White Paper

Robust, rapid discrimination of SARS-CoV-2 variants of concern from heterogeneous RNA samples by NGS using the QIAseq® DIRECT SARS-CoV-2 Kit

Nicolò Musso[1], Paolo Bonacci[1], Dalida Bivona[2], Carmelo Bonomo[2], Dafne Bongiorno[2] and Stefania Stefani[2]

[1] Department of Biomedical and Biotechnological Sciences (BIOMETEC), University of Catania, 95123 Catania, Italy
[2] Laboratory of Molecular and Resistant Antibiotic Medical Microbiology (MMAR), Department of Biomedical and Biotechnological Sciences (BIOMETEC), University of Catania, 95123 Catania, Italy

Abstract: The emergence of SARS-CoV-2 variants of concern (VOCs) demanded the development of robust next-generation sequencing (NGS) library preparation methodologies for identifying and discriminating SARS-CoV-2 sequence variants. The QIAseq DIRECT SARS-CoV-2 Kit was developed to address the distinct needs for SARS-CoV-2 NGS, and offers a workflow that reduces preparation time by half while maintaining high library quality. SARS-CoV-2 NGS performance data obtained using QIAseq DIRECT SARS-CoV-2 libraries demonstrate that (1) depth and breadth of sequence coverage is sufficient to accurately identify individual genomic mutations, (2) VOCs can be discriminated in heterogeneous real-world samples, (3) RNA quality is critical for library generation and (4) threshold cycle value (C_T) is not necessarily the best indicator of library sequencing success.

Introduction

Developing new kits and methods to sequence the SARS-CoV-2 viral genome is vital to keep up with the virus, or better yet, stay ahead of it during the COVID-19 pandemic. With the emergence of VOCs*, the ability to identify and discriminate among different sequence variants of SARS-CoV-2 has become paramount in the fight against the virus. NGS is the preferred approach to identify these sequence variants, especially when it is necessary to rapidly report on a multitude of samples. Early in the pandemic, QIAGEN developed the QIAseq SARS-CoV-2 Primer Panel, which uses optimized ARTIC-based primers[1] for the preparation of viral genome NGS libraries from nasopharyngeal swabs and other sample types. These libraries are compatible with Illumina® and Oxford Nanopore® NGS platforms. However, ARTIC-based protocols come with challenges. The protocol is time-consuming, with numerous steps for library preparation. In addition, a number of these steps require ethanol purification, which could alter the quality and quantity of the final product. ▷

* In September 2021, the US government reclassified several previous Variants of Concern to "Variants Being Monitored" (VBM), and further reclassifications may occur through the course of the COVID-19 pandemic. More recently (November 2021), the new Omicron VOC has emerged. For further details, see **https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html**.

To overcome these challenges and for other reasons, QIAGEN introduced a new kit – the QIAseq DIRECT SARS-CoV-2 Kit. This workflow reduces the NGS library preparation time by half, condenses some steps into one, and reduces the number of purifications required, while maintaining high library quality.

In this White Paper, we test the QIAseq DIRECT SARS-CoV-2 Kit protocol on real-world samples with heterogeneous RNA quality, and discuss the advantages and limitations of the kit for the preparation of SARS-CoV-2 libraries, with an emphasis on the ability to detect and discriminate among different VOCs circulating in the population at the time of this study.

