Grant agreement No. 101046314

END-VOC

ENDING COVID 19 VARIANTS OF CONCERN THROUGH COHORT STUDIES: END-VOC

HORIZON-HLTH-2021-CORONA-01-02

D3.1

Sequencing pipeline optimisation

WP 3 – Sequencing and phylogenetics

Due date of deliverable	Month 6 – October 2022
Actual submission date	07/12/2022
Start date of project	01/05/2022
Duration	36 months
Lead beneficiary	UCL
Last editor	Damien Richard (UCL)
Contributors	François Balloux, Lucy van Dorp (UCL)

Dissemination Level		
PU	Public	X
SEN	Sensitive	

END-VOC has received funding from the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Disclaimer

The content of this deliverable does not reflect the official opinion of the European Union. Responsibility for the information and views expressed herein lies entirely with the author(s).

Copyright

History of the changes

Version	Date	Released by	Comments
0.2	08.11.2022	D. Richard	First draft
0.4	21.11.2022	Lucy van Dorp	Edits
0.6	30.11.2022	François Balloux	Deliverable version

Table of contents

Definitions and acronyms

Acronyms	**Definitions**
NGS	Next generation sequencing
SNP	Single nucleotide polymorphism

1. Introduction

Next generation sequencing (NGS) encompasses diverse approaches all aiming at generating high-throughput genomic data from a biological sample. The approach is used to determine the order of nucleotides within a DNA or RNA sequence, which can range from a targeted short sequence to entire genomes. There is considerable variation both in the sequencing technologies currently in use and the post-processing methods to analyse raw sequence data.

In the context of a global and collaborative setting in which multiple laboratories are involved in SARS-CoV-2 surveillance and sequencing, it is critical to effectively integrate and standardise the results produced by different cohorts each using subtly different sequencing methods and protocols. This document describes the challenge of integrating different sources of raw genomic data and proposes a series of recommendations as to how to minimise biases and problems when analysing datasets combining raw data drawn from heterogeneous sources.

Specifically, we aim at delivering a synthetic document providing a clear and accessible overview of the key steps of SARS-CoV-2 NGS to including methodological approaches. Recognising the specificities of each approach is useful to (i) integrate data from multiple sources - hence likely generated using multiple approaches - and (ii) optimize the method to be applied according to the scientific objectives of the study (deliverable “Optimize sequencing pipeline”). Moreover, we ambition to foster a collaborative environment in which methodological aspects of SARS-CoV-2 NGS will be openly, critically and productively discussed and the impact of specific methodological choices collectively assessed and endorsed (deliverable “provide support to cohorts”).

1.1. General context

The sequencing effort of SARS-CoV-2 is unprecedented. To date, over 14 million SARS-CoV-2 genomes assemblies have been deposited on the GISAID database. As a consequence, every step of the generation of SARS-CoV-2 genomic data has been assessed by dozens of experts since the beginning of the pandemic. This has led to the development of tools and methods specifically designed to handle the peculiarities of the SARS-CoV-2 global dataset, which represents an exceptionally large number of relatively small genomes, characterised by low genetic diversity. The scrutiny by the scientific community has also led to detailed characterization of the effect each methodological step can have on downstream results, such as SNP calling artefacts induced by specific primer sets or read depth required for confident variant calling. Fine-tuning of the sequencing approach parameters helps with cost-effectiveness when producing new raw sequencing data suitable to address the questions of a specific research project. Additionally, the fine characterization of SARS-CoV-2 sequencing artifacts broadens the scale of data (re)usability, as positions in the genome known to be prone to errors can be omitted even if only consensus sequences are available.

1.2. Deliverable objectives

In this document, we first highlight the key parameters that must be adequately set at each step of the bioinformatic pipeline used to transform raw sequencing data into genome

assemblies. We then provide non-exhaustive guidelines for downstream analyses of genome assemblies most commonly used in SARS-CoV-2 genetics research.

