

1 **Comparison of different sequencing techniques with multiplex real-**
2 **time PCR for detection to identify SARS-CoV-2 variants of concern**

3 **Short titled: Comparison of different sequencing techniques for SARS-CoV-2**

4 Diyanath Ranasinghe^{1#}, Tibutius Thanesh Pramanayagam Jayadas^{1#}, Deshni Jayathilaka^{1#},
5 Chandima Jeewandara^{1#}, Osanda Dissanayake¹, Dinuka Guruge², Dinuka Ariyaratne¹, Dumni
6 Gunasinghe¹, Laksiri Gomes¹, Ayesha Wijesinghe¹, Ruwan Wijayamuni¹, Gathsaurie Neelika
7 Malavige^{1*}

8
9 ¹ Allergy Immunology and Cell Biology Unit, Department of Immunology and Molecular
10 Medicine, University of Sri Jayewardenepura, Nugegoda, Sri Lanka; ² Colombo Municipal
11 Council, Colombo, Sri Lanka

12
13
14 #contributed equally

15
16 Correspondence should be addressed to:

17 Prof. Neelika Malavige DPhil (Oxon), FRCP (Lond), FRCPath (UK)

18 Allergy Immunology and Cell Biology Unit, Department of Immunology and Molecular
19 Medicine, Faculty of Medical Sciences, University of Sri Jayawardanapura, Sri Lanka.

20 Tel +94 (0) 772443193; Fax: +94 (0) 112802026, Email: gathsaurie.malavige@ndm.ox.ac.uk

21

22

1 **Abstract**

2 As different SARS-CoV-2 variants emerge and with the continuous evolution of sub-lineages of
3 the delta and other variants, it is crucial that all countries carry out sequencing of at least >1% of
4 their infections, in order to detect emergence of variants with higher transmissibility and with
5 ability to evade immunity. However, as many resource-poor countries are unable to sequence
6 adequate number of viruses, we compared to usefulness of a commercially available multiplex
7 real-time PCR assay to detect important single nucleotide polymorphisms (SNPs) associated
8 with the variants and compared the sensitivity, accuracy and cost effectiveness of the Illumina
9 sequencing platform and the Oxford Nanopore Technologies' (ONT) platform. 138/143 (96.5%)
10 identified as the alpha and 36/39 (92.3%) samples identified as the delta variants due to the
11 presence of lineage defining SNPs by the multiplex real time PCR, were assigned to the same
12 lineage by either of the two sequencing platforms. 34/37 of the samples sequenced by ONT had
13 <5% ambiguous bases, while 21/37 samples sequenced using the Illumina generated <15%
14 ambiguous bases. However, the mean PHRED scores averaged at 32.35 by Illumina reads but
15 10.78 in ONT. Sub-consensus single nucleotide variations (SNV) were highly correlated
16 between both platforms ($R^2=0.79$) while indels showed a weaker correlation ($R^2=0.13$).

17 Although the ONT had a slightly higher error rate compared to the Illumina technology, it
18 achieved higher coverage with a lower number of reads, generated less ambiguous bases and was
19 significantly cheaper than Illumina sequencing technology.

20 **Key words:** SARS-CoV-2; COVID-19, Illumina; Nanopore; variant PCR; Single nuclear
21 polymorphism

22

1 **Introduction**

2 Since the emergence of the SARS-CoV-2 virus two years ago, it continues to evolve giving rise
3 to many variants with higher transmissibility and immune evasion abilities [1]. Despite a lower
4 mutation rate of SARS-CoV-2 compared to many other RNA viruses such as HIV and influenza,
5 [2], many SARS-CoV-2 variants of concern (VOCs) have emerged, which continue to drive the
6 pandemic. Although the ongoing COVID-19 vaccine drive and other control measures have
7 resulted in a reduction in the number of cases and deaths, many countries still report intense
8 transmission rates [3]. This high rate of continued transmission of the virus is likely to give rise
9 to new variants, which may evade immunity induced by vaccines, or may have a higher
10 transmissibility than the dominant delta variant. This is already evident with the emergence of
11 many delta-sub lineages [4], of which some are under investigation for higher transmissibility. In
12 order to detect such possible new variants emerging, the WHO has recommended that all
13 countries carry out genomic sequencing in at least 1% of their infections [5].

14

15 Although there are many sub-lineages of SARS-CoV-2 emerging [4], so far the WHO has named
16 four SARS CoV-2 variants as variants of concern (VOC), which are they are alpha (B.1.1.7),
17 beta (B.1.351), gamma (P.1) and delta (B.1.617.2)[5]. Even though sequencing is the gold
18 standard to identify VOCs, it is time consuming and too expensive for many lower income and
19 lower middle-income countries. In addition, although many developed countries sequence over
20 5% of their positive cases, many lower income countries are unable to sequence 1% and many
21 sequence <0.2% [6]. Therefore, in order to carry out surveillance to monitor the current VOCs
22 and to identify mutations of concern than may occur in such variants, cheaper and rapid methods

1 are required that can be used in resource poor settings. The B.1.1.7 variant (alpha) was identified
2 initially in the UK due to the changes in the SARS-CoV-2 real-time PCR results that occurred in
3 the primers targeting the spike protein (S-drop) when using the Taq Path (Thermo Fisher
4 Scientific Inc.)[7]. Subsequently, many multiplex real-time PCR assays were developed to
5 identify Single Nucleotide Polymorphisms (SNPs) that associate with these four VOC such as
6 the HV69-70del, K417N, N417T, W152C, E484K, N501Y, L452R, D614G, P681H or
7 V1176F[8]. The combination of these different SNPs in the SARS-CoV-2 which can be
8 identified by multiplex real-time PCR can therefore be used to identify these four VOCs [9].

9
10 Many sequencing platforms are currently used to sequence the SARS-CoV-2 virus, which have
11 their advantages and disadvantages [10]. Comparison of different sequencing protocols have
12 shown that they have significant differences in sensitivity, reproducibility, and precision for
13 detection of SNPs [11]. However, with the continued emergence and spread of new delta sub
14 lineages, in addition to carrying out surveillance for VOCs, it is crucial to carry out whole
15 genomic sequencing to identify variants which develop mutations of concern, that may evade
16 immunity or may associate with higher transmissibility. However, as many lower income
17 countries such as Sri Lanka, do not have adequate resources to carry out sequencing of >1% of
18 the positive samples, we investigated the sensitivity and accuracy of a commercially available
19 multiple real-time PCR assays for detection of different SNPs associated with VOCs. In addition,
20 we compared the sensitivity, reproducibility, precision of identification SNPs and the cost of
21 sequencing using the Illumina and Oxford Nanopore sequencing platforms.

