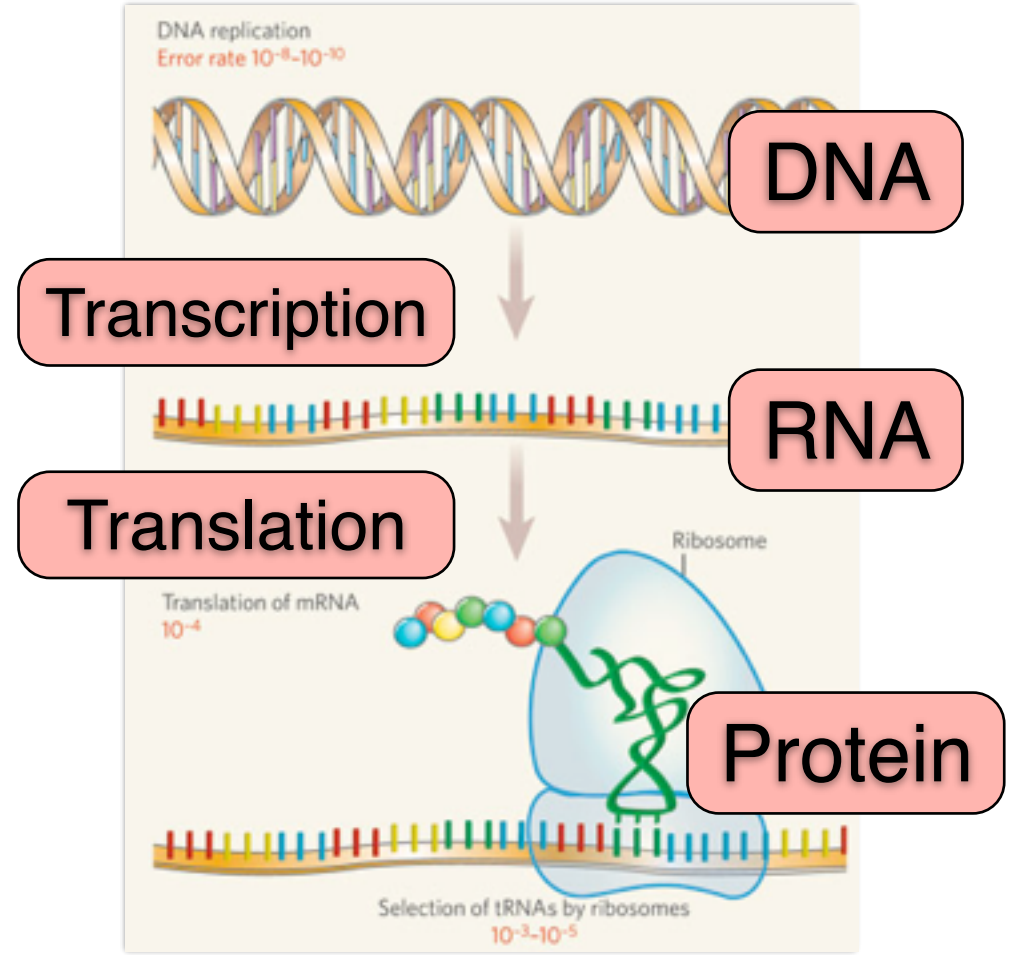
Long version:

DNA molecules contain information about how to create proteins; this information is *transcribed* into RNA molecules, which, in turn, direct chemical machinery which *translates* the nucleic acid message into a protein.

Hunter, Lawrence. "Life and its molecules: A brief introduction." *AI Magazine* 25.1 (2004): 9.

Links genotype and phenotype

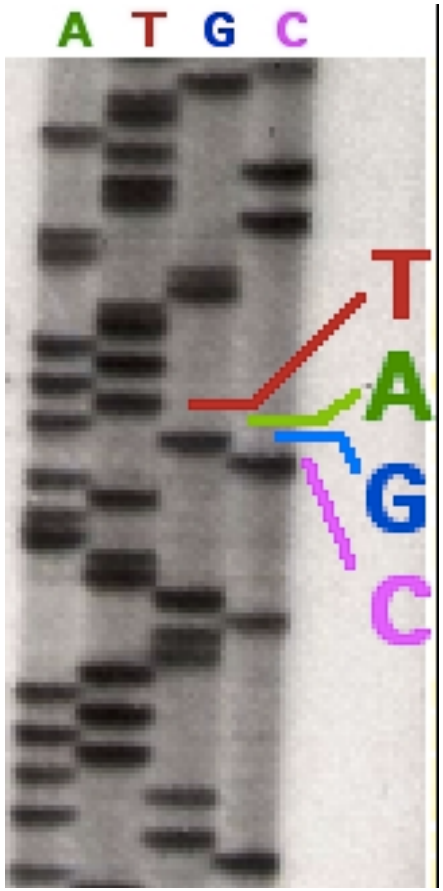
First stated by Francis Crick in 1958



Picture from: Roy H, Ibba M. Molecular biology: sticky end in protein synthesis. Nature. 2006 Sep 7;443(7107):41-2.

Timeline of DNA Sequencing

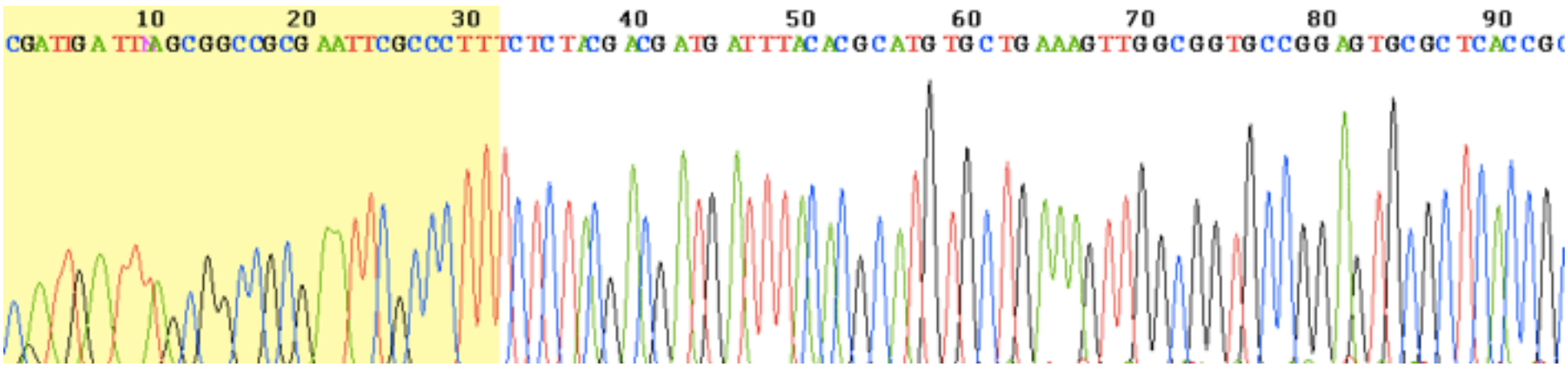
1970s
Sanger sequencing invented
Manual process; low throughput



en.wikipedia.org/wiki/Sanger_sequencing

Timeline of DNA Sequencing

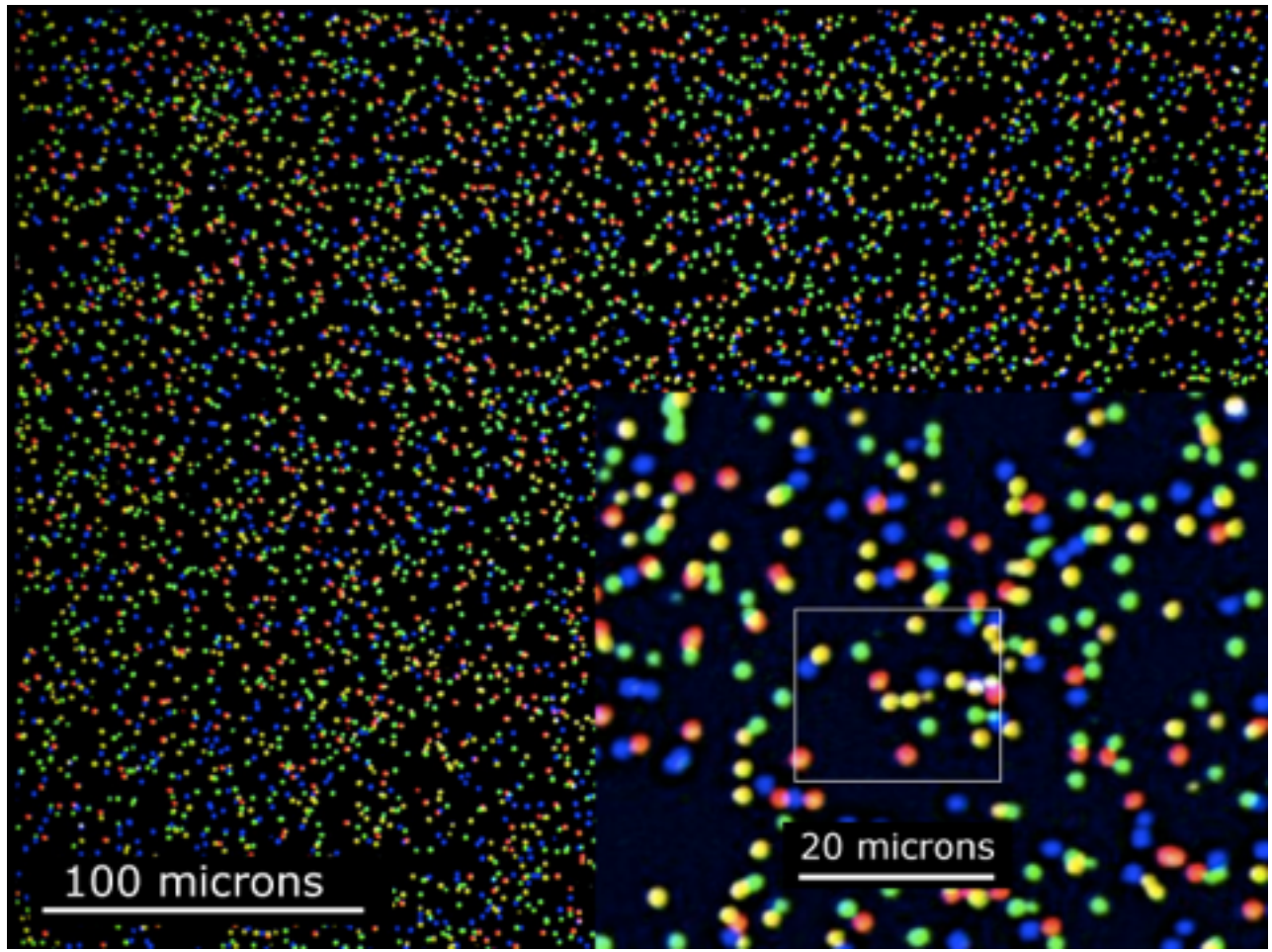
1980-1990s
Automation of Sanger sequencing
10-100 Megabases of data per run