Materials and methods

Samples

Samples that had tested positive for SARS-CoV-2 were delivered to our laboratory from public and private institutions in Eastern Sicily in mid-2021. For samples delivered as extracted and purified viral RNA, threshold cycles at which the SARS-CoV-2 viral genome was detected by quantitative PCR (qPCR) were very heterogeneous, which may also have been influenced by the use of different RNA extraction kits. Some samples were delivered as nasopharyngeal swabs; one sample was delivered as lung tissue from which the viral RNA of interest was extracted[2]. Information on prior handling (e.g., good laboratory practices, cold chain details, etc.) and other secondary data about the samples were not available to us. Therefore, these samples were considered a suitable real-world set to test the robustness of the QIAseq DIRECT SARS-CoV-2 Kit. Many of these samples were first subjected to fluorometric analysis with Qubit™ RNA HS Assay Kit (Cat. No. Q32852, Thermo Fisher, Waltham, MA, USA) to verify the presence of starting material. Given the independence of cDNA synthesis from the sample threshold cycles, we replaced 8 µl of Nuclease Free Water specified in the DIRECT Kit protocol with 8 µl of template RNA to ensure a sufficient amount of RNA in the reaction. Thermal cycling was conducted using a Mastercycler® Nexus thermal cycler (Eppendorf AG, Barkhausenweg 1, 22339 Hamburg, Germany) and all reactions were initiated at the recommended protocol temperature to avoid temperature gradients.

Nucleic acid purification

During the purification phase with QIAseq Beads, a modification to the DIRECT Kit protocol was made: samples with beads were centrifuged at 3000 rpm for one minute instead of for two. Sterile 5 ml syringes were used to remove the supernatant and the ethanol. The use of a syringe, instead of a micropipette, reduces the time in which the beads are exposed to air (a precaution also reported in the original protocol) and shortens the length of the purification step.

Normalization

Following QIAGEN's recommended protocol, a fluorometric quantification using Qubit is required. A further quantification step was added by us to improve reliability and accuracy. Thus, before normalization at 100 ng, all samples were quantified with both a Qubit and an Eppendorf BioPhotometer® D30. We noted that the readings obtained from the BioPhotometer were overestimated compared to those from the Qubit. For this reason, and by taking readings of known concentrations, we introduced a correction factor of 38.45% to address the excess in BioPhotometer readings. Thus:

Effective Concentration =
BioPhotometer Concentration x 0.81

Applying this correction factor was necessary since some samples (Figure 1) showed high Qubit quantification values but a low quality/quantity ratio from Bioanalyzer readings (Agilent, Santa Clara, CA 95051, USA). Applying the correction factor resulted in the resolution of this discrepancy and provided notable savings in terms of time and costs. Our experience suggests it could be good manufacturer practice to calculate a new correction factor for every sample set, to account for the discrepancy between fluorometric and photometric values.

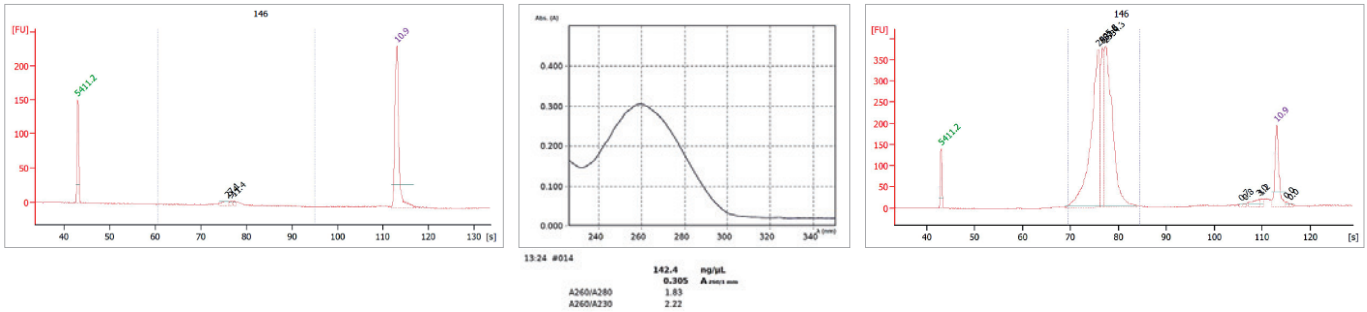

Figure 1. Example library profile from a Bioanalyzer obtained without (left panel) and with (right panel) the application of the correction factor. The original quantification of the sample is reported in the middle panel (142.4 ng/µl). The lack of a correction factor (calculated with the fluorescence value obtained from the Qubit) led to an overestimation of nucleic acid concentration, which resulted in the lack of library amplification.

