Long Read Data Analysis

Richard Orton
Richard.Orton@glasgow.ac.uk
MRC-University of Glasgow Centre for Virus Research

Long Reads

Nanopore reads

not PacBio reads (I've never analysed them)

Practical

- **HCMV**
  - Not amplicons – normal reads
  - Downloaded from SRA
  - Align the reads using minimap2, call medaka consensus
  - Show how to process first sample (urine)
  - Up to you to adapt commands for the other sample (lung)

- **SARS-CoV-2**
  - Amplicons
  - ARTIC pipeline (conda)
  - Show how to process first sample (barcode06)
  - Up to you to adapt commands for another sample

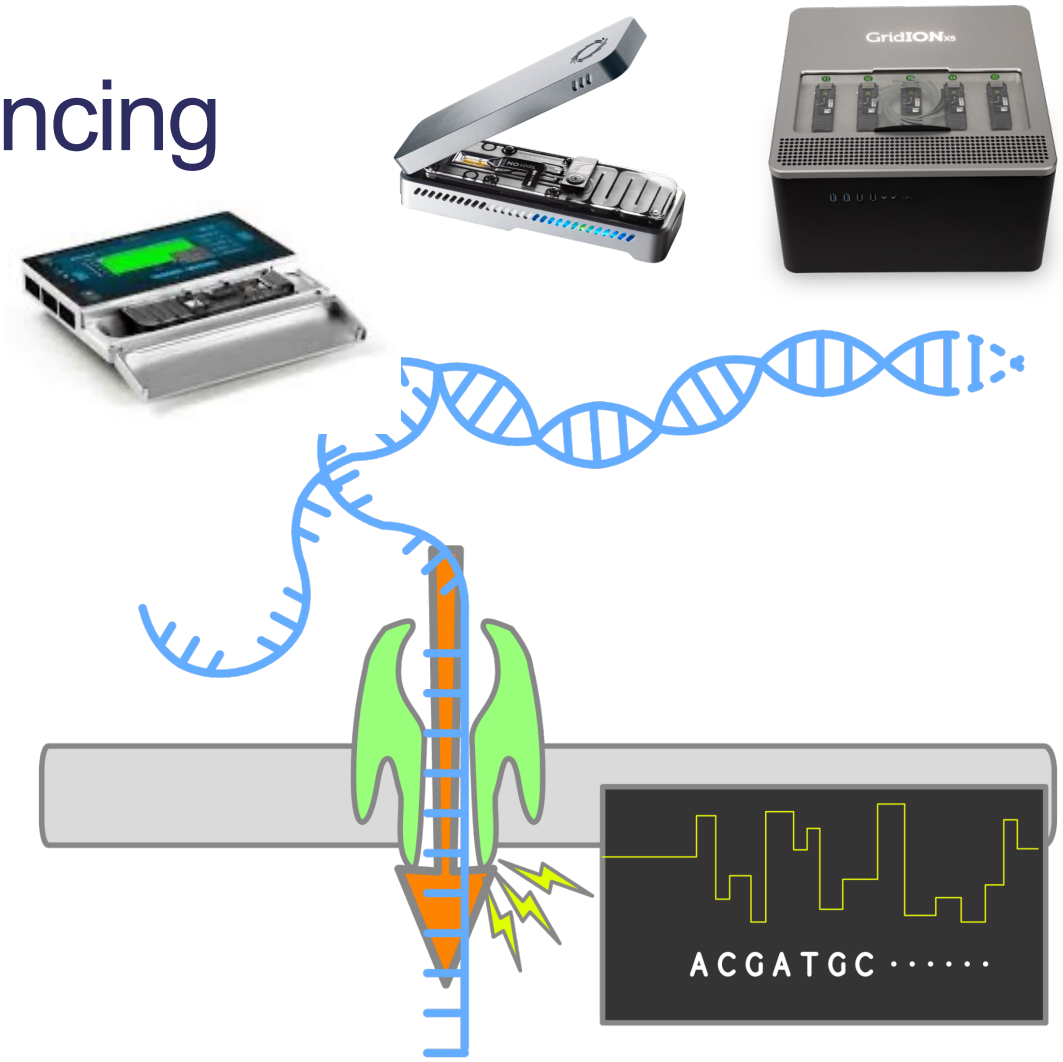
Nanopore videos

https://nanoporetech.com/platform/technology

Medical Research Council
University of Glasgow
CVR Centre for Virus Research

Oxford Nanopore Sequencing

- All Oxford Nanopore sequencing devices use **flow cells** which contain an array of tiny holes — **nanopores** — embedded in an electro-resistant membrane.
- Each nanopore corresponds to its own **electrode** connected to a **channel** and **sensor** chip, which measures the **electric current** that flows through the nanopore.
- When a molecule passes through a nanopore, the current is disrupted to produce a characteristic '**squiggle**'.
- The squiggle is then decoded using **basecalling** algorithms to determine the DNA or RNA **sequence** in real time.
- https://nanoporetech.com/how-it-works