en.wikipedia.org/wiki/Sanger_sequencing

Timeline of DNA Sequencing

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



http://library.cshl.edu/Meetings/sequencing/slide-pages/Balasubramanian.php

Timeline of DNA Sequencing

2010s

Centralisation of sequencing

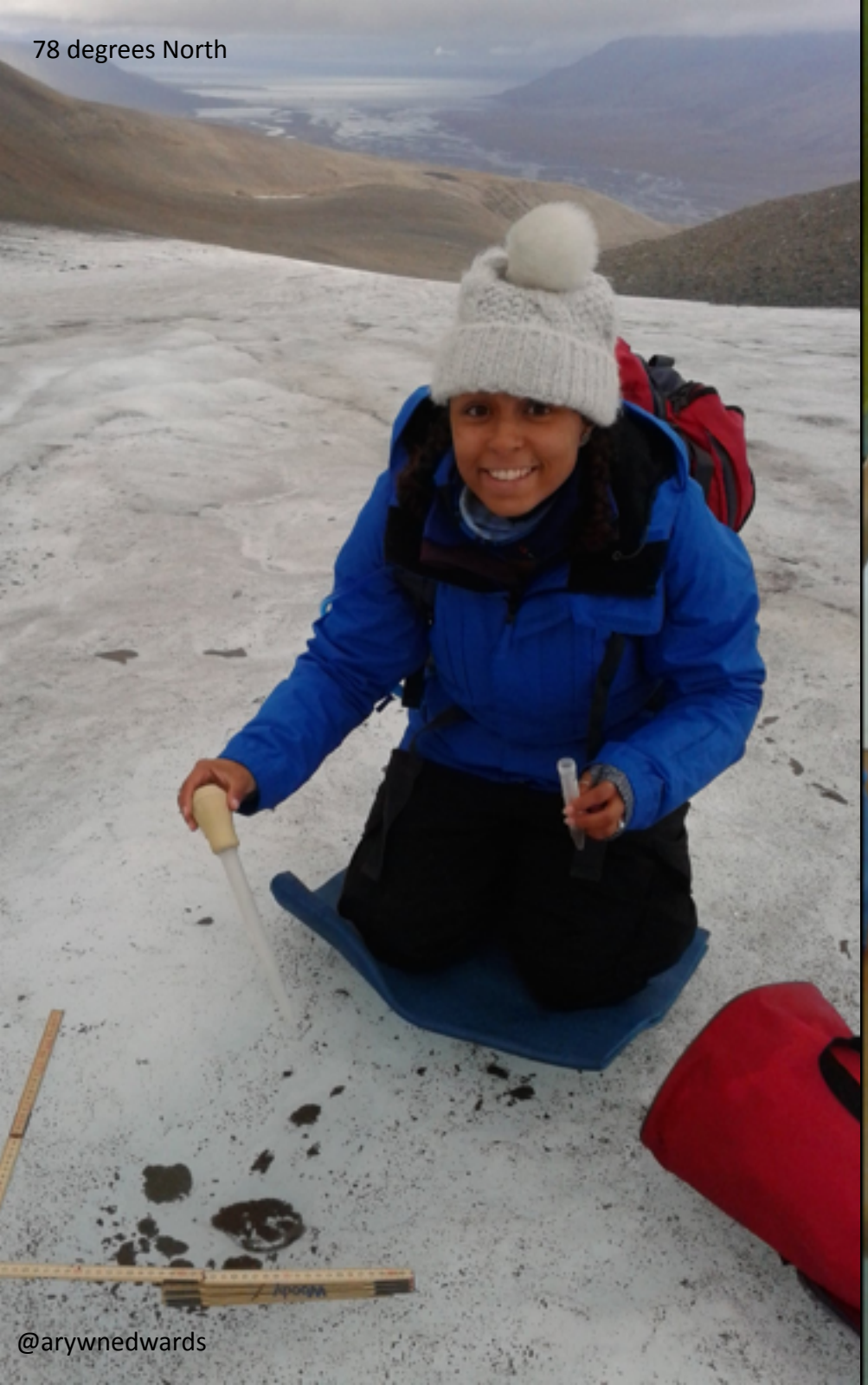
18,000 human genomes/year

https://hudsonalpha.org/collaborate/sequencing-projects

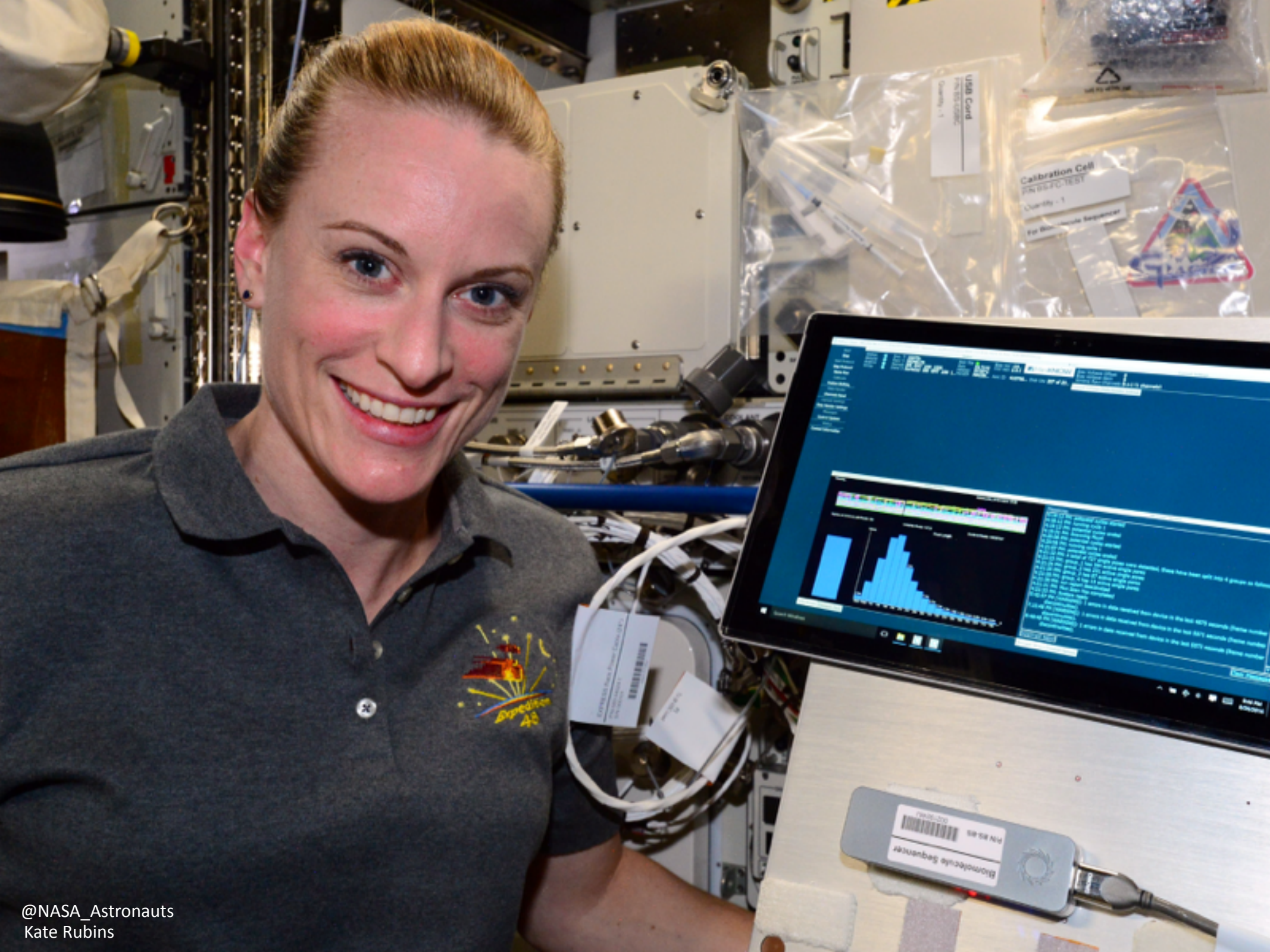
Miniature, Portable Sequencing

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


@aphilosof
porecamp.github.io


78 degrees North
@arywnedwards


MinION



@NASA_Astronauts
Kate Rubins

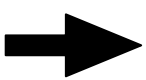
Computational Problems for DNA Sequencing

- Basecalling: infer sequence of nucleotides from instrument measurements
- Genome Assembly: reconstruct genome from sequencing data
- Phylogenetics: reconstruct evolutionary history
- Cancer Genetics: discover mutations that cause tumours
- …and many others!

Sequencing for infectious disease

Diagnosis	Pathogen detection & discovery Polymicrobial infections
Phenotypic prediction	Antimicrobial resistance Prediction of virulence
Source attribution	Human or animal reservoir? Place?
Cluster and transmission chain inference	Outbreak or pseudo-outbreak? How are cases linked?
Response to infection & treatment	Evolution of resistance Host response to infection

Sequencing for infectious disease



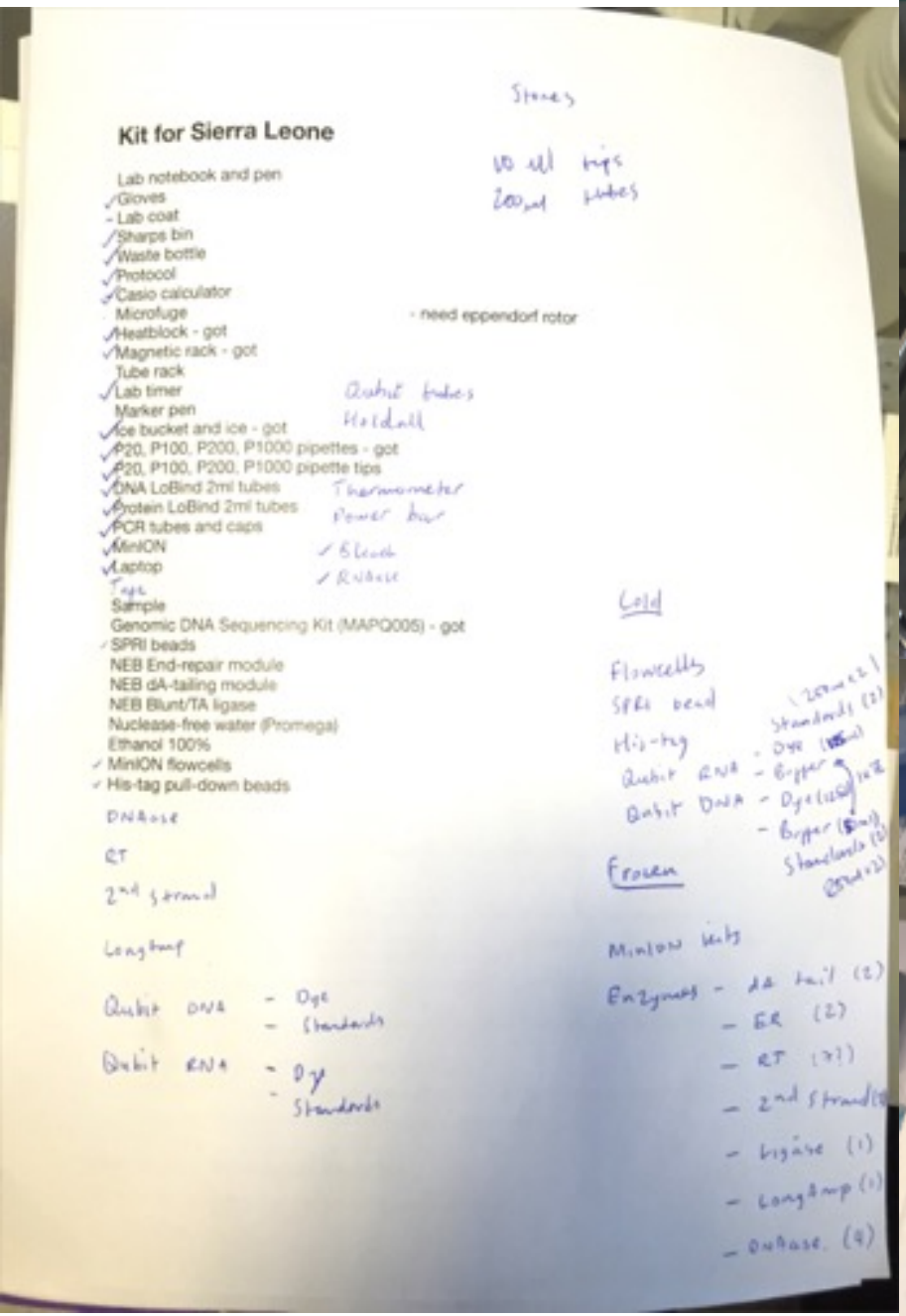
Diagnosis	Pathogen detection & discovery Polymicrobial infections
Phenotypic prediction	Antimicrobial resistance Prediction of virulence
Source attribution	Human or animal reservoir? Place?
Cluster and transmission chain inference	Outbreak or pseudo-outbreak? How are cases linked?
Response to infection & treatment	Evolution of resistance Host response to infection

How can we sequence outbreaks?