Next-generation sequencing

NGS was performed according to the manufacturer's instructions on a MiSeq® platform (Illumina, San Diego, CA 92122, USA) in the Molecular Biology laboratory of the University of Catania. Libraries were quantified and their quality evaluated using both the fluorometric Qubit dsDNA HS Assay Kit (Ref. Q32851, Invitrogen, Carlsbad, CA 92008, USA) and the Agilent® High Sensitivity DNA Kit (Ref. 5067-4626). Libraries were denatured and diluted following the 'MiSeq System Denature and Dilute Libraries Guide' (Illumina, Document #15039740 v10). The libraries were multiplexed with different barcodes and pooled at 4 nM in equimolar amounts. The pooled libraries were sequenced at a final concentration of 7.5 pM using the MiSeq v2 reagent Kit (Ref. 15033624), reporting an average cluster density of about 1400K/mm^2, and an average cluster passing filter of about 90%.

Data analysis

Data were analyzed using QIAGEN CLC Genomics Workbench software and following the User Manual for software v21.0.3, released on January 25, 2021 (QIAGEN, Aarhus, 8000 Denmark). All samples were analyzed with the SARS-CoV-2 workflow, using the 'Identify QIAseq SARS-CoV-2 Low Frequency and Shared Variants (Illumina)' pipeline. The alignment and mutation detection were carried out using the complete genome sequence of SARS-CoV2 isolate Wuhan-Hu-1 (MN908947.3) as reference.

Results

Quality checks

Threshold cycle ranges for all 250 samples tested are shown in Table 1. Evaluation of library profiles obtained using the Bioanalyzer led us to conclude that the threshold cycle of the starting sample is less relevant than the quality of the initial RNA in predicting the success of generating NGS libraries that yielded good sequence information. As can be seen in Figure 2, the threshold cycles do not correlate with the position and quality of the bands in the Bioanalyzer chromatograms, and the more significant determining factor of library success is the quality of the RNA.

Table 1. Threshold cycle (C_T) values for all samples in this study

C_T Range	# Samples
14.0 – 16.9	40
17.0 – 19.9	50
20.0 – 22.9	20
23.0 – 25.9	30
26.0 – 28.9	20
29.0 – 31.9	5
32.0 – 34.9	5
>35	5
Not available	70
Total	250