Medical Research Council
University of Glasgow
CVR Centre for Virus Research

Nanopore Basecalling

- The electrical signals – or squiggles – for each read are stored in the FAST5 format – **NOW POD5 format**
- Typically, each **.fast5** file has the data for 4,000 reads in it
- Basecalling converts FAST5 into **FASTQ**
- **guppy** basecaller (available to download from Oxford Nanopore after registration) – **NOW DORADO**
- https://nanoporetech.com/how-it-works/basecalling
- You need a Graphical Processing Unit (**GPU**) on the computer running the MinION in order to perform basecalling in a reasonable amount of time
- Needs to be NVidia graphics card - **CUDA**
- GridION & Mk1c have a GPU built in

By Marrabbio2 - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1785530

By https://www.nvidia.com/object/io_1221568471314.html, Fair use, https://en.wikipedia.org/w/index.php?curid=53650345

UKRI Medical Research Council | University of Glasgow | CVR Centre for Virus Research

Basecalling Modes

- Basecalling converts FAST5/POD5 into **FASTQ -** modes:

- DNA **fast**: q-score 8
- DNA **hac** (high accuracy): q-score 9
- DNA **super** hac: q-score 10

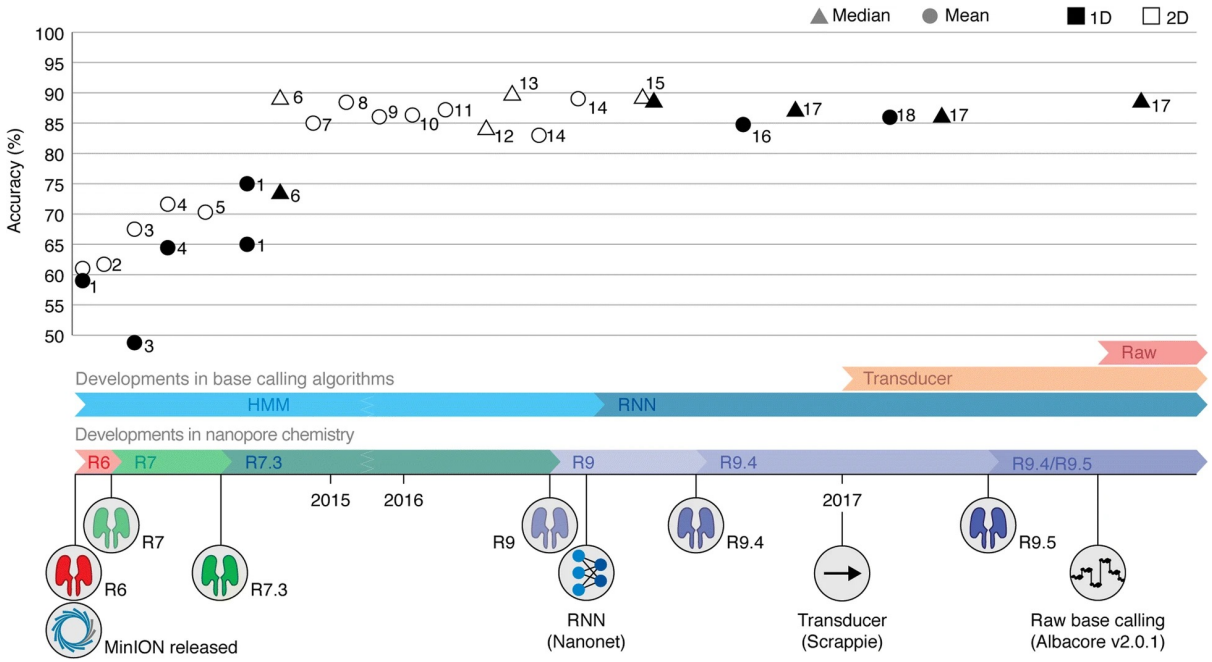
- If sequencing error is random, then if you have enough depth, you will get the correct consensus
  - Does compound low frequency variants