22

1 **Materials and methods**

2 **Samples**

3 Quantitative SARS-CoV-2 real-time PCR (qRT-PCR) was carried out on 190 samples of sputum
4 or nasopharyngeal swabs sent for diagnostic purposes to our laboratory from patients with
5 COVID-19. Of the 190 samples, all analysed by the multiplex real-time PCR to detect SNPs of
6 VOCs. 97 were subjected to sequencing by the Illumina platform, 57 by the Oxford nanopore
7 sequencing technology and 37 by both methods. Those with a cycle threshold (Ct) values of <30
8 were subjected to genomic sequencing and further multiplex q RT-PCR for identification of
9 SNPs associated with VOCs. Briefly, viral RNA was extracted using QIAamp viral RNA mini
10 kit (Qiagen, USA), SpinStar™ Viral Nucleic Acid Extraction kit 1.0 (ADT Biotech, Malaysia)
11 or FastGene RNA Viral Kit, (Nippon Genetics, Germany) according to manufacturer's
12 instructions.

13 Ethical approval for the study was obtained by the Ethics Review Committee of the University of
14 Sri Jayewardenepura.

15

16 **Multiplex qRT-PCR for identification of SNPs in VOCs**

17 Multiplex qRT-PCR to detect SNPs was carried out using the SARS-COV-2 variant PCR
18 Allplex™ assay 1 and assay 2 in all 190 samples. The variant assay 1 detects SNPs for 501Y,
19 E484K and 69-70 deletion, whereas variant assay 2 detects the SNPs L452R, W152C, K417T
20 and K17N. The variant PCR assays were carried out according to the manufacturer's
21 instructions using 5µl of the extracted RNA from samples (Supplementary methods). All the

1 thermal cycling steps were carried out in a CFX96 Deep Well, Real time System (Bio-Rad,
2 USA).

3 **Genomic sequencing using Oxford Nanopore (ONT) platform**

4 The extracted RNA of 57/190 samples were subjected to nanopore sequencing according to the
5 manufacture's instruction using the SQK-RBK110.96 rapid barcoding kit (ONT, Oxford, UK).
6 1200 bp tiled PCR amplicons were generated with midnight primers as described in Freed et al.,
7 2020[12]. All the thermal cycling steps were carried out in a QuantStudio™ 5 Real-Time PCR
8 Instrument (Applied biosystems, Singapore). Barcodes were attached to resulting amplicons and
9 pooled together before the clean-up step. Finally, 800ng of library was loaded into R9.4.1 flow
10 cell and sequenced on the Oxford Nanopore Minion Mk1C. The run was stopped once desired
11 number of reads (~15,000 reads per sample) were achieved.

12

13 **Genomic sequencing using the Illumina platform**

14 97 were subjected to sequencing by the Illumina sequencing platform. Library preparation
15 Illumina sequencing was carried out using the AmpliSeq for Illumina SARS-CoV-2 Community
16 Panel, in combination with AmpliSeq for Illumina library prep, index, and accessories (Illumina,
17 San Diego, USA). The methodology is further described in supplementary methods.

18

19 **Bioinformatics and statistical analysis**

20 Data from Illumina Nextseq 550 platform were base called using inbuilt bcl2fastq and the
21 resulting data were analyzed using the Dragen somatic pipeline on Basespace sequencing hub

1 [13]. Resulting Binary alignment files (BAM), variant call files (VCF) and consensus fasta
2 sequences were taken further for comparison. ONT fast5 data were base called and
3 demultiplexed using guppy version 4.4.2 occupying the fast model with a 60 single end barcode
4 score. Resultant fastq files were analyzed using Nextflow wf-artic version v0.3.2
5 (<https://github.com/epi2me-labs/wf-artic>) with “fast_min_variant_c507” model. Read length
6 cutoff values obtained were in the range of 150bp and 1200bp, without barcode trimming.
7 Resulting BAM, VCF and consensus fasta sequences were carried forward. Fastq and Bam
8 statistics were generated using fastp 2 [14], Nanocomp 3 [15], samtools version 1.9 and in-house
9 scripts. VCF data were processed using vcftools version 0.1.15 and bcftools version 1.7[16].
10 Multi-way pileup (mpileup) was made from ONT and Illumina alignments before analyzing with
11 VarScan2[17] for detection of low frequency variants.

12
13 Fasta sequences were run through the Covsurver variant annotation pipeline
14 (<https://www.gisaid.org/epiflu-applications/covsurver-mutations-app/>) and for lineage
15 assessment each sequence was assessed using Pangolin version v3.1.11 [18] with the lineage
16 version of 2021-08-09 and Serious constellations of reoccurring phylogenetically-independent
17 origin (Scorpio) version 0.3.12 (<https://github.com/cov-lineages/scorpio>). All the consensus fasta
18 sequences were aligned to the reference sequence MN908947.3 using MAFFT version 7.487 and
19 a maximum likelihood phylogenetic tree was created using IQ-TREE version 1.6.1 [19]
20 occupying the best fit TIM2+F+R2 model with 1000 bootstrap replicates. Raw fastq files for 74
21 datasets were deposited to NCBI Sequence read archive (SRA) while the fasta sequences were
22 uploaded to GISAID (Supplementary table 1). Plots were generated in R version 4.0.1 using
23 ggplot2, dplyr and tidyverse.

1

2 **Results**

3 Multiplex real-time PCR to detect SNPs of VOCs was carried out initially in 190 unblinded
4 samples. These 190 samples were then sequenced either on Illumina or ONT (depending on the
5 availability of sequencing reagents) to evaluate the accuracy of the variant PCR. 37/190 of the
6 samples were sequenced with both the Illumina and ONT platforms to compare the quality and
7 accuracy of the data generated by these two platforms.

8

9 **Comparison between next generation sequencing with multiplex real-time PCR assays for** 10 **detection of SNP (variant PCR assay)**

11 143/190 samples were positive for the SNPs N501Y and HV69/70 deletion in the spike protein
12 by the multiplex PCR and therefore, considered to be as B.1.1.7. The samples in which these two
13 SNPs were not detected (n=47) were re-screened using the variant assay 2, which detected the
14 SNPs L452R, W152C, K417T and K17N. Based on the results of the variant assay 2, 39 samples
15 were positive only for the spike L452R mutation and were therefore, considered to be B.1.617.
16 One sample with spike K417N mutation along with N501Y and HV69/70 deletion from the
17 assay was considered to be B.1.351. 7/190 samples were not assigned to be any VOC as they did
18 not have any of the SNPs associated with the VOCs.

19

20 138/143 (96.5%) of the sample which were considered to be B.1.1.7 by variant PCR assay were
21 also assigned the same Pangolin lineage based on the sequencing by either sequencing platforms.

