

Third Generation Sequencing: A Zika Virus Case Study

daniel.ward1@lshtm.ac.uk

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



Third Generation Sequencing Platforms

	Second Generation sequencing	PacBio: Single Molecule Real-Time Sequencing SMRT	Oxford Nanopore Technology: MinION
Read Length (bp)	150	15,000	90,000
Read output	2.5-5,000,000	~500,000	1,000,000
Cost Per Run (\$)	3000	850	700
Hardware Cost (\$)	900,000	350,000	1000
Observed Error Rate	<1%	~12%	~12%

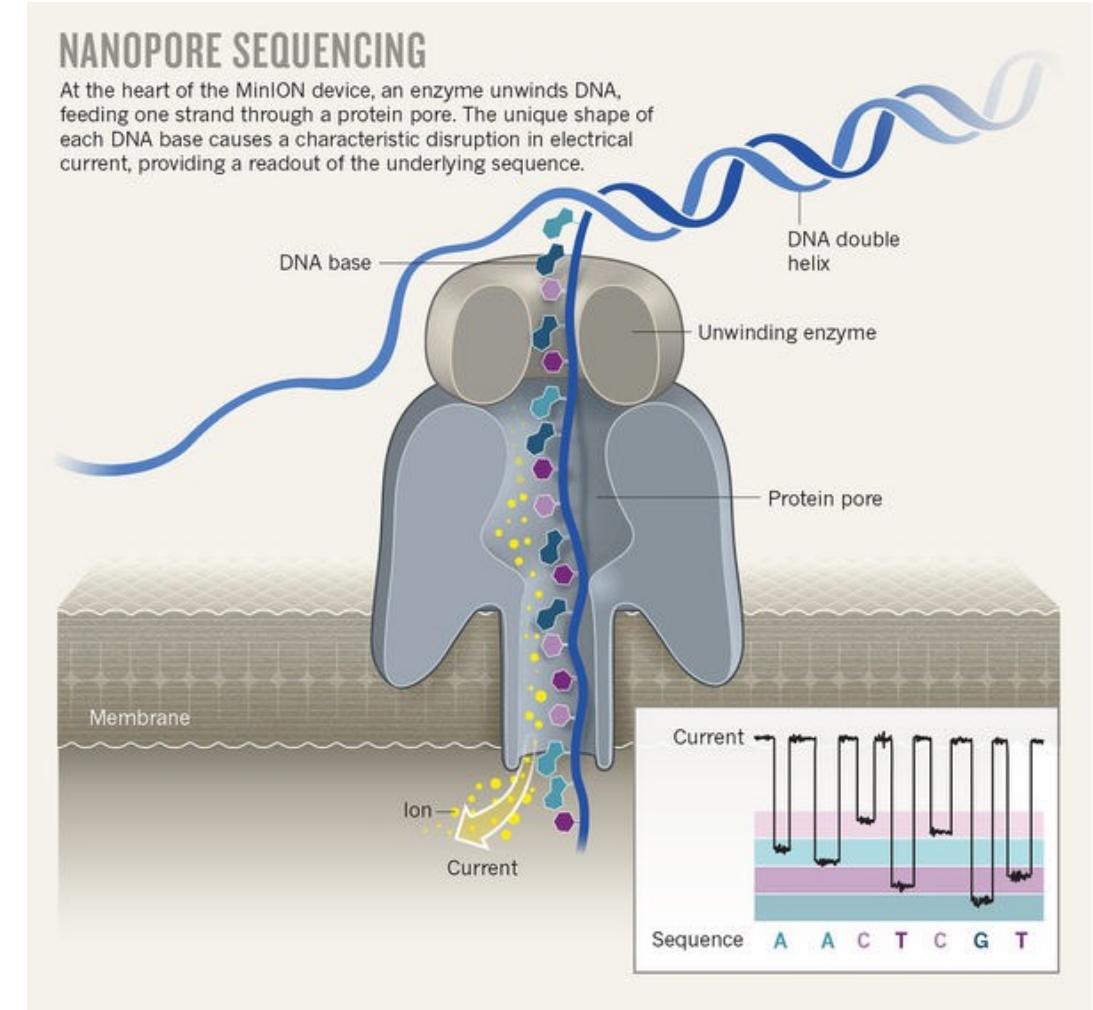
Table 2: Illumina WGS Solutions.