- **Central Model**: take local samples, send to lab in Europe/North America, sequence, return results to local epidemiologists
- **Local model**: use portable sequencers directly at the location they are needed

Image credit: Genome Research Limited

Image credit: Josh Quick



Kit for Sierra Leone

- Lab notebook and pen
- ✓ Gloves
- Lab coat
- ✓ Sharps bin
- ✓ Waste bottle
- ✓ Protocol
- ✓ Casio calculator
- Microfuge
- ✓ 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 (MAP006) - 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

DNAse

RT

2nd strand

Longamp

Qubit DNA - Dye
- Standards

Qubit RNA - Dye
- Standards

Stores

10 ml tips

200 µl tubes

- need eppendorf rotor

Qubit tubes

Holdall

Thermometer

Power bar

✓ Bleach

✓ RNAse

Cold

Flowcells

SPRI bead

His-tag

Qubit RNA - Dye

Qubit DNA - Dye

- Buffer

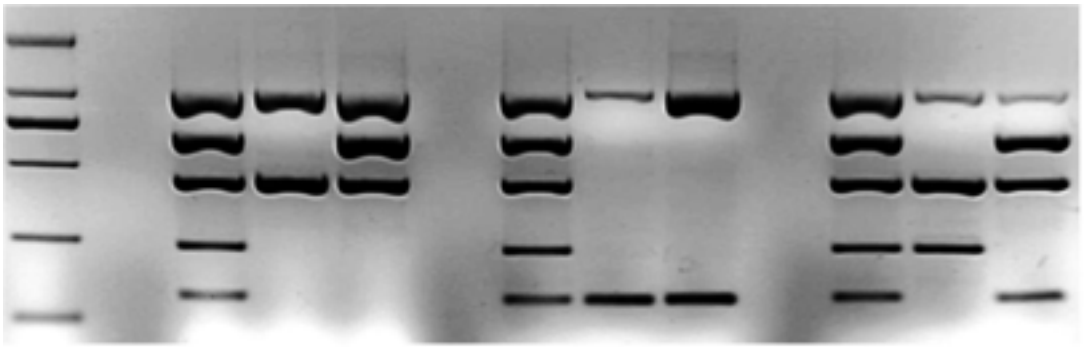
Standards (2)

Frozen

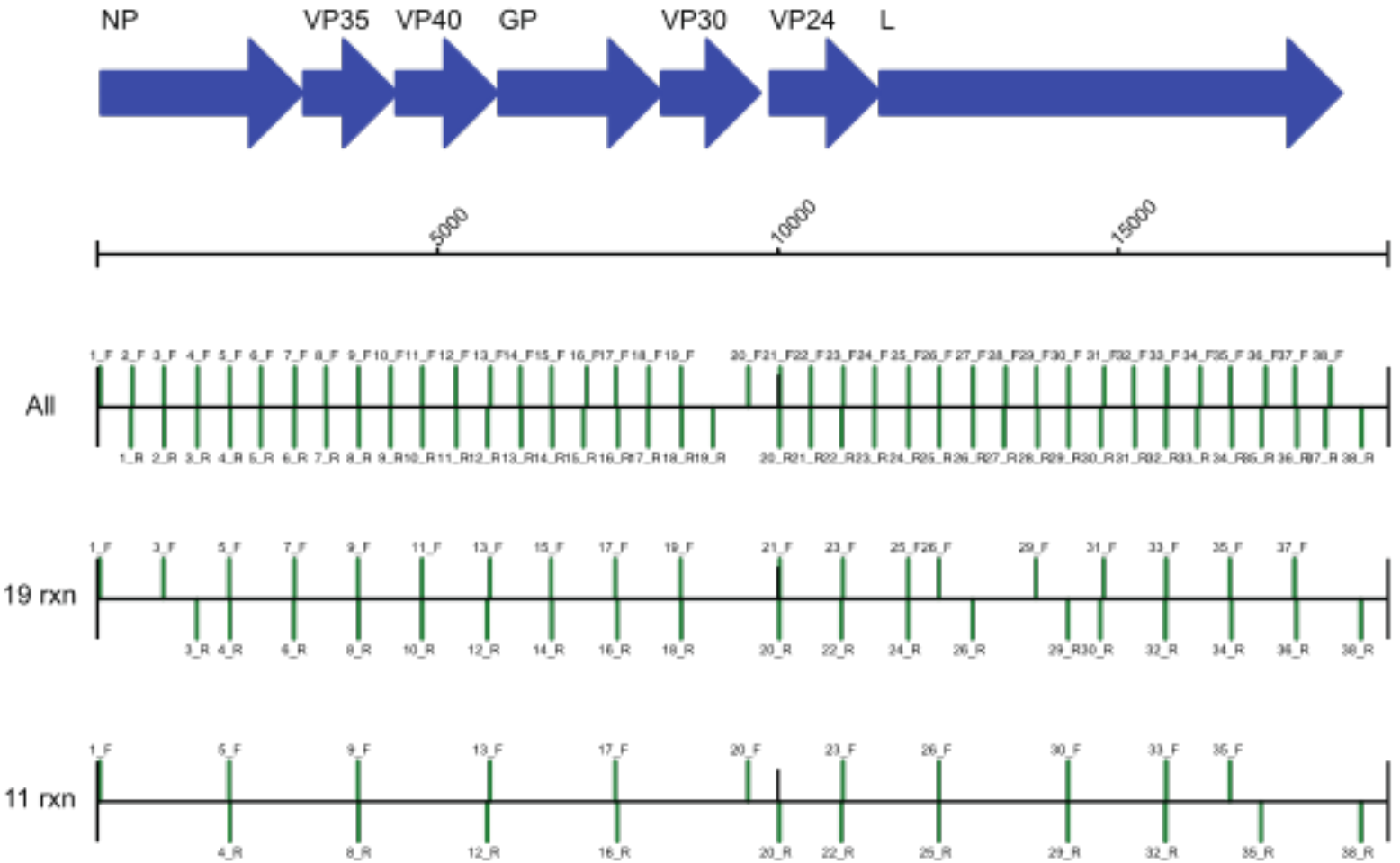
MinION kits

Enzymes - dA tail (2)
- ER (2)
- RT (x1)
- 2nd strand
- ligase (1)
- Longamp (1)
- DNAse (4)

Image credit: Josh Quick



A.



B. 19 reactions

Forward	Reverse	Length
1_F	3_R	1426
3_F	4_R	973
5_F	6_R	952
7_F	8_R	941
9_F	10_R	940
11_F	12_R	958
13_F	14_R	906
15_F	16_R	974
17_F	18_R	969
19_F	20_R	1445
21_F	22_R	906
23_F	24_R	958
25_F	26_R	947
26_F	29_R	1898
29_F	30_R	946
31_F	32_R	901
33_F	34_R	963
35_F	36_R	977
37_F	38_R	975

C. 11 reactions

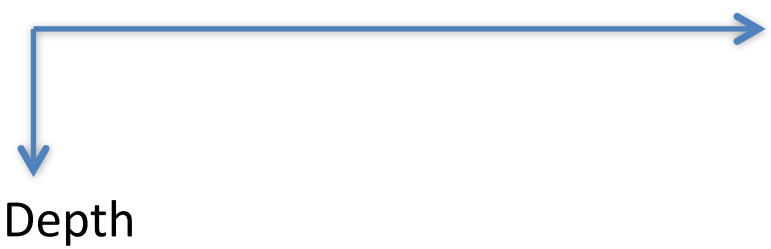
Forward	Reverse	Length
1_F	4_R	1911
5_F	8_R	1901
9_F	12_R	1896
13_F	16_R	1874
17_F	20_R	2406
20_F	22_R	1371
23_F	25_R	1410
26_F	29_R	1898
30_F	32_R	1427
33_F	35_R	1396
35_F	38_R	1921