1 The remaining 5/143 could not be assigned to a lineage, due to the low sequencing coverage.
2 3/39 samples which were considered to be B.1.617 due to the presence of the SNP spike L452R,
3 were re-classified into C.36 lineage, while the rest (36/39) were classified into Pangolin lineage
4 B.1.617.2 (92.3%). 7/7 samples in which the SNPs were not detected, were classified into either
5 B.1, B.1.1 or B.1.411 (Supplementary table 2). Of the 305/332 individual mutations detected by
6 the variant PCR assay were confirmed by at least one of the sequencing methods. Rest of the
7 unconfirmed spike mutations were masked by N bases in the consensus sequences due to either
8 low coverage or base call accuracy.

9

10 **Comparison between Illumina and ONT sequencing platforms**

11 Sequencing statistics

12 37 SARS-CoV-2 samples sequenced using both approaches had more than 45% of the genome
13 coverage and were included in the analysis for comparison of both techniques. Ampliseq SARS-
14 CoV-2 primers used for Illumina library contained 247 amplicons in 2 pools covering >99% of
15 the genome (reference positions between 30 to 29842). The Midnight primer scheme used in the
16 ONT, with just 29 primers in two pools created 1200bp long amplicons and typically covered 54
17 to 29903 positions. The 150bp Illumina library yielded ~50 million bases, whereas 1200bp ONT
18 library yielded only ~9 million. However, despite the lower number of reads with ONT, it was
19 able to achieve twice the coverage depth of which was achieved by the Illumina platform (Table
20 1). With the Illumina sequencing technology, the samples required an average of 372,654 reads
21 per sample, whereas ONT achieved even higher coverage depth with an average 9741 reads per
22 sample due to its ability to score ultra-long reads. The higher proportion of ambiguous bases

1 generated by the Illumina platform, resulted in non-detection of several mutations in each
2 sample. We observed one consistent amplicon drop at third amplicon in ONT dataset due to
3 K634N mutation in ORF1a, which is prevalent in Sri Lankan delta sub-lineage,
4 B.1.617.2.AY.28.

5
6 Illumina raw reads had an average PHRED quality score (probability of error per base call in a
7 log scale) of 32.35 with a highest score of 36 which is between 1 in 1000 to 1 in 10,000
8 probability of calling an incorrect base (99.9 – 99.99% accuracy). ONT on the other hand had
9 recorded an average score of 10.78 with a 16.2 highest PHRED quality score indicating between
10 1 in 10 to 1 in 100 probability of error (90% - 99% accuracy). This explains the average identity
11 to the reference genome of 99.6% in Illumina reads and 91.4% in ONT reads (Table 1 and Figure
12 1).

13
14 Consensus accuracy

15 We used iVar [20] and Medaka workflows to call consensus for Illumina and ONT respectively.
16 ONT detected more average mutations per sample compared to Illumina (36 vs 31) and majority
17 of them were non-synonymous mutations scoring average of 22 and 25 amino acid substitutions
18 in Illumina and ONT respectively. Some known amino acid mutations such as Spike N501Y and
19 NSP12 P323L were constantly missed in Illumina consensus sequences which was caused by
20 masking of those regions due to drop of coverage (Figure 2). However, average number of
21 deletions per sample were between 14 to 15 range for both approaches (Figure 3). Average
22 percentage of ambiguous (N) bases was higher in Illumina sequences compared to ONT (9% vs

1 6%), whereas samples with less than 10% ambiguous bases were higher for ONT compared to
2 Illumina (29 vs 25). The ONT consensus sequences had 2 frameshift mutations at nucleotide
3 positions of 1634 and 21992, whereas no frameshift mutations were found in the Illumina
4 dataset.

5 Due to improved coverage over the SARS-CoV-2 genome, more samples which were sequenced
6 using the ONT platform were lineage call-able by both Pango and Scorpio nomenclature. Pango
7 lineage calls of 16/37 were identical in both platforms while 7/37 lineage calls were different
8 between the two sequencing platforms. Two B.1.1.7 lineage calls were classified as B.1.1 in
9 Illumina dataset and the remaining 5 calls were different only at the sub-lineage level
10 (Supplementary table 2). In the combined phylogenetic tree of consensus sequences, 28/37
11 samples paired with their counterpart sequences, while those that were not paired together had
12 moderate to high (3% - 31%) ambiguous bases in one of the counterpart sequences. Higher than
13 98% bootstrap support can be observed in 21/37 samples which had > 90% genome coverage
14 from both sequencing technologies (Figure 4).

15

16 Intra-sample variation

17 We further analyzed BAM alignments using VarScan2 to determine the rate of both consensus
18 and sub-consensus variants between two technologies. Collectively, 191 single nucleotide
19 variants (SNVs) and 9 indels were detected in the read pileup of Illumina dataset, while 226
20 SNVs and 304 indels were detected in ONT read pileup. However, 175 SNVs were concordant
21 across both datasets suggesting they are true variants. 15 SNVs detected only in the Illumina
22 dataset are of very low allele frequency (5%-0.1%) whereas 51 SNVs detected only in the ONT

1 dataset are of allele frequency ranging between 52%-0.1% (Supplementary table 2). 8 indels
2 were concordant between both platforms while a significant number (296) of false positive indels
3 were detected in the ONT dataset with allele frequency between 52% - 0.1%. We also observed a
4 strong correlation between the SNV frequencies ($R^2=0.79$) between the two platforms, while
5 there was a weak correlation between indel frequencies ($R^2=0.13$) of Illumina and ONT datasets
6 (Figure 5).

7 **Discussion**

8 In this study we compared the accuracy and sensitivity of a commercial multiple real-time PCR
9 assay for detection of different SNPs with two sequencing technologies in identifying SARS-
10 CoV-2 VOCs. 190 samples which were tested by the variant PCR technique showed 100%
11 concordance with the results of either Illumina or ONT sequencing platforms as far as the VOC
12 assignment. However, 3 samples which were considered to be B.1.617 due to the presence of the
13 SNP L452R, were assigned to the C.36 Pangolin lineage following sequencing. Since there are
14 many variants with the L452R [21], when multiple SARS-CoV-2 variants are co-circulating,
15 classifying a virus into the B.1.617 lineage based on the L452R mutation alone, would result in
16 inaccuracies. However, for the SNPs E484K, K417T, L452R, K417N, W152C and 69/70
17 deletion no false positive or false negatives were detected. Therefore, the VOCs assigned to
18 either B.1.1.7 or B.1.351 showed 100% accuracy with both sequencing platforms. Therefore, the
19 variant PCR appears to be a relatively inexpensive and rapid technique to carry out surveillance
20 for SARS-CoV-2 variants in resource poor settings. However, with the dominance of the delta
21 variant globally [22], these variant PCR assays have limited value in detection of new mutations
22 of concern arising in the delta variant that may give rise to higher transmissibility and immune
23 evasion.