Illumina Sequencing System	Read Length	Output	No. of samples
NextSeq 500 System	Mid Output Flow Cell (130 M)	2 × 150 bp	40 Gb up to 24 small genomes
	High Output Flow Cell (400 M)	2 × 150 bp	120 Gb 1 × 30× human genome
HiSeq 2500 System	Rapid Run Mode	2 × 100 bp	100–120 Gb 1 × 30× human genome
	High Output Mode	2 × 125 bp	900 Gb–1 Tb 8 × 30× human genome

MinION – Nanopore Sequencing Technology

"Portable, real-time sequencing analyses"

- Very portable – can take to the field*
- Produces long reads compared to NGS platforms
- Can produce very high depth and coverage – high data yield.
- Still requires sophisticated lab techniques for library preparation
- Has a very high error rate compared to Illumina
- Is still under development – high dependency on community contributions
- Data is very challenging to process and analyse



Nanopore technology

PromethION



GridION



MinION

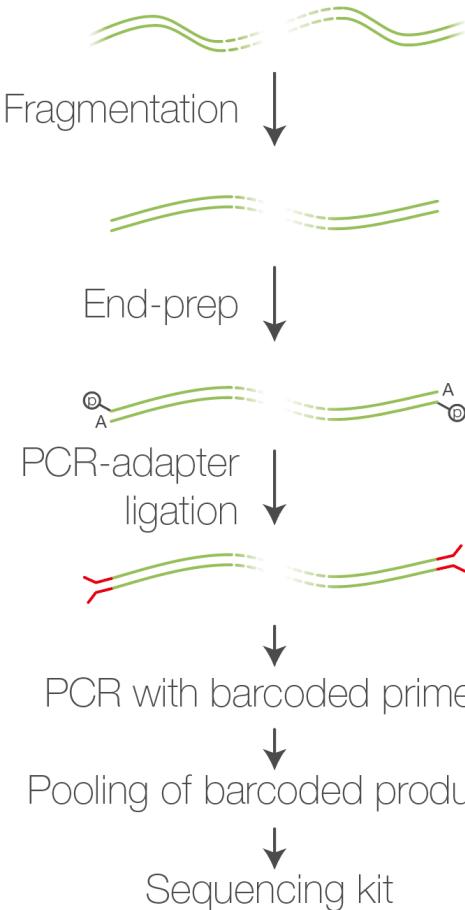
Flongle

HPC + GPU
(Nvidia > 1080)

