SeqAfrica SARS-CoV-2 Whole Genome Analysis

Raw FastQ sequencing files obtained from Illumina NextSeq sequencing using the ARTIC V3 protocol were obtained for whole genome assembly of full-length SARS-CoV-2 genomes. Whole genome assembly was conducted for all 7 samples, by Dr Cathrine Scheepers using the Exatype SARS-CoV-2 (https://sars-cov-2.exatype.com/) and Galaxy COVID-19 project workflows (usegalaxy.eu; https://covid19.galaxyproject.org/artic/#live-resources). The use of both these platforms allows one to compare outputs and have confidence in the final result.

The majority (6/7) of the samples have >1 million reads sequenced. With one sample (N4330) having less than 1 million reads (Figure 1). For 3/7 sequence we observed a high level of sequences not being aligned to the reference genome (orange).

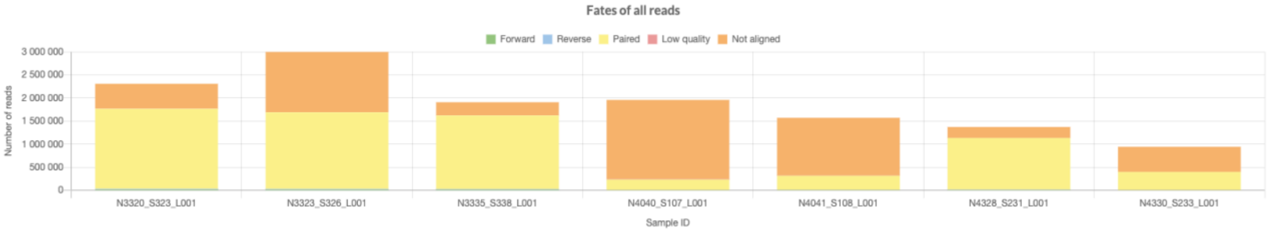

Figure 1: Overall read information per sample

Coverage across the SARS-CoV-2 genome using the Wuhan reference (Genbank: NC_045512.2) showed evidence of “drop-off” or lack of coverage over certain parts of the genome. (Figure 2).

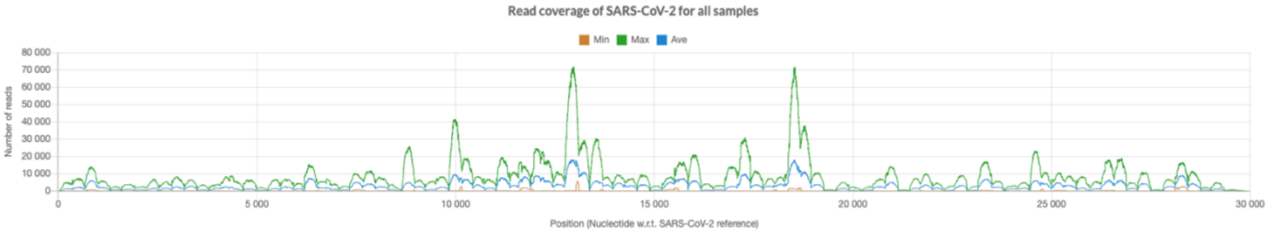

Figure 2: Coverage across the genome

Overall, for the 7 samples, 3 had good quality data, 2 with mediocre quality and 2 with poor quality data, highlighted in red in Table 1.

Table 1: Quality metrics for samples sequenced

Sample	Galaxy-Quality	Exatype-Quality	Galaxy- >30x Coverage	Exatype-100x Coverage
N3320	good	good	99.6%	93.8%
N3323	good	good	99.8%	99.5%
N3335	good	good	99.6%	93.6%
N4040	mediocre	none	80.1%	59.6%
N4041	mediocre	none	84.8%	64.2%
N4328	bad	none	47.6%	38.2%
N4330	bad	none	45.7%	39.4%

Figure 3 shows a phylogenetic tree of the 7 samples compared to global isolates with 4 clustering with 501Y.V2 (yellow) and 3 within the clade 20B.