1
2 In this study we also compared to accuracy and ease of use of two sequencing platforms. The
3 ONT rapid barcoding workflow occupies transposase-based library preparation, which does not
4 require individual sample washes and allows samples to be processed uniformly without
5 quantification or normalization [23]. For Illumina, traditional ligation-based library preparation
6 was used which required extended preparation time and effort. Run time per 96 samples on ONT
7 is nearly half the time required for Illumina (14 hours compared to 26 hours) mainly due to the
8 ability of real-time data analysis with ONT. The rapid barcoding library preparation method used
9 for the ONT platform also required less reagents and therefore, was cheaper than Illumina
10 sequencing. Although the single ended barcoding of the transposase-based libraries is thought to
11 result in improper demultiplexing and sample crossover, it has shown to rarely affect variant
12 calling and consensus generation in ONT [12].

13
14 The Illumina sequencing produced ~200bp long amplicons, whereas the ONT sequencing
15 platform produced 1200bp. The long amplicons generated by the ONT sequencing achieved
16 higher coverage and less ambiguous bases even with 10,000 reads (after filtering). As a result of
17 this, certain mutations were not detected by the Illumina platform due to a larger proportion of
18 ambiguous bases. For instance, 34/37 of the ONT samples had < 5% ambiguous bases while
19 29/37 samples sequenced using Illumina generated <15% ambiguous bases. However, the mean
20 PHRED scores averaged at 32.35 in Illumina reads but 10.78 in ONT. This difference results in a
21 base call error probability of 1 in 10 by the ONT and 1 in 1000 for Illumina sequencing platform.
22 The reduced accuracy in base calls and increase the frequency of erroneous bases can be
23 minimized by using high accuracy or super accuracy models of guppy basecaller[24]. Therefore,

1 ONT sequencing detected more nucleic acid and amino acid substitutions compared to Illumina,
2 possibly due to the improved coverage. While short read sequencing is considered as the gold
3 standard for sequencing of viral genomes [25], despite ONT having a slightly higher error rate,
4 ONT appears to generate high quality data at a very affordable cost. Therefore, ONT appears to
5 be the most cost effective, high throughout sequencing technology, especially suited for
6 countries with limited resources for genomic surveillance and for identification of emerging
7 variants of concern.

8

9 **Conclusions**

10 We have compared the usefulness, accuracy and reliability of two sequencing technologies and
11 also compared the usefulness of a commercial multiplex real-time PCR for surveillance of
12 VOCs, by identification of SNPs associated with the VOCs. We found that the multiplex real-
13 time PCR assay detected the alpha variant with 96.5% accuracy (5/147 could not be confirmed
14 due to low coverage by NGS) and the delta variant with 92.3% accuracy, when compared with
15 either sequencing technology. Although the ONT had a slightly higher error rate compared to the
16 Illumina technology, it achieved higher coverage with a lower number of reads, generated less
17 ambiguous bases and was significantly cheaper than Illumina sequencing technology.

18

19 **Funding information**

20 World Health Organization; World bank, Sri Lanka Covid 19 Emergency Response and Health
21 Systems Preparedness Project (ERHSP) of Ministry of Health Sri Lanka funded by World Bank.

1

2 **Conflicts of interest**

3 None of the authors have any conflicts of interest.

4

5 **References**

- 6 1. Dyson L, Hill EM, Moore S, Curran-Sebastian J, Tildesley MJ, Lythgoe KA, et al.
7 Possible future waves of SARS-CoV-2 infection generated by variants of concern with a range of
8 characteristics. *Nat Commun.* 2021;12(1):5730. Epub 2021/10/02. doi: 10.1038/s41467-021-
9 25915-7. PubMed PMID: 34593807; PubMed Central PMCID: PMCPMC8484271.
- 10 2. Luo R, Delaunay-Moisan A, Timmis K, Danchin A. SARS-CoV-2 biology and variants:
11 anticipation of viral evolution and what needs to be done. *Environ Microbiol.* 2021;23(5):2339-
12 63. Epub 2021/03/27. doi: 10.1111/1462-2920.15487. PubMed PMID: 33769683; PubMed
13 Central PMCID: PMCPMC8251359.
- 14 3. Coronavirus Pandemic (COVID-19) [Internet]. OurWorldInData.org. 2021 [cited 28th
15 June 2021]. Available from: <https://ourworldindata.org/coronavirus>.
- 16 4. Alaa Abdel Latif JLM, Manar Alkuzweny, Ginger Tsueng, Marco Cano, Emily Haag,
17 Jerry Zhou, Mark Zeller, Emory Hufbauer, Nate Matteson, Chunlei Wu, Kristian G. Andersen,
18 Andrew I. Su, Karthik Gangavarapu, Laura D. Hughes. Global variant report: Center for Viral
19 Systems Biology; 2021 [cited 2021 7th November 2021]. Available from:
20 <https://outbreak.info/location-reports?loc=LKA>.

- 1 5. WHO. Guidance for surveillance of SARS-CoV-2 variants: Interim guidance, 9 August
2 2021. WHO Headquarters (HQ): 2021 9th August 2021. Report No.: WHO/2019-
3 nCoV/surveillance/variants/2021.1.
- 4 6. SARS-CoV-2 sequence entries with complete collection date information shared via
5 GISAID [Internet]. 2021 [cited 13 September 2021]. Available from:
6 <https://www.gisaid.org/test5/submission-tracker-global/>.
- 7 7. Borges V, Sousa C, Menezes L, Goncalves AM, Picao M, Almeida JP, et al. Tracking
8 SARS-CoV-2 lineage B.1.1.7 dissemination: insights from nationwide spike gene target failure
9 (SGTF) and spike gene late detection (SGTL) data, Portugal, week 49 2020 to week 3 2021.
10 Euro Surveill. 2021;26(10). Epub 2021/03/13. doi: 10.2807/1560-7917.ES.2021.26.10.2100130.
11 PubMed PMID: 33706862; PubMed Central PMCID: PMC7953529.
- 12 8. WHO. Methods for the detection and identification of SARS-CoV-2 variants, March
13 2021. World Health Organization. Regional Office for Europe.: 2021 WHO/EURO:2021-2148-
14 41903-57493.
- 15 9. Harper H, Burridge A, Winfield M, Finn A, Davidson A, Matthews D, et al. Detecting
16 SARS-CoV-2 variants with SNP genotyping. PloS one. 2021;16(2):e0243185. Epub 2021/02/25.
17 doi: 10.1371/journal.pone.0243185. PubMed PMID: 33626040; PubMed Central PMCID:
18 PMC7904205 This does not alter our adherence to PLOS ONE policies on sharing data and
19 materials. All remaining authors have declared that no competing interests exist.
- 20 10. Hourdel V, Kwasiborski A, Baliere C, Matheus S, Batejat CF, Manuguerra JC, et al.
21 Rapid Genomic Characterization of SARS-CoV-2 by Direct Amplicon-Based Sequencing
22 Through Comparison of MinION and Illumina iSeq100(TM) System. Front Microbiol.

