



Optimizing the Illumina COVIDSeq laboratorial and bioinformatics pipeline on thousands of samples for SARS-CoV-2 Variants of Concern tracking



Sara Donzelli ^{a,1}, Ludovica Ciuffreda ^{b,1}, Martina Pontone ^{c,1}, Martina Betti ^d, Alice Massacci ^d, Carla Mottini ^b, Francesca De Nicola ^b, Giulia Orlandi ^e, Frauke Goeman ^b, Eugenia Giuliani ^e, Eleonora Sperandio ^d, Giulia Piaggio ^b, ISG COVID Team, Aldo Morrone ^e, Gennaro Ciliberto ^f, Maurizio Fanciulli ^b, Giovanni Blandino ^a, Fulvia Pimpinelli ^{c,1}, Matteo Pallocca ^{d,*}

^a Oncogenomics and Epigenetics, IRCCS Regina Elena National Cancer Institute, Rome, Italy

^b SAFU Unit, IRCCS Regina Elena National Cancer Institute, Rome, Italy

^c Microbiology and Virology Unit, IRCCS San Gallicano Dermatological Institute, Rome, Italy

^d Biostatistics, Bioinformatics and Clinical Trial Center, IRCCS Regina Elena National Cancer Institute, Rome, Italy

^e Scientific Direction, IRCCS San Gallicano Dermatological Institute, Rome, Italy

^f Scientific Direction, IRCCS Regina Elena National Cancer Institute, Rome, Italy

ARTICLE INFO

Article history:

Received 22 February 2022

Received in revised form 6 May 2022

Accepted 16 May 2022

Available online 19 May 2022

Keywords:

Illumina COVID-seq

SARS-CoV-2 genome

SARS-CoV-2 mutation

COVID mutations

SARS-CoV-2 Variants of Concern

Bioinformatics workflow

Oncology

Oncology Metagenomics

ABSTRACT

The SARS-CoV-2 Variants of Concern tracking via Whole Genome Sequencing represents a pillar of public health measures for the containment of the pandemic. The ability to track down the lineage distribution on a local and global scale leads to a better understanding of immune escape and to adopting interventions to contain novel outbreaks. This scenario poses a challenge for NGS laboratories worldwide that are pressed to have both a faster turnaround time and a high-throughput processing of swabs for sequencing and analysis. In this study, we present an optimization of the Illumina COVID-seq protocol carried out on thousands of SARS-CoV-2 samples at the wet and dry level. We discuss the unique challenges related to processing hundreds of swabs per week such as the tradeoff between ultra-high sensitivity and negative contamination levels, cost efficiency and bioinformatics quality metrics.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Background

The SARS-CoV-2 virus pandemic has posed many novel challenges to worldwide health care and laboratorial processes. On the one hand, the logistics in dealing with hundreds of thousands of novel infected individuals per day [1,2], of which a small but steady percentage require hospitalization and intensive care treat-

ment. Among the many epi processes involved with pandemic resilience, an effective viral genomic surveillance is one that heavily challenged laboratories and genomic facilities worldwide.

The importance of a robust and resilient genomic and bioinformatics workflow is of utter importance since it permits to 1. discover putative novel Variants of Concern (VoC) 2. Monitor variant distribution at the geographical and population scale. Coupling these capabilities with an international infrastructure of public datasets where sequences are promptly downloaded and shared [3,4] allows to generate a valuable amount of data that can share light on human-viral interactions with regard to public containment measures and vaccine distribution efficacy.

Several Next Generation Sequencing (NGS) protocols and library standards have been released in order to deal with VoC tracking, such as the ARTIC which is supported by a wealth of open

Abbreviations: BAM, Binary Alignment Map; BED, Browser Extensible Data; FDA, Food and Drug Administration; HPC, High Performance Computing; LIMS, Laboratory Information Management System; NGS, Next Generation Sequencing; RBD, Receptor-Binding Domain; SARS-CoV-2, Severe Acute Respiratory Syndrome Coronavirus; TAT, Turnaround Time; VoC, Variants of Concern.

