COVseq is a cost-effective workflow for mass-scale SARS-CoV-2 genomic surveillance

Michele Simonetti
Karolinska Institute
Ning Zhang
Karolinska Institute https://orcid.org/0000-0002-6430-4236
Luuk Harbers
Karolinska Institute https://orcid.org/0000-0003-3910-6497
Maria Grazia Milia
Ospedale Amedeo di Savoia
Thi Nguyen
Karolinska Institute
Silvia Brossa
Istituto di Candiolo FPO - IRCCS
Magda Bienko
Karolinska Institutet, SciLifeLab https://orcid.org/0000-0002-6499-9082
Anna Sapino
Istituto di Candiolo FPO - IRCCS
Antonino Sottile
Istituto di Candiolo FPO - IRCCS
Valeria Ghisetti
Microbiology and Molecular Biology Laboratory, Amedeo di Savoia Hospital, ASL Città di Torino
Nicola Crosetto (✉ nicola.crosetto@scilifelab.se)
Karolinska Institute https://orcid.org/0000-0002-3019-6978

Article

Keywords: genomic surveillance, SARS-CoV-2, virology

Posted Date: January 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-147464/v1

Version of Record: A version of this preprint was published at Nature Communications on June 23rd, 2021. See the published version at https://doi.org/10.1038/s41467-021-24078-9.

Abstract

Genomic surveillance of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is critical to monitor the spread and evolution of the virus across different populations, geographical regions and species. Here, we present a streamlined workflow—COVseq—based on the CUTseq method that we previously described, which can be used to generate highly multiplexed sequencing libraries compatible with Illumina platforms, from hundreds of SARS-CoV-2 samples in parallel, in a rapid and cost-effective manner. We validated COVseq on RNA extracted from the supernatant of a SARS-CoV-2 culture as well as from 85 RNA samples from nasopharyngeal swabs, demonstrating the ability of COVseq to achieve near complete genome coverage, including the S region encoding the spike protein. A cost analysis showed that COVseq could be used to sequence thousands of samples per week at less than 10 USD per sample, including library preparation and sequencing costs. COVseq is a versatile and scalable method that can be readily applied for genomic surveillance of the ongoing pandemic and easily adapted to other pathogens such as influenza viruses.

Introduction

Since the identification of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as the causative agent of coronavirus disease 2019 (COVID-19)[1], thousands of SARS-CoV-2 genomes have been sequenced worldwide and the sequences have been made publicly available (https://www.gisaid.org/). This has enabled a phylogenetic reconstruction of the viral spread and evolution across different countries and continents at an unprecedented scale[2], allowing the rapid identification of genomic variants of potential epidemiological concern, such as the A23403G mutation in the S region causing the D614G amino acid substitution in the spike protein S, which enhances infectivity[3]. SARS-CoV-2 whole genome sequencing (WGS) is being increasingly applied in epidemiological surveillance to track infections in hospitals and community settings, thus informing public health decisions[4,5]. SARS-CoV-2 WGS was also recently deployed to monitor outbreaks of the virus in mink farms[6], which represent a potential source of viral variants with increased pathogenicity. In genomic surveillance, the availability of rapid and cost-effective methods for sequencing hundreds or even thousands of samples per week would be greatly beneficial.

Various approaches have been adopted for SARS-CoV-2 WGS on different sequencing platforms. These include standard RNA sequencing[7–9], amplicon-based approaches[8–15], oligonucleotide capture-based methods[8,10–12], and direct RNA sequencing[13]. In addition, direct tagmentation of retro-transcribed RNA extracted from patient samples was shown to enable high SARS-CoV-2 genome coverage and simultaneous metagenomic analysis of other viruses, bacteria and yeast present in the same sample[14]. Recently, RNA-mediated oligonucleotide annealing selection and ligation coupled with next-generation sequencing (RASL-seq) was applied to detect SARS-CoV-2 in clinical samples, without the need for nucleic acid extraction and reverse transcription[15]. However, this method can only cover a small fraction of the SARS-CoV-2 genome. More generally, an important limitation of existing SARS-CoV-2 WGS

methods is that they cannot be scaled up in a cost-effective manner. This is due to the fact that, typically, one sequencing library must be prepared for each individual sample and multiple indexed libraries must be carefully quantified and balanced before pooling them together prior to sequencing. In the case of transposase-based approaches, such as Illumina's Nextera, the preparation of indexed libraries from multiple samples can be performed rapidly, but the cost per sample is very high. Here, we describe a versatile workflow—COVseq—based on the CUTseq method that we previously described[16], which allows preparing multiplexed sequencing libraries from a large number of SARS-CoV-2 samples in parallel, in a streamlined and cost-effective manner. We technically validate COVseq on RNA extracted from a SARS-CoV-2 viral culture and on 85 SARS-COV-2 positive left-over RNA samples collected in two different phases of the 2020 pandemic at two hospitals in Italy. Lastly, we perform a cost analysis to demonstrate that COVseq is a highly cost-effective method that can be adopted for mass-scale genome surveillance of the ongoing pandemic.

Results

The CUTseq method, which we previously described[16], enables a cost-effective preparation of highly multiplexed DNA sequencing libraries, by using restriction enzymes to barcode multiple samples before pooling them together into a single library. When the SARS-CoV-2 pandemic started, we wondered whether we could adapt CUTseq to sequence a large number of SARS-CoV-2 genomes in parallel, at an affordable cost. We first determined the SARS-CoV-2 genome breadth of coverage that could be achieved theoretically using restriction enzymes compatible with CUTseq[16]. *In silico* double digestion with two 4-base cutters, MseI and NlaIII, predicted that 74.1%, 94.4% and 98.8% of the SARS-CoV-2 genome would be covered at least once using 75, 150 and 300 nucleotides (nt) single-end sequencing (SE75/150/300), respectively (Fig. 1a, **Supplementary Fig. 1a**, Table 1 **and Methods**). Furthermore, *in silico* analysis showed that 18 of the 19 most frequent SARS-CoV-2 single-nucleotide variants (SNVs) reported worldwide to date[1], including the A23403G variant, as well as 12 of the 13 SNVs characterizing the new SARS-CoV-2 lineage recently emerged in the UK[2], could be theoretically covered using SE150, while 31 out of 32 of them could be covered by SE300 (**Supplementary Fig. 1b**).

Table 1
Theoretical fraction (%) of the SARS-CoV-2 genome covered by CUTseq using one or two restriction enzymes (MseI and NlaIII) and different sequencing read lengths (nt).