- BUT – errors (in particular) indels are much more likely at homopolymers – can still affect the consensus sequence

15	0.0316228
14	0.0398107
13	0.0501187
12	0.0630957
11	0.0794328
10	0.1000000
9	0.1258925
8	0.1584893
7	0.1995262
6	0.2511886
5	0.3162278
4	0.3981072
3	0.5011872
2	0.6309573
1	0.7943282
0	1.0000000

Medical Research Council | University of Glasgow | CVR Centre for Virus Research

Nanopore is continually improving



- Software, hardware, chemistry continually improving

https://nanoporetech.com/platform/accuracy

From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy
Rang et al. 2018
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1462-9/figures/1

Flow cell	Kit	Sequencing & basecalling parameters	Sample	Raw read accuracy	Output
R10.4.1	Ligation Sequencing Kit V14	400 bps, 5 kHz, HAC basecalling	Human HG002	99.0% (Q20)	●●●
R10.4.1	Ligation Sequencing Kit V14	400 bps, 5 kHz, SUP basecalling	Human HG002	99.5% (Q23)	●●●
R10.4.1	Ligation Sequencing Kit V14	400 bps, 5 kHz, Duplex basecalling	Human HG002	>99.9% (Q30)	●

Medical Research Council
University of Glasgow
CVR Centre for Virus Research

MinION Run Data Folder

- **After basecalling**
- **fast5_pass**
  - this folder contains all the raw FAST5 reads that **PASSED** the basic quality control filters of the guppy basecaller.
- **fast5_fail**
  - this folder contains all the raw FAST5 reads that **FAILED** the basic quality control filters of the guppy basecaller
- **fastq_pass**
  - this folder contains all the FASTQ reads that were converted from the those within the fast5_pass fiolder
- **fastq_fail**
  - this folder contains all the FASTQ reads that were converted from the those within the fast5_fail fiolder
- **sequencing_summary_FAO14190_ad60b376.txt**
  - the sequencing_summary file is produced by the basecaller and contains a summary of each read such as it's name, length, barcode and what FAST5 and FASTQ files it is located in.

MinION Barcoding and demultiplexing

- ONT Native Barcoding Expansion kit allows up to 96 samples to be sequenced at once on a single flow cell.
- If given the appropriate information, the guppy basecaller will **demultiplex** reads into their different **barcodes**
- **fastq_pass**
  - **barcode06**
    - barcode06_0.fastq, barcode06_1.fastq, barcode06_2.fastq, …
  - **barcode07**
    - barcode07_0.fastq, barcode07_1.fastq, barcode07_2.fastq, …
  - **barcode12**
    - barcode12_0.fastq
  - **unclassified**
    - unclassified_0.fastq, unclassified_1.fastq, unclassified_2.fastq, unclassified_3.fastq, …
- Typically, each FASTQ files contains 4,000 reads
- Unclassified contains reads whose barcode could not been determined – a large proportion of reads can end up here

1 - Combine reads

- **fastq_pass**
  - **barcode06**
    - barcode06_0.fastq, barcode06_1.fastq, barcode06_2.fastq, …
- Typically, each FASTQ files contains 4,000 reads

- The initial step is often to combine all the FASTQ read files into one file:

```
cat barcode06*.fastq > barcode06.fastq

zcat barcode06*.fastq.gz > barcode06.fastq
```

2 - QC

- Average read quality filtering already applied during base calling
  - DNA **fast**: q-score 8
  - DNA **hac** (high accuracy): q-score 9
  - DNA **super** hac: q-score 10

- Quality doesn't tend to decrease along the read length like illumina, so trimming is not normally done

- Often you want an overview of read lengths – and also to know what the longest read length is:
  - NanoPlot
  - prinseq
  - Assembly-stats

- Size filtering is sometimes applied
  - Amplicons – filter for expected size range

3 - Alignment

- Reads in **FASTQ** format, reference sequence to align to
- We need to use a nanopore capable aligner, that can cope with the elevated error rate:
  - **minimap2**
- Alignment creates a **SAM** file

```
minimap2 -x map-ont -a -o my.sam ref.fasta reads.fastq

samtools sort my.sam –o my.bam

samtools index my.bam

rm my.sam
```