2. Summary of activities and research findings

2.1. From raw data to results: key bioinformatics steps

After generating raw DNA reads (or RNA under the form of cDNA in the case of SARS-CoV-2), several bioinformatics steps are required before interpretable results are generated. The widespread use of NGS technology has led to the development of mature and well-described data analysis pipelines reviewed for instance in (Carriço et al. 2018) and (Pereira et al. 2020). However, certain steps of the pipeline vary according to (i) the nature of the input data and (ii) the intended downstream use that will be made of the genomic data. We herein focus on identifying those steps and how one should adapt them to suit (i) and (ii) in the case of SARS-CoV-2. Numerous tools can perform each step, for an overview see (Hu et al. 2021). We choose to only cover mapping-based approaches, as *de novo* assembly (for which specific pipelines and methods are compared in (Islam et al. 2021)) is rarely used for SARS-CoV-2 sequencing.

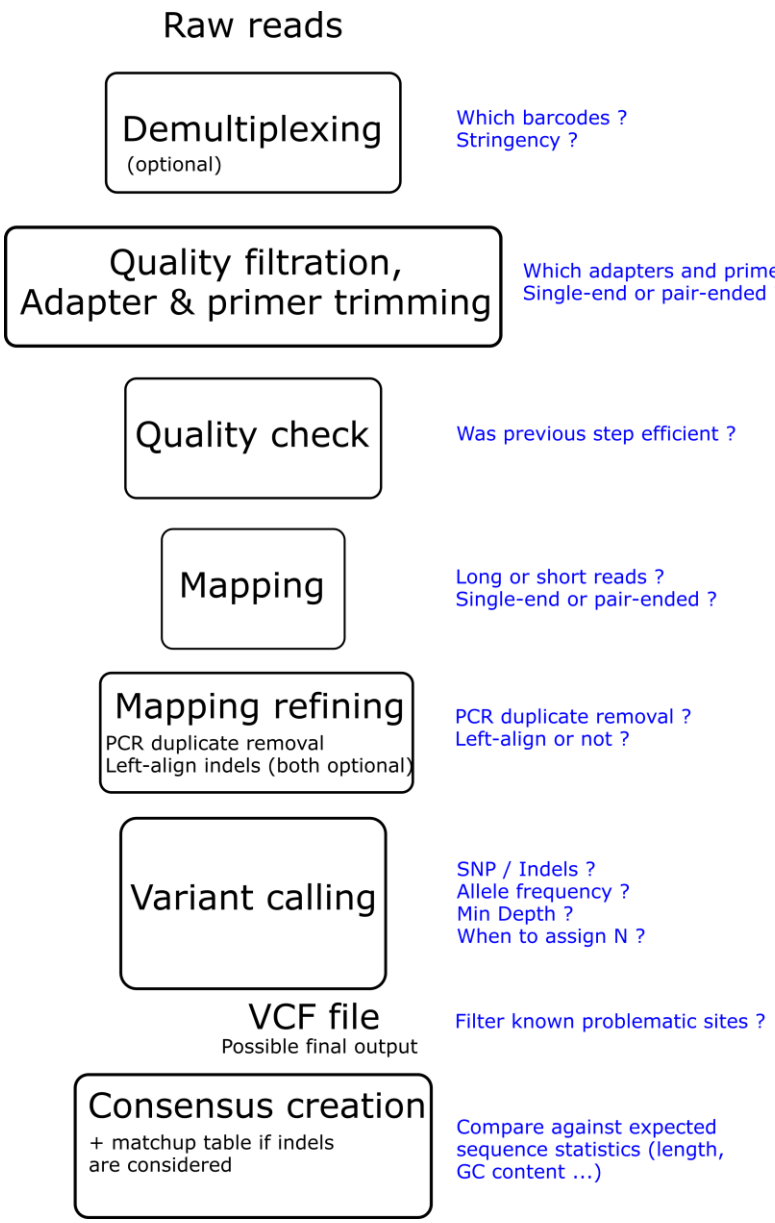

Figure 1. *Schematics of a standard bioinformatics pipeline with a focus on the key variable parameters that one must set / steps that can be performed in several ways*

Raw data quality and quantity

The combination of viral load, sequencing approach and sequencing intensity condition the quantity and the quality of the raw data that will used as input of the bioinformatics pipeline. While wet-lab steps fall outside of the scope of the present report, they are a prerequisite for successful further data processing (for more information, see Additional information). An elegant study evaluating the SARS-CoV-2 sequencing results of several laboratories that were provided with the same samples (and only having in common the use of an amplicon-based sequencing approach) showed that insufficient quantity of raw data was strongly correlated with poor sequencing reports (Wegner et al. 2022). However, after a certain quality and quantity of raw data is achieved, further gain through increased read depth is not obvious and final assembly quality is likely more dependent upon the bioinformatics choices taken.