Read length	Region	NlaIII	MseI	NlaIII + MseI
75	Whole genome	34.9%	61.8%	74.1%
150	Whole genome	61.2%	87%	94.4%
300	Whole genome	84.4%	96.5%	98.8%
75	S	22.4%	63.7%	75.1%
150	S	40.6%	92.1%	98.5%
300	S	67.5%	100%	100%

In order to achieve high SARS-CoV-2 genome coverage even in samples containing a small fraction of viral RNA, we implemented a workflow that begins with a multiplexed PCR step in order to selectively amplify the whole SARS-CoV-2 genome (Fig. 1a, b). To this end, we adopted an assay developed by the U.S. Centers for Disease Control and Prevention (CDC) based on a previously published protocol for WGS of the Zika virus[3] (Fig. 1a, **Supplementary Table 1 and Methods**). We first tested the performance of the multiplexed PCR assay, by preparing a sequencing library from one sample consisting of RNA extracted from the supernatant of a cell culture inoculated with SARS-CoV-2, using a standard library preparation kit (NEBNext) (**Methods**). Sequencing of this library on Illumina's NextSeq 500 resulted in complete genome coverage (**Supplementary Fig. 2a, b and Supplementary Table 2**), although at non-uniform depth, most likely due to the uneven distribution of the amplicons along the SARS-CoV-2 genome (Fig. 1a). We then prepared a COVseq library from the same RNA sample, using MseI and NlaIII either alone or in combination to digest the pre-amplified SARS-CoV-2 genome (**Supplementary Table 3 and Methods**). In line with our theoretical expectations (Table 1), single enzyme digestion and SE150 resulted only in partial SARS-CoV-2 genome coverage, including the S region (**Supplementary Table 2**). However, double digestion resulted in almost complete genome coverage, with ~ 99% of the whole genome and S region covered at least once and ~ 95% covered at least ten times (Fig. 1c, d **and Supplementary Table 2**). Based on our calculations (Table 1), we anticipate that COVseq followed by SE300 would provide full coverage of the SARS-CoV-2 genome and S region.

We then sought to validate COVseq while implementing a low-input version of the method in order to reduce the cost per sample (see COVseq workflow #1 in **Methods**). To this end, we leveraged on the I-DOT nanodispensing device that we previously utilized for high-throughput CUTseq[16] and used it to prepare a single multiplexed library using as little as 50 ng of each of 30 SARS-CoV-2 positive RNA samples (samples 1–30 in **Supplementary Table 4 and Methods**). As a comparison, we prepared NEBNext libraries from a recommended higher amount (250 ng) of each of the same 30 samples individually and sequenced all the libraries on NextSeq 500 (**Supplementary Table 2**). The number of reads per sample

inversely scaled with the corresponding Ct value, both in the case of COVseq and NEBNext, and it was well correlated (Pearson's correlation coefficient: 0.78) between the two methods (Fig. 1e, f **and Supplementary Fig. 2c**), suggesting that both methods quantitatively capture inter-sample differences in similar ways. Moreover, the proportions of reads aligned to the SARS-CoV-2 genome *vs.* other genomes and the fraction of unmapped reads were very similar between matched COVseq and NEBNext samples (Fig. 1g). In samples with high Ct (> 35)[17], a sizable fraction of the reads were aligned to the human genome, independently of the method used to prepare the libraries (Fig. 1g). This reflects the fact that these samples were not treated with DNase during the RNA extraction procedure (**Methods**), thereby allowing for human genomic DNA present in the sample to become incorporated into the library when the viral load is low. To further validate COVseq, we assessed how many SNVs identified by NEBNext in the above samples were also detected by COVseq (**Methods**). In 21 samples with low Ct (≤ 35) and consequently a high percentage of sequencing reads aligned to the SARS-CoV-2 genome (Fig. 1g), the number of SNVs per sample was highly correlated between COVseq and NEBNext (Pearson's correlation coefficient: 0.93). In total, 159 out of 228 (70%) SNVs identified by NEBNext in these samples were also detected by COVseq (Fig. 1h, i). Importantly, the five most frequent SNVs in these samples—including 4 out of the 19 most frequent SNVs reported worldwide[2]—were detected by both COVseq and NEBNext (**Supplementary Fig. 3a**). Only three frequent variants affecting three consecutive bases (G28881A, G28882A and G28883C) inside the N gene were not detected by COVseq, most likely due to the fact that they are located inside one of two 'dark' genomic regions more than 300 nt away from the closest NlaIII or MseI recognition site (**Supplementary Fig. 3b**). However, using one additional restriction enzyme (BfaI) that can cut within these regions is expected to result in complete coverage of these variants (**Supplementary Fig. 3b**).

Having demonstrated the feasibility and analytical validity of COVseq, we then sought to assess its scalability and reproducibility. To this end, we omitted the time-consuming step of purification of the multiplexed PCR amplicons (see COVseq workflow #2 in **Methods**) and prepared three replicate (Rep) COVseq libraries, each derived from 55 additional SARS-CoV-2 positive RNA samples (samples 31–85 in **Supplementary Table 4 and Methods**). This faster workflow allowed us to prepare COVseq libraries from a total of 165 samples in only two days with less than two hours hands-on time. To confirm the validity of this approach, we generated a reference (Ref) library from the same 55 samples, using COVseq workflow #1 (**Methods**). SE150 sequencing of the Rep and Ref libraries on NextSeq 500 confirmed the ability of COVseq to quantitatively capture inter-sample differences in the number of reads as well as percentage of reads aligned to the SARS-CoV-2 genome, and to do so in a reproducible manner (**Supplementary Fig. 4a-d, Supplementary Fig. 5a-e, and Supplementary Table 2**). Comparison of each Rep library with the Ref library showed a high degree of overlap (mean: 89.4%; range: 88.6– 90.0%) between the number of SNVs detected in samples with low Ct (≤ 30) (Fig. 1j **and Supplementary Fig. 6a, c, e**), confirming the validity of COVseq workflow #2. Pair-wise comparisons of the three Rep libraries showed that the breadth of genome coverage was highly correlated (Pearson's correlation coefficient > 0.88) between replicates (**Supplementary Fig. 6a-c**), highlighting the reproducibility of COVsEq. This was further confirmed by the observation that 81.5% of all the SNVs identified in samples with low Ct (≤ 30)

were detected in all three replicates, and 86.2% were detected in at least four of them (Fig. 1j **and Supplementary Fig. 6d-j**).

We then assessed whether COVseq could be potentially utilized for genomic surveillance of the ongoing pandemic. To this end, we examined the depth of coverage at 32 genomic sites encompassing the locations of the 19 most prevalent SNVs detected thus far worldwide[2], and of the 13 SNVs in the recently emerged UK lineage[18], in the 74 low Ct (≤ 35) samples sequenced by COVseq (**Methods**). On average, these sites were covered at a relative depth of 2,930 reads per million reads per sample (2,930 ± 7,291, mean ± s.d.) (Fig. 2a), suggesting that COVseq would be able to detect SNVs at these locations. One site inside the S region (C23709T) was covered at low depth in all the 74 samples (2 ± 3 reads per million reads per sample, mean ± s.d.) (Fig. 2a). The low coverage at this location can be explained by its distance from the nearest MseI and NlaIII site, which is larger than the sequencing read length (130 nt). However, sequencing with longer reads or using an additional restriction enzyme (BfaI) is expected to result in a higher coverage at this genomic position.

Next, we explored whether COVseq data are compatible with Nextstrain[19], a popular tool for phylogenetic analysis of viral genomes. We applied Nextstrain to the same 74 low Ct samples sequenced by COVseq together with 991 samples randomly selected among 285,390 samples available for download from GISAID (https://www.gisaid.org/) as of 24 December 2020, including 4 sequences from the original Wuhan lineage and 292 from Italy (**Methods**). COVseq samples formed two distinct clusters within clades 20A and 20B, which also contained most of the GISAID sequences from Italy (Fig. 2b, c). Each cluster corresponds to the different phase of the 2020 pandemic when the samples were collected at two close-by hospitals in Italy (Mar-Apr and Oct-Nov, respectively, see **Supplementary Table 4**), and therefore possibly reflects the evolution of the virus in Italy during that period. Notably, none of the two clusters included the variants characterizing the recently emerged UK lineage[18].