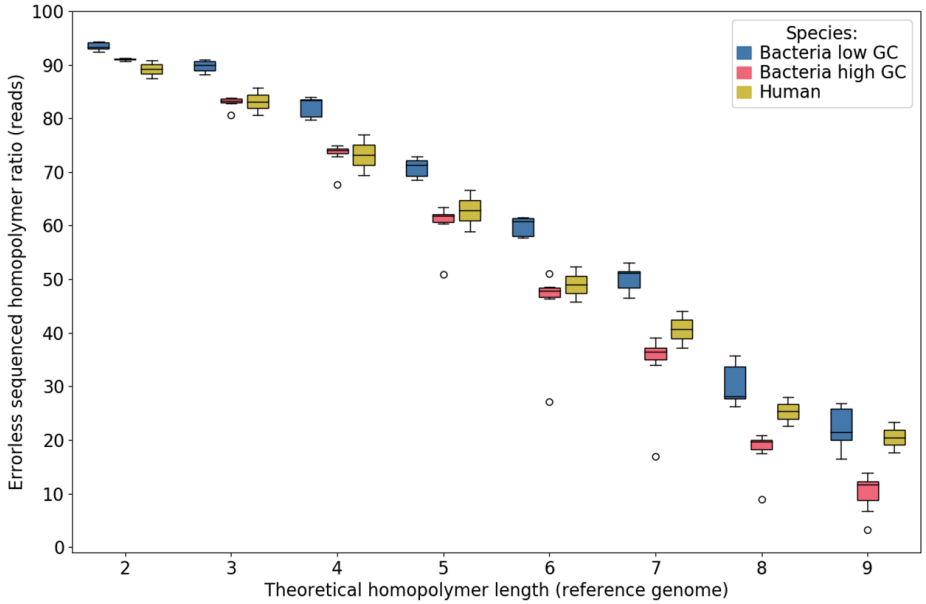
Medical Research Council | University of Glasgow | CVR Centre for Virus Research

As we now have a BAM most things from yesterday still apply:

- samtools to count the number of mapped/unmapped reads
- Coverage plots
- Tablet

4 – consensus

- As there is an elevated error rate, and as nanopore has a systematic bias (covered yesterday) around homopolymers
- We typically have to use a specialised consensus caller that takes nanopore error rates into account
  - nanopolish
  - **medaka**
- However, the latest r10 chemistry is improving things greatly



Delahaye C, Nicolas J (2021) Sequencing DNA with nanopores: Troubles and biases. PLoS ONE 16(10): e0257521. https://doi.org/10.1371/journal.pone.0257521

```
medaka_consensus -i reads.fastq -d ref.fasta -m r941_min_high_g360
```

-m tells medaka the nanopore chemistry (r941), the hardware (minion or promethion), the base calling mode (high) and the guppy version (g360)

The –g (don't use ref seq to replace regions of 0 cov) and –r (use gap-filling character) options can be used to further optimize the consensus