Demultiplexing the data

Multiple biological samples are often sequenced in a single sequencing run. To differentiate reads of each different sample, unique small DNA sequences called barcodes are added to the sequencing reads. This mutiplexing procedure requires a bioinformatic step to reallocate the reads to the samples according to the barcode they carry. Sequencing service providers often supply already demultiplexed raw reads. If not, dedicated methods have been developed based on the sequencing approach. Particular attention must be taken as to which barcode set has been used to multiplex the data, as this information must be provided to the demultiplexing algorithm. Negative and non-SARS-CoV-2 controls should be used in order to estimate misassignment rates. After the process, most methods place the reads of each sample in a distinct folder as well as the unassigned reads. A high proportion of unassigned reads can reveal problems in the demultiplexing process, mistakes in the set of barcode set provided, or incorrect barcode strand orientation (i.e., non-reverse-complemented). Particular attention to the demultiplexing step is critical when working with low depth samples (<20x) or when minor allele frequencies are of interest, because misattribution of even a small proportion of reads can strongly affect the results in such situations. In these cases, one could consider requiring perfect adapter matching (on both sides) for read allocation and/or use the union of two different demultiplexing algorithms to confidently allocate the reads.

Removing sequencing adaptors and low-quality base pairs

Raw sequencing data often contain sequencing adapters and stretches of low-quality bases located at the read ends. Knowledge of which adapters and primers were used for the sequencing is useful for that step, although it can be performed without that information, by checking which sequencing adapters are overrepresented in the dataset (tools exist to provide this, for instance fastQC). In the case of amplicon sequencing, primer pairs are used to amplify overlapping SARS-CoV-2 genomic regions. Loci covered by those primers can harbour mutations or indels in the sample, so it is important to remove the primers at read ends not to create heterozygous SNPs. In addition, depending on the sequencing approach used, a step that removes rRNA and host reads might be needed. Quality-checking the raw data after the use of a trimming program (checks can also be conducted before, although typically less usefully) ensures that raw data is of high enough quality to continue processing.

Mapping refining

Further mapping refinements can be performed after the main mapping step. A first refinement is referred to as “PCR duplicate removal”. The PCR step creates multiple raw reads from the same RNA template, which are identified after the mapping step by their exactly matching mapping coordinates and subsequently removed. Of note, in an amplicon-based sequencing, this step should be omitted because all reads originating from a primer pair would map at the same place even if they are from distinct RNA templates. A second common refinement is termed “local realignment” where, during the mapping process, each read is treated independently. As a consequence, the mapper could report one or a few SNPs instead of an indel genuinely located near read ends. Finally, the “indel left alignment” refinement is an arbitrary convention. Indels are often located within tandem repeat loci, though they could be placed anywhere in the tandem. Left-aligning them homogenises the call across reads spanning the same position (Sehn 2015). The latter two steps are now

commonly included in the variant-caller (for example in GATK HaplotypeCaller and Freebayes) in which case there is no need to perform it as stand-alone step.

Filtering the variants

The below sections show the parameters that can be taken into account to differentiate between what one considers genuine variations from what one considers artefactual (or not needed in the given study). Most of those filters can be applied **at** the variant calling step itself, so that variants not meeting the criterion are not even reported, or **after** the variant calling step, so that the user can keep track of how many variants the user excluded. Parameters are not all independent, e.g., a dataset filtered for high quality SNPs will contain a few or no SNP displaying a low depth [Note: Herein, "depth" refers to what is sometimes called depth of coverage, which represents the mean number of reads that cover single nucleotide positions in the genome. This is to avoid confusion with another useful metric, coverage percent, which measures the proportion of positions covered by at least one read] and/or a low allele frequency.