Lastly, we assessed the cost effectiveness of applying COVseq for genomic surveillance. Assuming to use COVseq workflow #1 (see **Methods**) and pool 96 samples into the same library, the average library preparation cost per sample would reach $21.48 and $20.28 for 10,000 and 100,000 samples processed, respectively (Fig. 2d **and Supplementary Notes**). Pooling 384 samples per library would result in a modest decrease in the cost per sample, with a further decrease when COVseq workflow #2 is used (Fig. 2d **and Supplementary Notes**), similar to what we did to prepare the replicate libraries described above. According to our analysis, the most cost-effective approach would be performing all reactions in nanoliter volumes, starting from the RT-PCR step until pooling all the samples for *in vitro* transcription (IVT) (COVseq workflow #3 in **Methods**). Indeed, a proof-of-principle experiment using two different amounts of synthetic SARS-CoV-2 RNA (5,000 and 10,000 copies) as input and SE150 sequencing, showed that near complete SARS-CoV-2 genome coverage can be achieved also with this workflow (**Supplementary Fig. 7a, b, Supplementary Table 2**). We then compared these costs to those that would have to be faced if sequencing libraries were prepared from individual samples using commercially available kits compatible with Illumina platforms (**Supplementary Notes**). Using the NEBNext library preparation kit or the transposase based Nextera kit would result, respectively, in one and two orders of magnitude higher cost

per sample compared to any of the three COVseq workflows (Fig. 2e **and Supplementary Notes**). The cost would be one extra order of magnitude higher using the TruSeq kit from Illumina, even though this does not require a pre-amplification step by multiplexed PCR (Fig. 2e **and Supplementary Notes**). Collectively, our results indicate that COVseq is a sensitive, reproducible, scalable and highly cost-effective method that is suitable for mass-scale SARS-CoV-2 genomic surveillance.

Discussion

As the COVID-19 pandemic continues to rage, the use of genomic surveillance to monitor SARS-CoV-2 outbreaks in healthcare settings as well as in farms where thousands of potentially susceptible animals live in close proximity has become increasingly important[4–6]. Moreover, with the appearance of new lineages with potentially higher infectivity and/or pathogenicity, such as the recently emerged UK lineage[18], there is an urgent need for streamlined and cost-effective approaches that could be deployed for sequencing thousands of viral samples per week. This is particularly relevant for monitoring and quickly responding to the possible emergence of variants conferring resistance to the SARS-CoV-2 vaccines that have been developed and are being administered worldwide. The main bottleneck towards this goal is that existing commercial solutions for preparing sequencing libraries from SARS-CoV-2 samples are costly and time-consuming, mainly because the reagent volumes used are high (microliter range) and because a single library must be prepared from each sample and quantified before sequencing. In contrast, the COVseq method that we have described here allows constructing highly multiplexed sequencing libraries starting from small volumes of purified RNA samples, and it only requires a nanodispensing device in order to drastically reduce reagent volumes and therefore the cost per sample. The I-DOT nanodispensing device that we have used here is a versatile bench-top instrument that requires minimal maintenance and training. However, any other device with similar characteristics could be deployed instead. Our cost analysis indicates that COVseq is an extremely cost-effective approach that can be harnessed to conduct genomic surveillance of the virus on a global scale. In the **Supplementary Information** we provide a detailed step-by-step COVseq protocol, which can be readily adopted by national and international public health agencies, in order to massively scale their sequencing capacity.

The main limitation of COVseq is that, since it relies on restriction enzymes that cut the genome non-randomly, some parts of the SARS-CoV-2 genome may not be covered at a depth sufficiently high to reliably call mutations in these regions. However, our data demonstrate that with two restriction enzymes and sufficiently long sequencing reads (150 nt), the vast majority of SNVs reported so far worldwide[2], as well as all the SNVs marking the recently emerged UK lineage[18], can be detected by COVseq (Fig. 2a). Most importantly, 93.4% (93.4% ± 0.051%, mean ± s.d.) of the S region—which mediates the entry of the virus into the cell and is the target of all the SARS-CoV-2 vaccines developed so far—was covered at least 10 times in all the 74 low Ct samples sequenced by COVseq, indicating that our method can be used to track the emergence of new variants of potential epidemiological concern. One possibility to achieve full coverage with the same read length would be to incorporate one additional restriction enzyme, such as

BfaI, to cover those regions in the SARS-CoV-2 genome that are too far from MseI or NlaIII restriction sites (**Supplementary Fig. 3b**). It is also conceivable that, once they have been identified through standard library preparation methods applied to a limited number of samples, new variants of interest could be monitored by COVseq on a mass scale, only using a restriction enzyme and the minimum sequencing read length sufficient to cover them, thus enabling highly cost-effective targeted genomic surveillance.

Although we have developed COVseq for SARS-CoV-2, the CUTseq method on which COVseq is based could be easily adapted to other RNA viruses, such as influenza viruses, as well as to DNA viruses. Indeed, a quick survey of existing multiplexed PCR assays for viruses other than SARS-CoV-2 shows that CUTseq could be readily implemented to sequence the genome of Influenza A and B viruses as well as Dengue virus, using the same restriction enzyme combination as in COVseq (**Supplementary Fig. 8 and Supplementary Table 6**). Different enzyme combinations could also be tested to achieve optimal genome coverage, depending on the pathogen of interest. In conclusion, we envision that CUTseq will play an important role in the genomic surveillance of the ongoing and future pandemics.

Methods

Samples

To test the feasibility of COVseq, we used RNA extracted from the supernatant of a SARS-CoV-2 viral culture, previously established at the 'Amedeo di Savoia' hospital in Turin, Italy. In addition, to technically validate our method, we used 85 fully anonymous, left-over SARS-CoV-2 positive RNA samples that were collected during the Phase 1 (diagnostic samples) or Phase 2 (screening samples) of the 2020 pandemic, at the 'Amedeo di Savoia' hospital and Candiolo Cancer Institute in Turin, Italy, respectively (see **Supplementary Table 4**). The RNA samples were extracted either using the EasyMag extraction kit (Biomérieux, cat. no. 280133-280134-200292-280130-280131-280132-280135-280146) and the NUCLISENS easyMAG instrument (Biomérieux) (samples 1–30 in **Supplementary Table 4**) or the MagMax Viral/Pathogen Nucleic Acid Extraction kit (Applied Biosystems, cat. no. A42352) and the KingFisher instrument (Thermo Fisher Scientific) (samples 31–85 in **Supplementary Table 4**). In all cases, there was no DNase treatment step in the RNA extraction procedure. The samples encompassed a broad range of cycle threshold (Ct) values based on real-time PCR (see **Supplementary Table 4**). Since the study was conducted on anonymous left-over samples and no clinical and personal information was collected, no informed consent subscription was required. The study was approved by the competent Ethical Committee of the participating institutions as per Italian regulations. All the original left-over samples have been used up and are no longer available.