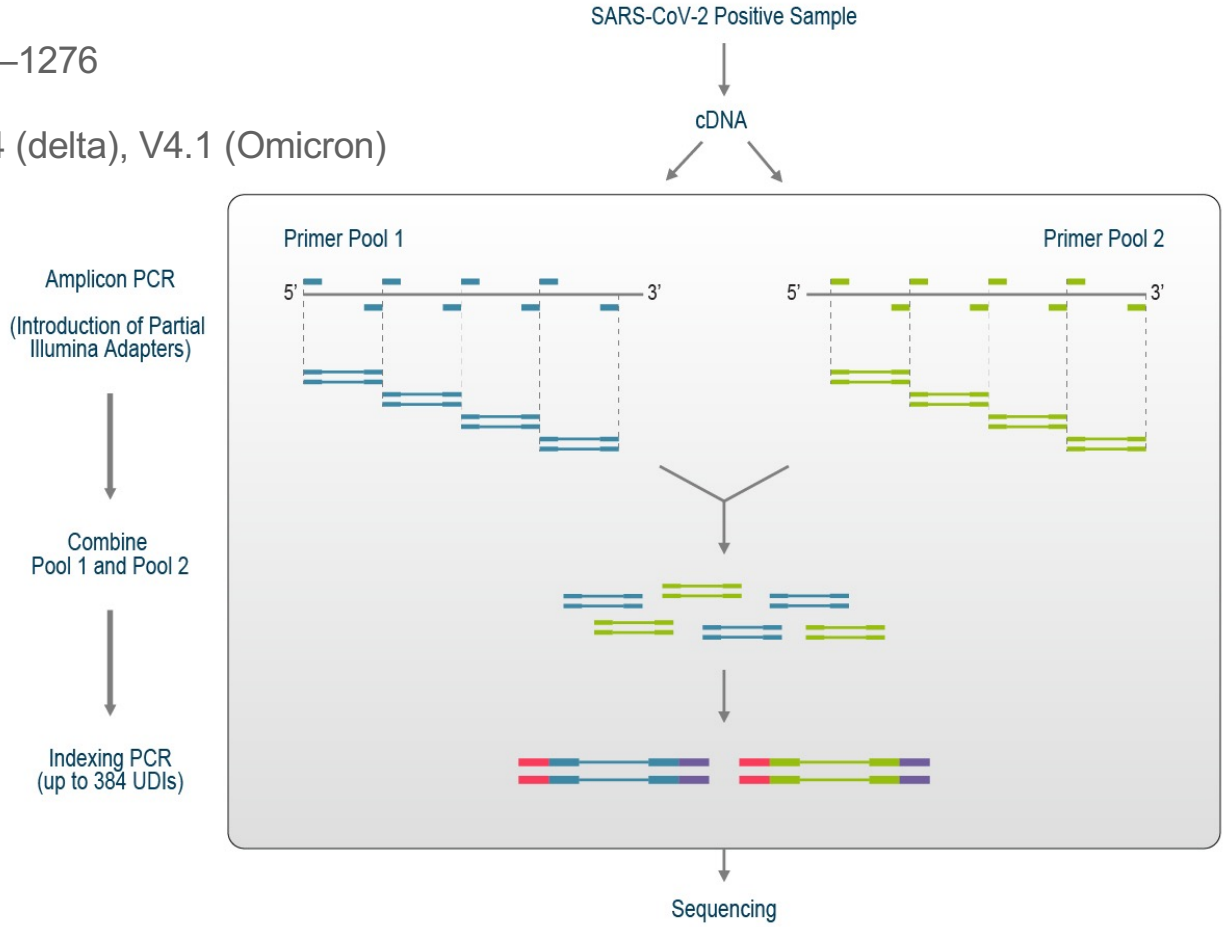
Medical Research Council | University of Glasgow | CVR Centre for Virus Research

ARTIC Amplicons – https://artic.network

- 98 ~400 base pair overlapping amplicons across the genomes
- Sequenced in two non-overlapping pools
- Quick et al (2017) Nat Protoc. 2017 Jun; 12(6): 1261–1276
- Tyson et al (2020) PMID: 32908977
- SARS-CoV-2 ARTIC Primer Versions: V1, V2, V3, V4 (delta), V4.1 (Omicron)

Pool 1 Pool2

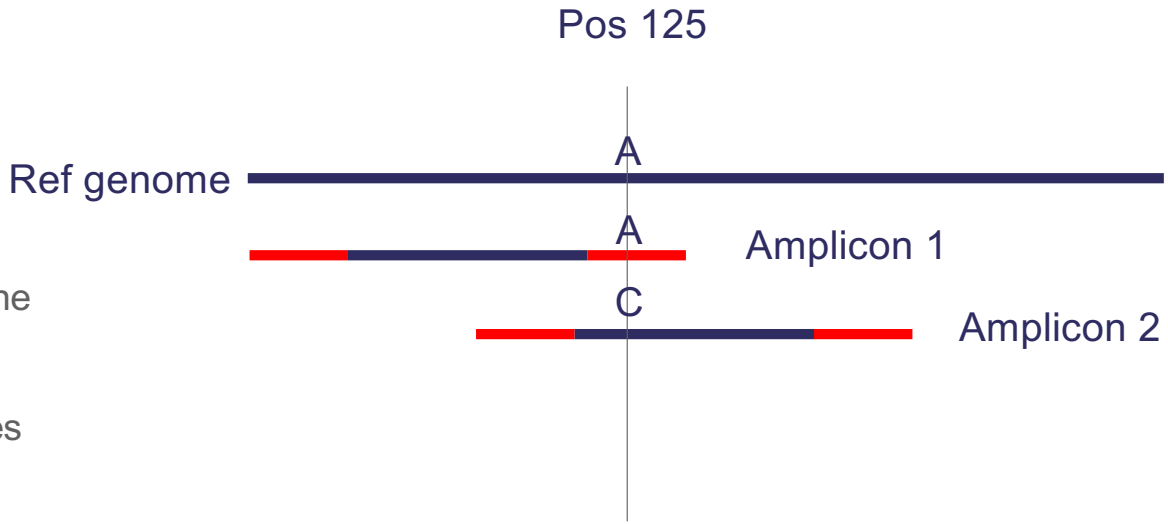
98 ~400bp overlapping amplicons



https://www.lexogen.com/sars-cov-2-whole-genome-sequencing-artic-panel/

UKRI Medical Research Council | University of Glasgow | CVR Centre for Virus Research