Variant type

SNP refers to the replacement of one nucleotide by another one and Indels to **In**sertion and **del**etions that can be of one or several nucleotides long. When multiple SNPs associated to Indels or not are next to each other on the genome, some variant callers (Freebayes, bcftools) can refer them to as MNPs for "Multiple nucleotide polymorphisms". MNPs are important to consider for protein annotation because the effect of multiple combined variant positions is different from the annotation obtained if considering them separately. However, one should only use this for annotation purposes as downstream programs are rarely compatible with this format. It is possible to either parametrize the caller to split MNPs as one nucleotide long variants (*-no-mnps* options in Freebayes for example) or to convert MNPs-containing files to SNP and Indels files (the vcfallelicprimitives function of Freebayes, *norm -a* option in vcftools). Even if the focus of SARS-CoV-2 genetics has largely been on SNPs, the SARS-CoV-2 alpha VoC has drawn back attention to Indels as it carried the spike 'PCR-dropout' H69/V70 deletion. Very generally, SARS-CoV-2 studies should not overlook indels, even if their analysis is less straightforward than for SNPs.

Allele frequency

For most research purposes where the goal is to obtain a consensus sequence, a phylogeny or a lineage assignment, minor alleles will not be used per se. It therefore makes sense to consider an SNP where the vast majority of reads support a non-reference allele, relative to the reference allele, and to consider as missing data any heterozygous allele (i.e., SNP with intermediate allele frequency). In the special case of studies focused on co-infection and/or intra-host genetic diversity, multiple alleles at one sole site provide usable information. In that case, the filter on minor allele frequency must be set with great care, taking into account the sequencing depth distribution along the genome.

Depth

A minimum depth cut-off prevents the study of variants only supported by a few reads. In the case of SARS-CoV-2, a maximum depth cut-off is not generally necessary because we do not expect the presence of repetitive genomic regions, which could cause the reads from those regions to all map at the same place and hence create artefactual SNPs. However, it is

always good practice to check evenness of read depth over the entire genome either with ad-hoc statistics or with user-interface tools such as Tablet (Milne et al. 2012).

Quality

In VCF files produced by most variant callers, the quality is given by a phred score which is logarithmically linked to the probability of the variant being an error. A quality of 20 (the commonly used cut-off) means that there is 99% probability of the allele being variant, 30 for 99.9% and so on. This quality is computed using multiple input metrics including mapping quality and allele frequency at the locus.

Known problematic sites

In addition to all of the above, some SARS-CoV-2 loci have been identified as being particularly prone to errors, and are frequently masked whatever their quality (see below, section "using third party consensus sequences"). If the decision is taken not to mask them, one should at least consider checking those sites meticulously.

When to assign Ns?

"Ns" should be used for uncovered and low-quality loci. Low quality loci should not be attributed the reference allele: in fact one should keep in mind that one should be as stringent for using the reference allele than for using an alternate allele at a loci. Deletions as identified by variant callers should use "-".

Missing data

Missing data often creates problems in science and sequencing makes no exception. There is a trade-off between the number of loci, the number of samples and the percentage of missing data tolerated. Missing data shouldn't just comprise reference bases not covered by any reads but all loci not meeting the quality criterion defined above and therefore attributed a N. In the case of SARS-CoV-2 genome sequencing, the cut-off for the per-sample missing data can be quite stringent (as low as a few percent, depending upon whether known problematic sites were masked) because we usually deal with high depth sequencing and a high proportion of missing data in a sample reveals very low quality probably associated with problems in the earlier steps of the process (wet lab for example).

Computing infrastructure needed

The amount of computing infrastructure required may depend on the number of samples processed and the quantity of data generated for each sample. Here, two variables are to be considered: the size of the raw data for storage and processing and the computing time needed to process the samples.

First, as a rough rule of thumb for uncompressed sequencing file sizes, 1ko of file stores ~1,000 nucleotides. By virtue, one SARS-CoV-2 genome requires ~30ko and a set of fastq reads corresponding to a mean depth of 1000 requires ~60Mo (30ko*1000*2, the factor two corresponding to the quality line of the fastq raw data file). The above proxy does not take into account reads that are filtered out and do not participate in the mean depth (duplicated reads, host reads, low quality reads etc …).