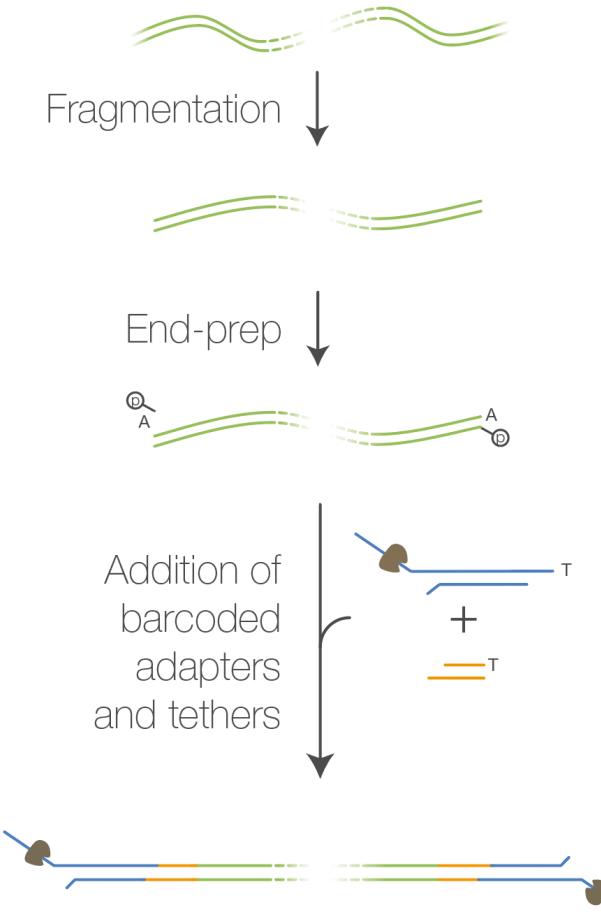


Library preparation and loading

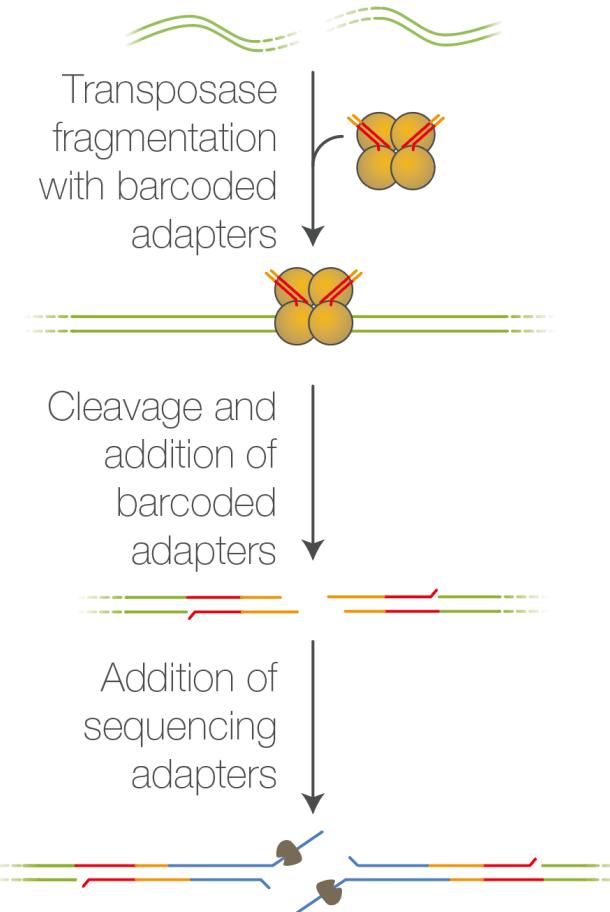
a) PCR barcoding



b) PCR-free barcoding



c) Rapid barcoding

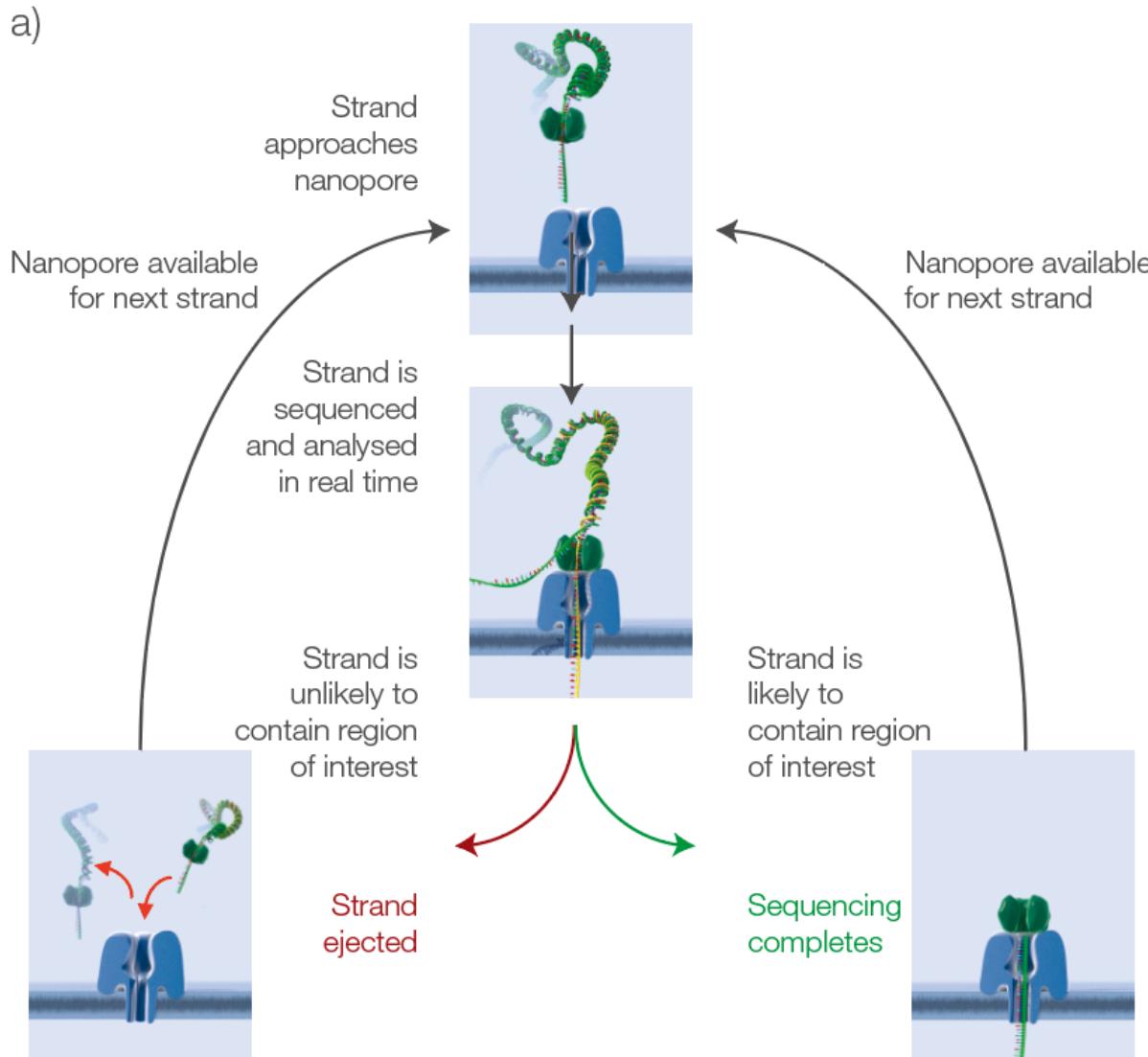


Approx. 4 hours

Approx. 1.5 hours

Approx. 10 minutes

Adaptive sampling

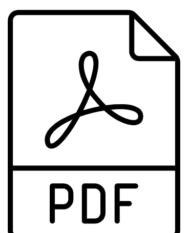


Adaptive sampling allows the automatic selection of an unlimited number of regions of interest during a sequencing run, without the need for additional library-preparation steps

MinION run QC



Remember to save
your PDF report in
MinKNOW



Analysis Pipeline

Basecalling

Converting the electrical signals generated by a DNA or RNA strand passing through the nanopore (**fast5**) into the corresponding base sequence of the strand (**fastq**).