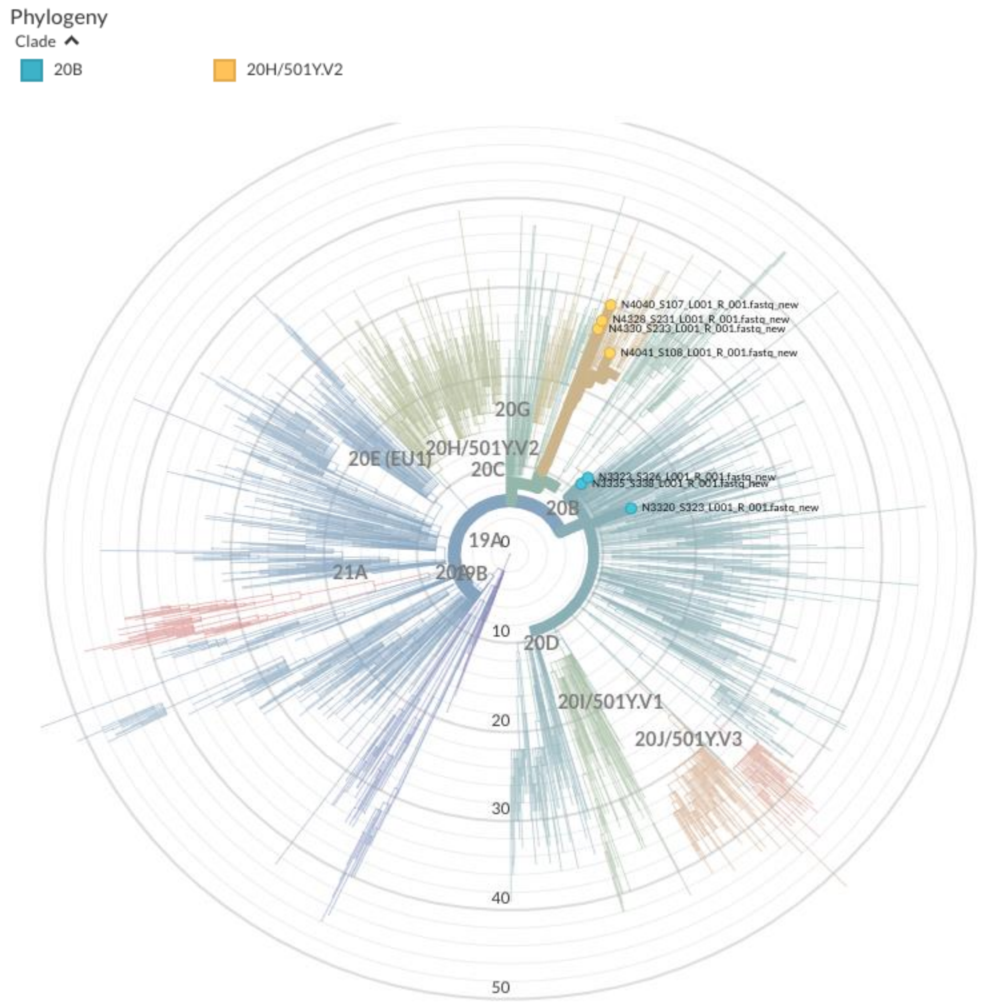
Phylogeny
Clade
20B
20H/501Y.V2
N4040_S107_L001_R_001.fastq_new
N4328_S231_L001_R_001.fastq_new
N4330_S233_L001_R_001.fastq_new
N4041_S108_L001_R_001.fastq_new
N3323_S326_L001_R_001.fastq_new
N3335_S338_L001_R_001.fastq_new
N3320_S323_L001_R_001.fastq_new
20G
20H/501Y.V2
20E (EU1)
20C
20B
19A
0
21A
20A
19B
10
20D
20
20I/501Y.V1
20J/501Y.V3
30
40
50

Figure 3: Phylogenetic tree of samples compared to global SARS-CoV-2 isolates
Circles represent sequenced samples with those in yellow clustering with 501Y.V2 and those associated with 20B in blue.