Porton Down validation set, 89.1% coverage

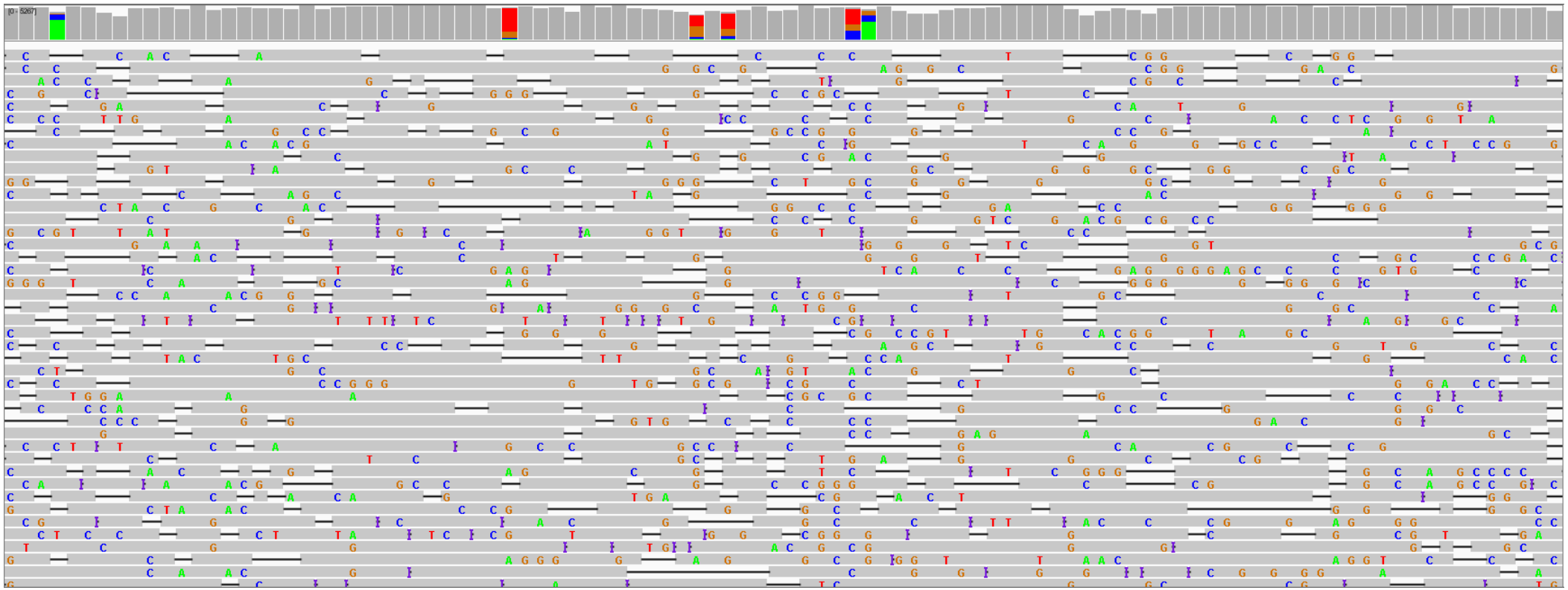
Guinea 19 reactions v1, 98.1% coverage

Guinea 11 reactions v1, 95.9% coverage

Guinea 11 reactions v2, 98.4% coverage



Computational Challenges of Nanopore Analysis



Main computational challenge: distinguishing mutations from sequencing errors

Computational Challenges of Nanopore Analysis

- 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

Nanopore Sequencing

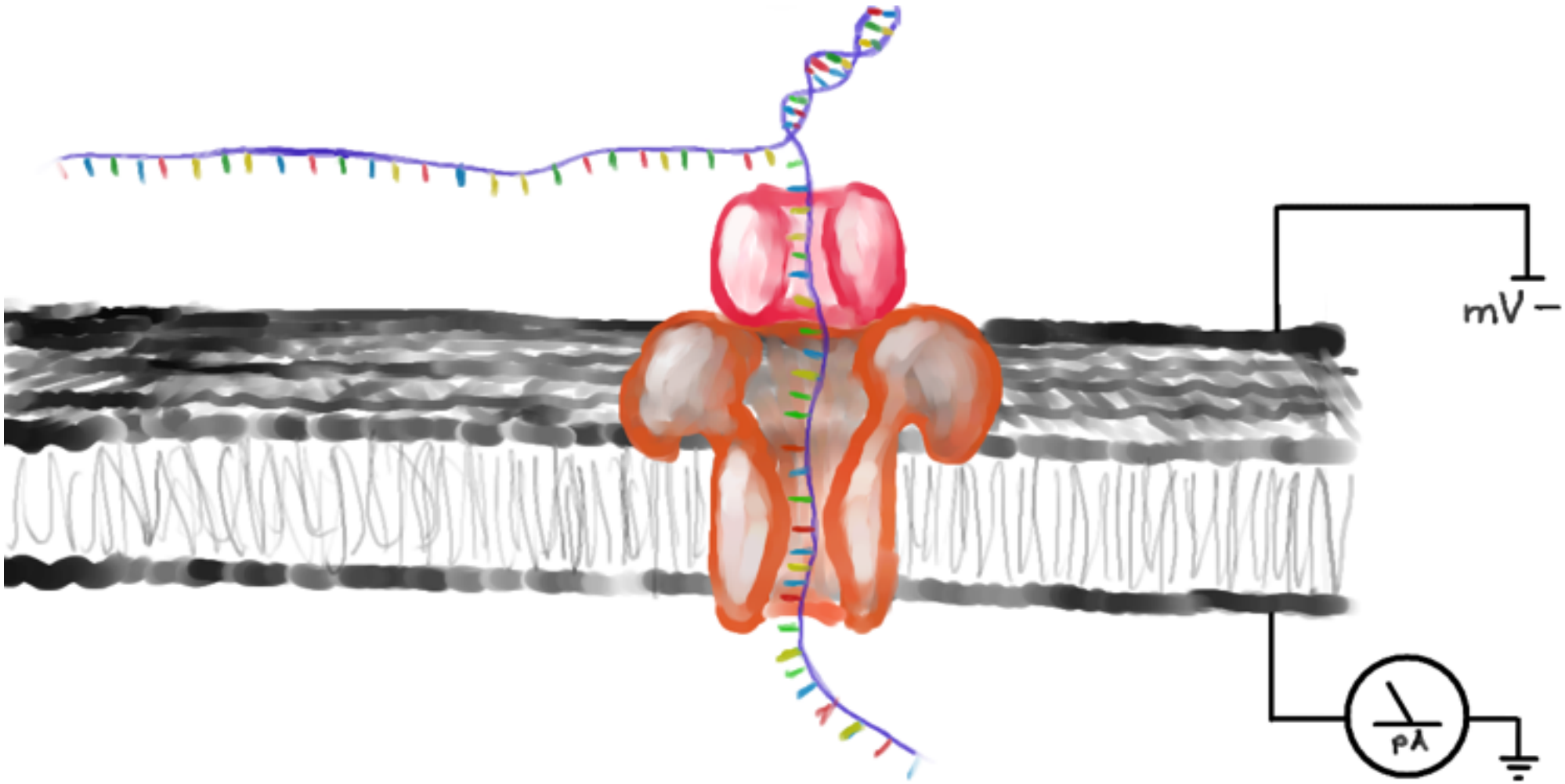
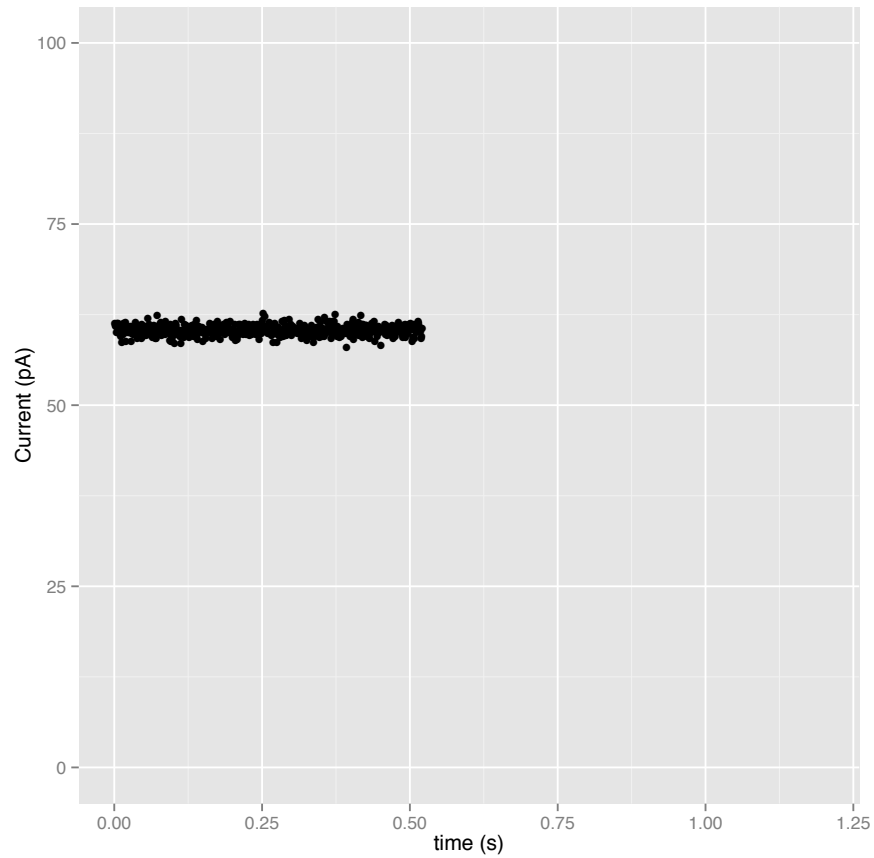
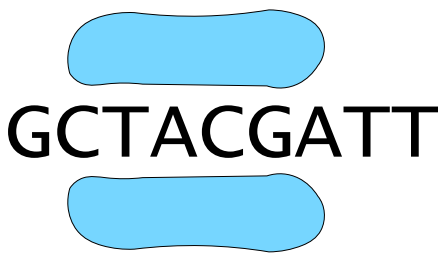