- 1 2020;11:571328. Epub 2020/10/27. doi: 10.3389/fmicb.2020.571328. PubMed PMID:
2 33101244; PubMed Central PMCID: PMC7546329.
- 3 11. Liu T, Chen Z, Chen W, Chen X, Hosseini M, Yang Z, et al. A benchmarking study of
4 SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples. *iScience*.
5 2021;24(8):102892. Epub 2021/07/27. doi: 10.1016/j.isci.2021.102892. PubMed PMID:
6 34308277; PubMed Central PMCID: PMC78294598.
- 7 12. Freed NE, Vlkova M, Faisal MB, Silander OK. Rapid and inexpensive whole-genome
8 sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid
9 Barcoding. *Biol Methods Protoc*. 2020;5(1):bpaa014. Epub 2020/10/09. doi:
10 10.1093/biomethods/bpaa014. PubMed PMID: 33029559; PubMed Central PMCID:
11 PMC7454405.
- 12 13. Illumina. BaseSpace Sequence Hub. 2021.
- 13 14. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
14 *Bioinformatics*. 2018;34(17):i884-i90. Epub 2018/11/14. doi: 10.1093/bioinformatics/bty560.
15 PubMed PMID: 30423086; PubMed Central PMCID: PMC6129281.
- 16 15. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack:
17 visualizing and processing long-read sequencing data. *Bioinformatics*. 2018;34(15):2666-9.
18 Epub 2018/03/17. doi: 10.1093/bioinformatics/bty149. PubMed PMID: 29547981; PubMed
19 Central PMCID: PMC6061794.
- 20 16. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant
21 call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-8. doi:
22 10.1093/bioinformatics/btr330. PubMed PMID: 21653522; PubMed Central PMCID:
23 PMC3137218.

- 1 17. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2:
2 somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome*
3 *Res.* 2012;22(3):568-76. Epub 2012/02/04. doi: 10.1101/gr.129684.111. PubMed PMID:
4 22300766; PubMed Central PMCID: PMC3290792.
- 5 18. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C, et al. A dynamic
6 nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat*
7 *Microbiol.* 2020. doi: 10.1038/s41564-020-0770-5. PubMed PMID: 32669681.
- 8 19. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective
9 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.*
10 2015;32(1):268-74. Epub 2014/11/06. doi: 10.1093/molbev/msu300. PubMed PMID: 25371430;
11 PubMed Central PMCID: PMC4271533.
- 12 20. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An
13 amplicon-based sequencing framework for accurately measuring intrahost virus diversity using
14 PrimalSeq and iVar. *Genome Biol.* 2019;20(1):8. Epub 2019/01/10. doi: 10.1186/s13059-018-
15 1618-7. PubMed PMID: 30621750; PubMed Central PMCID: PMC6325816.
- 16 21. Tchesnokova V, Kulasekara H, Larson L, Bowers V, Rechkina E, Kisiela D, et al.
17 Acquisition of the L452R Mutation in the ACE2-Binding Interface of Spike Protein Triggers
18 Recent Massive Expansion of SARS-CoV-2 Variants. *Journal of clinical microbiology.*
19 2021;59(11):e0092121. Epub 2021/08/12. doi: 10.1128/JCM.00921-21. PubMed PMID:
20 34379531; PubMed Central PMCID: PMC8525575.
- 21 22. Genomic epidemiology of hCoV-19 [Internet]. 2020 [cited 16.09.1010]. Available from:
22 <https://www.gisaid.org/epiflu-applications/hcov-19-genomic-epidemiology/>.

- 1 23. Radukic MT, Brandt D, Haak M, Muller KM, Kalinowski J. Nanopore sequencing of
2 native adeno-associated virus (AAV) single-stranded DNA using a transposase-based rapid
3 protocol. *NAR Genom Bioinform.* 2020;2(4):lqaa074. Epub 2021/02/13. doi:
4 10.1093/nargab/lqaa074. PubMed PMID: 33575623; PubMed Central PMCID:
5 PMCPMC7671332.
- 6 24. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for
7 Oxford Nanopore sequencing. *Genome Biol.* 2019;20(1):129. Epub 2019/06/27. doi:
8 10.1186/s13059-019-1727-y. PubMed PMID: 31234903; PubMed Central PMCID:
9 PMCPMC6591954.
- 10 25. Bull RA, Adikari TN, Ferguson JM, Hammond JM, Stevanovski I, Beukers AG, et al.
11 Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat*
12 *Commun.* 2020;11(1):6272. Epub 2020/12/11. doi: 10.1038/s41467-020-20075-6. PubMed
13 PMID: 33298935; PubMed Central PMCID: PMCPMC7726558.

14

15

16

17

18

19

20

21

1
2
3
4
5
6 **Tables**

Sequencing metrics	Illumina	ONT
Number of samples	37	37
Alignment start and end positions	30 - 29842	54 - 29903
Mean Coverage depth	109	266
Total Number of reads	372,654	9741
Yielded bases	50,576,802	9,307,884
Fraction of bases aligned	0.928	0.897
Mean Read length	140	945
Average identity	99.6	91.4
Average PHRED score	32.35	10.78
No of SNPs	31	36
No of amino acid substitutions	22	25
No of deletions	14	15
No of Amino acid substitutions	22	25
No of frameshift mutations	0	2

% of ambiguous bases	9%	6%
No of samples with < 10% ambiguous bases	25	29
Successful Pangolin calls	31	34
Successful Scorpio calls	25	31
Run time (h) for 96 samples	26	14
Cost per sample (USD)	~150-250	~10-40

1 **Table 1.** Basic sequencing matrices for Illumina and Oxford Nanopore (ONT) outputs of 37
2 samples.

3

4

5

6

7

8

9

10

11

12

13

14

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

Figure legends

Figure 1. PHRED base call quality score distribution of samples sequenced by Illumina and ONT. Distribution plot of PHRED (probability of error per base call in a log scale) quality score (x axis) and error probability (secondary x axis) derived from the PHRED score for the data set sequenced from Illumina (n=37) and ONT (n=37). The scores of ONT are shown in blue and Illumina in red. The mean PHRED scores/error probability are shown with the dashed line for each technology. The mean PHRED scores averaged at 32.35 in Illumina reads and 10.78 in ONT.

Figure 2. Comparison of amino acid changes detected in SARS-CoV-2 genomes by both sequencing technologies. Annotated amino acid substitutions and deletions detected in each sample (n=37). Mutations colored in green indicates they are synonymously detected by both sequencing technologies, whereas yellow and red indicate mutations detected exclusively by only one technology. The X axis indicates each amino acid, which is denoted by the original amino acid, its position in the protein and the substitution/deletion. Amino acid deletions are denoted by “del”.