Guppy:

- Integration with MinKNOW GUI
- Demultiplexing and kit selection built in
- GPU optimised

Bonito

- Experimental with high performance
- Model training

Quality control

The PDF from **MinKNOW** and tools such as **pycoQC** can be useful in understanding metrics.

- Demultiplex and adapter trimming (**Guppy or Porechop**)
- Removal of contaminant reads (**Kraken and centrifuge**)
- Quality filtering (**Guppy and Filtlong**)

Different requirements for mapping or assembly.

Mapping, assembly and variant calling

There are specialised tools for assembling and mapping nanopore reads.

Mapping:

- Minimap2 (-ax map-ont)
- BWA (-x ont2d)

Assembly:

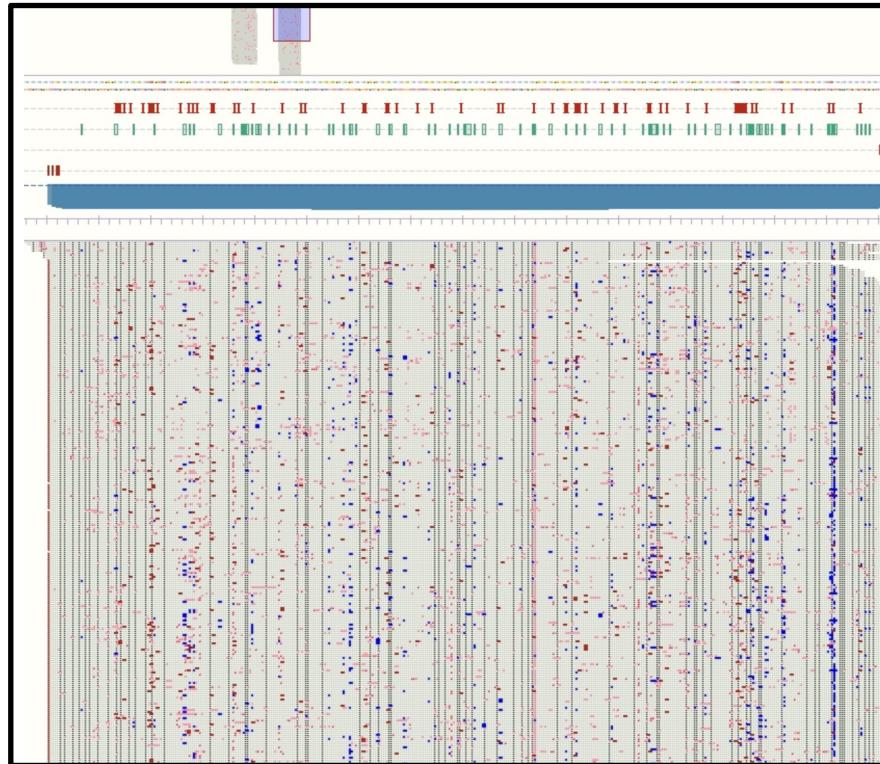
- Flye
- Unicycler
- SPAdes
- Canu

Sequencing errors

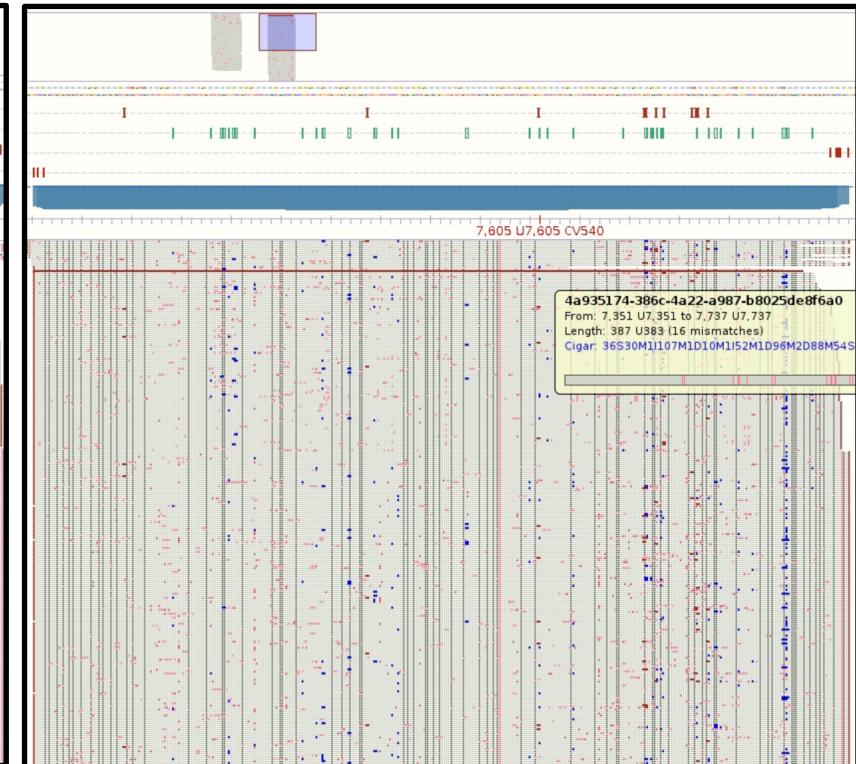
Illumina



Nanopore: Guppy

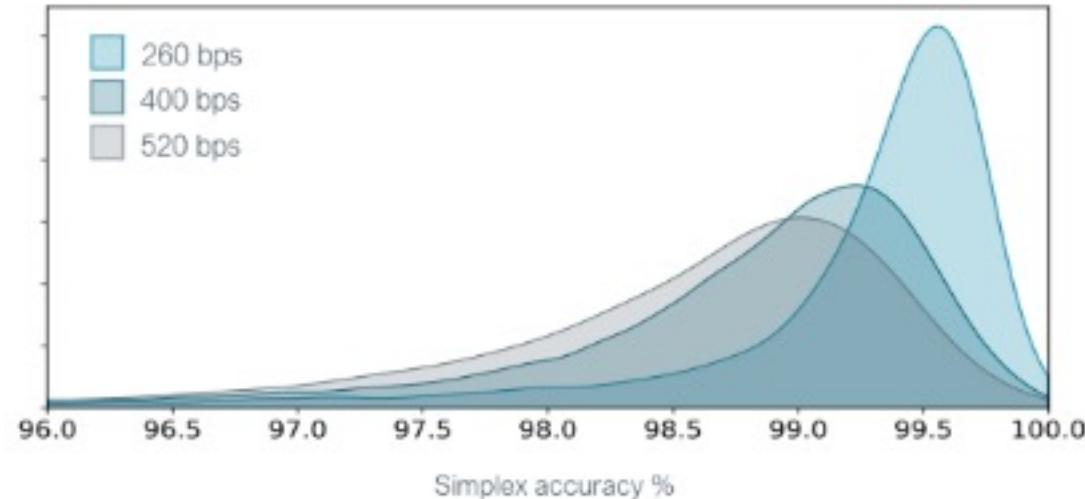


Nanopore: Bonito



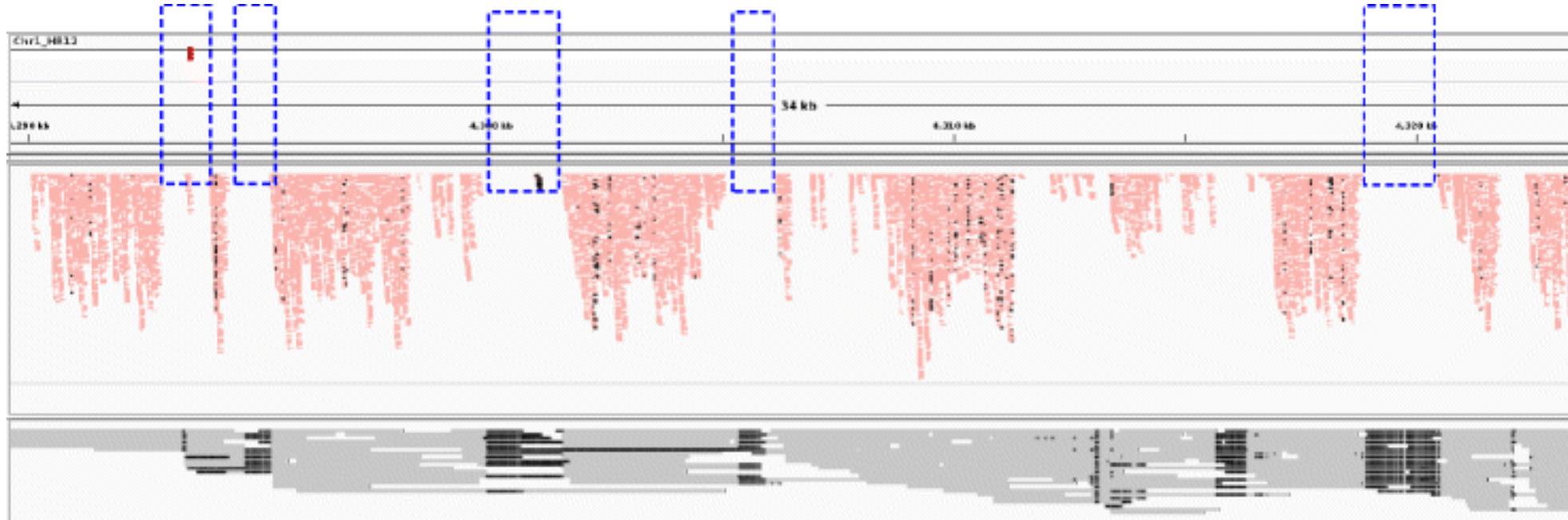
Sequencing errors

- The major caveat of nanopore sequencing is the high error rate.
- It is a common misconception that nanopore errors are all random.
Some are systematic.
- ONT have been working on Q20+ sequencing chemistries, which decrease the sequencing error rate significantly.
- Duplex basecalling combines the data from the complementary strand to increase basecall confidence.