* Corresponding author.

E-mail address: matteo.pallocca@ifo.it (M. Pallocca).

¹ These authors contributed equally to this work.

<https://doi.org/10.1016/j.csbj.2022.05.033>

2001-0370/© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

resources. Biotech companies such as Illumina and Thermo Fisher developed specific kits optimized for their platforms, a few of which have obtained FDA emergency approval for clinical-grade diagnostics [5,6]. While for the ARTIC amplicon a wealth of literature exists [7–10], where several protocols and reports were published for the Illumina setting, there is a lack of material for specific high-level optimization of the mentioned protocols, aiming to optimize Turnaround Times (TAT), sensitivity in sequencing low-abundance RNA samples, and specificity for accurate variant calling and noise-contamination reduction [11,12].

In this scenario, we present an optimization of the COVIDSeq Test vv3 protocol based on our experience of 3416 total sequences, representing 3% of Italian sequences and 31% of the Lazio region at the time of writing this paper.

2. Methods

The overall protocol is depicted in Fig. 1. Data stemming from local laboratory LIMS was finally merged with the lineage report for both patient-level reporting and cohort-level description for region network surveillance (rete Coronet) and Oncology Health Care Surveillance.

2.1. Molecular testing

Bosphore® Novel Coronavirus (2019-nCoV) Detection Kit v4 (Anatoliagenetworks, Istanbul, Turkey) was used to detect and characterize 2019-nCoV in human respiratory samples. Fluorescence detection was accomplished using FAM, HEX, Texas RED and Cy5 filters. SARS-CoV-2 was detected by three regions of the virus in a single reaction: E-gene assays are specific for bat-related betacoronaviruses, i.e. they detect both SARS-CoV-1 and SARS-CoV-2 and the ORF1ab target and N gene regions were used to discriminate SARS-CoV-2 specifically (Corman et al., 2020). Before amplification, a fast extraction was performed, which does not require a separate extraction but only a pre-treatment step that takes less than 10 min. Real-time PCR was performed with Montania 4896® thermal cycler. Swabs were tested within 4 h from collection.

2.2. RNA extraction

Samples with a *ct* value < 25 were re-extracted for the NGS analysis by using the total automated protocol of QIASymphony technology (Qiagen, Hilden, Germany) with a final elution of 60ul. This technology allows an automated purification of viral nucleic acids and combines the speed and efficiency of silica-based purification of viral nucleic acids with the convenient handling of magnetic particles.

3. Results

3.1. Contamination level reduction

During the first phase of the delta wave in Italy (August–September 2021), we reported a run with a particularly high contamination level, reaching a dropout rate of 82%. Not only was the Illumina protocol threshold of 5 amplicons covered in the negative controls higher with >50% of the SARS-CoV-2 genome covered, but the absolute coverage of contamination was higher than the positive samples i.e., the noise exceeding the actual signal, as shown from a test run containing 10 negative water samples that were highly enriched with viral coverage (Fig. 2A).

In order to tackle this issue, and having already double-checked all the contamination procedures and the lineage calling in previ-

ous known clusters and families, we reasoned that: (1) With current standards we could be over-sequencing SARS-CoV-2 for positive samples, being a ~30 kb genome and spending >2 million reads on each sample on the standard setups, having a high duplication level (~60%) due to saturation; (2) Ultra-high coverage coupled with high-PCR amplification cycles could result in over-amplification and sequencing of spurious, non-specific, single viral molecules; (3) The dropout level for swab concentration at 35 cycles was low (median 0.23%, Table S1) in all previous test runs, making it possible to reduce amplification without compromising sensitivity. Furthermore, the initial design of the CovidSeq protocol was designed to detect viral positivity, with hyper-sensitivity in mind, despite the fact that this information was always included with previous rtPCR for all our samples.

We thus introduced a variation in the Illumina CovidSeq protocol, namely the reduction of PCR cycles for the cDNA amplification and amplicon tagmentation steps, from 35 to 22 cycles and from 7 to 6 cycles, respectively. Finally, we had to align the final concentration of the positive control from 25 to 1500 copies to a novel amplification level.