Real-time PCR

Supernatant and samples 1–30 (see **Supplementary Table 4**). We performed SARS-CoV-2 real-time PCR on these samples using the Liferiver Novel Coronavirus (2019-nCoV) Real-Time Multiplex RT-PCR kit (Shangai ZJ Bio-Tech CO. ltd. Liferiver, cat. no. RR-0479-02-ZJ) following the manufacturer's instructions. The kit allows simultaneous detection of three genes: ORF1ab, N and E. For each sample, we prepared a

25 µL reaction containing 5 µL of purified RNA and 20 µL of PCR master mix. PCR conditions were as following: (i) 45 °C for 10 min; (ii) 95 °C for 3 min; (iii) 45 cycles of 95 °C for 15 sec and 58 °C for 30 sec. Ct ≤ 43 was set as a cut-off for SARS-CoV-2 positivity.

Samples 31–85 (see **Supplementary Table 4**). We performed SARS-CoV-2 real-time PCR on these samples using the TaqPath COVID-19 CE-IVD RT-PCR kit (Applied Biosystems, cat.no. A48067) following the manufacturer's instructions for RNA samples extracted from up to 200 µL of input material. The kit allows simultaneous detection of three genes: ORF1ab, N and S. Ct ≤ 37 was set as cut-off for SARS-CoV-2 positivity.

Reverse transcription (RT) and multiplexed PCR

In order to be able to sequence the SARS-CoV-2 genome even in high Ct value samples containing only small amounts of SARS-CoV-2 RNA, we adopted a SARS-CoV-2 multiplexed PCR protocol (v200325.2) developed by the U.S. Centers for Disease Prevention and Control. (https://github.com/CDCgov/SARS-CoV-2_Sequencing/tree/master/protocols). We performed all the following steps in a biosafety level 2 (BSL-2) lab using standard reagent volumes. Briefly, we first reversed transcribed each RNA sample, by preparing a mix containing 5 µL of purified RNA, 1 µL of 50 µM random hexamers (Thermo Fisher Scientific, cat. no. N80800127), 1 µL of 10 mM dNTPs (Thermo Fisher Scientific, cat. no. R0191) and 6 µL of nuclease-free water (Thermo Fisher Scientific, cat. no. AM9932) and incubating it for 5 min at 65 °C. To generate single-stranded cDNA, we added (in order) 4 µL of SuperScript IV buffer (Thermo Fisher Scientific, cat. no. 18090050), 1 µL of 0.1 M DTT (Thermo Fisher Scientific, cat. no. 18090050), 1 µL of RNaseOUT (Thermo Fisher Scientific, cat. no. 10777-019) and 1 µL of SSIV reverse transcriptase enzyme (Thermo Fisher Scientific, cat. no. 18090050) to the RT mix and incubated it in a thermocycler using the following program: 25 °C 10 min, 50 °C for 10 min, 85 °C for 10 min and hold at 4 °C. Afterwards, we added 1 µL of RNAse H (Thermo Fisher Scientific, cat. no. 18021071) to the sample and incubated it for 20 min at 37 °C. For multiplexed PCR, we first mixed equal volumes of the corresponding forward (F) and reverse (R) primers at 50 µM in nuclease-free water (Integrated DNA Technologies) (see **Supplementary Table 1** for all primer sequences). We then prepared six primer pools according to the aforementioned CDC protocol, by mixing an equal volume of each F + R primer pair in a pool (see **Supplementary Table 1** for the primer pairs contained in each pool). To perform the multiplexed PCR, for each primer pool, we aliquoted 3 µL of each cDNA prepared as described above in 6 separate PCR tubes pre-filled with the following reaction mix: 15 µL of NEBNext Q5 Hot Start HiFi PCR Master Mix (NEB, cat. no. M0543L), 9.2 µL of nuclease-free water (Thermo Fisher Scientific, cat. no. AM9932), 1 µL of 4X SYBR Green (Thermo Fisher Scientific, cat. no. S7563) and 1.8 µL of primer pool at 10 µM. We then performed the PCR reaction using a thermocycler (Biometra GmBH) and the following program: (i) 98 °C for 30 sec; (ii) 40 cycles of 98 °C for 15 sec and 65 °C for 5 min; (iii) hold at 4 °C. We then pooled an equal volume (20 µL) from each of the six amplicon pools into a 1.5 mL tube and purified DNA with a 1.0 vol/vol ratio of PCR product and Ampure XP (Beckman Coulter, cat. no. A63881) bead suspension and eluted the purified PCR in 80 µL of nuclease-free water. To measure the DNA concentration in the sample, we used the Qubit dsDNA BR kit (Thermo Fisher Scientific, cat. no. Q32850) according to the manufacturer's instructions.

COVseq

A detailed step-by-step COVseq protocol is available in the **Supplementary Information**. Below, we briefly describe three COVseq workflows, whose cost effectiveness is discussed in the Cost Analysis in the **Supplementary Notes**.

Workflow I. To test the feasibility of COVseq, we initially applied the standard CUTseq protocol to few RNA samples prepared as described above and processed in individual 0.5 mL tubes. Briefly, we mixed 7µL (300 ng) of purified pooled PCR product with 1µL of NlaIII (NEB, cat. no. R0125L), 1 µL of MseI (NEB, cat.no. R0525L) and 1µL of 10x CutSmart Buffer (NEB, cat. no. B7204S) and incubated the sample for 3 h at 37 °C followed by inactivation for 20 min at 65 °C. Afterwards, we added the following reagents to the same sample (without purifying it) to reach a final volume of 30 µl: 1 µL of NlaIII and 1 µL of MseI adapters (both at 0.33 µM and prepared as we previously described[16]), 1 µl of T4 ligase (Thermo Fisher Scientific, cat. no. EL0011), 3 µl of T4 ligase buffer 10x (Thermo Fisher Scientific, cat. no. EL0011), 2.4 µl of ATP 10 mM (Thermo Fisher Scientific, cat.no. R0441), 0.6 µl BSA 50 mg/ml (Thermo Fisher Scientific, cat.no. AM2616) and 11 µl of nuclease-free water. We incubated the sample for 16 h at 16 °C and the next day purified it with a 1.2 vol/vol ratio of sample and Ampure XP bead suspension and eluted the purified DNA in 10 µL of nuclease-free water. We performed *in vitro* transcription (IVT) with the MEGAscript® T7 Transcription kit (Thermo Fisher Scientific, cat. no. AM1334) using 8 µL of purified DNA in a final volume of 20 µl and incubated the reaction for 14 h at 37 °C. After IVT, we purified the amplified RNA with a 1.8 vol/vol ratio of sample and Ampure XP bead suspension and eluted the purified RNA in 10 µL of nuclease-free water. We then ligated the RA3 adaptors by pre-heating 1 µL of RA3 adapter for 2 min at 70 °C, followed by the addition of 7.8 µL of purified RNA, 1 µL of T4 RNA Ligase 2 truncated (Thermo Fisher, cat. no. M0242L), 1 µL of RNase OUT (Thermo Fisher Scientific, 10777-019) and 1.2 µL of RNA ligase buffer 10x (Thermo Fisher Scientific, cat. no. M0242L) and incubating the mix for 2 h at 37 °C. To reverse transcribe the RNA, we added to the same samples 2 µL of the RT primer (RTP) pre-heated for 2 min at 70 °C, 2 µL of SuperScript IV reverse transcriptase (Thermo Fisher Scientific, cat. no. 18090050), 5 µL of SuperScript IV buffer 5x (Thermo Fisher Scientific, cat. no. 18090050), 1 µL of 12.5 mM dNTPs (Thermo Fisher Scientific, cat.no. R1121), 2 µL 0.1 M DTT (Thermo Fisher Scientific, cat. no. 18090050) and 1 µL of RNase OUT (Thermo Fisher Scientific, cat.no. 10777-019) to reach a final volume of 25 µL, and incubated the mix for 20 min at 50 °C followed by an inactivation step of 10 min at 80 °C. After RT, we prepared a PCR mix containing 25 µL of cDNA, 16 µL of RP1 primer, 16 µL of the index primer RPI, 200 µL of NEBNext® Ultra™ II Q5® Master Mix 5x (NEB, cat. no. M0544S) and 143 µL of nuclease-free water. We split the PCR mix into 8 strips (50 µL each) and performed PCR in a thermocycler (Biometra GmBH) with the following program: (i) 98 °C for 30 sec; (ii) 10 cycles of 98 °C for 10 sec, 60 °C for 30 sec, 65 °C for 45 sec; (iii) 65 °C for 5 min; (iv) hold at 4 °C. We purified the final library with a 0.8 vol/vol ratio of sample and Ampure XP bead suspension and eluted the purified library in 30 µL of nuclease-free water. We measured the DNA concentration in the library using the Qubit dsDNA HS kit (Thermo Fisher Scientific, cat. no. Q32851) and analyzed the fragment size distribution on a Bioanalyzer 2100 (Agilent Technologies, cat. no. G2943CA) using the High Sensitivity DNA kit (Agilent Technologies, cat. no. 5067–4626).