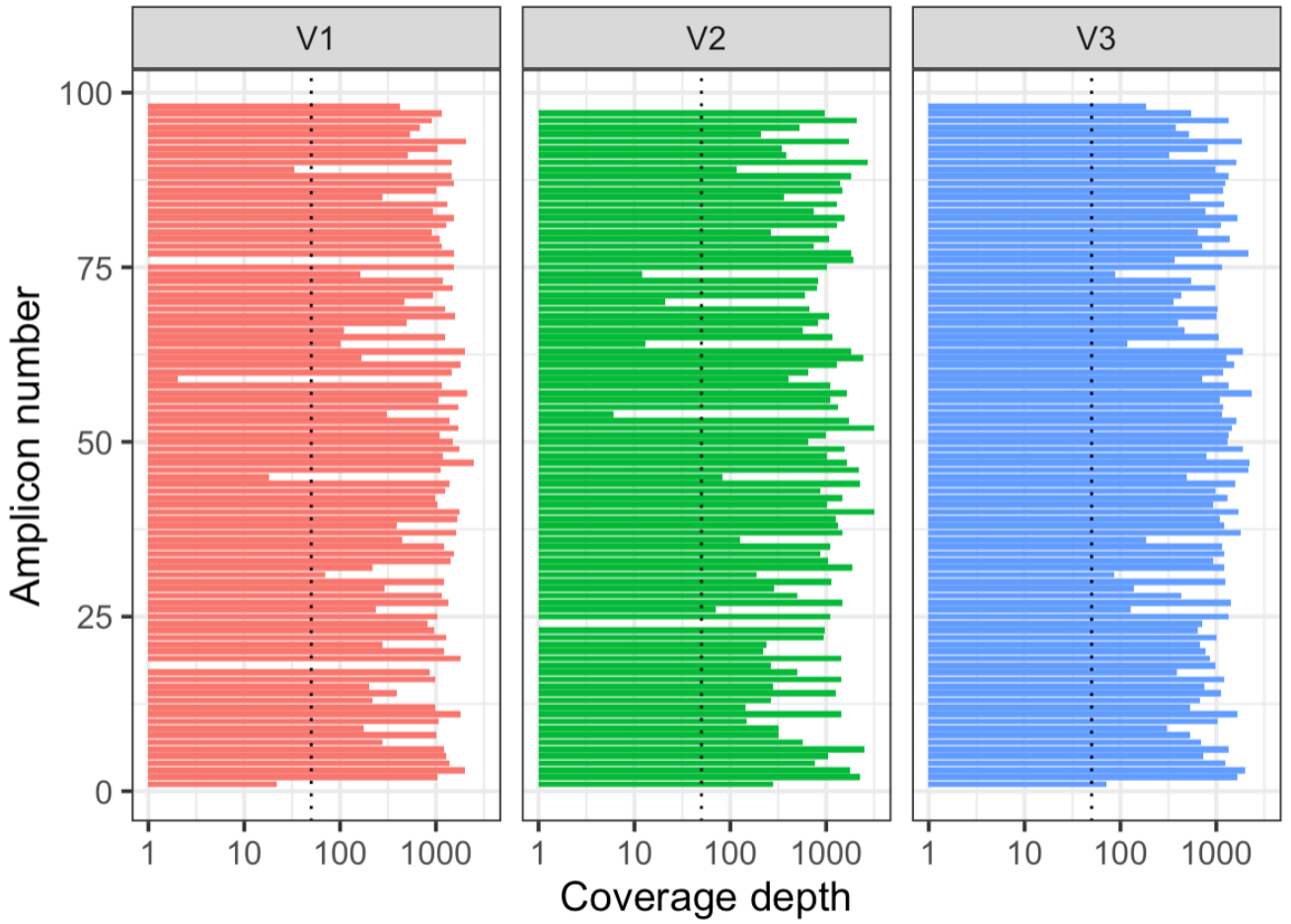
Primers need to be removed



- Primers are not the virus genome - they bind to the viral genome to initiative amplification
- Amplicons therefore contain the primer sequences themselves
- The viral genome may have a mutation where the primer binds - so primers need to be removed
- This can be done post-alignment – if you know where the primer binds to in relation to the genome