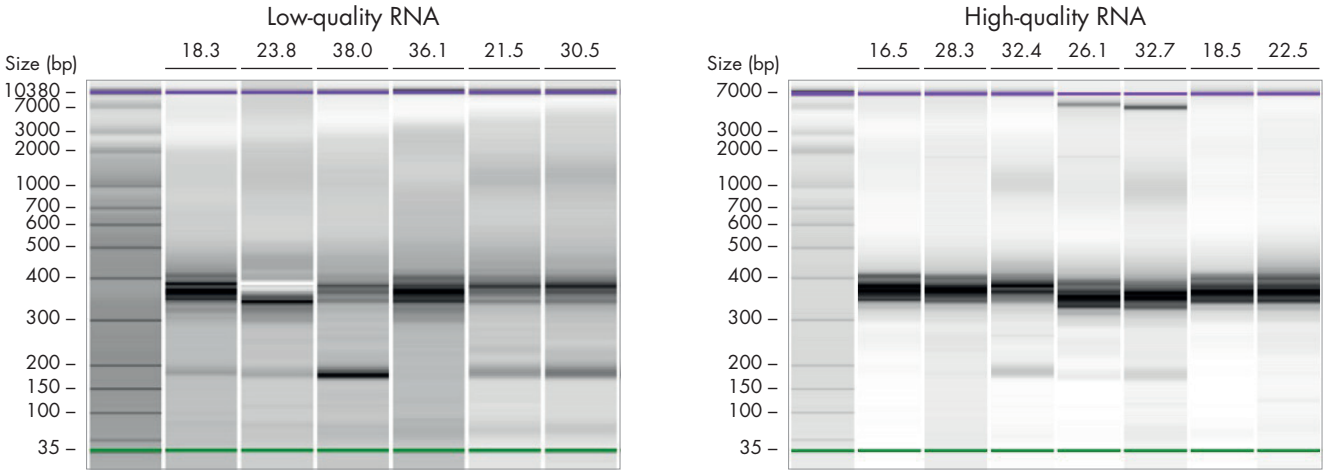

Figure 2. Bioanalyzer chromatograms of low-quality (left panel) and high-quality (right panel) RNA samples. The C_T value for each sample is shown at the top of each lane.

Sequencing results

Despite the heterogeneity of the starting samples, we obtained good read quality (i.e., >Q30) in more than 90% of the runs performed during sequencing (Figure 3). The quality parameters, however, were not enough to attest to equally satisfactory breadth and depth of sequencing coverage (Figure 4). The comparison between the sequenced viral genomes highlighted the importance of reads and coverage, since samples with a high number of reads can still contain regions with extensive gaps in sequence coverage.