Illustration by David Eccles

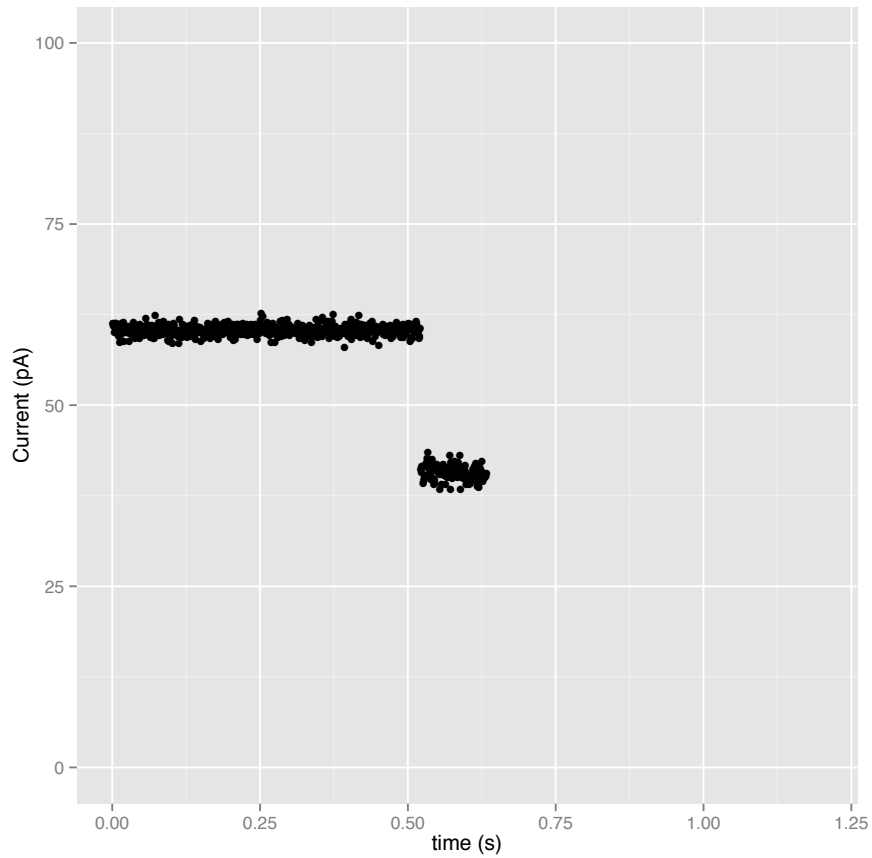
Nanopore Sequencing



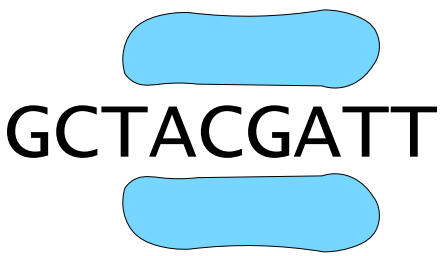
Nanopore Sequencing



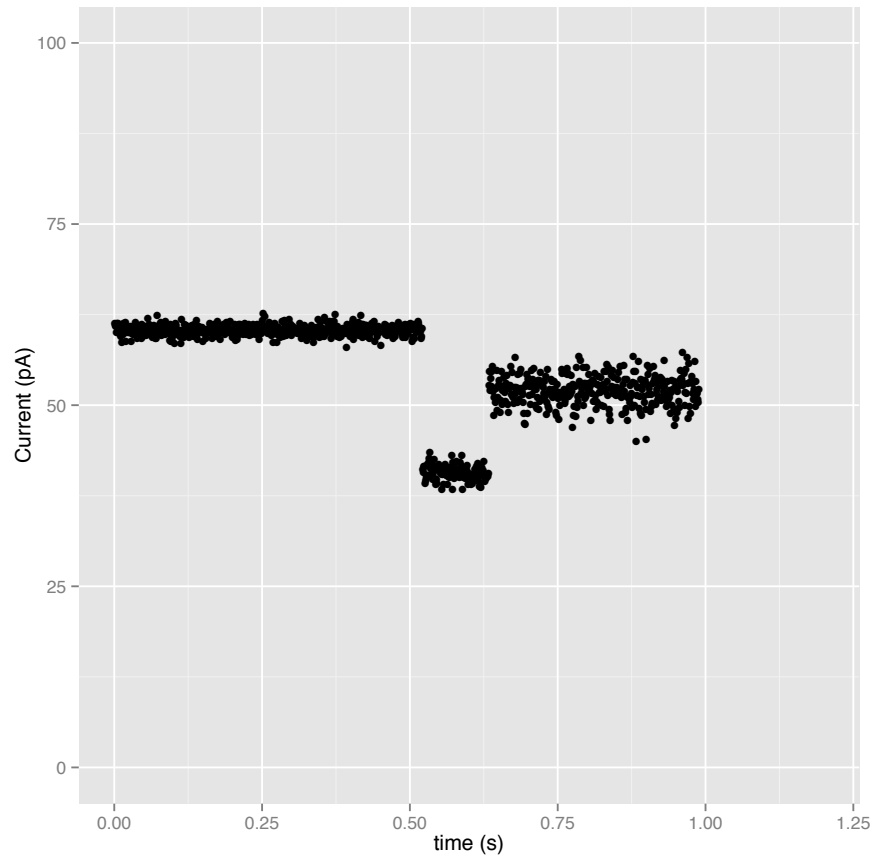
Sample Current



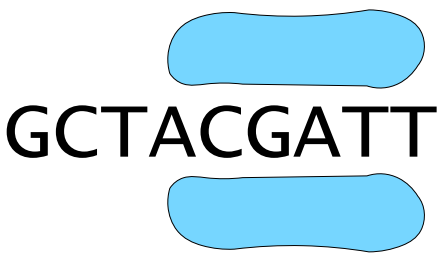
Nanopore Sequencing



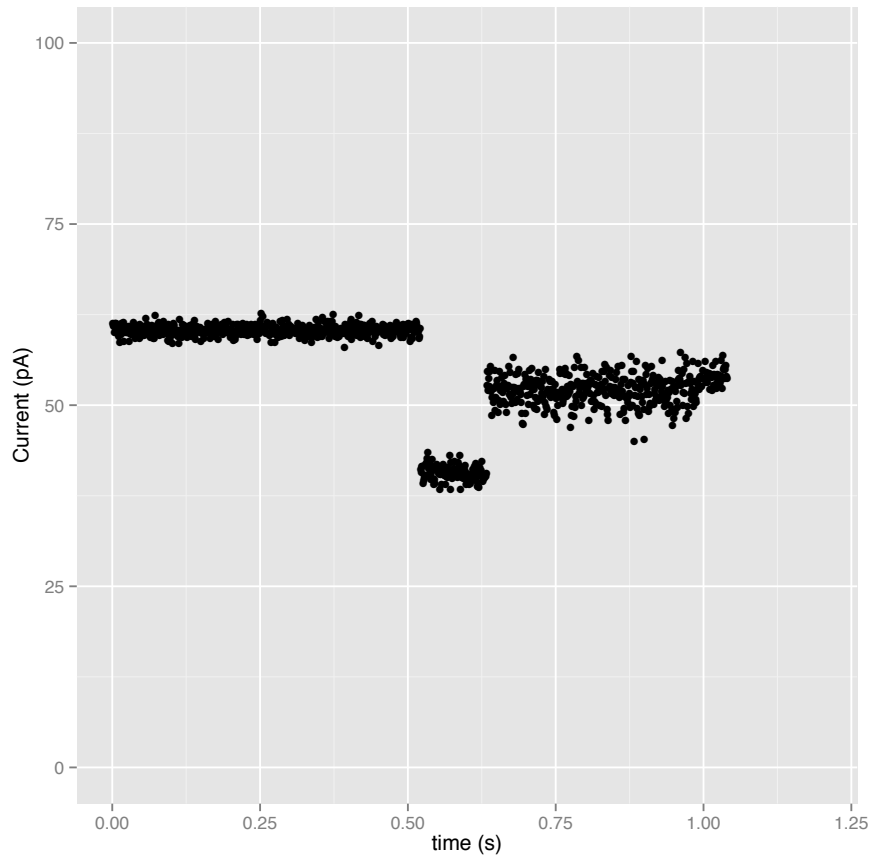
Sample Current



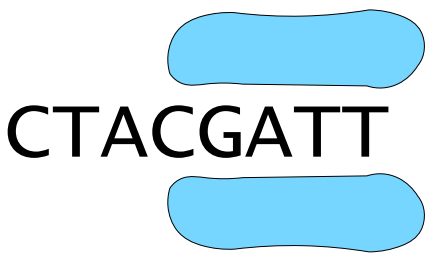
Nanopore Sequencing



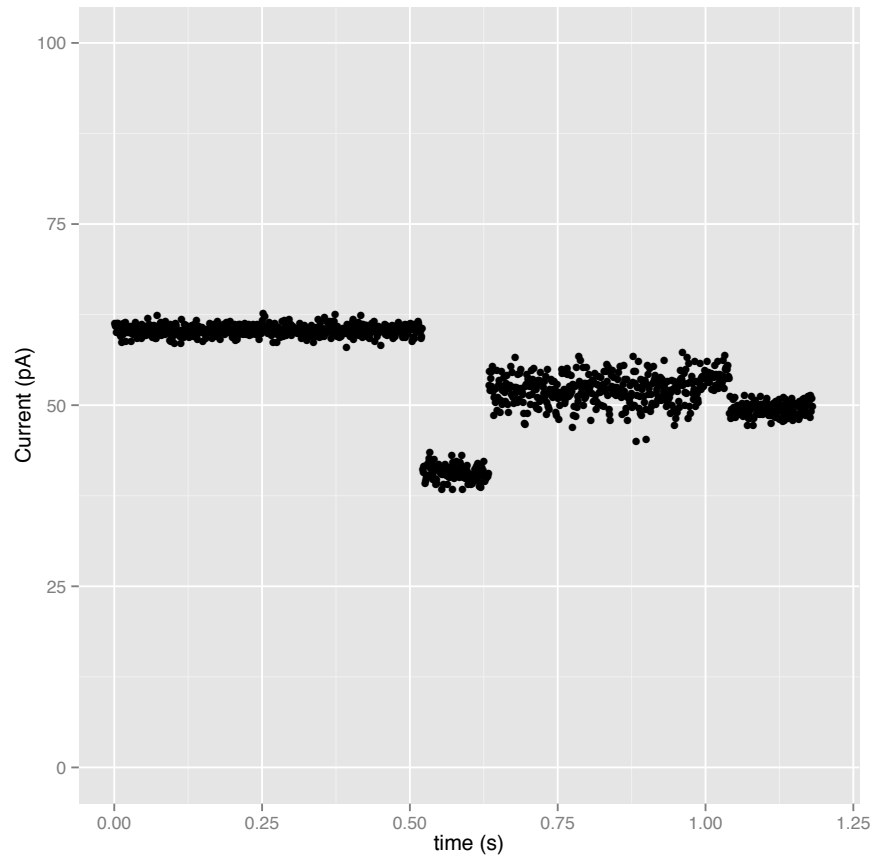
Sample Current



Nanopore Sequencing

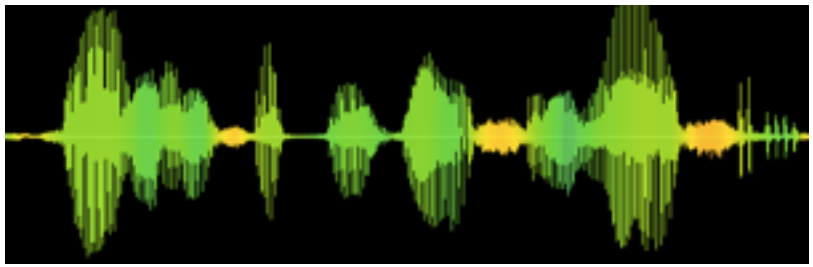


Sample Current



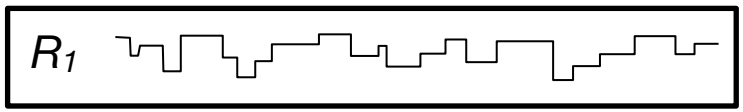
Base Calling

Speech recognition: predict a sentence from a recording of someone's voice



→ "Welcome to Computing for Medicine"

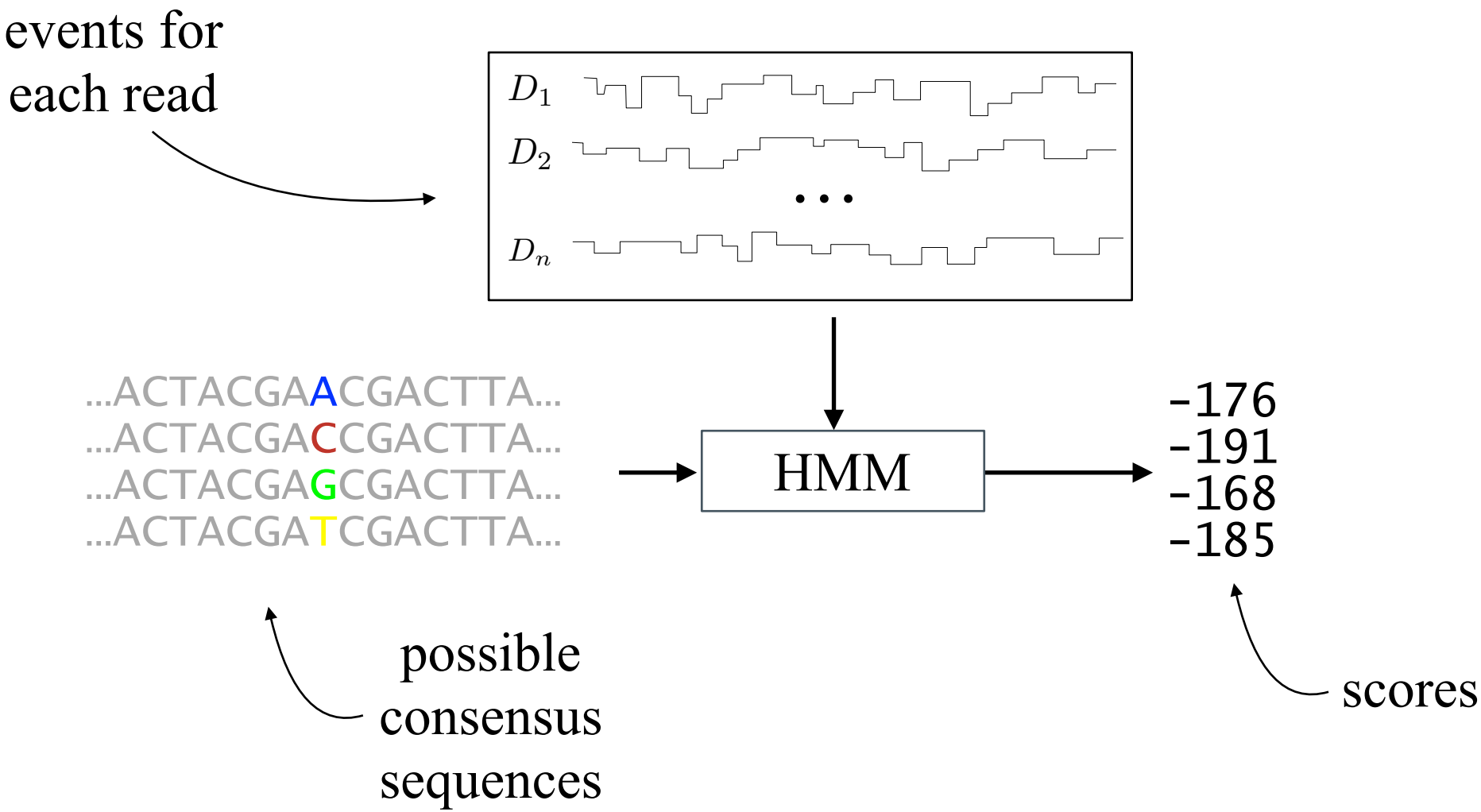
Base calling: predict a nucleotide sequence from a nanopore signal