ARTIC Primer Versions

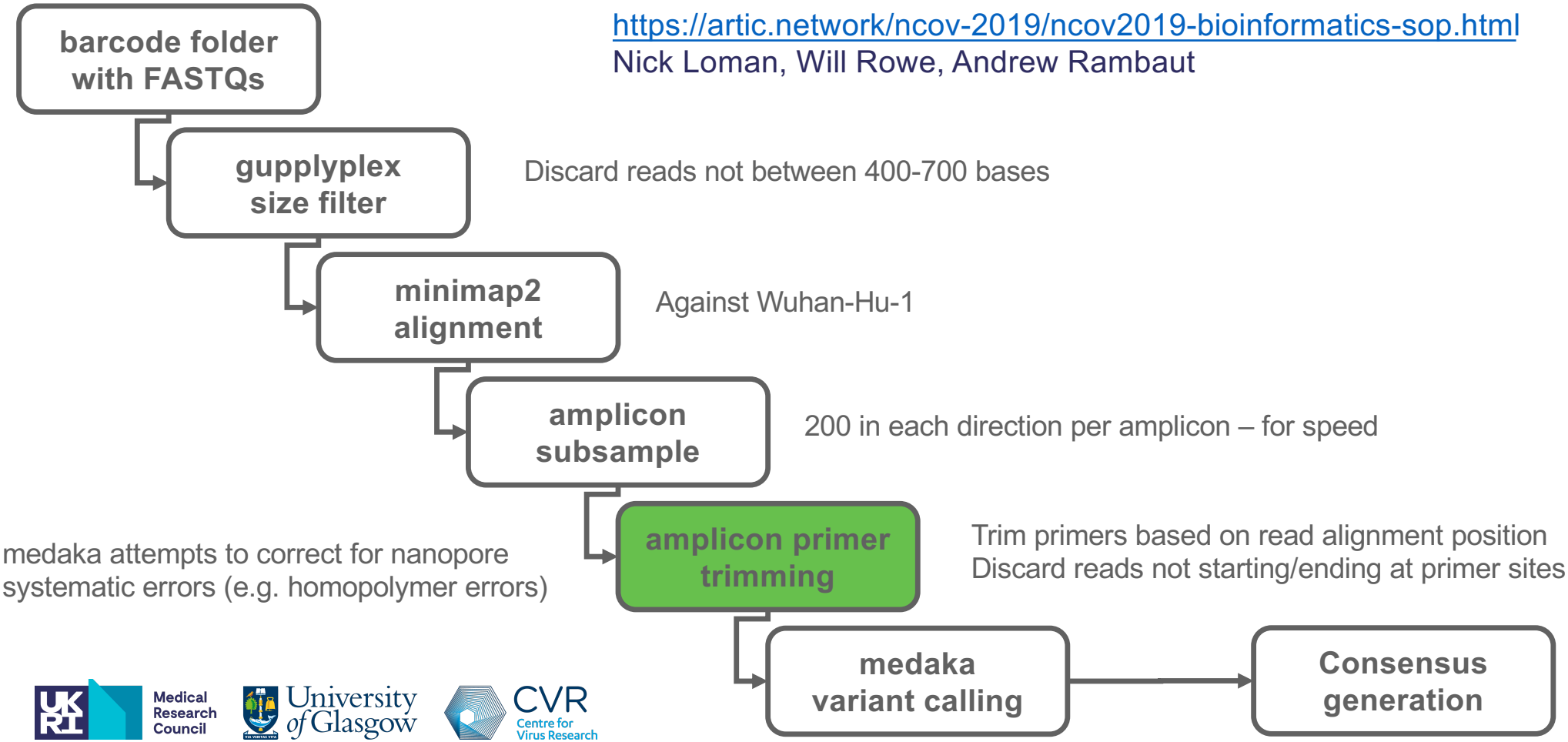
- SARS-CoV-2 ARTIC Primer Versions: V1, V2, V3, V4 (delta), V4.1 (Omicron)
- **V1:** systematic dropping out of amplicons 18 and 76: no more than 98% genome coverage
  - nCoV-2019_18_LEFT and nCoV-2019_76_RIGHT might form a dimer
- **V2:** substituting nCoV-2019_18_LEFT for nCoV-2019_18_LEFT_alt2.
  - Users quickly reported that this change caused other amplicons to drop out instead resulting in a sort of *amplicon whack-a-mole!*
- This illustrates the unpredictable interactions between primers within a multiplex PCR reaction.
- **V3:** addition of alternative primers (alts) for amplicons 7, 9, 14, 15, 18, 21, 44, 45, 46, 76 and 89
- **V4, V4.1:** deletions/mutations in delta and omicron