Run condition	Translocation speed	Theoretical max output (PromethION Flow Cell)	Modal accuracy (simplex)
Accuracy	260 bps	185 Gb	99.6% (Q24)
Default	400 bps	285 Gb	99.2% (Q21)
Output	520 bps	370 Gb	~99% (Q20)

Hybrid assembly

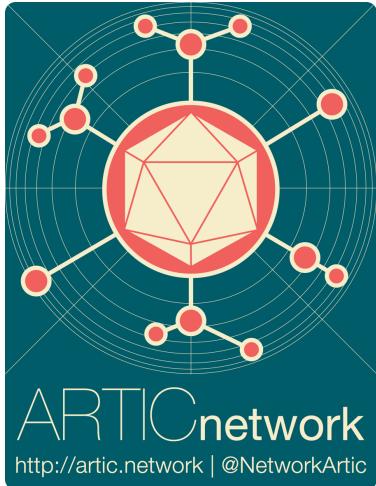


3

Hybrid assembly helps
resolve ambiguities with
higher coverage and
differing read lengths



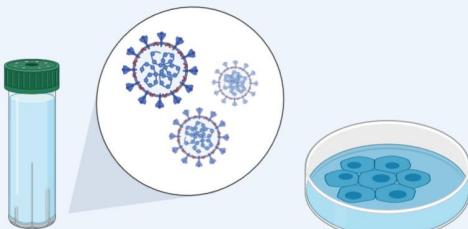
SARS-CoV-2 Sequencing



SARS-CoV-2 sequencing workflow

1 Specimens collected

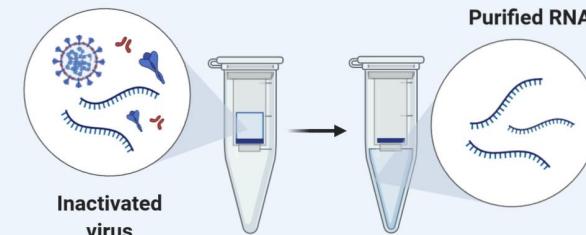
Clinical samples and isolates



2 Viral RNA extraction

NucleoMag Virus kit (Macherey-Nagel)