→ ACGATAGACCGTATTAG

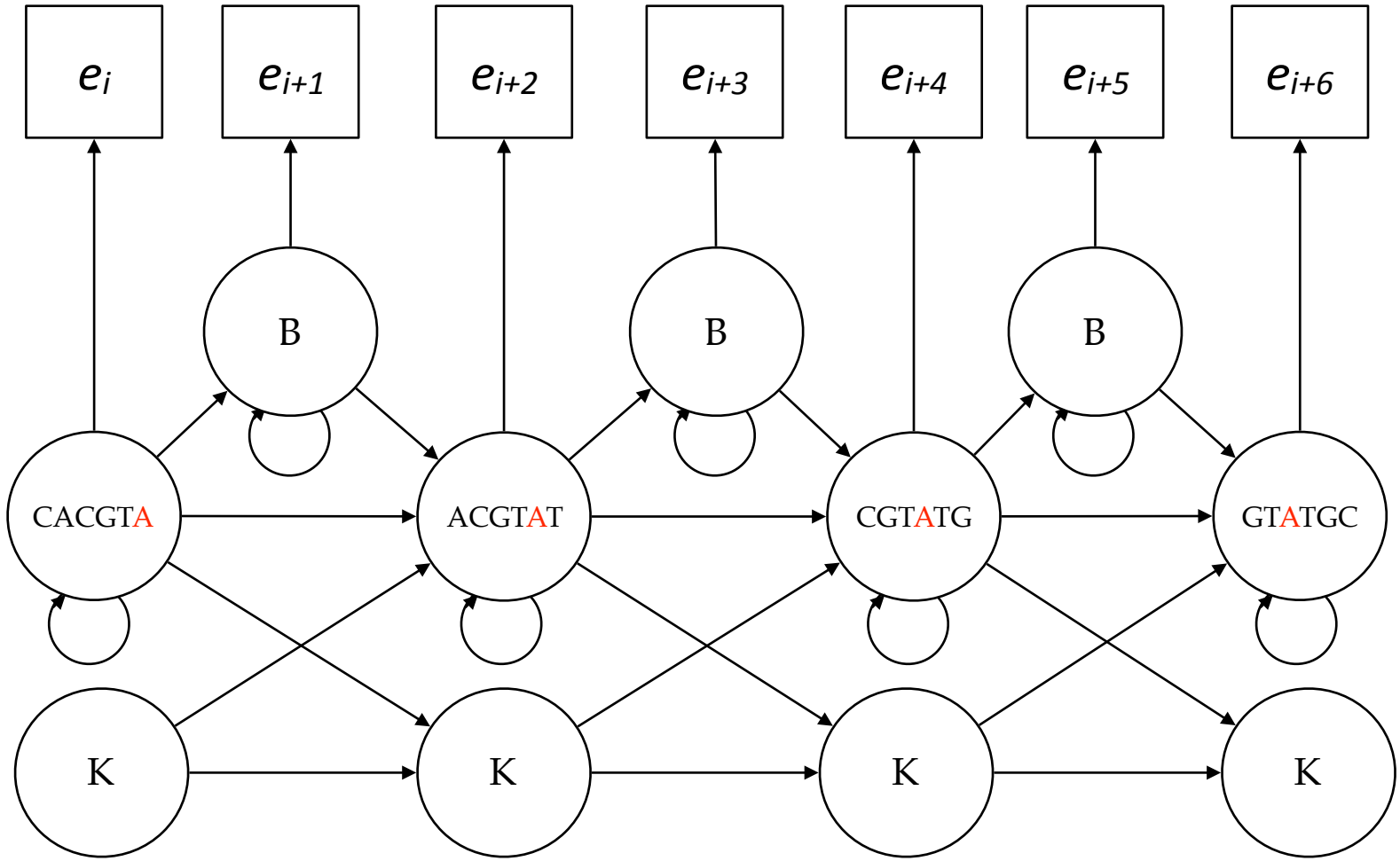
The same methods that are used to process speech (hidden Markov Models, deep neural networks) can be used for nanopore data.

Nanopore Consensus Calling



Nanopore consensus algorithms

We use nanopolish's HMM to calculate the probability of observing a sequence of events from an arbitrary sequence



Ebola Genomes

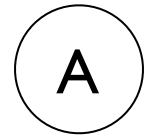
We sequenced and reconstructed 142 Ebola genomes using the MinION and our HMM-based analysis software (nanopolish)

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

Next computational challenge: What do these genomes tell us about the outbreak?
Approach: build a phylogenetic tree

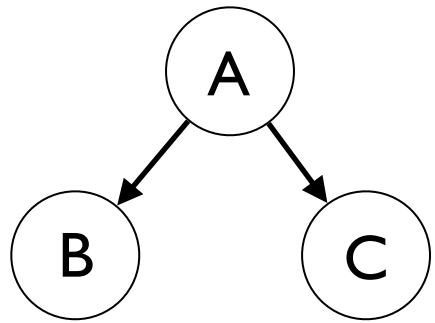
Ebola Surveillance

Ebola is passed through direct contact



Ebola Surveillance

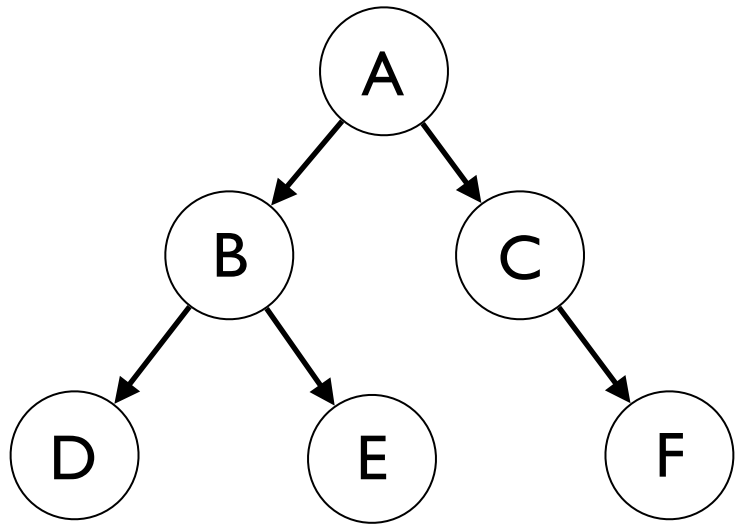
Ebola is passed through direct contact



Person A transmits the virus to B and C

Ebola Surveillance

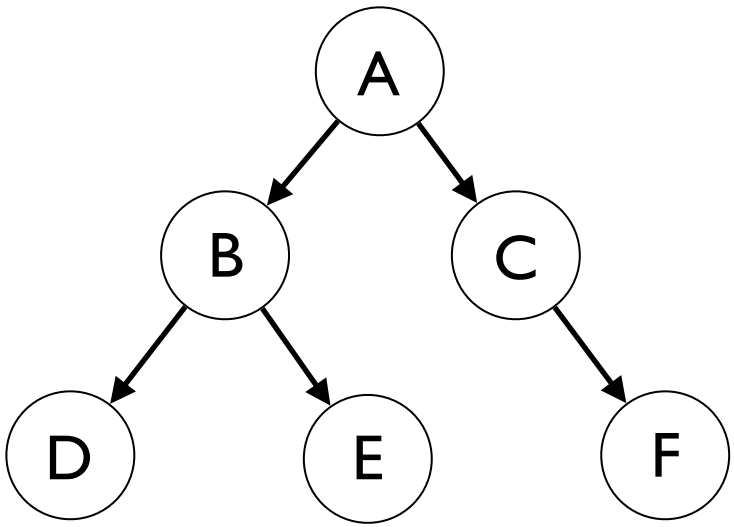
Ebola is passed through direct contact



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

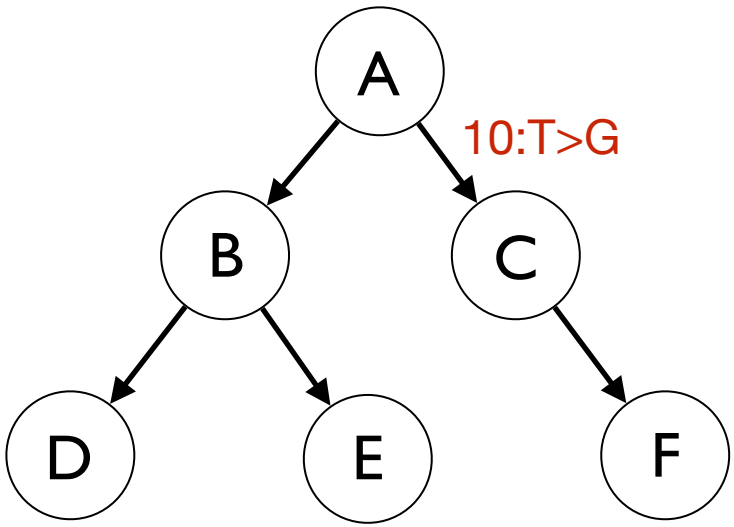
Ebola virus mutates at a rate of ~1.15 $\times 10^{-3}$ mutations/bp/year
These mutations allow us to track patterns of transmission




```
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

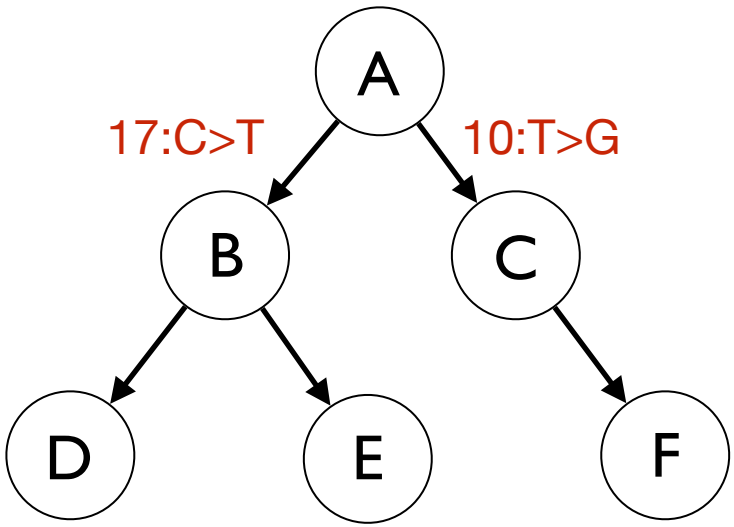



