# Summary of the paper

## (Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis)

الاسم :هدير اسماعيل محمد حسين إسماعيل

الفرقه :الثالثه

القسم :المعلوماتيه الطبيه

## Assignment 1: abstract and introduction summary

■We want to do whole genome sequencing for (SARS-COV-2) with the use of Long - read sequencing from Oxford Nanopore

■The problem is the short read sequencing platform takes a lot of time and more costly so we want to use long read sequencing devices because it is more efficient, it's response time little and the possibility of moving it easier but we have a problem in them also and this problem is the accuracy of the sequence

■In order to solve this problem here we perform viral WGS with ONT and illumina platforms on 157 matched SARS-COV-2 positive patient specimens and synthetic RNA control enabling rigorous evaluation of analytical performance

■after we use this solution We report that, despite the elevated error rates observed in ONT sequencing reads, highly accurate consensus-level sequence determination was achieved, with single nucleotide variants (SNVs) detected at >99%sensitivity and >99% precision above a minimum ~60-fold coverage depth, thereby ensuring suitability for SARS-CoV-2 genome analysis. ONT sequencing also identified a surprising diversity of structural variation within SARS-CoV-2 specimens that were supported by evidence from short-read sequencing on matched samples.

■Now we can say that however ONT is more better than illumina but there is a problem in it like that ONT sequencing failed to accurately detect short indels and variants at low read-count frequencies. as a result of the problem in two sequencing technologies and in order to address concerns regarding ONT sequencing accuracy and evaluate its analytical validity for SARS-CoV-2 genomics, we have performed amplicon-based nanopore and short-read WGS on matched SARS-CoV-2-positive patient specimens and synthetic RNA controls, allowing rigorous evaluation of ONT performance characteristics.

■(SARS-CoV-2) is the causative pathogen for COVID-19 disease1,2. SARS-CoV-2 is a positive-sense single-stranded RNA virus with a ~30-kb poly-adenylated genome Complete genome sequences published in January 2020

■Whole-genome sequencing (WGS) of SARS-CoV-2 provides additional data to complement routine diagnostic testing. Viral WGS informs public health responses by defining the phylogenetic structure of disease outbreaks5. Integration with epidemiological data identifies transmission networks and can infer the origin of unknown cases

■WGS can be performed by 2 ways

■The first way is PCR amplification

■The second way is hybrid-capture of the reverse-transcribed SARS-CoV-2 genome sequence, followed by high-throughput sequencing. Short-read sequencing technologies (e.g., Illumina) enable accurate sequence determination and are the current standard for pathogen genomics. However, long-read sequencing devices from Oxford Nanopore Technologies (ONT) offer an alternative with several advantages. ONT devices are portable, cheap, require minimal supporting laboratory infrastructure or technical expertise for sample preparation, and can be used to perform rapid sequencing analysis with flexible scalability

■ONT has been used during Ebola, zika and other diseases outbreak

■And like what we said before ONT devices exhibit lower read - level sequencing accuracy than short - read platforms and This may have a disproportionate impact on SARS-CoV-2 analysis, due to the virus' low mutation rate ($8 \times 10^{-4}$ substitutions per site per year26), which ensures erroneous (false-positive) or undetected (false-negative) genetic variants have a strong confounding effect.

■and to solve this problem we have performed amplicon-based nanopore and short-read WGS on matched SARS-CoV-2-positive patient specimens and synthetic RNA controls.

## Related work(discussion)

Nanopore sequencing offers an alternative to established short-read platforms for viral WGS with several advantages. ONT devices: are relatively inexpensive, highly portable and require minimal associated laboratory infrastructure; enable rapid generation of sequencing data and even real-time data analysis; require comparatively simple procedures for library preparation and; offer flexibility in sample throughput, accommodating single , multiple or tens/hundreds of specimens per flow-cell16,18.

Due to the relatively low mutation rate observed in SARS-CoV-226, accurate sequence determination is vital to correctly define the phylogenetic structure of disease outbreaks. With ONT sequencing known to exhibit higher read-level sequencing error rates than short-read technologies23,24,25, reasonable concerns exist about suitability of the technology for SARS-CoV-2 genomics.

The present study resolves these concerns, demonstrating accurate consensus-level SARS-CoV-2 sequence determination with ONT data. Although SNVs alone are sufficient for routine phylogenetic analysis, small indels and large structural

variants can profoundly impact gene function and are, therefore, of interest to studies of virus evolution and pathogenicity15.

While short-read sequencing platforms remain the gold-standard for high-throughput viral sequencing, the advantages to portability, cost and turnaround-time afforded by nanopore sequencing imply that this emerging technology can serve an important complementary role in local, national and international COVID-19 response strategies.