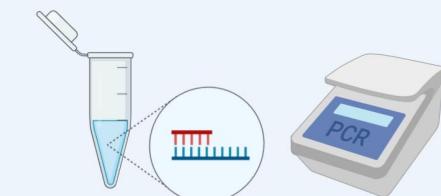
~45 min



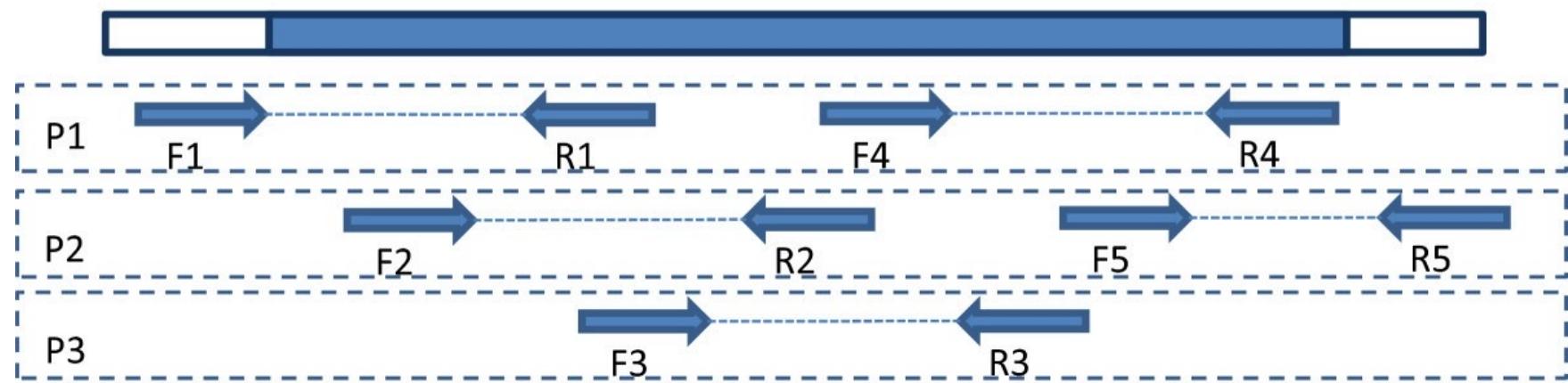
3 cDNA synthesis and multiplex PCR

ProtoScript II First Strand cDNA Synthesis Kit (NEB) ~25 min

Q5 Hot Start High-Fidelity DNA Polymerase (NEB) ~4h per primer set



Target region



Real-Time Sequencing = Real-Time Surveillance

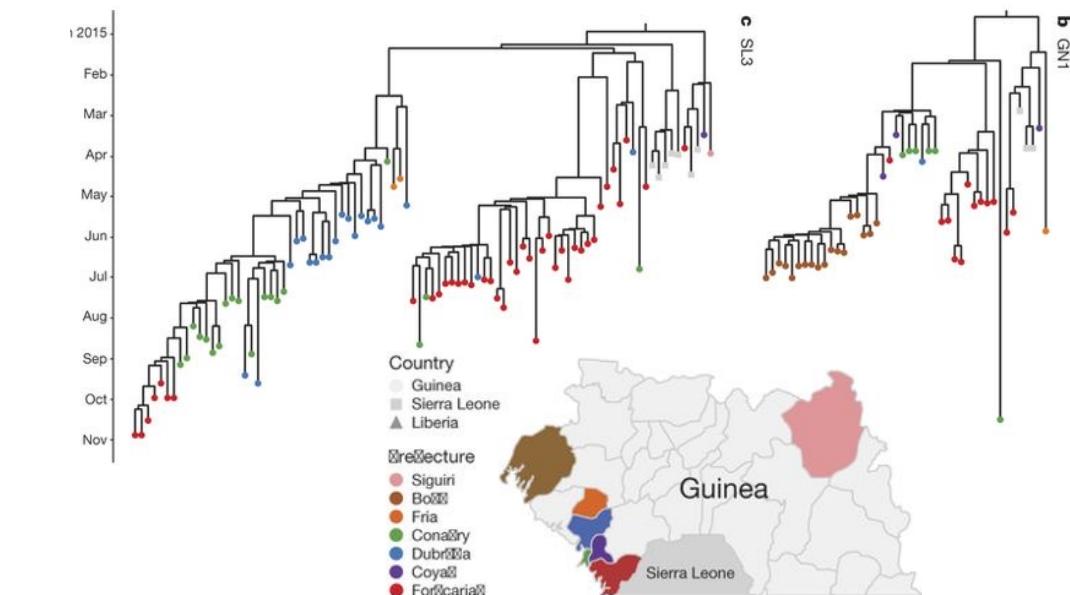
- Genome sequencing can provide a high-resolution view of pathogen evolution and is increasingly sought after for outbreak surveillance.
- Sequence data may be used to guide control measures, but only if the results are generated quickly enough to inform interventions.
- In April 2015 the MinION system was transported in standard airline luggage to Guinea and used for real-time genomic surveillance of the ongoing epidemic.
- They presented sequence data and analysis of 142 EBOV samples collected during the period March to October 2015.