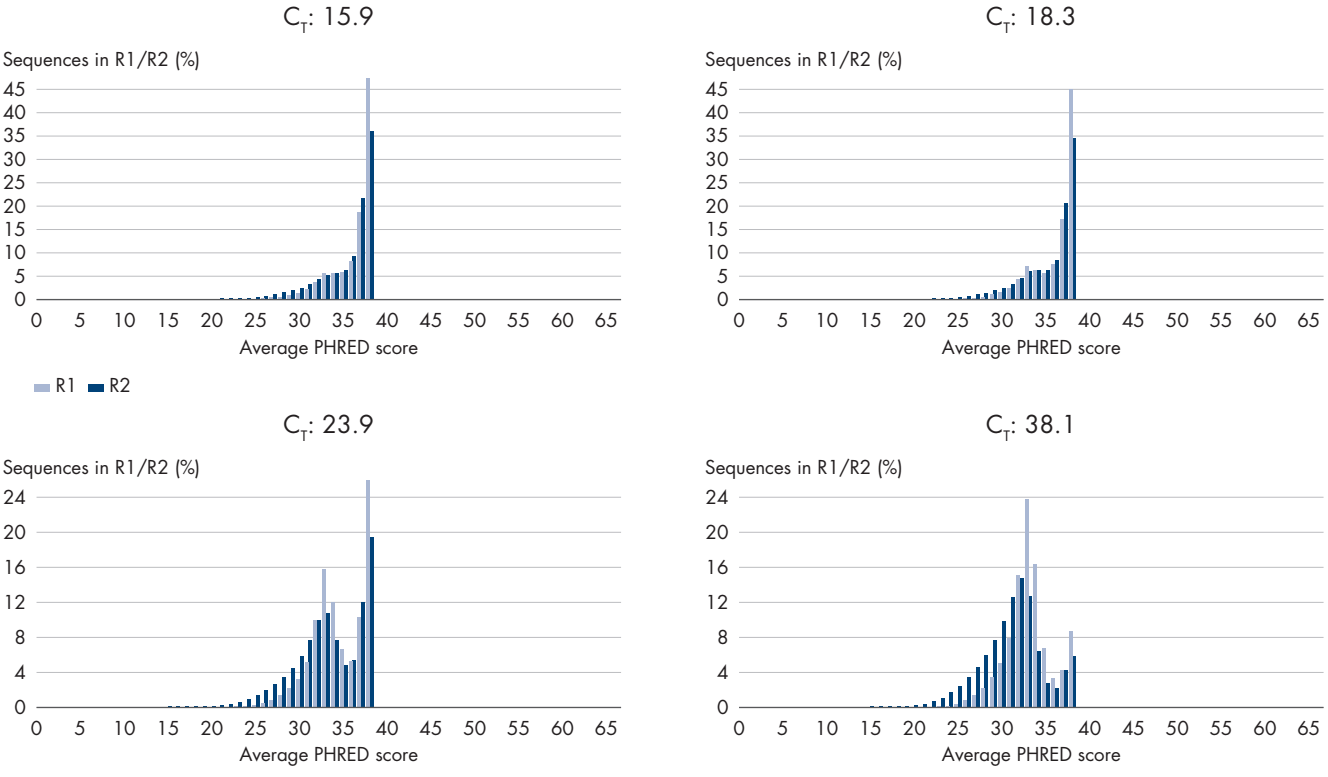

Figure 3. Examples of Q30 values reported from samples with different threshold cycles.

Figure 4 further highlights that the presence of a gap in terms of coverage is not necessarily correlated with the threshold cycle values of the analyzed samples. Non- or poorly-sequenced portions are particularly problematic in the case of viral sequencing, as the determination of variants is frequently associated with mutations that occur across multiple and distinct regions of the genome.