## The summary of the methodology and results

### Methodology

The controls comprise synthetic RNA generated by in vitro transcription of the SARS-CoV-2 genome sequence, representing the complete genome in 6 ~5 kb continuous sequences. We note that it is not possible to amplify the entire SARS-CoV-2 genome in this way, since amplicons that span boundaries of the 6 ~5-kb IVT products necessarily fail. Nevertheless, we were able to evaluate ~95% of the SARS-CoV-2 genome sequence.

SARS-CoV-2-positive extracts from 157 cases, tested at NSW Health Pathology East Serology and Virology Division , were retrieved from storage and included in this study. Wherever relevant, ethical regulations for work with human participants with informed consent were observed, with oversight by HREC at South Eastern Sydney Local Health District . Reverse-transcription was performed on viral RNA extracts using Superscript IV VILO Master Mix , which contains both random hexamers and oligo-dT primers. Prepared cDNA was then amplified separately with each of 14 ~2.5-kb amplicons tiling the SARS-CoV-2 genome, as described elsewhere6 . All 14x amplicon products from a given sample were then pooled at equal abundance and partitioned into separate aliquots for analysis by short-read and nanopore sequencing.

Trimmed alignments were converted to pileup format using samtools mpileup 38, with anomalous read pairs retained , base alignment quality disabled and all bases considered, regardless of PHRED quality . Variants were identified using bcftools call 38, assuming a ploidy of 1 , then filtered for a minimum read depth of 30 and minimum quality of 20. Variants were classified according to their read-count frequencies as consensus or sub-consensus variants, with the latter further divided into high , intermediate or low-frequency . Variants at read-count frequencies below 20% were considered to be potentially spurious and excluded on this basis.

Depending on the illumina pooled amplicons were prepped for short-read sequencing and each sequencing lane a blank samples was also prepared and sequenced and the resulting reads were aligned to Wuhan - Hu-1 reference genome using bwa mem (0.7.12-r1039)36. Primer sequences were trimmed from the termini of read alignments using iVar (1.0)37. Trimmed alignments were converted to pileup format using samtools mpileup

with anomalous read pairs retained (--count-orphans), base alignment quality disabled (--no-BAQ) and all bases considered, regardless of PHRED quality (--min-BQ 0). Variants were identified using bcftools call and after we filtered for minimum read depth of 30 and minimum quality of 20

And we classified the variants according to their read-count frequencies as consensus (>80% reads supporting the variant) or sub-consensus (20–80%) variants, with the latter further divided into high (60–80%), intermediate (40–60%) or low-frequency (20–40%). Variants at read-count frequencies below 20% were considered to be potentially spurious and excluded on this basis.

Up to 12 samples were multiplexed on a FLO-FLG001, FLO-MIN106D or FLO-PRO002 or flow-cell and sequenced on a GridION X5 or PromethION P24 device, respectively. In addition, a no-template negative control from the PCR amplification step was prepared in parallel and sequenced on each flow-cell . At this point, the run was terminated and the flow-cell washed using the ONT Flow Cell Wash kit , allowing re-use in subsequent runs.

The resulting reads were basecalled using Guppy and aligned to the Wuhan-Hu-1 reference genome using minimap2 40. Consensus-level variant candidates were identified using each of two workflows developed by ARTIC , using Nanopolish41 or Medaka to variants, respectively. Sub-consensus level variant candidates were identified using Varscan2 43.

For synthetic RNA controls, read-level quality metrics, such as sequencing error rates, were derived from read alignments using pysamstats, with any bases that differed from the Wuhan-Hu-1 reference sequence considered errors.

accuracy of variant detection by ONT sequencing was evaluated by comparison to the set of variants identified by Illumina sequencing in matched cases. Variant candidates identified by Illumina/ONT could then be considered concordant based on matching genome position, reference base and alternative base/s. The following statistical definitions were used to evaluate results

To identify structural variation, nanopore reads were re-aligned to the Wuhan-Hu-1 reference genome using the rearrangement-aware aligner NGMLR 44. Sniffles 44 was then used to detect candidate variants with a minimum length of 10 bp and 20 supporting reads. To validate SVs detected with ONT alignments, split short-read alignments and discordant read-pairs were extracted from matched Illumina libraries using lumpy45.