Second, the computing time mainly varies according to the quantity of raw reads, the number of samples and the number of parallel processes (i.e., number of CPU available for use) one can jointly use. The latter is the adapting variable. Most commonly used home

computers have between 4 and 8 computing threads available, while high performance computers have between 12 and 24, and computing clusters can have hundreds.

Table 1. *Disk space and computing time needed to process and store SARS-CoV-2 sequencing data*

Sequencing approach	Sequencing technology	Computing time per sample	Per sample raw data disk space needed
WGS	Illumina	to be assessed	to be assessed
WGS	Nanopore	to be assessed	to be assessed
Amplicon-based	Illumina	to be assessed	to be assessed
Amplicon-based	Nanopore	to be assessed	to be assessed
Hybrid capture-enrichment	Illumina	to be assessed	to be assessed
Hybrid capture-enrichment	Nanopore	to be assessed	to be assessed

All-in-one pipelines

Three years into the pandemic (late 2022), a vast number of all-in-one pipelines have been developed for the analysis of SARS-CoV-2 sequencing data. This makes the process less tedious and more homogeneous across sequencing entities. Before choosing to use an all-in-one pipeline, it is important to define (i) where the actual computing will be performed, (ii) whether one needs a user interface or has expertise in the use of command line tools and (iii) which pipeline is best able to handle the specificities of your study and data.

Very few options don't require command line expertise. One of them is the Galaxy infrastructure (https://galaxyproject.org/) (Community 2022). Galaxy provides a web interface for bioinformatics workflows linked to a computing facility that can be either public (with computing resources limitations, the European server for example allows 250Go of free space, see usegalaxy.eu the website for full specifications) or locally hosted and maintained by the IT team of your laboratory. Using the user interface (no coding at all is needed), one can import their data, make their own workflow comprising each of the tools needed for a complete bioinformatics pipeline, launch the analysis, and download the results (consensus sequence for example). One can also use existing workflows built by the community (all workflows dedicated to SARS-CoV-2, including the iVAR and the ARTIC pipelines, are listed at https://usegalaxy.fr/workflows/list_published?f-tags=covid-19).

Table 2. *Characteristics of all-in-one pipelines for the processing of SARS-CoV-2 sequencing data*

Pipeline	User interface	Compatible sequencing technology	Third-party computation available ?	outputs	ref	Link(s) of the actual tool
HAVoC	No	Illumina pair end	No	Lineage report, consensus sequence	https://doi.org/10.1186/s12859-021-04294-2	https://bitbucket.org/auto_cov_pipeline/havoc/src/master/
COVID-profiler	No	Illumina pair end, Illumina single end, Nanopore (?)	Yes	Consensus sequence, phylogenetic tree	https://doi.org/10.1186/s12859-022-04632-y	https://github.com/jodyphelan/covid-profiler http://genomics.lshtm.ac.uk/covid-profiler/
Artic SARS-CoV-2 protocol*	No	Nanopore	No	Consensus sequence		https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html
PipeCoV	No	Illumina pair end, amplicon and WGS	No	Lineage report, consensus sequence	https://doi.org/10.7717/peerj.13300	https://github.com/alvesrco/pipecov
ESCA pipeline	No	amplicon-based (Illumina paired-end and Ion Torrent only)	No	Consensus sequence	https://doi.org/10.2196/31536	https://github.com/cesaregruber/ESCA
ASPICov	No	capture or amplicon strategy and Illumina or Ion Torrent	No	Consensus sequence	https://doi.org/10.1371/journal.pone.0262953	https://gitlab.com/vtilloy/aspicov
DRAGEN COVID Lineage	Yes	amplicon-based	Yes	Lineage report, consensus sequence	https://doi.org/10.1101/2022.01.07.475443	https://emea.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/dragen-covid-lineage.html
iVAR*#	No	amplicon-based	No	Consensus sequence	https://doi.org/10.1186/s13059-018-1618-7	https://github.com/andersen-lab/ivar
Galaxy-hosted workflows	Yes	All£	Yes$	Various		https://usegalaxy.fr/workflows/list_published?f-tags=covid-19

*Pipelines available as Galaxy workflows; #iVAR doesn't include the mapping step os is not an "all-in-one"; £depending on the chosen pipeline; $quotas are in place on public servers, local installations aren't limited, see useful information at https://usegalaxy-eu.github.io/about.