Nextstrain clade and Pangolin lineage assignments for the assembled SARS-CoV-2 sequences identified a Variant of Concern (VOC) 501Y.V2/B.1.351 assigned to four of the samples (though 2 had poor quality data and therefore the assignment should be taken with caution). For the remaining samples the clade 20B was assigned being linked to either B.1.1 (for 2 samples) and B.1.1.99) for another (Table 2). Both exatype and galaxy gave the same results.

Table 2: Clade and Lineage assignments for assembled SARS-CoV-2 sequences

Sample	Galaxy-Nextclade	Exatype-Nextclade	Galaxy-Pangolin	Exatype-Pangolin
N3320	20B	20B	B.1.1	B.1.1
N3323	20B	20B	B.1.1	B.1.1
N3335	20B	20B	B.1.1.99	B.1.1.99
N4040	20H/501Y.V2/Beta	20H/501Y.V2/Beta	B.1.351/Beta	B.1.351/Beta
N4041	20H/501Y.V2/Beta	20H/501Y.V2/Beta	B.1.351/Beta	B.1.351/Beta
N4328	20H/501Y.V2/Beta	20H/501Y.V2/Beta	none	none
N4330	20H/501Y.V2/Beta	20H/501Y.V2/Beta	none	none

Using a variant frequency plot we can see that the four samples assigned to 501Y.V2/B.1.351 cluster together and show evidence of multiple mutations within the spike region that is typical of this VOC (Figure 4).

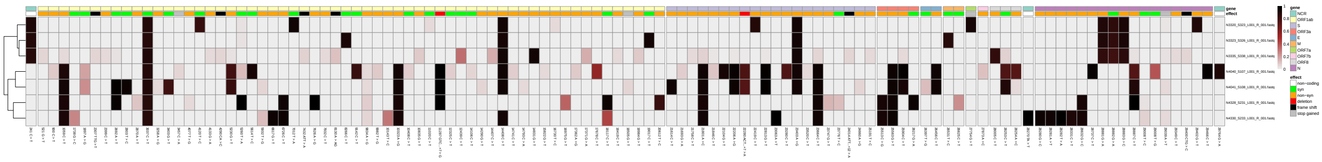

Figure 4: Variant frequency plot for all 7 samples

The two samples assigned to 20B/B.1.1 shared the mutation D614G, with N3320 having an additional T286I mutation within the NTD of the Spike protein (Figure 5).

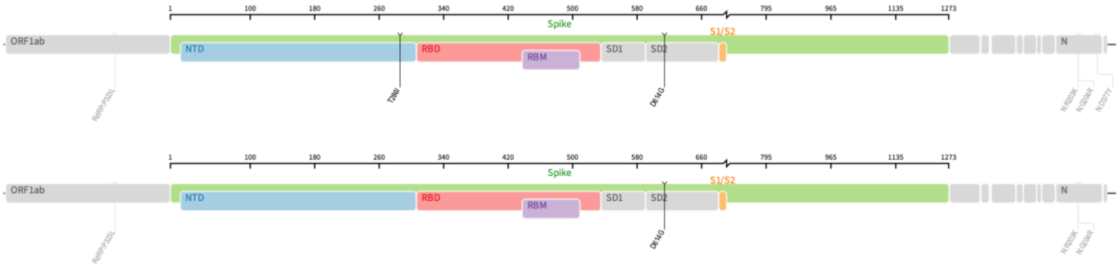

Figure 5: Genome Representation for N3320 and N3323 assigned to 20B/B.1.1

The sample N3335 assigned to 20B/B.1.1.99 also shared the D614G mutation within the spike but had different mutations outside of the spike protein which could be the reason for the different assignment (Figure 6).