Workflow II. To process multiple samples in parallel, we performed all reactions until IVT in 384-well plates, leveraging on the I-DOT One nanodispensing device (Dispendix GmBH), which we previously deployed for high-throughput CUTseq[5], to reduce the volume of each reagent and therefore the cost per sample. However, since our I-DOT machine could not be placed inside a biosafety level 2 (BSL-2) lab—which is required to safely handle potentially infectious RNA samples—we used it only for the CUTseq step, while the RT and multiplexed PCR steps were done in a BSL-2 lab using standard multi-channel pipettes. In principle, however, all the steps could be implemented on I-DOT, provided that the machine can be placed inside a BSL-2 lab. Briefly, we dispensed 50 nL of purified or 100 nL of non-purified pooled PCR amplicons in each well of a 384-well plate manually pre-filled with 5 µL per well of mineral oil (Sigma-Aldrich, cat. no. M5904), and then brought up to 350 nL with nuclease-free water to each well. After dispensing for each step, we briefly vortexed the plate on a thermomixer (Eppendorf) at 1,000 rpm for 1 min and centrifuged the plate at 3,220 g for 5 min before each incubation. For digestion, we dispensed 150 nL per well of a digestion mix containing 50 nL of NlaIII, 50 nL of MseI and 50 nL of 10x CutSmart Buffer, and incubated the plates at 37 °C for 1 h followed by 65 °C for 20 min to inactivate the enzymes. After digestion, we dispensed 150 nL of NlaIII adaptors and 150 nL of MseI adaptors (each at 33 nM and prepared as previously described[5]) into each well, followed by 700 nL of a ligation mix containing 200 nL of T4 rapid DNA ligase (ThermoFisher, cat. no. K1423), 300 nL of T4 ligase buffer (Thermo Fisher Scientific, cat. no. K1423), 120 nL of 10 mM ATP, 30 nL of BSA 50 mg/mL, and 50 nL of nuclease-free water. We incubated the plates at 22 °C for 30 min, after which we manually dispensed 5 µL of nuclease-free water into each well, pooled the contents of multiple wells in the same plate manually, and transferred the solution into a 1.5 mL tube. Lastly, we purified the pooled samples with a 1.2 vol/vol ratio of sample and Ampure XP bead suspension and eluted each pool in 20 µL of nuclease-free water. We prepared sequencing libraries in the same way as described above for COVseq in single tubes. The number of samples per library depends on the number of differently barcoded adapters available. We have designed 384 different NlaIII and MseI adapters (see **Supplementary Table 3**), allowing a maximum of 384 samples to be pooled into the same library. However, higher multiplexing could be easily achieved by using a larger number of sequence barcodes when designing COVseq adapters.

Workflow III. As a proof-of-principle we tested this COVseq workflow on SARS-CoV-2 synthetic RNA (Twist Bioscience, cat. no. 102019), which can be handled in a biosafety level 1 (BSL-1) lab, since it was logistically difficult for us to place our I-DOT nanodispenser in a BSL-2 lab. We performed RT, multiplexed PCR and barcoding by CUTseq sequentially in the same wells of a 384-well plate, without any intermediate purification step until the samples were pooled before IVT. In brief, we dispensed 50 nL of synthetic SARS-CoV-2 RNA containing either 5,000 and 10,000 genome copies into multiple wells of a 384-well plate pre-filled with 5 µL per well of mineral oil. For each sample, we set up six parallel reactions since we used six individual primer pools following the CDC multiplexed PCR protocol (see above). We then dispensed 15 nL per well of a primer-dNTP mix containing 5 nL of 50 µM random hexamers (Thermo Fisher Scientific, cat.no. N80800127), 5 nL of 10 mM dNTPs (Thermo Fisher Scientific, cat.no. R0191) and 5 nL of nuclease-free water (Thermo Fisher Scientific, cat. no. AM9932). We incubated the plate at 65 °C for 5 min and cooled it down immediately on ice. Afterwards, we added 35 nL per well of a first

strand synthesis mix containing 20 nL of SuperScript IV buffer (Thermo Fisher Scientific, cat.no. 18090050), 5 nL of 0.1M DTT (Thermo Fisher Scientific, cat.no. 18090050), 5 nL of RNAseOUT (Thermo, 10777-019) and 5 nL of SSIV reverse transcriptase enzyme (Thermo Fisher Scientific, cat.no. 18090050). We incubated the samples at the following temperatures: 25 °C for 10 min, 50 °C for 10 min, 85 °C for 10 min and hold on ice at 4 °C. After RT, we spun down the plate and performed the multiplex PCR by dispensing 250 nL per well of a PCR mix containing 42 nL of nuclease-free water, 12 nL of 4x SYBR Green (Thermo Fisher Scientific, cat.no. S7563), 175 nL of NEBNext Q5 Hot Start HiFi PCR Master Mix (NEB, cat.no. M0543L) and 21 nL of one of the six primer pools (10 μM). We performed PCR using the same settings as described above for workflow #1. We then performed digestion, ligation and library preparation exactly as described above for workflow #2. Notably, we dispensed the same sample barcode in each of the six wells corresponding to the same sample (each well containing a different pool of amplicons, as explained above).