```
Ebola-A ...AGTAGCCTACGATACTACGATCGACTTA...
Ebola-B ...AGTAGCCTACGATATTACGATCGACTTA...
Ebola-C ...AGTAGCCGACGATACTACGATCGACTTA...
Ebola-D ...AGTTGCCTACGATATTACGATCGACTTA...
Ebola-E ...AGTAGCCTACGATATTACGATCGACATA...
Ebola-F ...AGTAGCCGACGATACTACGATGGACTTA...
```

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

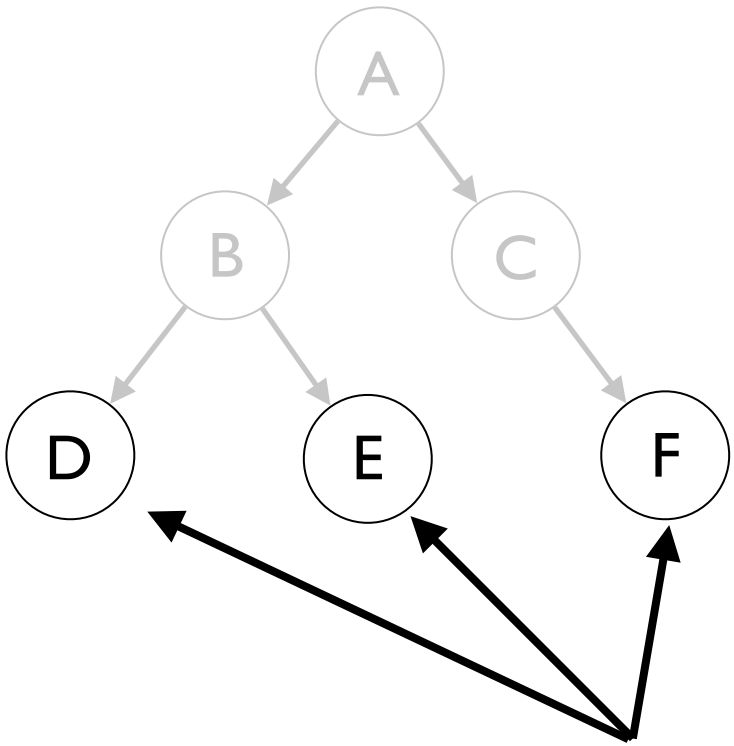



```
Ebola-A ...AGTAGCCTACGATACTACGATCGACTTA...
Ebola-B ...AGTAGCCTACGATATTACGATCGACTTA...
Ebola-C ...AGTAGCCGACGATACTACGATCGACTTA...
Ebola-D ...AGTTGCCTACGATATTACGATCGACTTA...
Ebola-E ...AGTAGCCTACGATATTACGATCGACATA...
Ebola-F ...AGTAGCCGACGATACTACGATGGACTTA...
```

C>T here indicates B/D/E lineage

Ebola Surveillance

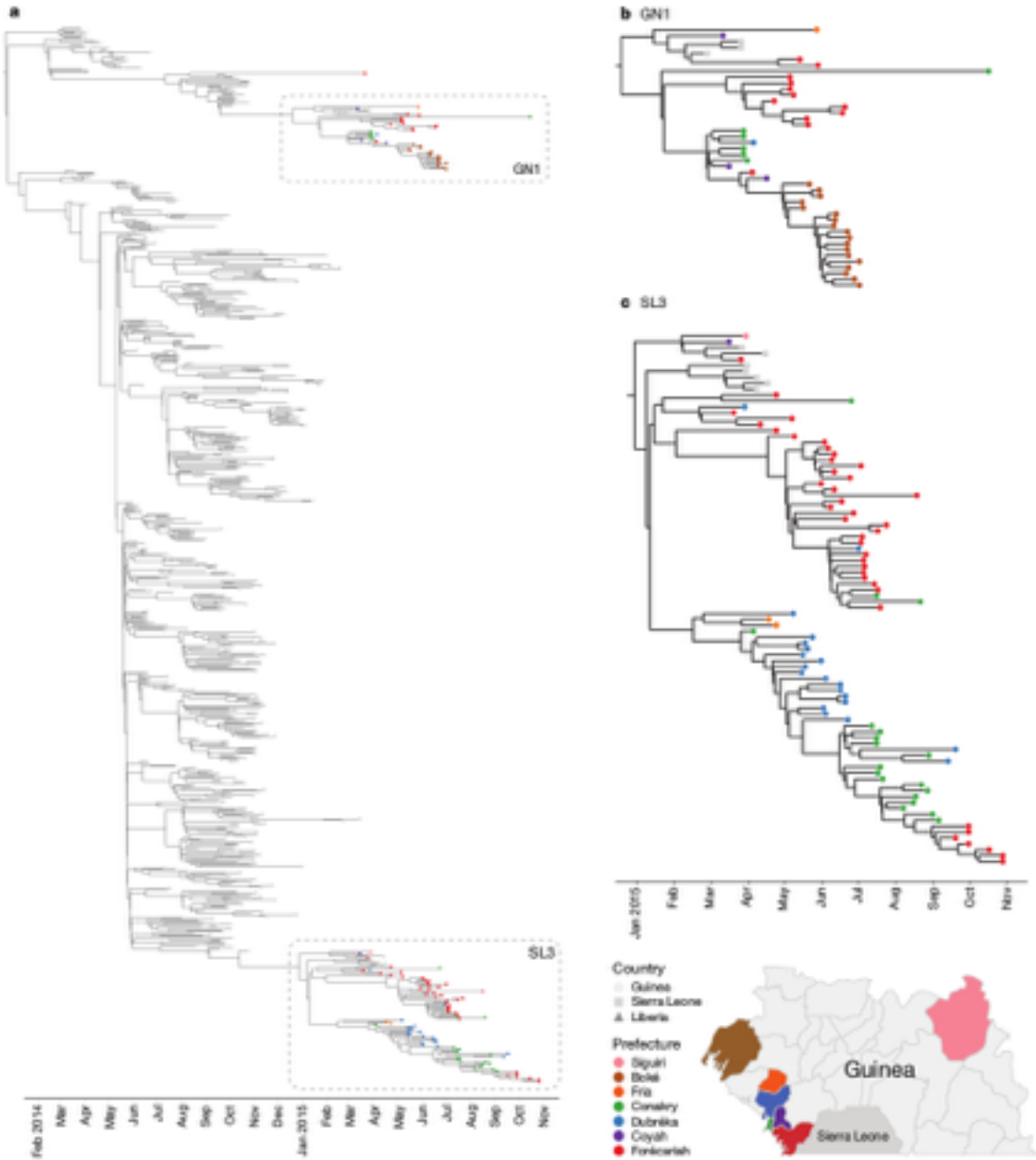
We can't sequence every case in the outbreak but by sampling enough cases we can build a picture of how the virus is spreading (e.g. geographically)

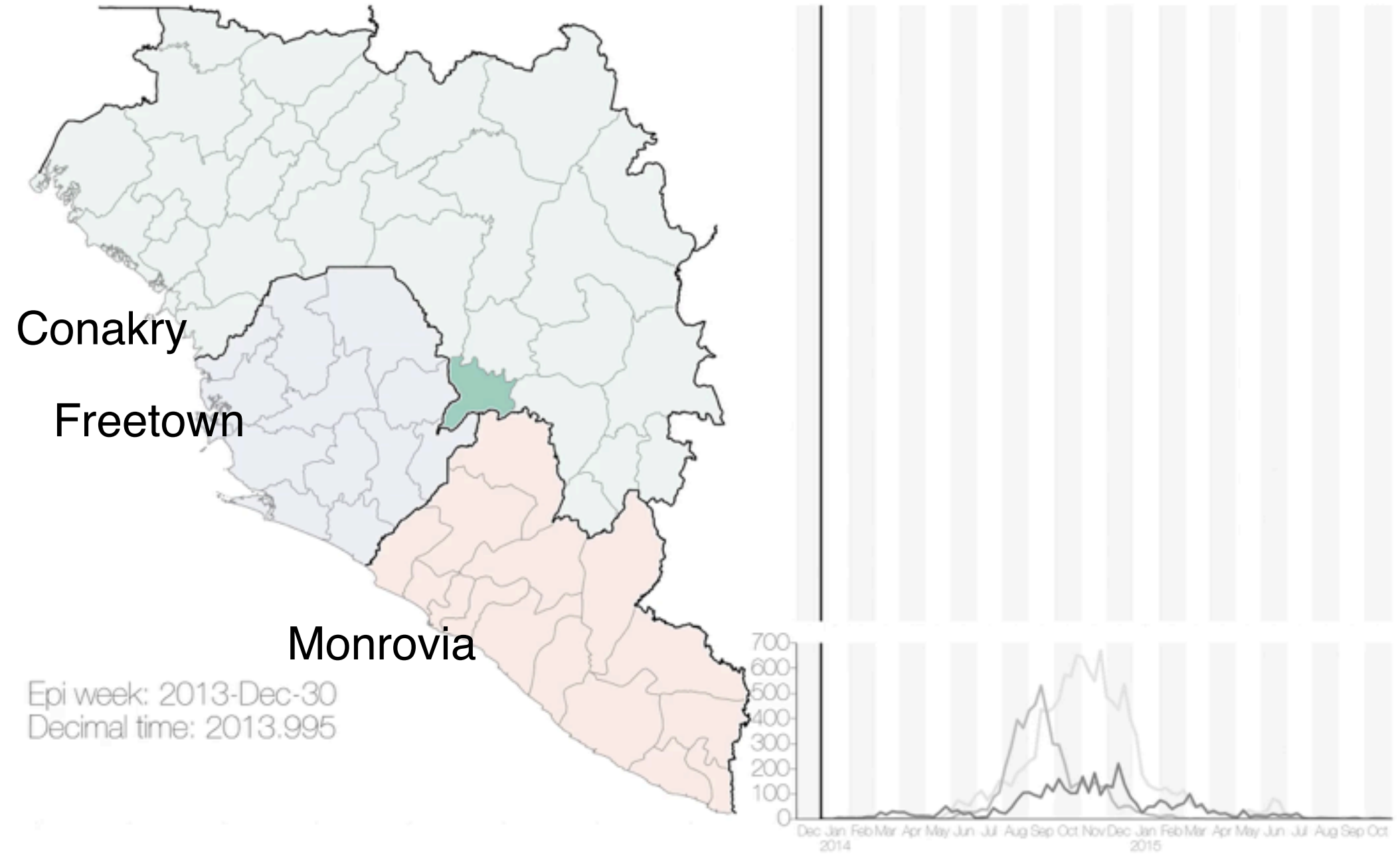