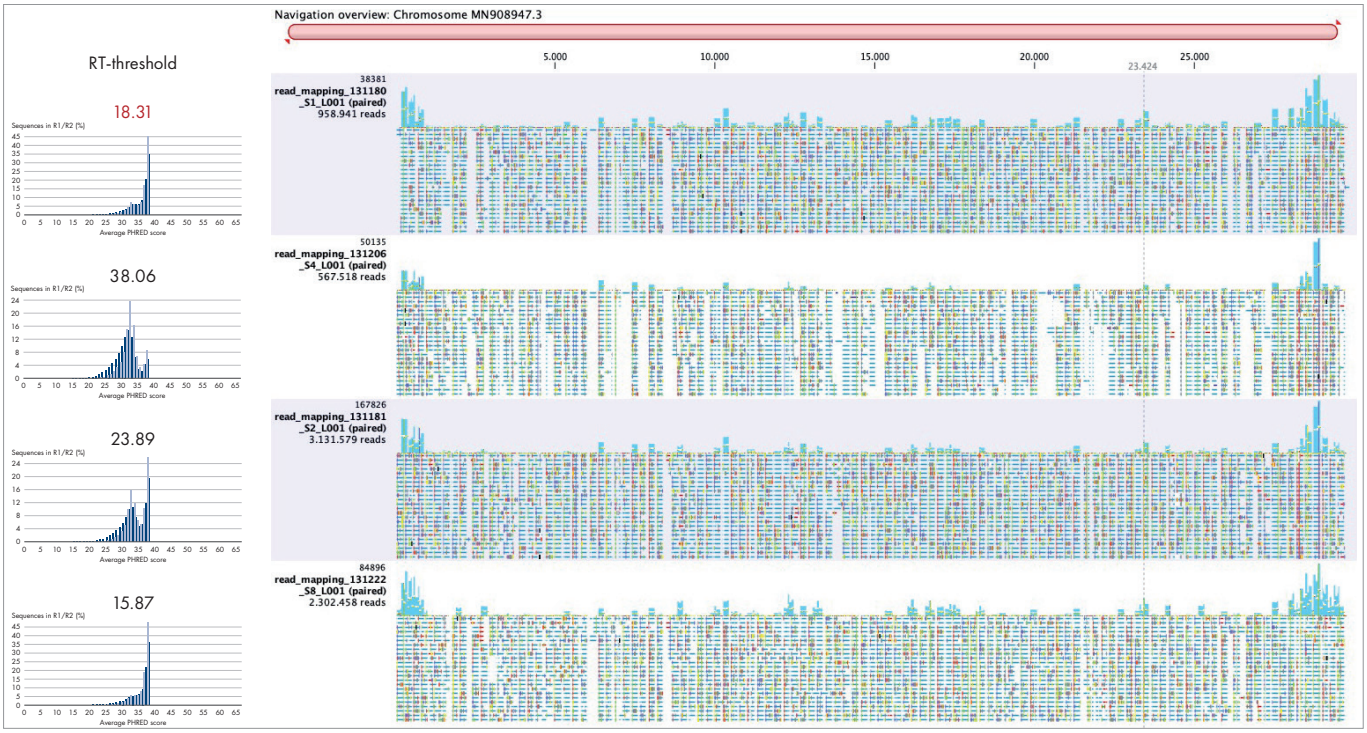

Figure 4. Reads and coverage of four samples with different C_T values.

Discrimination of SARS-CoV-2 Variants of Concern

The QIAseq DIRECT SARS-CoV-2 Kit allowed us to rapidly discriminate different variants of SARS-CoV-2 with high efficiency in terms of coverage and quality parameters. Figure 5 shows the percentage of VOCs and Variants of Interest (VOIs) that were identified across the 250 real-world samples tested in this study. Table 2 shows the specific mutations detected that define each variant, as well as the sequence read depth across each mutation for an example of a sample carrying that mutation.

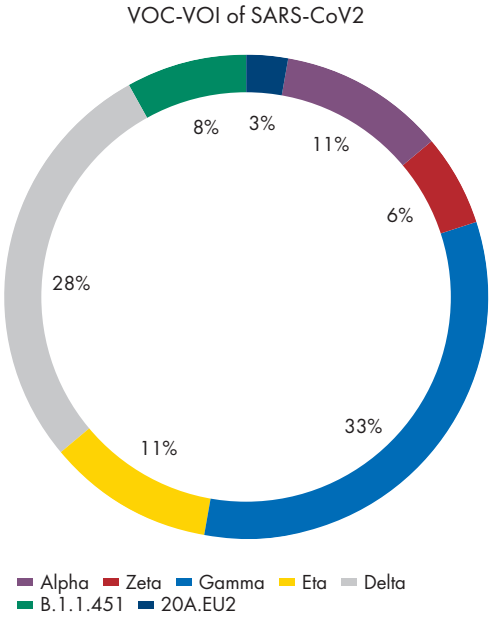

Figure 5. Percentage of VOCs and VOIs identified in our laboratory with the QIAseq DIRECT SARS-CoV-2 Kit across the 250-sample set.

Table 2. Mutations detected in this study, variant classifications, and sequence read depths by mutation for an example of a sample carrying that mutation

Alpha VOC (B.1.1.7)				Example coverage
S	H	69	–	1326
S	V	70	–	
S	Y	144	–	3707
S	N	501	Y	4212
S	A	570	D	3859
S	D	614	G	3892
S	P	681	H	1261
S	T	716	I	2106
S	S	982	A	3374
S	D	1118	H	2471
ORF1a	T	1001	I	3362
ORF1a	A	1708	D	2025
ORF1a	I	2230	T	4141
ORF1a	S	3675	–	5048
ORF1a	G	3676	–	
ORF1a	F	3677	–	
N	D	3	L	2352
N	R	203	K	6516
N	G	204	R	
N	S	235	F	10187
ORF1b	P	314	L	3282
ORF8	Q	27	–	1429
ORF8	R	52	I	2064
ORF8	Y	73	C	5508

Gamma VOC (P.1)				Example coverage
S	L	18	F	6498
S	T	20	N	2347
S	P	26	S	5655
S	D	138	Y	1248
S	R	190	S	6890
S	K	417	T	55
S	E	484	K	77
S	N	501	Y	77
S	D	614	G	10324
S	H	655	Y	9138
S	T	1027	I	116
S	V	1176	F	38
ORF3a	S	253	P	499
ORF1a	S	1188	L	658
ORF1a	K	1795	Q	195
ORF1a	S	3675	–	211
ORF1a	G	3676	–	
ORF1a	F	3677	–	
N	P	80	R	26765
N	R	203	K	64316
N	G	204	R	
ORF1b	P	341	L	234
ORF1b	E	1264	D	483
ORF8	E	92	K	181