Preparation of SARS-CoV-2 sequencing libraries using a standard approach

To validate COVseq, we generated individual libraries from 30 left-over samples (1–30 in **Supplementary Table 4**) using the NEBNext® Ultra™ II FS DNA Library Prep Kit (NEB, cat. no. E7805L) following the manufacturer's instructions. Briefly, we used 250 ng of the pooled purified amplicons from each sample as input to prepare a library. First, we enzymatically fragmented the amplicons for 7 min at 37 °C followed by incubation for 30 min at 65 °C to achieve a target size around 200 bp. After fragmentation, we performed end-repair and adapter ligation in the same tube followed by purification of the fragments using a 0.8 vol/vol ratio of Ampure XP beads (Beckman, cat. no. A63881), following the manufacturer's instructions. We amplified adaptor-ligated DNA fragments by 3 PCR cycles with barcoded primers (NEB, cat. no. E7500S) following the manufacturer's instructions and purified the PCR product with a 0.9 vol/vol ratio of Ampure XP beads. We assessed the size distribution and concentration of the libraries on a Bioanalyzer 2100 (Agilent Technologies, cat. no. G2943CA) using the High Sensitivity DNA kit (Agilent Technologies, cat. no. 5067–4626).

In silico coverage prediction

We extracted all the cut sites from the SARS-CoV-2 genome using a custom Python script. Following this, we predicted COVseq coverage of the most frequent mutations currently known for SARS-CoV-2 by extending known cut site locations in the SARS-CoV-2 genome by the effective read length (sequence read length minus 20 bp of adapter sequence) on both sides. Finally, we calculated how many times each mutation overlaps these locations.

Sequencing

We sequenced all the COVseq and NEBNext libraries on the NextSeq 500 system from Illumina using either the High Output Kit v2.5 (75 Cycles) (Illumina, Cat. No. 20024906) or the High Output Kit v2.5 (150 Cycles) (Illumina, Cat. No. 20024907) following the manufacturer's instructions.

Sequencing data pre-processing and variant calling

We demultiplexed raw sequence reads to fastq files based on index sequences using the BaseSpace® Sequence Hub cloud service of Illumina. We then processed individual libraries using a custom *snakemake* pipeline. In short, for each library we extracted the sample barcodes from the reads and assigned them to their corresponding sample while allowing one mismatch using a Python script. We aligned the reads to the SARS-CoV-2 reference genome (NC_045512.2) using *bwa mem*[20] (version 0.7.17-r1188) and discarded the reads that mapped further than 20 bases away from NlaIII and/or MseI cut sites. Following this, we used *ivar*[21] (version 1.3) to trim the primer sequences. To call variants, we used *bcftools mpileup* followed by *bcftools call* (version 1.10.2). We generated a consensus sequence using *bcftools mpileup* (version 1.10.2) followed by *ivar consensus*[21] (version 1.3). To determine the percentage of reads that mapped to different organisms and common contaminants we used *FastQ-screen*[22] (version 0.14.1). Briefly, 100,000 reads were sampled from the fastq files and aligned to 15 reference sequences using *bowtie2*[23] (version 3.5.1) (see **Supplementary Table 7**). We performed all subsequent analyses using custom R scripts.

Phylogenetic analyses

We downloaded sequences and sequence metadata from GISAID (https://www.gisaid.org/ 2020-12-24) and added all the COVseq generated libraries and relevant metadata from 74 left-over samples with low Ct (≤ 35). We then used the *ncov* tool (https://github.com/nextstrain/ncov) built on *nextstrain*[19] to generate a temporal and spatial phylogenetic tree. Additionally, we randomly selected from GISAID 300 samples from Italy and 700 samples from around the World. We analyzed and visualized the resulting *newick* tree in R using *ggtree*[24] (version 2.2.4).

Code availability

All the custom code used for processing COVseq sequencing data is available at https://github.com/ljwharbers/COVseq. The custom code in MATLAB used in the Cost Analysis (see **Supplementary Notes**) is available upon request to the corresponding author.

Declarations

Data availability

All reference sequences used in this study are listed in **Supplementary Table 7**. The BAM files used to generate all the plots in the main Figures and Supplementary Figures have been deposited at the SciLifeLab Data Repository and can be accessed through the following DOI: 10.17044/scilifelab.13560533.

Acknowledgements

We would like to acknowledge the thousands of research groups from all over the World who publicly shared their SARS-CoV-2 genome sequences through the GISAID initiative platform. This work was supported by funds from the National Natural Science Foundation of China (no. 81972475) and Chinese

Postdoctoral Science Foundation (2019T120593, 2018M630787) to N.Z.; by funds from the Fondazione Piemontese per la Ricerca sul Cancro (INTEGRAZIONI DIAGNOSTICA IN ONCOLOGIA – INTERONC FPRC 5x1000 MIUR 2017) to A.S. and A.SO.; by a SciLifeLab/KAW National COVID-19 Research Program project grant, Research Area Viral Sequence Evolution to N.C.; and by a private donation for COVID-19 research from Chiesi Pharma AB, also to N.C.

Author contributions

Conceptualization: N.C. *Samples*: A.S., A.SO., V.G. *Data curation*: L.H. *Formal analysis*: L.H., M.B., N.C. *Funding acquisition*: N.C. *Investigation*: M.S., N.Z., M.G.M., S.B., T.T.H.N. *Methodology*: M.S., N.Z., T.T.H.N. *Project administration*: N.C. *Software*: L.H. *Supervision*: N.C. *Visualization*: L.H., N.C. *Figure preparation and writing*: N.C. with contributions from all the authors.

Competing interests

The authors declare no competing interests.

References

1. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
2. Mercatelli, D. & Giorgi, F. M. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. *Front. Microbiol.* **11**, 1800 (2020).
3. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* (2020) doi:10.1038/s41586-020-2895-3.
4. Harilal, D. *et al.* SARS-CoV-2 Whole Genome Amplification and Sequencing for Effective Population-Based Surveillance and Control of Viral Transmission. *Clin. Chem.* **66**, 1450–1458 (2020).
5. Meredith, L. W. *et al.* Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect. Dis.* **20**, 1263–1272 (2020).
6. Oude Munnink, B. B. *et al.* Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* (2020) doi:10.1126/science.abe5901.
7. Pillay, S. *et al.* Whole Genome Sequencing of SARS-CoV-2: Adapting Illumina Protocols for Quick and Accurate Outbreak Investigation during a Pandemic. *Genes* **11**, (2020).
8. Nasir, J. A. *et al.* A Comparison of Whole Genome Sequencing of SARS-CoV-2 Using Amplicon-Based Sequencing, Random Hexamers, and Bait Capture. *Viruses* **12**, (2020).
9. Tyson, J. R. *et al.* Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv* (2020) doi:10.1101/2020.09.04.283077.
10. Xiao, M. *et al.* Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. *Genome Med.* **12**, 57 (2020).