Letter | Published: 03 February 2016

Real-time, portable genome sequencing for Ebola surveillance

Joshua Quick, Nicholas J. Loman  [...] Miles W. Carroll

Nature 530, 228–232 (11 February 2016) | Download Citation 

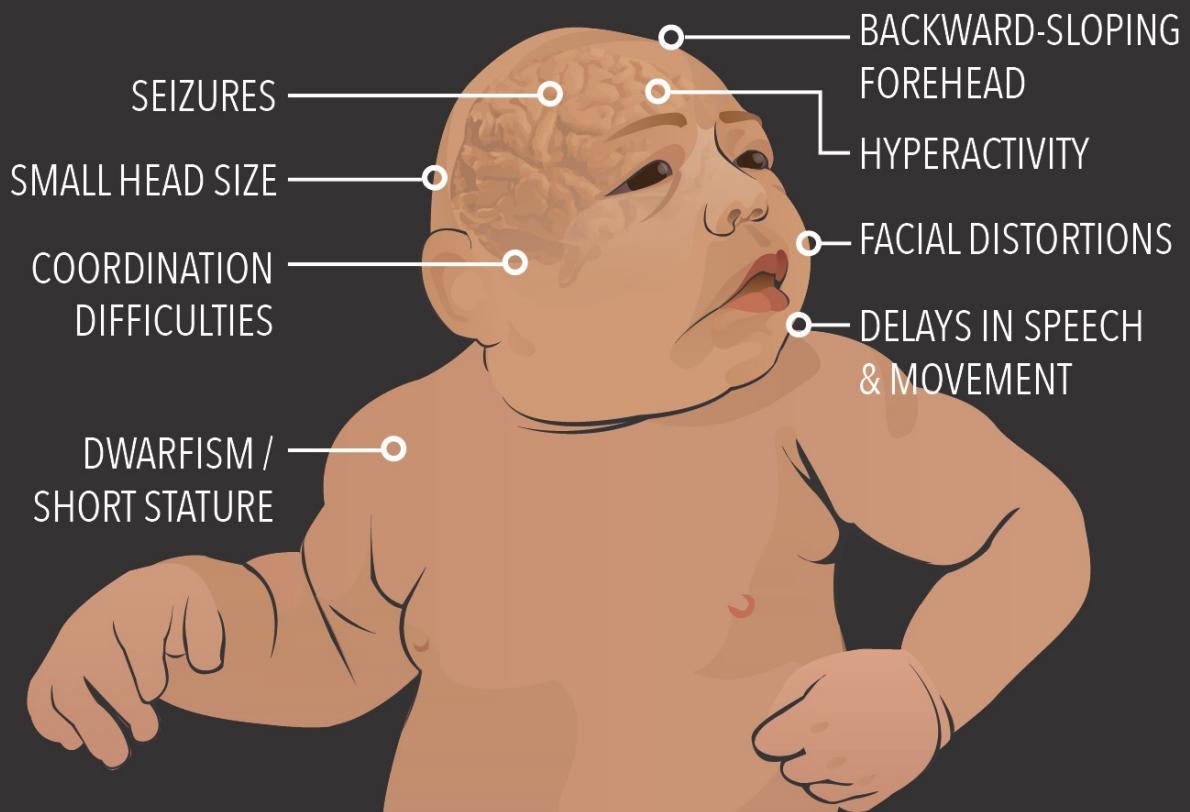


<https://nextstrain.org/ebola?dmax=2014-09-26&dmin=2014-08-10>

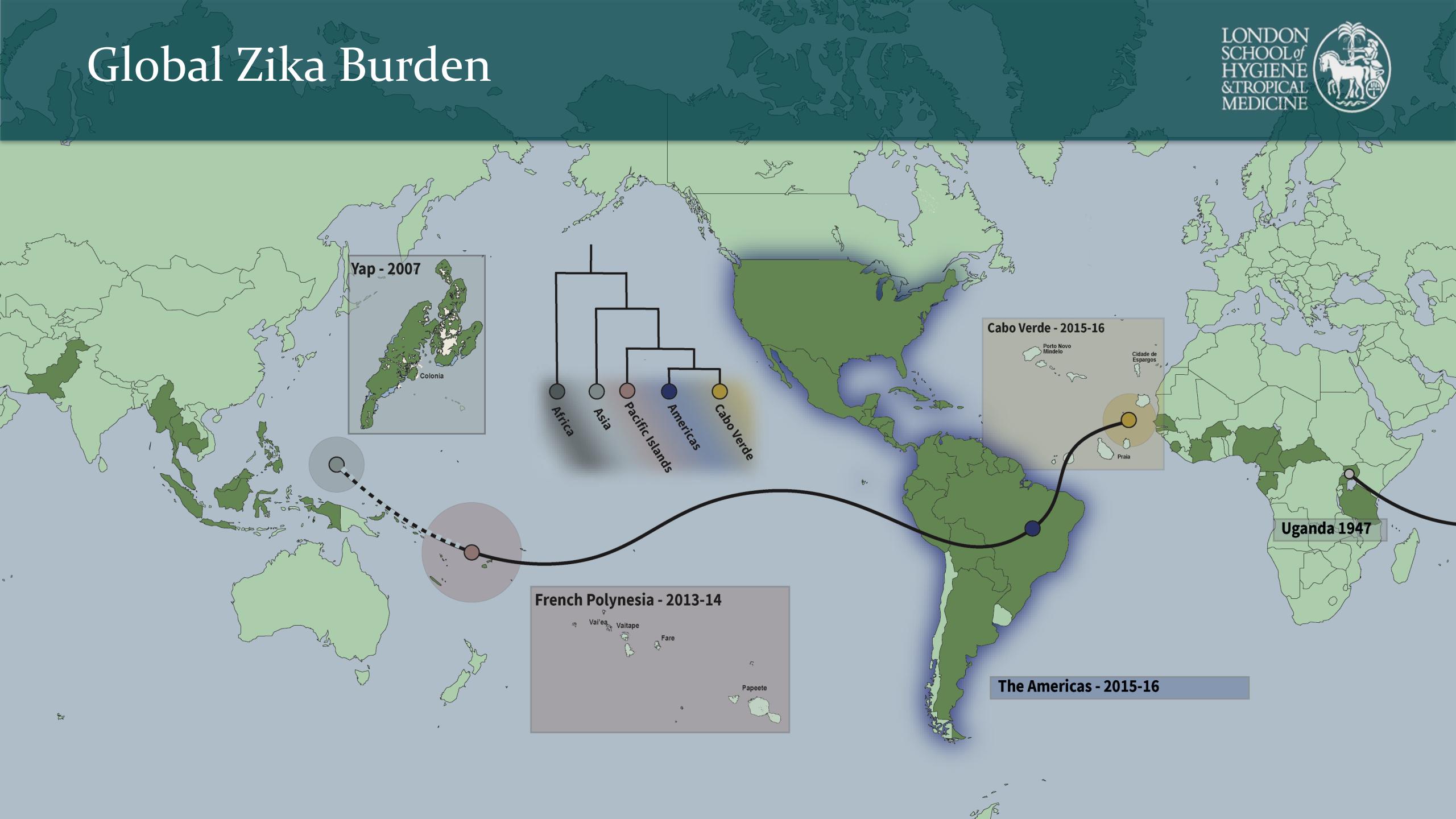
Zika Virus Morbidity

- Zika virus disease is caused by an RNA virus, a member of the *Flavivirus* genus.
- Other flavivirus include Dengue, yellow fever and West Nile Virus.
- The 2015-16 Zika outbreak in South America triggered the most recent WHO PHEIC, following association with microcephaly.
- The primary vector (transmitting agent) is the *Aedes Aegypti* mosquito.

SYMPTOMS OF MICROCEPHALY



Global Zika Burden



Nanopore Sequencing Data Activity