3.2. Protocol change validation

In order to test lineage calling reproducibility of the 35-cycles contamination level, we re-sequenced the same samples deriving from different pandemic time frames (March 2021 and August 2021). The most striking results depend on the signal-to-noise ratio shown in the 10 negative controls that illustrate a 70% median coverage reduction (Fig. 2A and B).

The reanalysis confirmed a strong overlap in lineage assignment, even if the negative control contamination always stemmed from the dominant variant in the period (Alpha to Delta), lineage assignment was identical in 87% of all resequenced samples and finally where super-lineage assignment showed 100% concordance with all high-coverage samples (Table S2). Discrepancies on low coverage samples involved lineage shift to a less specific variant in the same family (2.1%) or either a missing assignment (1.2%), which was caused by an even lower coverage on low-quality samples. Nonetheless, a subset of samples (1.2%) was observed to switch from a non-assignment to a low-coverage assignment (e.g. *none* to AY.57), resulting from noise made by a consensus cleaner albeit incomplete. Lineage specificity increased for 2.1% of high-coverage samples (e.g., B.1.177.75 from B.1.177), a byproduct of a dramatic reduction in the variability of negative background coverage (from $2748x \pm 32582x$ to $4704x \pm 3435x$, Fig. S1A). All these results stress the importance of using multiple metrics for low-quality samples lineage calling, composed of a mixture of technical (e.g., median and RBD coverage) and qualitative (lineage class) data.

When considering the reproducibility of the process at the mutational level, we found a 100% concordance in 37/82 (45.1%) and >95% for additional 26/82 (31.7%) (Table S4); the lower tailed concordance samples reflect the nature of lineage non-assignment or super-class shift as described in Table 2. Furthermore, we characterized the mutations specifically present only in negative control samples. These 31 mutations are listed in Table S5 which represent major environmental contamination. The number of *contaminating mutations* does not change with the PCR cycle shift, since the main value of the protocol change relies on the lower background noise level (Fig. 2A).

In addition, PCR cycle reduction saved 1.5 h hours for each amplification process (4 h for 35 cycles required against 2.5 h for 22 cycles). This validation encouraged us to keep this amplification level as an acceptable signal-to-noise threshold, with a dropout level even lower thanks to cleaner consensus (Fig. S1B). To further reduce the possibility of lineage miscalling, we further con-