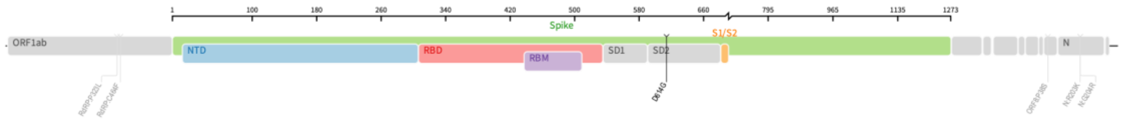

Figure 6: Genome Representation for N3335 assigned to 20B/B.1.1.99

Though the assignment of the VOC 501Y.V2/B.1.315 has been assigned to potentially 4/7 samples sequenced, all of these samples had missing data across the genome, including within the spike region (shown as red lines, Figure 7). Nevertheless many of the mutations associated with B.1.351 have been detected, particularly D80A, K417N, E484K and N501Y. These mutations have been associated with loss of neutralization and are therefore cause for concern.

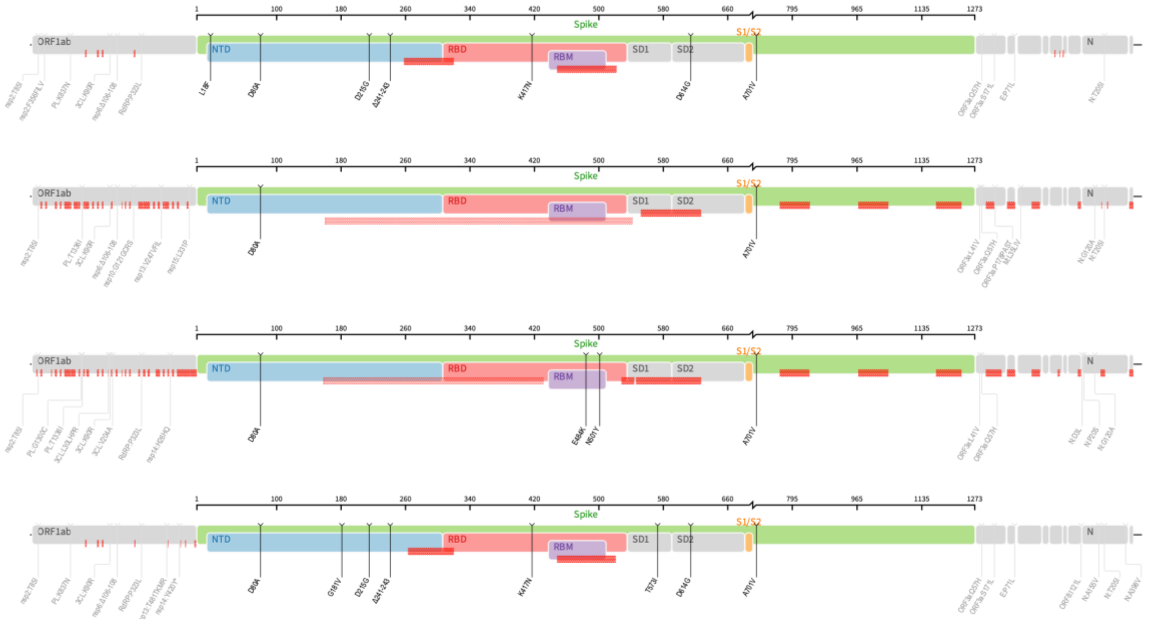

Figure 7: Genome Representation for N4040, M4041, N4328 and N4330 assigned to 20H/501Y.V2/B.1.351

It is our recommendation that samples from these donors (N4040, N4041, N4328 and N4330) be collected once again for resequencing in the hope of improving coverage and quality of the sequencing. Furthermore contact training from these donors should be carried out, contacts should be tested and if positive, sequencing analysis should be carried out. All contacts should undergo isolation for a period of 14 days to prevent further transmission of this variant.