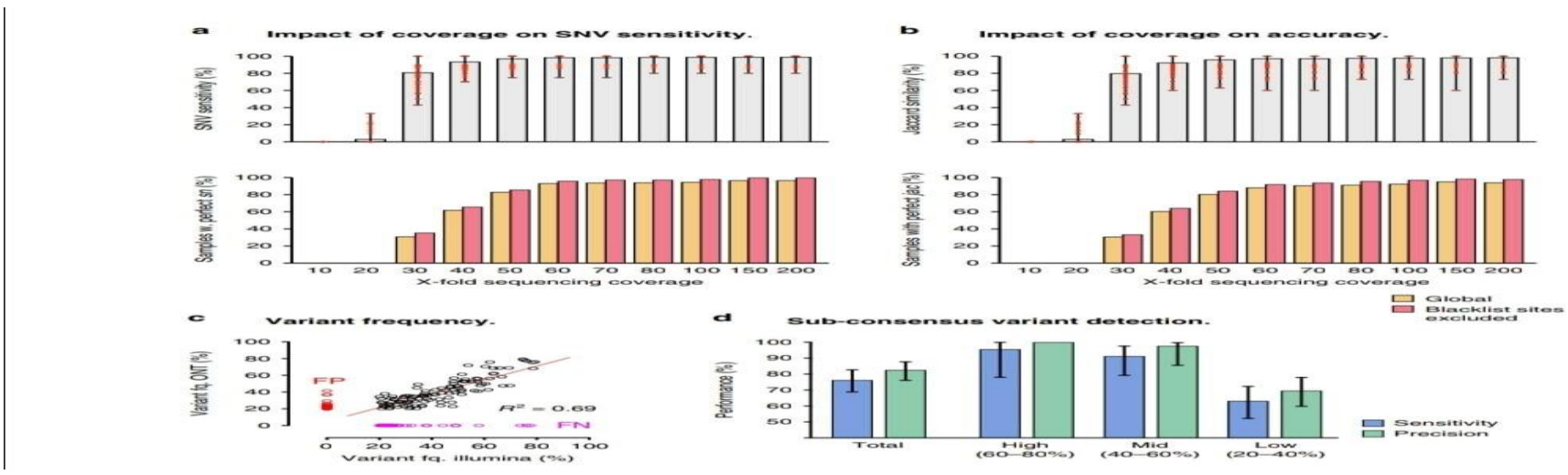
Results:

This indicates that ONT sequencing errors are not entirely random but are influenced by local sequence context. To further evaluate the suitability of ONT sequencing for SARS-CoV-2 genomics, we conducted rigorous proficiency testing using bona fide clinical specimens. By comparison to the Wuhan-Hu-1 reference strain, Illumina sequencing identified 7.6 consensus single-nucleotide variants and 0.04 indels, on average, per sample. A further 1.0 SNVs and 0.2 indels per sample were detected at sub-consensus read-count frequencies , indicative of intra-specimen genetic diversity .

We used each of two best-practice bioinformatics pipelines developed by the ARTIC network to identify consensus variants with ONT sequencing data. In general, ONT variant candidates identified by both pipelines were highly concordant with the Illumina comparison set. Illumina variants were detected with 99.17% sensitivity and 99.58% precision by Nanopolish, compared to 98.33% sensitivity and 99.24% precision by Medaka . Only 1/7 of consensus indels in the Illumina comparison set was detected by either Nanopolish or Medaka, while a further five and nine false-positive indels were detected by the respective pipelines .

Inspection of false-positive and false-negative variant candidates detected with ONT sequencing data showed that these tended to occur in low-complexity sequences, which are known to be refractory to ONT base-calling algorithms23. We identified 15 problematic low-complexity sites in the SARS-CoV-2 genome ranging in size from 9 to 42 bp in length that showed elevated read-level sequencing error rates . Consensus SNVs detected with the Nanopolish workflow were identical between ONT and Illumina data in 155/157 of samples . To do so, we down-sampled nanopore sequencing reads from a uniform 200-fold coverage across the SARS-CoV-2 genome and repeated variant detection across a range of coverage depths .

Both sensitivity and precision of variant detection were strongly influenced by sequencing coverage, showing a sharp decline below ~50-fold coverage depth, with minimal improvement observed above ~60-fold .




Sensitivity with which Illumina comparison SNVs at consensus-level variant frequencies were detected via ONT sequencing on matched SARS-CoV-2 specimens . Data are plotted separately for genome-wide variant detection and variant detection with error-prone ‘blacklist’ sites excluded . b Same as in a but Jaccard similarity scores for all variant candidates are plotted instead of SNV sn. c Correlation of variant frequencies observed for SNV candidates detected at sub-consensus frequencies with Illumina and ONT sequencing.

Only one variant, a 328-bp deletion in ORF8 , was detected in multiple specimens, although highly similar 28 bp and 29 bp deletions were also detected in S in two unrelated specimens .