Eta VOI (B.1.525)				Example coverage
S	Q	52	R	203
S	A	67	V	254
S	H	69	–	311
S	V	70	–	
S	Y	144	–	159
S	E	484	K	177
S	D	614	G	5605
S	Q	677	H	9454
S	F	888	L	1512
ORF1b	P	314	F	139
N	S	2	–	8479
N	D	3	Y	
N	A	12	G	21412
N	T	205	I	73404
M	I	82	T	332
ORF1a	T	2007	I	2433
ORF1a	S	3675	–	149
ORF1a	G	3676	–	
ORF1a	F	3677	–	
E	L	21	F	273
ORF6	F	2	–	65

Delta VOC (B.1.617.2)				Example coverage
S	T	19	R	115
S	E	156	–	78
S	F	157	–	
S	R	158	G	
S	L	452	R	165
S	T	478	K	123
S	D	614	G	20026
S	P	681	R	261
S	D	950	N	457
ORF1b	P	314	L	22814
ORF1b	P	1000	L	6548
M	I	82	T	193
N	D	63	G	21439
N	R	203	M	1656
N	D	377	Y	8017
ORF3a	S	26	L	3567
ORF7a	V	82	A	2416
ORF7a	T	120	I	1987

B.1.1.451 Lineage				Example coverage
S	E	156	D	648
S	D	614	G	20341
S	Q	1208	H	4879
N	R	203	K	3398
N	G	204	R	
N	T	296	I	213
N	T	362	I	5487
N	V	392	V	5978
ORF1b	P	314	L	1579

20A.EU2 VOI				Example coverage
S	S	477	N	6958
N	M	234	I	12495
N	A	376	T	5545
ORF1b	A	176	S	7591
ORF1b	V	767	L	5168
ORF1b	K	1141	R	4910
ORF1b	E	1184	D	3731

Zeta VOI (P.2)				Example coverage
S	E	484	K	31
S	D	614	G	8192
S	V	1176	F	1547
N	R	203	K	74001
N	G	204	R	

Sample sequence quality

Sequence quality parameters of two samples with varied sequence quality are shown in Table 3. Both samples were sequenced using the QIAseq DIRECT SARS-CoV-2 Kit on the MiSeq platform and evaluated using the QIAGEN CLC Genomics Workbench software. The quality parameters of sample 1 were very satisfactory while that of sample 2 were less desirable, although a significant portion of the whole genome was sequenced correctly from both samples (Table 3).

- 'Number target regions' and 'total length of targeted regions' are two numerically identical parameters that refer to the effective length of the SARS-CoV-2 genome and not to the quality of the sequencing.
- 'Average coverage' is a parameter that indicates the average coverage of each base of the genome but does not take into account sequence gaps and regions that are more deeply covered than others. However, it can give a rough estimate of the coverage obtained in the run; sample 1 has a much higher average coverage than sample 2 (5671 vs 611).
- 'Median coverage' function similar to 'average coverage'; sample 1 has a much higher median coverage than sample 2 (2410 vs 246).
- The parameters 'Total length of target regions containing positions with coverage </ ≥30' are meaningful quality parameters that refer to the single base sequence quality along the genome. The fewer the target regions with coverage <30, the higher the percentage of positions with coverage ≥30, and therefore, the higher the quality of the run. Sample 2 has 1/6 of the genome with a sequencing quality of less than 30.
- The 'Percentage of target region positions with coverage ≥ 30' expresses the ratio between the number of base pairs sequenced with quality greater than 30 and the length of the total genome. It is a more intuitive parameter than the previous two, which allows us to immediately estimate the quality of the sequencing. Indeed, sample 1 shows a higher value than sample 2 (98.3% vs 83.5%).