```
Ebola-D ...AGTTGCCTACGATATTACGATCGACTTA...
Ebola-E ...AGTAGCCTACGATATTACGATCGACATA...
Ebola-F ...AGTAGCCGACGATACTACGATGGACTTA...
```

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

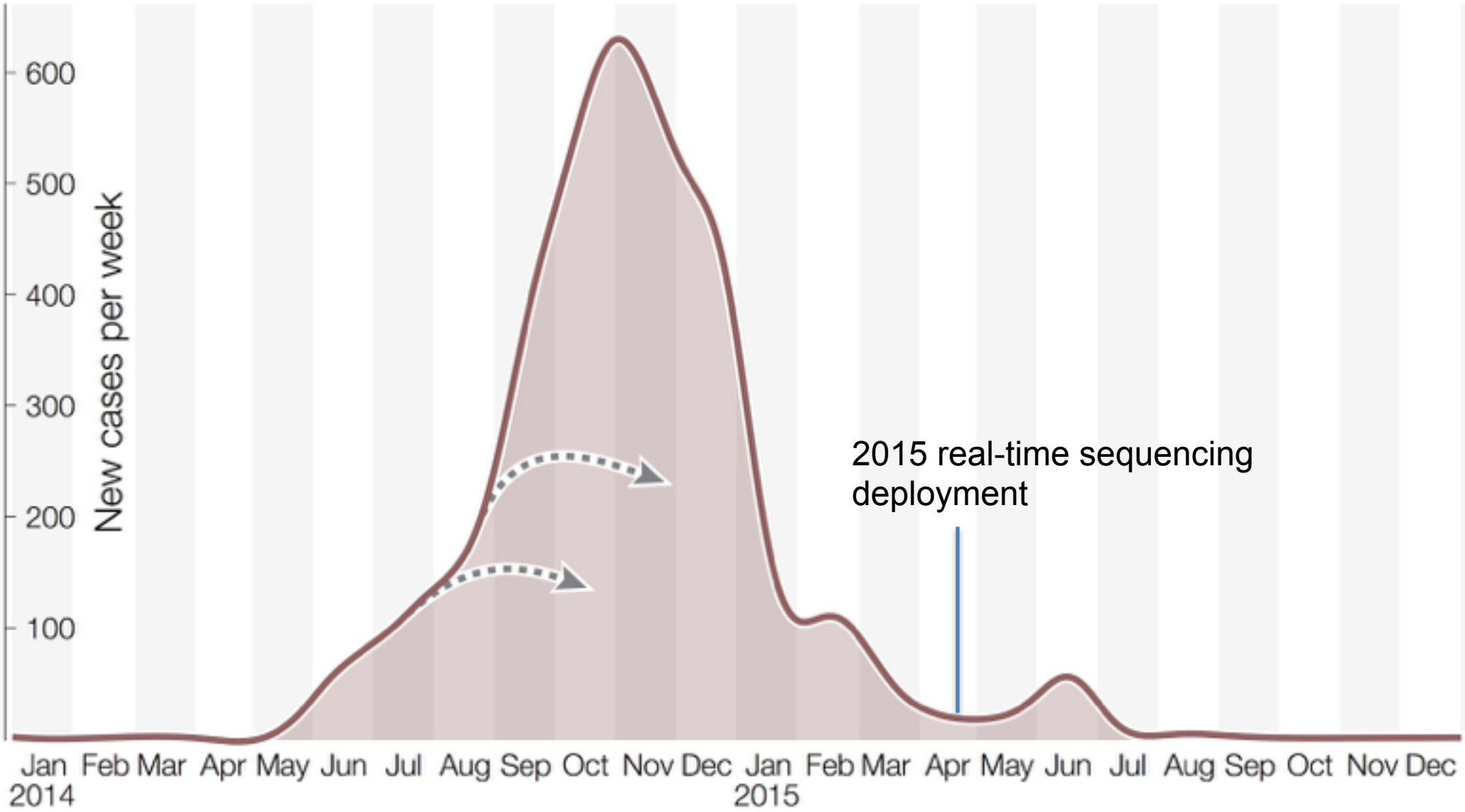
Ebola Phylogeny



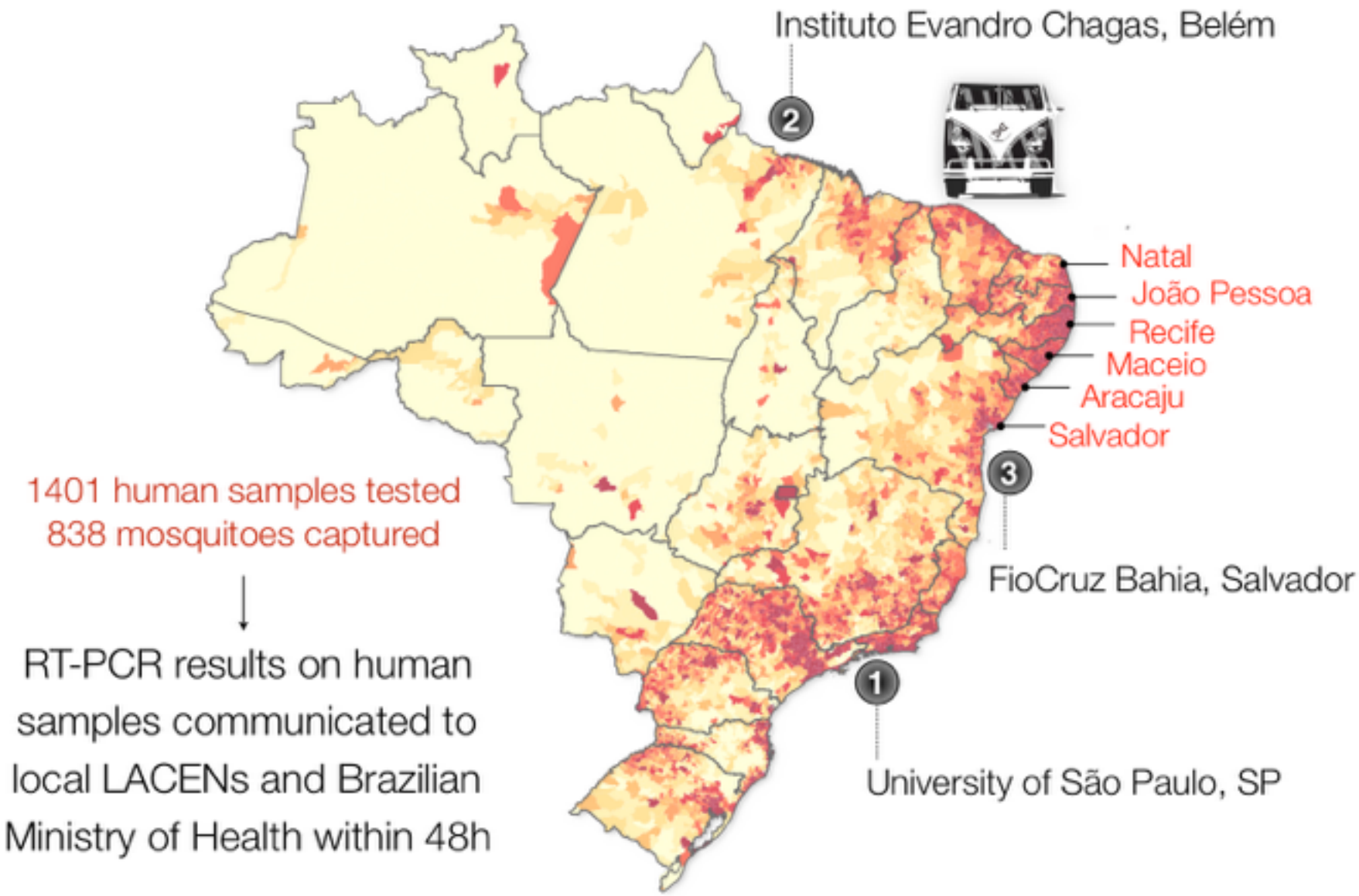


Dudas et al, Nature 2017

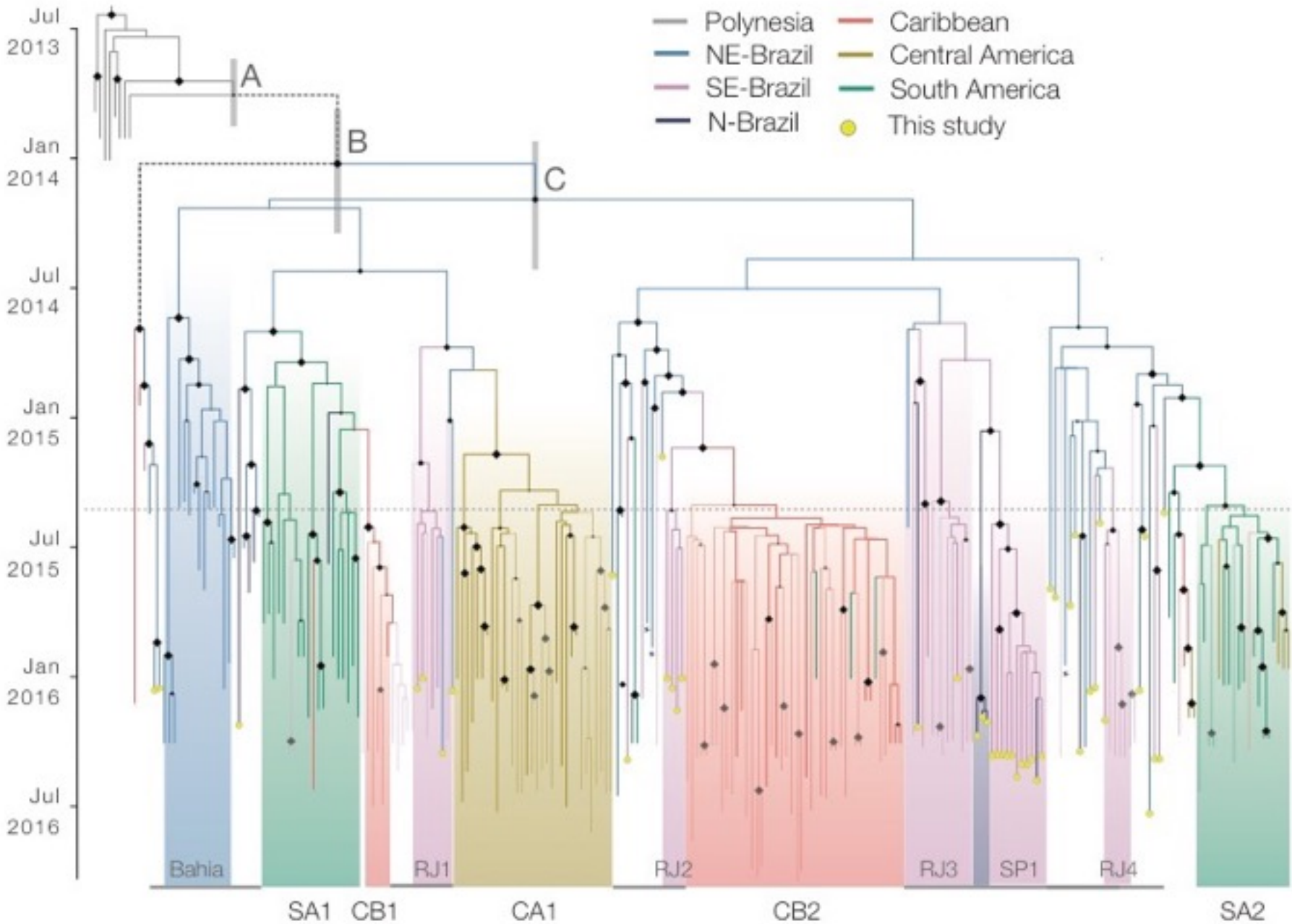
Ebola Timeline



ZIKV sequencing road trip



Post-road trip molecular surveillance in SP and Bahia



Faria et al. Nature 2017;

Discussion/Outlook

- Portable genome sequencing enables new applications
- Data is challenging to work with but similar to classic machine learning problems
- Sequencing is likely to become a larger and larger part of diagnostics (e.g. sequencing an infection to discover anti-microbial resistance genes)
- Many applications to human genomes: detecting structural variation, methylation, phasing
- If you're interested in discussing further contact me: jared.simpson@oicr.on.ca