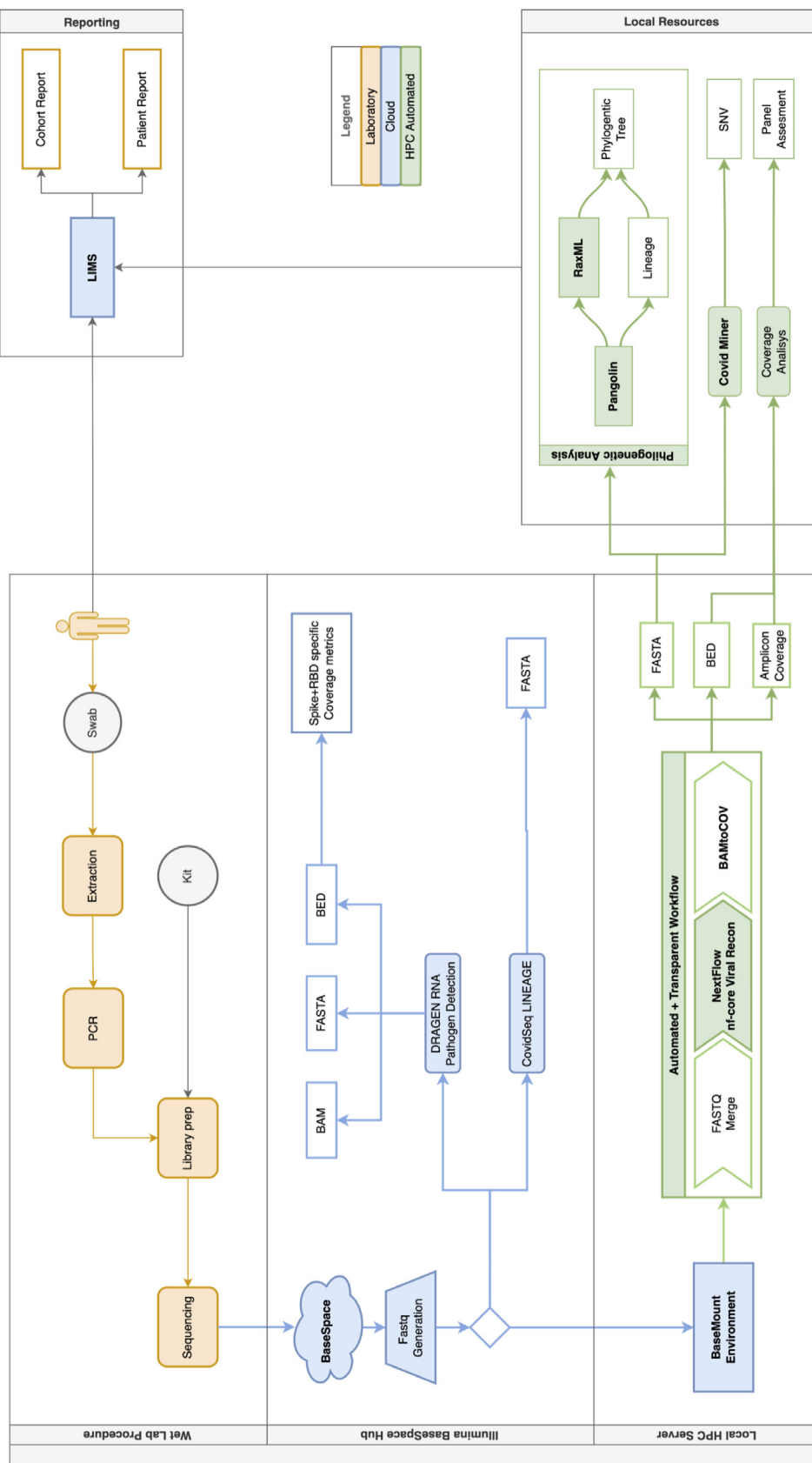


Fig. 1. A: High-level abstraction of wet laboratory and bioinformatics procedures. The overall workflow is divided into three major sections, including laboratorial (wet), cloud, and in-house HPC-automated analyses. Every section underwent prompted specific line of engineering and research, from LIMS automated extraction and reporting (Clinical Informatics) to library prep optimization, to bioinformatics pipeline parameter tuning, in order to achieve the required amount of information in the shortest timeframe possible.