https://community.artic.network/t/ncov-2019-version-3-amplicon-release/19

Medical Research Council
University of Glasgow
CVR Centre for Virus Research

ARTIC SARS-CoV-2 Bioinformatics Protocol



https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html

Nick Loman, Will Rowe, Andrew Rambaut

Medical Research Council

University of Glasgow

CVR Centre for Virus Research

ARTIC Consensus Sequences

- **Medaka/ARTIC** called and filtered variants – **pass VCF** file
- **Wuhan-Hu-1** Reference Sequence
- **Modify** the reference based on the passed variants (**SNPs, indels**) in the pass VCF
- Calculate **depth** – flag genome positions with depth less than 40
- Replace all sites in the consensus that have a **depth less than 40** with an **N**
- **Ambiguity codes** are not used

- Coverage/depth of 40 is the threshold used to create a reliable consensus with nanopore data
- **Failed amplicon** = genome positions with a depth less than 40
  - As nanopore sequences the whole amplicon on a single read – this means there were less than 40 reads of that amplicon

ARTIC Commands – One line each

artic guppyplex
--skip-quality-check
--min-length 400
--max-length 700
--directory ./barcode06
--prefix cvr124a

Creates: cvr124a_barcode06.fastq

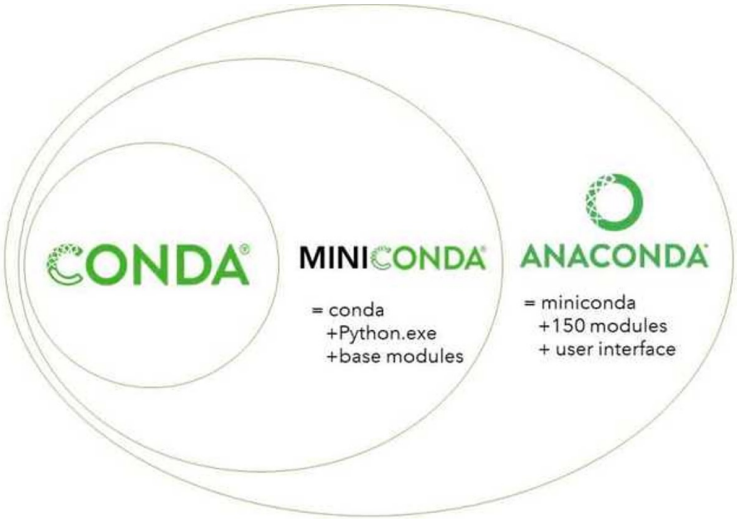
artic minion
--normalise 200 **(how much to subsample each amplicon)**
--threads 4
--scheme-directory ~/artic-ncov2019/primer_schemes
--read-file cvr124a_barcode06.fastq
--medaka
--medaka-model `r941_min_high_g360`
nCoV-2019/V2 **(primer scheme to use, within the scheme directory)**
barcode06 **(output name to use)**

Creates: barcode06.consensus.fasta, barcode06.sorted.bam, barcode06.primertrimmed.rg.sorted.bam, barcode06.pass.vcf.gz, …

Medical Research Council
University of Glasgow
CVR Centre for Virus Research

Conda - https://conda.io

- **Conda** is an open source package management system and environment management system
- A conda environment is a directory that contains a specific collection of conda packages that have been installed.
- The ARTIC pipeline comes as a conda environment and has been pre-installed on the course Ubuntu virtual machine, which installs all the other tools it needs (such as minimap2, nanopolish etc).
- conda activate artic-ncov2019
- conda deacticate



https://www.educative.io/answers/anaconda-vs-miniconda

Mamba

https://mamba.readthedocs.io/en/latest/

Medical Research Council

University of Glasgow

CVR Centre for Virus Research

Practical

- **HCMV**
  - Not amplicons – normal reads
  - Downloaded from SRA
  - Align the reads using minimap2, call medaka consensus
  - Show how to process first sample (urine)
  - Up to you to adapt commands for the other sample (lung)

- **SARS-CoV-2**
  - Amplicons
  - ARTIC pipeline (conda)
  - Show how to process first sample (barcode06)
  - Up to you to adapt commands for another sample