END-VOC has received funding from the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

2.2. SARS-CoV-2 genetic variations

Sources of SARS-CoV-2 mutations

***Table 3**. Biological, technical and computational processes from which a SNP(s) may arise*

SNP origin	Should be considered guenuine ?	How to treat it	Consequence on downstream analysis	Reference
Error in the replication process	Yes			
Host induced	Yes			
Adapter sequencing	No	Adapter trimming	False SNP calls, recombination, homoplasies or heterozygous SNPs	
Primer sequenced	No	Primer trimming		
Off-target primer matching	No	Primer trimming		https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/16
Biological contamination	No	Minor allele frequency cut-off (Depending on ratio of contamination)	Recombination, homoplasies	
RNA degradation	No	Minor allele frequency cut-off (Depending on ratio of degradated reads)		https://virological.org/t/gained-stops-in-data-from-the-peter-doherty-institute-for-infection-and-immunity/486/10
Contamination from the human transcriptome		Minor allele frequency cut-off (Depending on ratio of degradated reads)		https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009175
Co-infection		Minor allele frequency cut-off (Depending on ratio of co-infecting lineages)	Heterozygozity	https://www.nature.com/articles/s41467-022-33910-9 and not clear if this is coinfection : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8962901/

Using third party consensus sequences

Most of the SARS-CoV-2 genomes available for analysis are shared as consensus sequences. Consensus sequences take up less memory and are easier and faster to reuse while also being more readily shared than raw reads. Moreover, with the global number of sequences on GISAID alone exceeding 14 million at the end of 2022, it has become unfeasible to process the raw reads of all the genomes in the context of one study. As a result most publications rely on consensus sequences shared on GISAID (Shu and McCauley 2017), on NCBI GenBank (Benson et al. 2012) and/or on the European Nucleotide Archive (Leinonen et al. 2010). Common research practice for SARS-CoV-2 has even gone one step further by often relying not only on consensus sequences generated by others but even on periodic releases of phylogenies produced by groups that specialize in SARS-CoV-2 high-throughput sequencing analysis (Turakhia et al. 2021). In any case, using third party consensus sequences comes at the cost of having no control over the approach used to convert raw reads to consensus sequences.

Vast community-based efforts including by members of the online community of virological.org has led to the identification of spurious SNP distribution patterns in consensus sequence sets. For example, some SNPs have been detected as associated to particular laboratories while some are present with high amounts of stop codons (https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473 and https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480). These observations have led to the establishment of a list of ~500 sites which might be problematic (https://github.com/W-L/ProblematicSites_SARS-CoV2/blob/master/problematic_sites_sarsCov2.vcf) and are frequently excluded from analysis of SARS-CoV-2 that rely on global datasets shared as consensus sequences. Similarly, tools aiming at quality-checking SARS-CoV-2 consensus sequences have been developed and released (Turakhia et al. 2020) and https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/16).

2.3. Metadata

Genomic sequence data without associated information such as collection data and place (metadata) is largely useless, and the richer the metadata, the higher the (re-)usability potential of the genomic sequence data. Despite being of crucial importance to most SARS-CoV-2 research studies, metadata is often overlooked. Crucial metadata is often incomplete or even lacking for submissions on publicly available sequence repositories, and GISAID is no real exception to this pattern. (Ling-Hu et al. 2022) showed that less than 5% of the sequences from the UK (as of September 2022) have accompanying data for both host age and sex. This is just an example that illustrates how the need for metadata have not been formalized, standardized and accepted to the point of being widely used by the scientific community yet (Schriml et al. 2020; Gozashti and Corbett-Detig 2021).