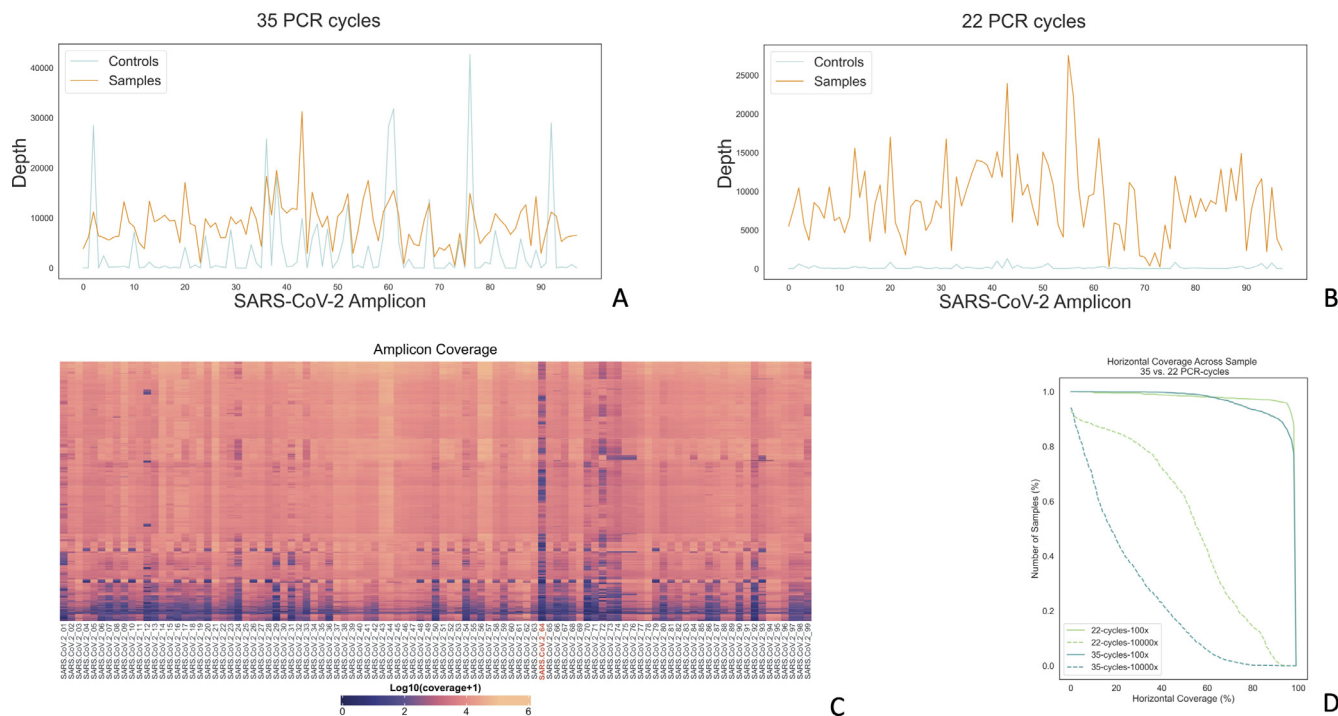


Fig. 2. A-B: Absolute coverage showing signal-to-noise reduction from 35 to 22 PCR cycles. Noise is represented by negative control samples (water) resequenced after the protocol change. Positive samples represent real swabs from March to September 2021, containing alpha and delta variants. **C:** Complete coverage analysis per amplicon on 3232 SARS-CoV-2 COVID-seq samples. Highlight on outlier amplicon 64. The overall coverage is consistently high across all samples (median >7000). **C:** Vertical coverage across samples on the whole genome. Highlight on Spike protein region, lower than the median coverage but consistently above the 200x threshold. **D:** Horizontal coverage across samples with different thresholds, showing that > 80% of samples have a median horizontal coverage > 100x across the whole SARS-CoV-2 genome, regardless of the number of PCR cycles. Overall, better horizontal coverage is achieved with the adapted protocol.

ducted a Bioinformatics check list at the single mutation level as described in the subsequent section.

3.3. Sequencing run setups

Sequencing runs had different setups based on the number of samples (Table 1). From 2 plates (188 samples), the Novaseq setup brought a reduction in cost-per-sample and sequencing time (from 24 h to 29 h).

3.4. Bioinformatics analysis

The DRAGEN COVIDSeq Test (RUO) App is considered the standard analysis workflow, following the Illumina guidelines. We benchmarked a two-way pipeline to have a wealthier amount of quality check and technical information on each test run. On the one hand, we employed the BaseSpace Hub cloud environment that allows high scalability and the power to process dozens or hundreds of samples in less than 10 h. We combined various output files from the DRAGEN RNA Pathogen Detection, the DRAGEN COVID Lineage and the DRAGEN COVIDSeq Test (RUO) Apps [13–15]. Specifically, global BAM files, FASTA consensus and BED files with coverage metrics were extracted to acquire an overview of the horizontal coverage. The FASTA consensus was then locally processed with an updated version of Pangolin/PangoLearn [16] to check for the latest changes in variant assignment. In addition, with the aim to have better control and governance of the analytical processes, we started looking for alternatives for the private cloud. We found a powerful workflow in the *viral-recon* pipeline of the NF-CORE environment, which was highly supported by the scientific community at the time of writing this paper [17]. We employed the *viral-recon* workflow on our NGS systems with 1,2,4,8 PCR plates setups, obtaining results in our HPC node of

Table 1
Sequencing instruments and flow cell/cartridge combos for desired throughput.