Table 3. Comparison of parameters (from QIAGEN CLC Genomics Workbench software) of two samples with higher (sample 1) and lower (sample 2) sequence quality

Parameter	Sample 1	Sample 2
Number target regions	1	1
Total length of target regions	29,837	29,837
Average coverage	5671.8	611.8
Median coverage	2410.0	246.0
Number of target regions with coverage < 30	1	1
Total length of target regions containing positions with coverage < 30	29,837	29,837
Total length of target region positions with coverage < 30	509	4910
Total length of target region positions with coverage ≥ 30	29,328	29,927
Percentage of target region positions with coverage ≥ 30	**98.3**	**83.5**

Conclusion

The QIAseq DIRECT SARS-CoV-2 Kit was effective in generating viral genome NGS libraries from 250 real-world samples of varying starting RNA quality. These libraries yielded sufficient high-quality sequence information to discriminate among several known SARS-CoV-2 VOCs and VOIs. The quality and integrity of the sample RNA are critical for obtaining optimal NGS libraries. RT-qPCR threshold cycle values are insufficient and sometimes misleading predictors of sequencing success, as the amplification in RT-qPCR involves only small and unrepresentative fragments of the genome. This issue can be successfully tackled using the highly robust QIAseq DIRECT SARS-CoV-2 Kit for NGS library preparation followed by a careful and thorough normalization of RNA quantities before sequencing. Using the kit, libraries can be produced from fragmented or poorly represented RNA. Optimal sequencing inputs for each sample can be determined through normalization, if necessary using both fluorometric and photometric quantification methods, and through the application of an appropriate correction factor. A careful evaluation of the quality parameters after sequencing is also important. In ideal conditions, a correct experiment would yield high coverage spread evenly along the entire genome, without any unsequenced gaps. This ensures the coverage necessary for the assignment of new variants, often characterized by the presence of different mutations distributed along the entire genome. Sequencing on NGS platforms, even using non-optimal samples, is fundamental for the identification of new variants, especially when these 'defining' mutations occur in the context of other background mutations in the viral genome.

Acknowledgments

We would like to thank the C.S. BRIT for the use of the advanced genomics platform.

Notes

References

1. ARTIC network; **https://artic.network/**
2. Musso, N. et al. (2021) Post-mortem detection of SARS-CoV-2 RNA in long-buried lung samples. Diagnostics, 11, 1158.

Ordering Information

Product	Contents	Cat. no.
QIAseq DIRECT SARS-CoV-2 Kit A	Single-box solution containing all materials for reverse transcription and library prep with 96 Unique Dual Indices (Set A)	333891
QIAseq DIRECT SARS-CoV-2 Kit B	Single-box solution containing all materials for reverse transcription and library prep with 96 Unique Dual Indices (Set B)	333892
QIAseq DIRECT SARS-CoV-2 Kit C	Single-box solution containing all materials for reverse transcription and library prep with 96 Unique Dual Indices (Set C)	333893
QIAseq DIRECT SARS-CoV-2 Kit D	Single-box solution containing all materials for reverse transcription and library prep with 96 Unique Dual Indices (Set D)	333894
QIAseq DIRECT SARS-CoV-2 HT (A–D)	Single-box solution containing all materials for reverse transcription and library prep with 384 Unique Dual Indices (Sets A, B, C, D)	333898
QIAseq DIRECT SARS-CoV-2 HT (E–H)	Single-box solution containing all materials for reverse transcription and library prep with 768 Unique Dual Indices (Sets E, F, G, H)	333899
QIAGEN CLC Genomics Workbench	A comprehensive analysis package for the analysis and visualization of data supporting all typical NGS workflows	832021

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN kit handbook or user manual. QIAGEN kit handbooks and user manuals are available at **www.qiagen.com** or can be requested from QIAGEN Technical Services or your local distributor.