1 **Figure 3. Percentage of ambiguous bases (% of N) compared to the no of mutations (SNP)**
2 **detected in each sample.** The trend between percentage of ambiguous bases is shown in the x
3 axis and number of mutations projected onto the consensus sequences is shown in the y axis.
4 The numbers displayed on each shape denotes the sample identification number (n=37)
5 sequenced by Illumina (red) or ONT (blue). Both technologies had detected a maximum SNPs
6 with 1% - 5% ambiguous bases. The consensus sequences generated by Illumina had a varying
7 percentage of ambiguous bases between 1% - 30%, whereas ONT sequencing generated either <
8 5% or more than 40% ambiguous bases due to its longer read lengths. Altogether ONT had
9 detected more SNPs than Illumina between of 1% - 5% ambiguous bases.

10

11 **Figure 4. Combined maximum likelihood phylogenetic tree created using sequence pairs of**
12 **37 the samples.** The ML tree was generated using the consensus sequences of each sequencing
13 technology with 1000 bootstrap replicates using TIM2+F+R2 model. Tree is rooted on SARS-
14 CoV-2 reference MN908947.3 and with samples sequences by Illumina coloured red and those
15 sequenced by ONT coloured blue. Bootstrap support values are shown on each branch. 21/37
16 samples coupled together with < 98% bootstrap support had > 90% genome coverage from both
17 Illumina and ONT datasets while 7/37 samples coupled together with less than 98% bootstrap
18 support. 9/37 of the samples which failed to couple with their counterpart from ONT or Illumina
19 had moderate to high (3% - 31%) ambiguous bases in either sequences.

20

21 **Figure 5. Correlation of sub-consensus allele frequencies observed for SNV and Indels**
22 **between two sequencing technologies.** The Correlation between sub-consensus single

1 nucleotide substitution frequencies observed for Illumina and ONT (a). Nucleotide substitutions
2 detected exclusively by one technology only are indicated in green (Illumina) or blue (ONT)
3 whereas the substitutions detected by both technologies are colored red. Even though more
4 nucleotide substitutions exclusive to ONT were observed, there was a positive correlation
5 ($R^2=0.79$) between two sequencing technologies. **b)** Correlation between sub-consensus indel
6 frequencies observed for Illumina and ONT. More indels exclusive to ONT were seen with a
7 weak correlation ($R^2=0.13$) between the indel frequencies between two technologies suggesting
8 ONT tend to result in more false-positive indels.