Samples	Instrument and Flow cell/cartridge
96	NextSeq 500/550 Mid Output Kit v2.5 (300 Cycles)
192–384	NextSeq 500/550 High Output Kit v2.5 (300 Cycles)
192–768	NovaSeq 6000 SP Reagent Kit v1.5 (300 cycles) with NovaSeq XP 2-Lane Kit v1.5 #20043130

48CPU/354 GB from 4 to 24 h. In order to achieve these performances, several non-critical steps of the pipeline were skipped via `--skip_kraken2 --skip_fastp --skip_nextclade --skip_snpeff --skip_variants_quast --skip_plasmidid --skip_blast --skip_bandage --skip_asciigenome`. An additional performance comparison is provided in Table 2, however it is important to mention that it is biased by our fixed local HPC environment when compared to the flexible and scalable Illumina Basespace/Amazon AWS combo. However, the local fixed HPC node more realistically represents the setup at the Research Hospital/Genomics Facility environment which came along privacy concerns.

Finally, the FASTA consensus produced from both cloud and local analysis processes are re-analyzed via our in-house developed COVID-miner framework (<https://covid-miner.ifo.it>) [18], specifically for the in-depth analysis of single mutations representative of unique variants. This approach is useful for either reconstructing putative cross-sample contamination, that is, how many Delta-specific mutations (e.g., R158G) are enriched in Omicron-assigned samples or vice-versa, and at which allele frequency as well as in calculating the level of viral similarity among known outbreaks or family clusters.

However, this setup leads to yet another challenge: the constant updating of bioinformatics databases and resources employed.

Table 2

Time performance metrics of several sample batches and bioinformatics workflows.

plates	samples (controls included)	Nfcore ViralRecon on 32 CPU/256 GB RAM	Illumina DRAGEN COVID Lineage	Illumina DRAGEN RNA Pathogen Detection
1	96	3–4 h	1.5 h (4 nodes)	3.5 h (4 nodes)
2	192	6–8 h	5 h–6 h (4 nodes)	3 h–4 h (9 nodes)
4	384	10–12 h	5 h–6 h (4 nodes)	5 h (10 nodes)
8	768	20–24 h	15.5 h (4 nodes)	9 h (16 nodes)

Even though the use of virtualized workflows permits more reproducible and deterministic behavior on the one hand, it does pose an issue when specific sub-routines and libraries are not updated on a weekly basis and where the virtualized pipeline is used as a black box. This issue pertains to both the *nf-core* and the *Basespace* approaches. A notable example of this is the *Scorpio* module employed along with the *Pangolin* tool: if only the second one is updated, the *Scorpio* module will not return a specific assignment metric since the novel variant is not found in its lookup table.

Finally, in regards to Health Care surveillance, we report only lineages using samples with high technical quality metrics (5x_genome_covered > 95% and Median Dedup Coverage > 100x and RBD locus minimum coverage > 10x).

3.5. Panel performance and quality metrics

Whole Genome Sequencing performed from April 2021 to January 2022 produced 3232 high quality samples. As expected, total coverage analysis showed that amplicons 43, 44, 56 and 57 are specifically highly enriched (coverage > 15000x) (Fig S1C), whilst in contrast amplicon 64 is consistently lower (median coverage = 800x) than the rest of the viral genome (Fig. 2C).

Amplicons spanning the Spike protein region [72–83], which are critical for VoC calling (Fig. S1D), have sufficient coverage on most samples (Median Undedup Coverage > 8000x, Minimum Undedup Coverage > 20x on 100% samples); nevertheless, their median coverage is lower than the rest of the genome overall (Fig. S1E).