11. Paden, C. R. *et al.* Rapid, Sensitive, Full-Genome Sequencing of Severe Acute Respiratory Syndrome Coronavirus 2. *Emerg. Infect. Dis.* **26**, 2401–2405 (2020).
12. Doddapaneni, H. *et al.* Oligonucleotide capture sequencing of the SARS-CoV-2 genome and subgenomic fragments from COVID-19 individuals. *bioRxiv* (2020) doi:10.1101/2020.07.27.223495.
13. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914-921.e10 (2020).
14. Chen, C. *et al.* MINERVA: A facile strategy for SARS-CoV-2 whole genome deep sequencing of clinical samples. *bioRxiv* 2020.04.25.060947 (2020) doi:10.1101/2020.04.25.060947.
15. Credle, J. J. *et al.* Highly multiplexed oligonucleotide probe-ligation testing enables efficient extraction-free SARS-CoV-2 detection and viral genotyping. *BioRxiv Prepr. Serv. Biol.* (2020) doi:10.1101/2020.06.03.130591.
16. Zhang, X. *et al.* CUTseq is a versatile method for preparing multiplexed DNA sequencing libraries from low-input samples. *Nat. Commun.* **10**, 4732 (2019).
17. Mina, M. J., Parker, R. & Larremore, D. B. Rethinking Covid-19 Test Sensitivity - A Strategy for Containment. *N. Engl. J. Med.* **383**, e120 (2020).
18. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *Virological* https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (2020).
19. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinforma. Oxf. Engl.* **34**, 4121–4123 (2018).
20. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).
21. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).
22. Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research* **7**, 1338 (2018).
23. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
24. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinforma.* **69**, e96 (2020).

Figures

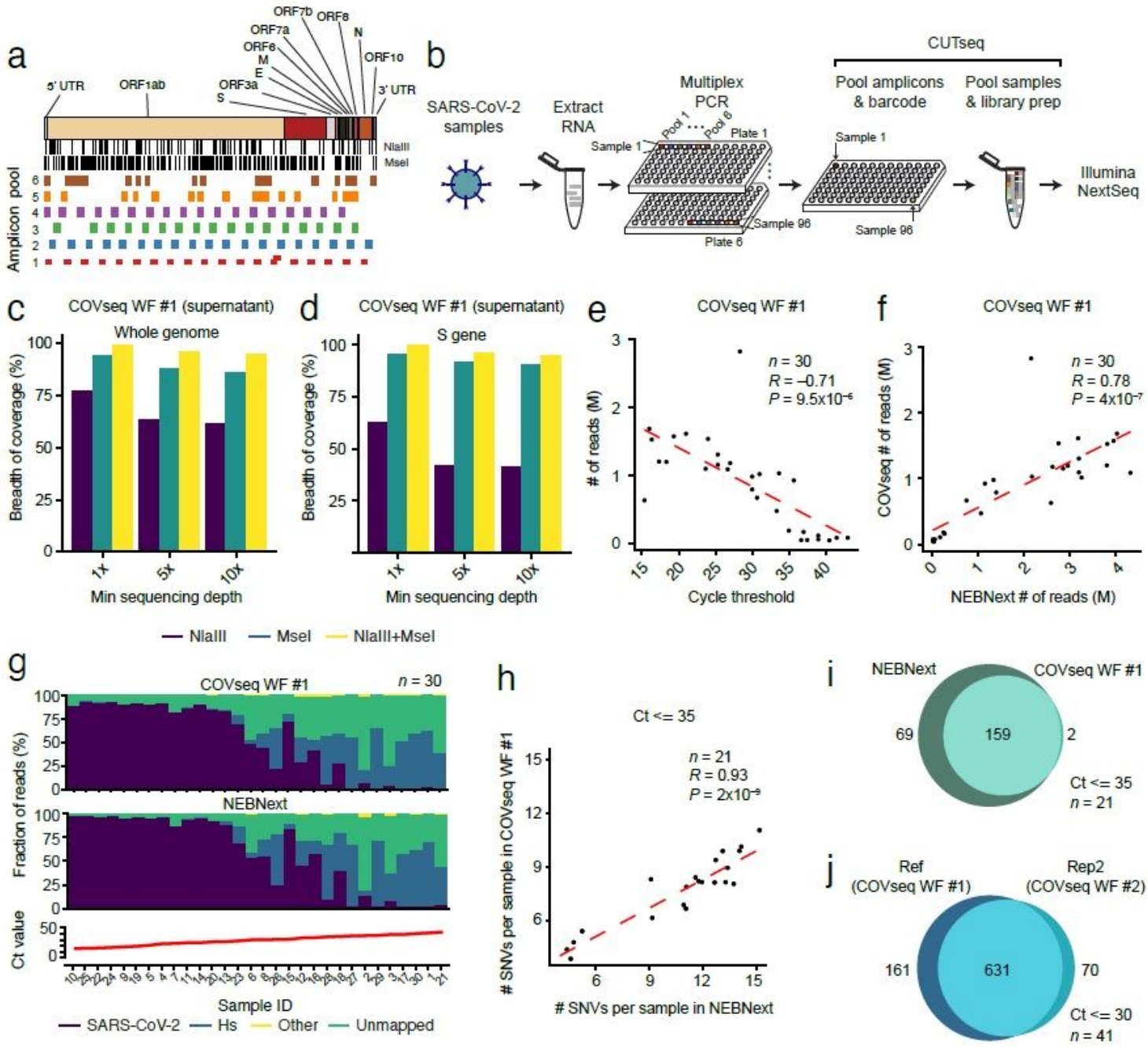

Figure 1

COVseq implementation and validation. (a) Location of MseI and NlaIII recognition sites (vertical black bars) and amplicons (colored rectangles) in the US CDC multiplexed PCR assay along the SARS-CoV-2 genome. Gene names (top) are according to the reference SARS-CoV-2 sequence NC_045512.2. (b) Schematic high-throughput COVseq workflow. Purified RNA samples (e.g., extracted from nasal- or oro-pharyngeal swabs) are first distributed in six wells of a multi-well plate and amplified using six different PCR primer pools to amplify the amplicons shown in (a). After PCR, the contents of the six wells are merged into a single well of a new plate (with or without purification) and CUTseq16 is used to barcode each sample. After barcoding, up to 384 samples are pooled together in the same sequencing library. (c) Percentage of bases in the SARS-CoV-2 reference genome covered by COVseq at varying sequencing

depths (SE150 sequencing) for three different libraries prepared from RNA extracted from the supernatant of a viral culture, using genome digestion with one or two restriction enzymes (MseI and NlaIII). The dashed red line represents the theoretical coverage at 1 using both enzymes. (d) Same as in (c), but for the S gene encoding the spike protein. (e) Inverse correlation between the cycle threshold determined by RT-PCR and the number of reads, for 30 samples (samples 1–30 in Supplementary Table 4) sequenced by COVseq. M, millions. Each dot represents a sample. Dashed red line: linear regression fit. R, Pearson's correlation coefficient. P, t-test, two-tailed. (f) Correlation between the total number of reads obtained by applying COVseq vs. a standard library preparation method (NEBNext) to the same 30 samples shown in (e). M, millions. Each dot represents a sample. Dashed red line: linear regression fit. R, Pearson's correlation coefficient. P, t-test, two-tailed. (g) Percentage of sequencing reads aligned to the SARS-CoV-2 reference genome, human reference genome (Hs), other genomes or unmapped, for the same samples shown in (f). The bottom plot shows the cycle threshold (Ct) value of each sample. Sample names are the same as in Supplementary Table 4. (h) Correlation between the number of single-nucleotide variants (SNVs) per sample detected by COVseq vs. NEBNext, in 21 (n) out of 30 samples (samples 1–30 in Supplementary Table 4) with low Ct (≤ 35). Each dot represents one sample. Dashed red line: linear regression fit. R, Pearson's correlation coefficient. P, t-test, two-tailed. (i) Venn diagram showing the extent of overlap between the SNVs identified by COVseq and NEBNext in the same 21 samples shown in (h). (j) Venn diagram showing the extent of overlap between the SNVs identified in 41 (n) out of 55 samples (samples 31–85 in Supplementary Table 4) pooled in the reference (Ref) and replicate 2 (Rep2) libraries.