9

10

11

12

13

14

15

16

17

18

19

20

1

2

3

4

Basecall accuracy

90%

99%

99.9%

Density

0.4

0.3

0.2

0.1

0.0

10

20

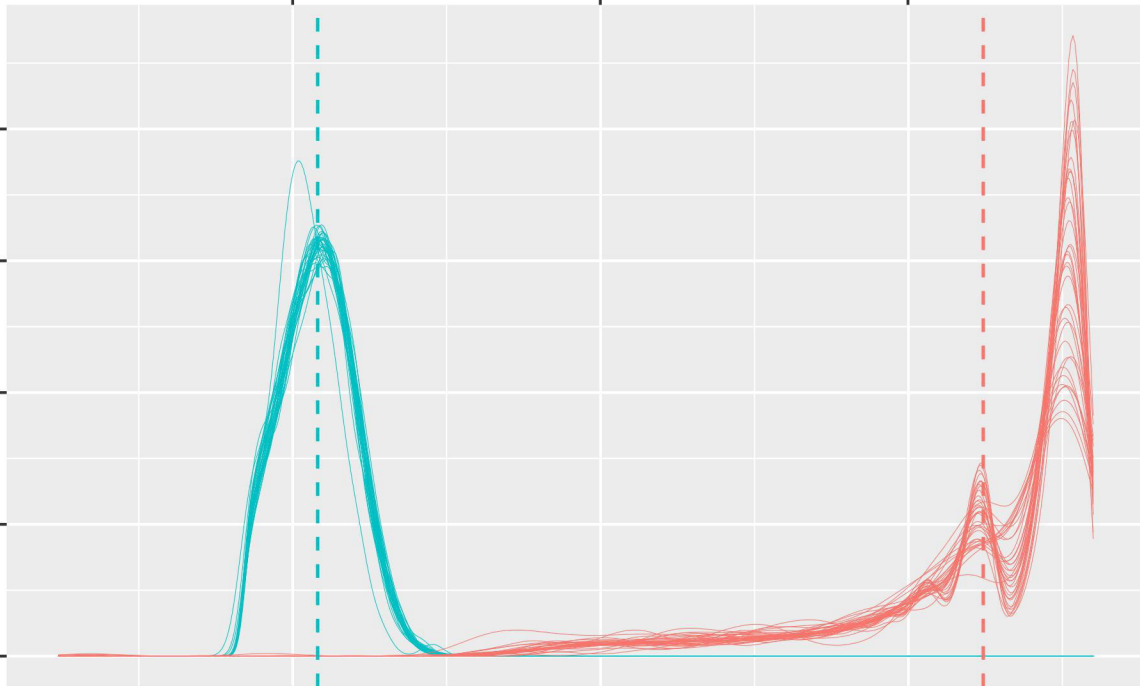
30

PHRED quality score

Technology

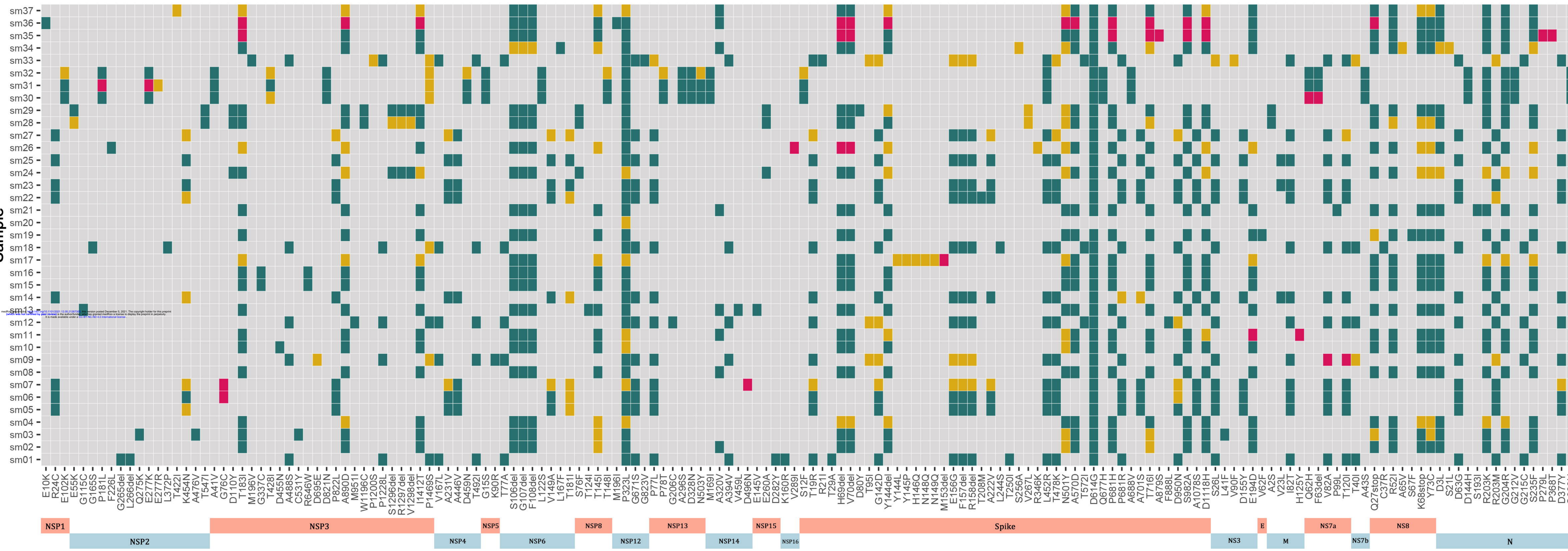
— Illumina

— ONT



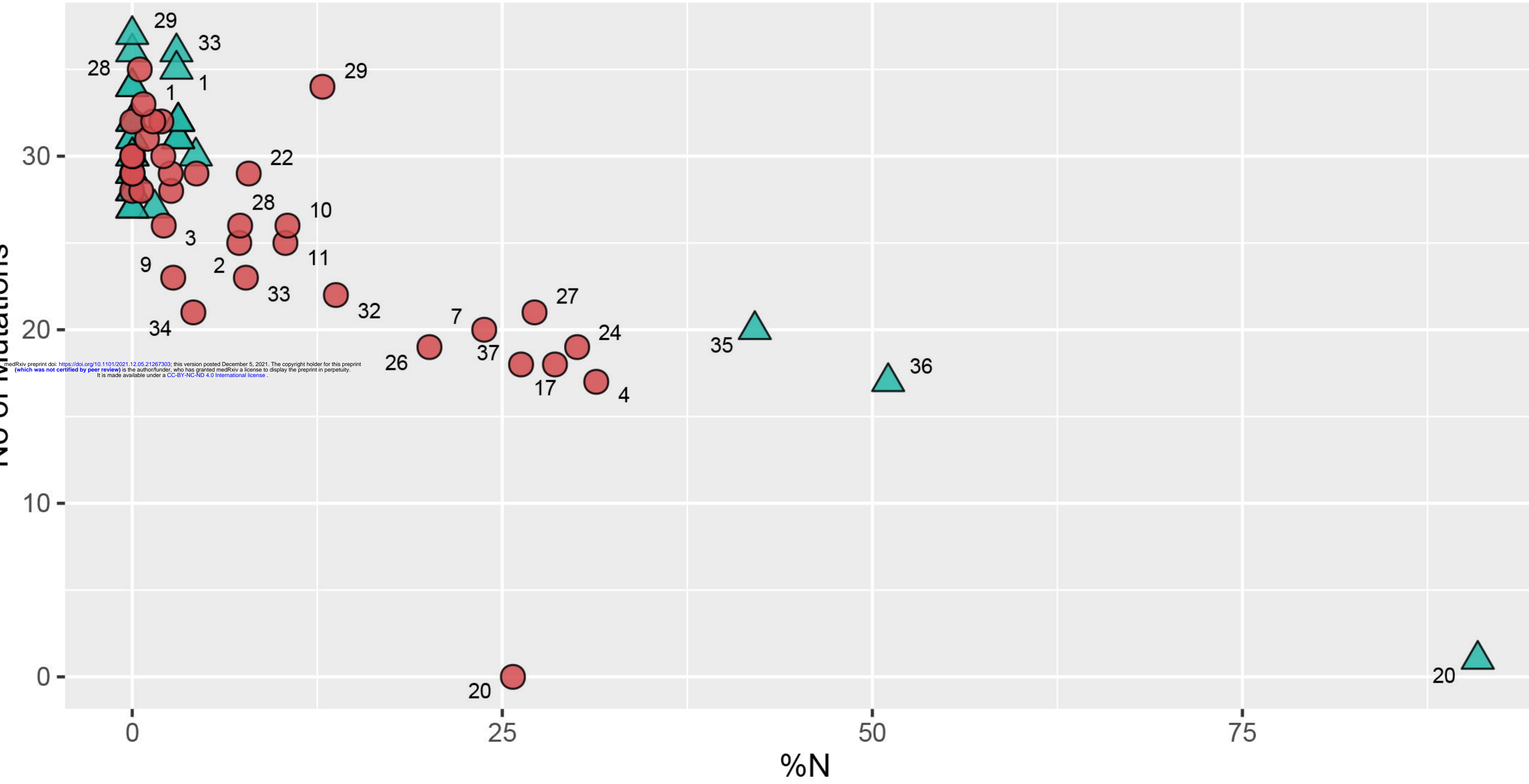
Sequencing technology

Both Illumina ONT

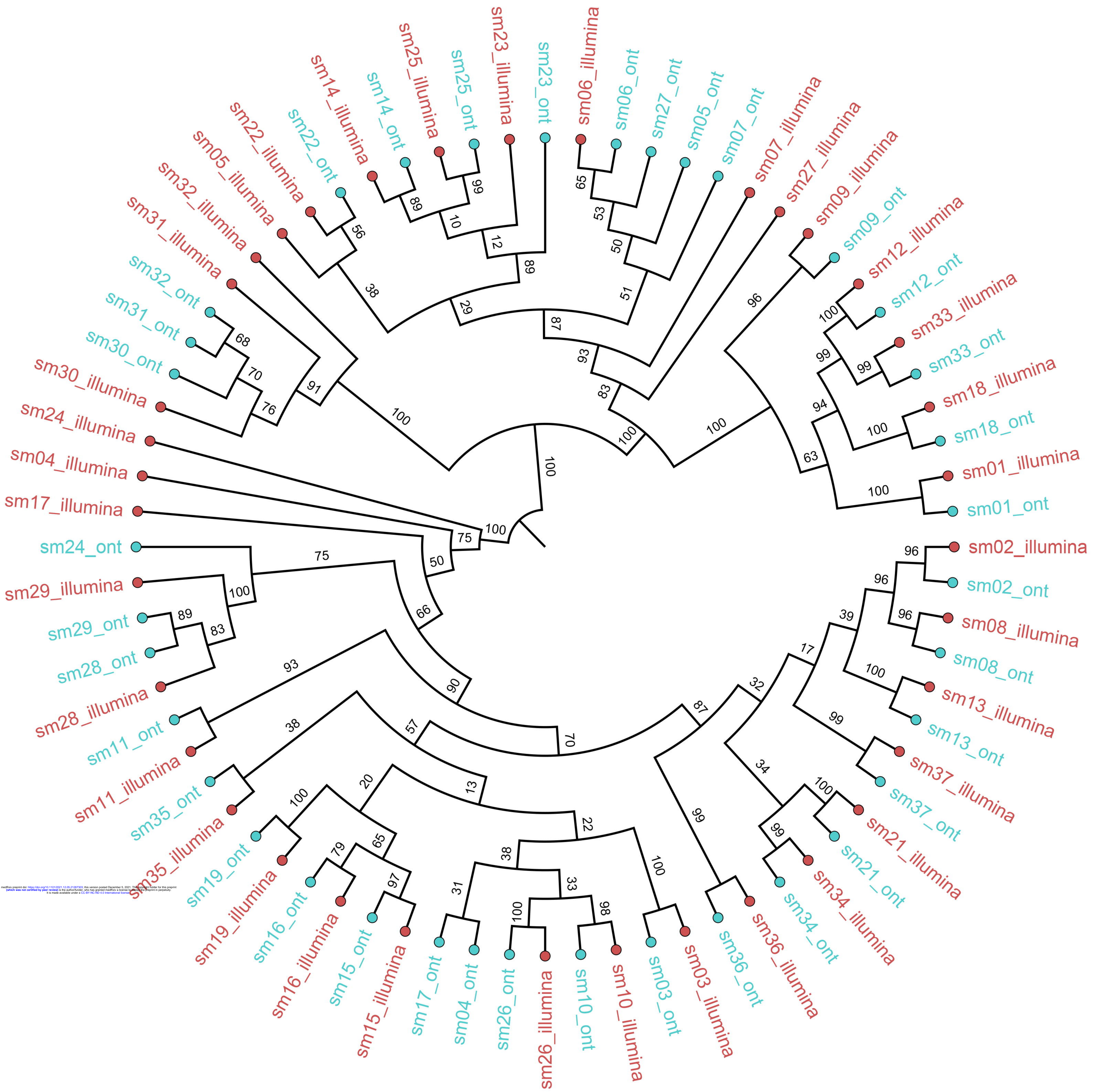


medRxiv preprint doi: <https://doi.org/10.1101/2021.12.05.21267320>; this version posted December 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