These outlier regions call for greater attention to be paid to both over-sequencing (false positives) and putative low-coverage in outlier samples. It is important to note that all these metrics are in deduplicated coverage suitable for amplicon sequencing, while many DRAGEN/Illumina thresholds are indicated in coverage including duplication.

When considering the performance of the panel, we found that over 80% samples have a minimum coverage above 100x across the whole genome and that the PCR change does not pan an issue for horizontal coverage (Fig. 2D). Lineage analysis showed a switch from the beta-delta dominated scenario towards an omicron increase (Table S3), consistent with what is reported in the international data mining (GISAID) and literature. Accuracy of lineage calling, computed as the proportion of the defining variants having alternative alleles, is consistently high across all samples (Median ScorpioSupport = 0.875).

3.6. Logistics and turnaround time

The whole process was split up into four separate units at IFO: (1) multi-center swab collection, processing and extraction carried out at the Virology Unit, 1 to 2 days (2) library prep at the Oncogenomics Laboratory, 1 to 2 days (3) library quantification, cartridge loading, and sequencing run at the Genomics Facility, 1 to 2 days (4) primary and secondary analysis at the Bioinformatics Unit, 1 day. Many steps were processed overnight such as sequencing runs and cloud or High-Performance Computing analyses in order

to achieve a turnaround time of seven days. Every sub-process had a group of 2 to 3 people for high redundancy, except for the Virology Unit that features a large with >10 technicians on a 24 h-shift (*ISG COVID Team*).

4. Conclusions

The coronavirus pandemic proved once again how protocol sharing and Open Science approaches lead to process optimization and massive savings on human and capital resources.

One typical question arises when a VoC tracking optimization is presented: would it be possible to detect Novel viral Variants better? In theory yes it is possible, bearing in mind that a Variant is defined via a mixture of analyses comprising of Viral, Epidemiological and Geographical data. Viral sequences are not the only resources used for detection, the results from the lab need to be continuously compared to public resources such as Nextstrain or the COVID-miner, in order to avoid irreproducible announcements caused by technical artifacts such as inter-sample contamination [19]. An alert module can easily be added to the workflow when lineage assignment scores are particularly low, or the set of single mutations associated with a sample are poorly represented.

In conclusion, here we presented an optimized protocol that led to a severe reduction in background noise due to contamination, a modest gain in turnaround time and an integrated bioinformatics metrics system for Variant of Concern tracking. We believe that all this fine-grained tweaking is particularly important for those laboratories lacking strong expertise in NGS, that would significantly benefit from sharpening their analytic expertise without having to simply follow emergency protocols.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank dr. Cesare Camma and the whole group at the Istituto Zooprofilattico Sperimentale dell' Abruzzo e del Molise for their valuable insight during our discussions and sharing of relevant ideas on SARS-CoV-2 Sequencing. We thank Harshil Patel and *nf-core* community for their comments and all the effort made regarding the *viralrecon* pipeline. We acknowledge the CINECA award under the ISCRA initiative, for their availability of high performance computing resources and support (project code: VoCAT). We thank Tania Merlino for editorial assistance.

ISG Virology Covid Team: Cavallo Ilaria, Cazzani Andrea, Celesti Ilaria, D'Agosto Giovanna, Diano Martina, Federico Antonio, Fratelli Fulvia, Furzi Lorenzo, Maione Francesca, Mastrofrancesco Arianna, Obregon Francisco, Paluzzi Silvia, Petrolo Sara, Prignano Grazia, Ricca Valentina, Salvo Serena, Tatangelo Miriam, Trento Elisabetta, Zucchiatti Marco.

Funding

This work was supported by the Italian Ministry of Health (Ricerca Corrente 2019, Ricerca Corrente 2020).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.csbj.2022.05.033>.