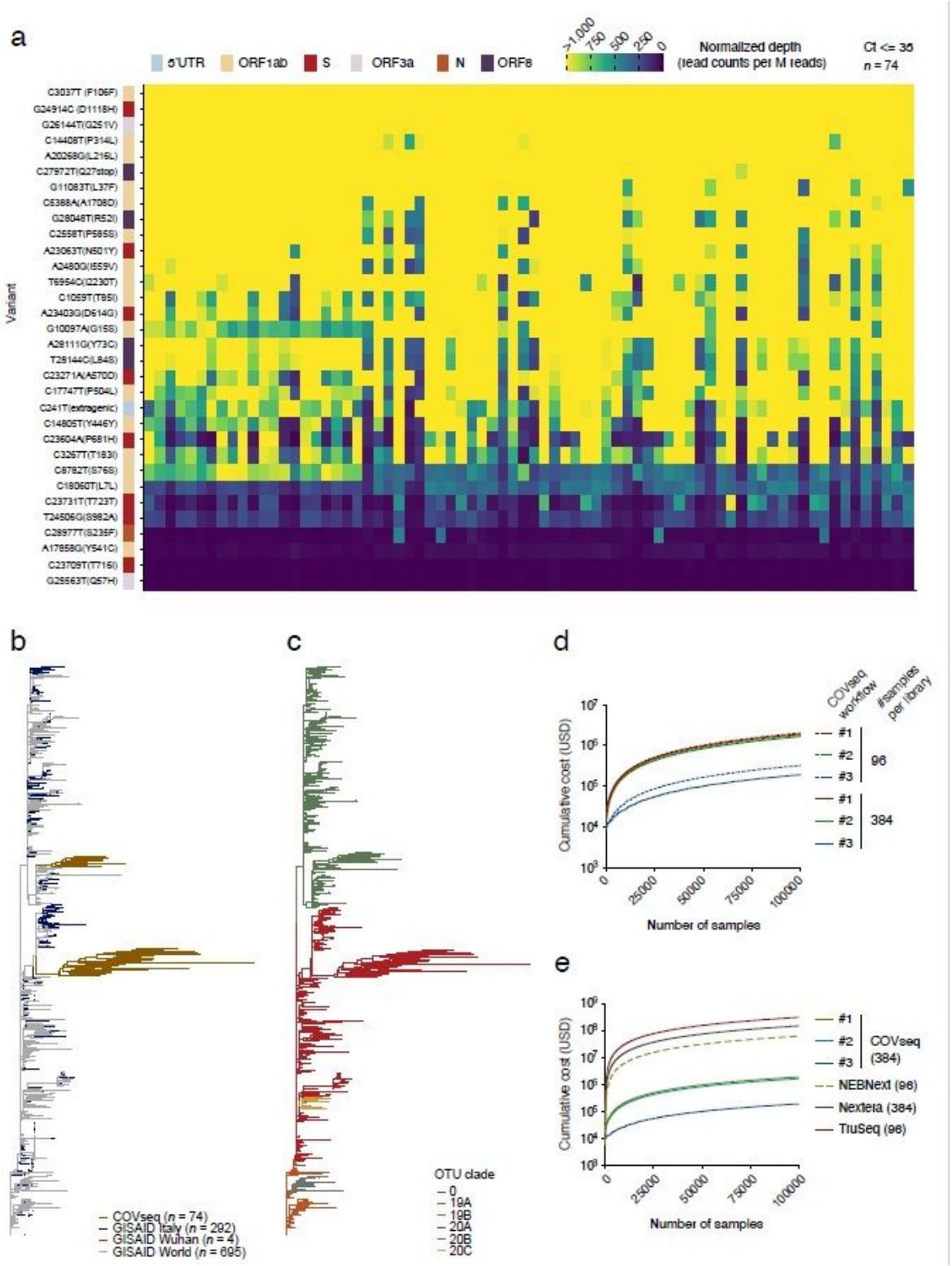

Figure 2

COVseq implementation and validation. (a) Location of MseI and NlaIII recognition sites (vertical black bars) and amplicons (colored rectangles) in the US CDC multiplexed PCR assay along the SARS-CoV-2 genome. Gene names (top) are according to the reference SARS-CoV-2 sequence NC_045512.2. (b) Schematic high-throughput COVseq workflow. Purified RNA samples (e.g., extracted from nasal- or oro-pharyngeal swabs) are first distributed in six wells of a multi-well plate and amplified using six different

PCR primer pools to amplify the amplicons shown in (a). After PCR, the contents of the six wells are merged into a single well of a new plate (with or without purification) and CUTseq16 is used to barcode each sample. After barcoding, up to 384 samples are pooled together in the same sequencing library. (c) Percentage of bases in the SARS-CoV-2 reference genome covered by COVseq at varying sequencing depths (SE150 sequencing) for three different libraries prepared from RNA extracted from the supernatant of a viral culture, using genome digestion with one or two restriction enzymes (MseI and NlaIII). The dashed red line represents the theoretical coverage at 1 using both enzymes. (d) Same as in (c), but for the S gene encoding the spike protein. (e) Inverse correlation between the cycle threshold determined by RT-PCR and the number of reads, for 30 samples (samples 1–30 in Supplementary Table 4) sequenced by COVseq. M, millions. Each dot represents a sample. Dashed red line: linear regression fit. R, Pearson's correlation coefficient. P, t-test, two-tailed. (f) Correlation between the total number of reads obtained by applying COVseq vs. a standard library preparation method (NEBNext) to the same 30 samples shown in (e). M, millions. Each dot represents a sample. Dashed red line: linear regression fit. R, Pearson's correlation coefficient. P, t-test, two-tailed. (g) Percentage of sequencing reads aligned to the SARS-CoV-2 reference genome, human reference genome (Hs), other genomes or unmapped, for the same samples shown in (f). The bottom plot shows the cycle threshold (Ct) value of each sample. Sample names are the same as in Supplementary Table 4. (h) Correlation between the number of single-nucleotide variants (SNVs) per sample detected by COVseq vs. NEBNext, in 21 (n) out of 30 samples (samples 1–30 in Supplementary Table 4) with low Ct (≤ 35). Each dot represents one sample. Dashed red line: linear regression fit. R, Pearson's correlation coefficient. P, t-test, two-tailed. (i) Venn diagram showing the extent of overlap between the SNVs identified by COVseq and NEBNext in the same 21 samples shown in (h). (j) Venn diagram showing the extent of overlap between the SNVs identified in 41 (n) out of 55 samples (samples 31–85 in Supplementary Table 4) pooled in the reference (Ref) and replicate 2 (Rep2) libraries.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf
- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.xlsx
- SupplementaryTable5.xlsx