When one makes some genetic data publicly available, one usually has to fill in a form in which the metadata has to be entered. Whatever the platform, these forms are likely to have been well thought through and perhaps even follow a published standard as GISAID does (Griffiths et al. 2022). “Why would the metadata generally be of so poor quality then?” One could ask. In the context of SARS-CoV-2, we believe that the two major barriers to good metadata are the researcher’s good will and interests and legal restrictions linked to data privacy and anonymization. Those two points might be addressed with anticipation alone. At the time of sample collection, creating a metadata table with the fields corresponding to those of the database to which we plan to submit the processed data at a later stage will inevitably make us face legal problems and/or find the relevant information before it is lost (researcher’s end of contract, patient left, lost document …). This will make the submitting process smoother and less time consuming.

2.4. Additional information

Sequencing approaches

Three main sequencing approaches are commonly used for SARS-CoV-2 genome reconstruction: shotgun metatranscriptomics (WGS), hybrid capture-enrichment and amplicon sequencing, the latter being the most widely used (a quick overview of SARS-CoV-2 amplicon sequencing is provided at https://artic.network/quick-guide-to-tiling-amplicon-sequencing-bioinformatics.html and (Kubik et al. 2021)). Parameters dictating the choice of

an approach include cost-effectiveness, viral load, the need for timely results and of course the expected outputs of the study. (Chiara et al. 2020) provide valuable characteristics of each approach that help choosing an approach that suits one's needs. Choosing the right sequencing technology adds up to this complexity (see Table 5 of the WHO report https://www.who.int/publications/i/item/9789240018440 for characteristics of the mainly used technologies), although Illumina and Nanopore are mostly used (Bull et al. 2020; Tshiabuila et al. 2022).

Effect of viral load

Viral load directly determines the number of distinct viral genomes sequenced. Low viral loads (ct value > ~25) affect most subsequent analyses. For example, a low quantity of RNA needs higher number of amplification cycles which might induce artefactual SNP (Heguy et al. 2022). Additionally, low viral load increases the risk of having dimer formation between the primers used in the case of amplicon-based sequencing, which in turn creates bias is sequencing depth across amplicons (Itokawa et al. 2020).

Moreover, all sequencing methods do not perform equally especially when dealing with low viral loads (Charre et al. 2020; Lam et al. 2021; Liu et al. 2021). At last, low viral loads might not be compatible with some expected results. For example, using amplicon-based sequencing, 1000 viral genomes are necessary to confidently detect minor alleles with a frequency >= 10% (Kubik et al. 2021).

3. Conclusions and future steps

The present document is intended to be used by the partners as a support guiding their bioinformatics practices without any formal obligation to follow it. It will be shared among all partners and will be navigated back and forth in an interactive manner so that the support is improved at each step to better fit the needs of all. This is the first version.

4. References

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EWJNar. 2012. GenBank. **41**: D36-D42.

Bull RA, Adikari TN, Ferguson JM, Hammond JM, Stevanovski I, Beukers AG, Naing Z, Yeang M, Verich A, Gamaarachchi H et al. 2020. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nature Communications* **11**: 6272.

Carriço JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. 2018. A primer on microbial bioinformatics for nonbioinformaticians. *Clinical Microbiology and Infection* **24**: 342-349.

Charre C, Ginevra C, Sabatier M, Regue H, Destras G, Brun S, Burfin G, Scholtes C, Morfin F, Valette M et al. 2020. Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evolution* **6**.

Chiara M, D'Erchia AM, Gissi C, Manzari C, Parisi A, Resta N, Zambelli F, Picardi E, Pavesi G, Horner DS et al. 2020. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. *Briefings in Bioinformatics* **22**: 616-630.

Community TG. 2022. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research* **50**: W345-W351.

Gozashti L, Corbett-Detig R. 2021. Shortcomings of SARS-CoV-2 genomic metadata. *BMC Research Notes* **14**: 189.

Griffiths EJ, Timme RE, Mendes CI, Page AJ, Alikhan N-F, Fornika D, Maguire F, Campos J, Park D, Olawoye IB et al. 2022. Future-proofing and maximizing the utility of metadata: The PHA4GE SARS-CoV-2 contextual data specification package. *GigaScience* **11**.