References

- [1] Maslo C, Friedland R, Toubkin M, Laubscher A, Akaloo T, Kama B. Characteristics and Outcomes of Hospitalized Patients in South Africa During the COVID-19 Omicron Wave Compared With Previous Waves. *JAMA* Feb. 2022;327(6):583–4. <https://doi.org/10.1001/JAMA.2021.24868>.
- [2] Karim SSA, Karim QA. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *Lancet* (London, England) Dec. 2021;398(10317):2126. [https://doi.org/10.1016/S0140-6736\(21\)02758-6](https://doi.org/10.1016/S0140-6736(21)02758-6).
- [3] S. Khare et al. "GISAIID's Role in Pandemic Response," *China CDC Weekly*, 2021, Vol. 3, Issue 49, Pages: 1049-1051, vol. 3, no. 49, pp. 1049–1051, Dec. 2021, doi: 10.46234/CCDCW2021.255.
- [4] Harrison PW et al. The COVID-19 Data Portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. *Nucleic Acids Res* Jul. 2021;49(W1):W619–23. <https://doi.org/10.1093/NAR/GKAB417>.
- [5] "FDA Grants Emergency Use Authorization for Two Next-Generation COVID-19 Assays from Thermo Fisher Scientific – Aug 16, 2021." <https://thermofisher.mediaroom.com/2021-08-16-FDA-Grants-Emergency-Use-Authorization-for-Two-Next-Generation-COVID-19-Assays-from-Thermo-Fisher-Scientific> (accessed Feb. 08, 2022).
- [6] Illumina, "Illumina COVIDSeq Test- Instructions for Use", Accessed: Feb. 08, 2022. [Online]. Available: <https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-emergency-use-authorizations->.
- [7] "Arctic Network." <https://artic.network/1-about.html> (accessed Feb. 08, 2022).
- [8] Charre C et al. Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evol* Jul. 2020;6(2):75. <https://doi.org/10.1093/VE/VFAA075>.
- [9] Rambaut A et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* Nov. 2020;5(11):1403. <https://doi.org/10.1038/S41564-020-0770-5>.
- [10] "Phylogenetic Analysis | 176 genomes | 6 Mar 2020 – SARS-CoV-2 coronavirus/nCoV-2019 Genomic Epidemiology – Virological." <https://virological.org/t/phylogenetic-analysis-176-genomes-6-mar-2020/356> (accessed Feb. 08, 2022).
- [11] Bhojar RC et al. High throughput detection and genetic epidemiology of SARS-CoV-2 using COVIDSeq next-generation sequencing. *PLoS ONE* Feb. 2021;16(2):. <https://doi.org/10.1371/JOURNAL.PONE.0247115>e0247115.
- [12] Bhojar RC et al. An optimized, amplicon-based approach for sequencing of SARS-CoV-2 from patient samples using COVIDSeq assay on Illumina MiSeq sequencing platforms. *STAR Protocols* 2021;2(3):Sep. <https://doi.org/10.1016/J.XPRO.2021.100755>.
- [13] "DRAGEN RNA Pathogen Detection." <https://emea.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/dragen-rna-pathogen-detection.html> (accessed Feb. 08, 2022).
- [14] "DRAGEN COVID Lineage." <https://emea.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/dragen-covid-lineage.html> (accessed Feb. 08, 2022).
- [15] "DRAGEN COVIDSeq Test (RUO)." <https://emea.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/dragen-covidseq-test-ruo.html> (accessed Feb. 08, 2022).
- [16] O'Toole Á et al. "Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool", *Virus*. Evolution 2021;7(2):Dec. <https://doi.org/10.1093/VE/VFAB064>.
- [17] H. Patel et al., "nf-core/viralrecon: nf-core/viralrecon v2.3 – Copper Coatimundi," Feb. 2022, doi: 10.5281/ZENODO.5974693.
- [18] Massacci A et al. Design of a companion bioinformatic tool to detect the emergence and geographical distribution of SARS-CoV-2 Spike protein genetic variants. *Journal of Translational Medicine* 2020;18(1):Dec. <https://doi.org/10.1186/S12967-020-02675-4>.
- [19] F. Kreier. "Deltacron: the story of the variant that wasn't." *Nature*. vol. 602, no. 7895. 19–19. Feb. 2022. doi: 10.1038/D41586-022-00149-9.