No of Mutations

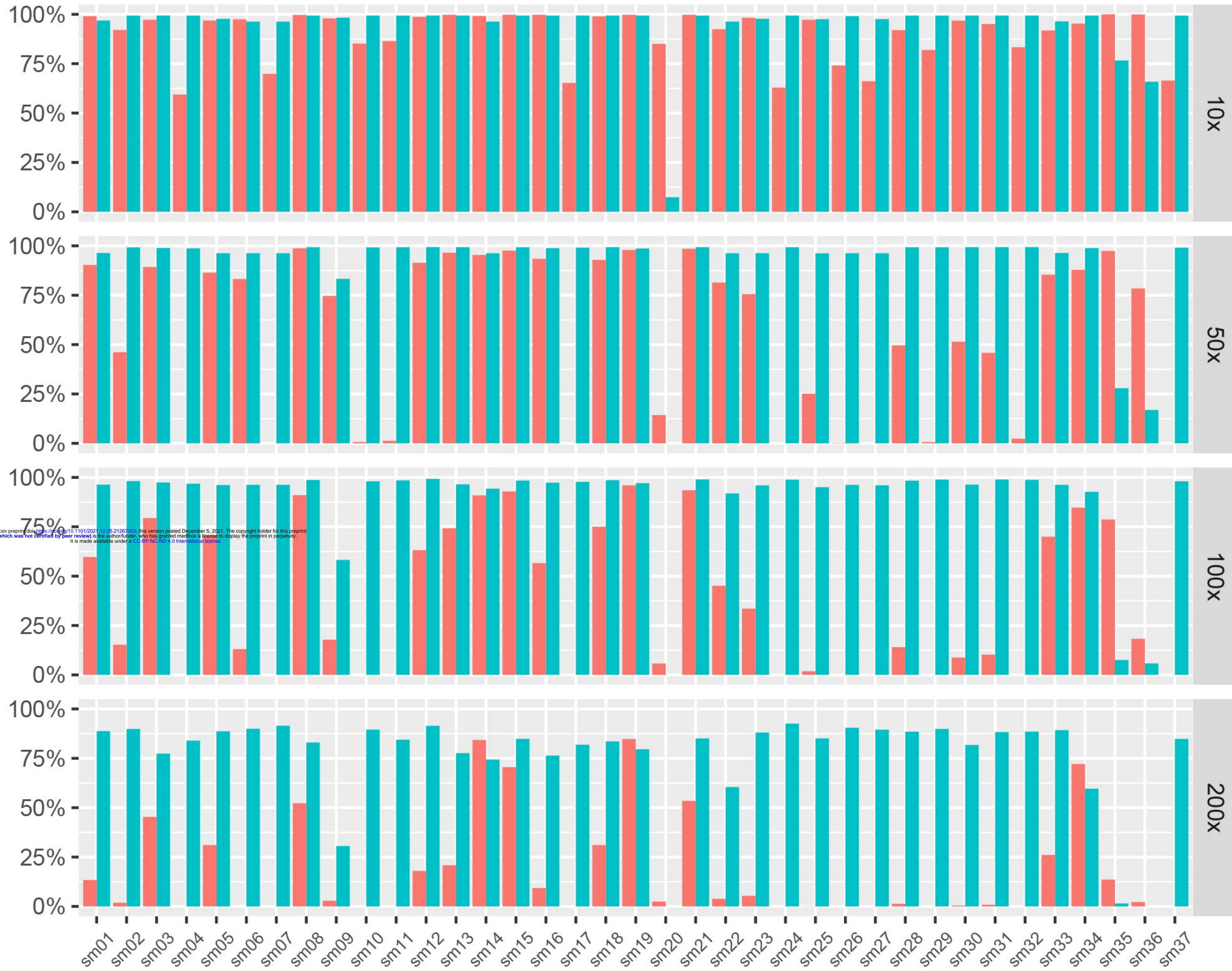


Technology  Illumina  ONT



medRxiv preprint doi: <https://doi.org/10.1101/2021.12.05.21077000>; this version posted December 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Genome coverage



Technology

- Illumina
- ONT

medRxiv preprint doi: <https://doi.org/10.1101/2021.12.15.21267303>; this version posted December 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).