Heguy A, Dimartino D, Marier C, Zappile P, Guzman E, Duerr R, Wang G, Plitnick J, Russell A, Lamson DM et al. 2022. Amplification Artifact in SARS-CoV-2 Omicron Sequences Carrying P681R Mutation, New York, USA. *Emerging infectious diseases* **28**: 881-883.

Hu T, Li J, Zhou H, Li C, Holmes EC, Shi W. 2021. Bioinformatics resources for SARS-CoV-2 discovery and surveillance. *Briefings in Bioinformatics* **22**: 631-641.

Islam R, Raju RS, Tasnim N, Shihab IH, Bhuiyan MA, Araf Y, Islam T. 2021. Choice of assemblers has a critical impact on de novo assembly of SARS-CoV-2 genome and characterizing variants. *Brief Bioinform* **22**.

Itokawa K, Sekizuka T, Hashino M, Tanaka R, Kuroda M. 2020. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. *PLOS ONE* **15**: e0239403.

Kubik S, Marques AC, Xing X, Silvery J, Bertelli C, De Maio F, Pournaras S, Burr T, Duffourd Y, Siemens H et al. 2021. Recommendations for accurate genotyping of SARS-CoV-2 using amplicon-based sequencing of clinical samples. *Clinical Microbiology and Infection* **27**: 1036.e1031-1036.e1038.

Lam C, Gray K, Gall M, Sadsad R, Arnott A, Johnson-Mackinnon J, Fong W, Basile K, Kok J, Dwyer DE et al. 2021. SARS-CoV-2 Genome Sequencing Methods Differ in Their Abilities To Detect Variants from Low-Viral-Load Samples. **59**: e01046-01021.

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson RJNar. 2010. The European nucleotide archive. **39**: D28-D31.

Ling-Hu T, Rios-Guzman E, Lorenzo-Redondo R, Ozer EA, Hultquist JF. 2022. Challenges and Opportunities for Global Genomic Surveillance Strategies in the COVID-19 Era. **14**: 2532.

Liu T, Chen Z, Chen W, Chen X, Hosseini M, Yang Z, Li J, Ho D, Turay D, Gheorghe CP et al. 2021. A benchmarking study of SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples. *iScience* **24**: 102892.

Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D. 2012. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* **14**: 193-202.

Pereira R, Oliveira J, Sousa M. 2020. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. **9**: 132.

Schriml LM, Chuvochina M, Davies N, Eloe-Fadrosh EA, Finn RD, Hugenholtz P, Hunter CI, Hurwitz BL, Kyrpides NC, Meyer F et al. 2020. COVID-19 pandemic reveals the peril of ignoring metadata standards. *Scientific Data* **7**: 188.

Sehn JK. 2015. Chapter 9 - Insertions and Deletions (Indels). In *Clinical Genomics*, doi:https://doi.org/10.1016/B978-0-12-404748-8.00009-5 (ed. S Kulkarni, J Pfeifer), pp. 129-150. Academic Press, Boston.

Shu Y, McCauley JJE. 2017. GISAID: Global initiative on sharing all influenza data–from vision to reality. **22**: 30494.

Tshiabuila D, Giandhari J, Pillay S, Ramphal U, Ramphal Y, Maharaj A, Anyaneji UJ, Naidoo Y, Tegally H, San EJ et al. 2022. Comparison of SARS-CoV-2 sequencing using the ONT GridION and the Illumina MiSeq. *BMC Genomics* **23**: 319.

Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, Hinrichs AS, Fernandes JD, Borges R, Slodkowicz G et al. 2020. Stability of SARS-CoV-2 phylogenies. *PLOS Genetics* **16**: e1009175.

Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, Haussler D, Corbett-Detig RJNG. 2021. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. **53**: 809-816.

Wegner F, Roloff T, Huber M, Cordey S, Ramette A, Gerth Y, Bertelli C, Stange M, Seth-Smith HMB, Mari A et al. 2022. External Quality Assessment of SARS-CoV-2 Sequencing: an ESGMD-SSM Pilot Trial across 15 European Laboratories. **60**: e01698-01621.