TASK3

High-throughput sequencing for COVID-19 pandemic**

Sample collection*

reachable The majority of SARS-CoV-2 sequence data comes from high-viral-load clinical diagnostic samples, allowing enough RNA to be collected for viral genome sequencing and reconstruction. According to the World Health Organization, COVID-19 can be detected using a variety of clinical specimens, the majority of which come from the upper or lower respiratory tract (WHO). Lower respiratory tract specimens, according to some studies, may have a higher viral load than upper respiratory tract specimens. However, the viral load varies between respiratory districts and respiratory groups during the infection phaseand non-respiratory tissues. SARS-CoV-2 genome assemblies have also been monitored using non-respiratory clinical specimens such as urine and faeces. They have never been made from blood or serum to our knowledge, most likely because of the low viral loads present in these samples. Although viral genetic material can be extracted from infected cell line supernatants, viral populations grown in cell lines often acquire novel genetic variants during laboratory passage and show substantial differences in viral quasi-species composition when compared to clinical samples for both SARS-CoV-2 and SARS-CoV-1.These factors have profound implications for the study of viral evolution and the suitability of laboratory-adapted viruses in downstream applications.Environmental specimens such as wastewater, air samples, and enigmatic "environmental swabs" have yielded a limited number of complete/nearly complete SARS-CoV-2 genomes. The low viral load, which results in a shortage and poor quality of viral RNA, has a big influence on the sequencing strategy and technologies used in these cases. In epidemiological studies, SARS-CoV-2 sequencing from wastewater is becoming more common, and it can be used to genotype the most common genomic variant circulating in a given geographic area as well as monitor viral prevalence in a population.Although not exhaustive, Supplementary Table S1 lists the isolation sources for the 23 791 SARS-CoV-2 genome sequences in the NCBI virus database (on 25 September 2020). While clinical respiratory specimens are the most common, many entries have unknown or unreported isolation sources, resulting in a .lack of viral genome metadata

RNA extraction*

additional details Figure S1 depicts a standard SARS-CoV-2 RNA extraction workflow in a wet lab. For viral RNA extraction, biosafety level 2 (BSL2) laboratories are needed. To extract and purify RNA from clinical specimens, cultured isolates, or environmental samples, any of a variety of commercially available kits for complete RNA extraction or viral RNA enrichment can be used. Guanidine salt prevents nucleases from killing viral RNA, and phenol denatures and dissolves protein, essentially inactivating the virus.To improve RNA recovery, most viral RNA extraction protocols recommend adding carrier RNA, such as poly-A RNA.While carrier RNA has no effect on SARS-CoV-2 amplicon or hybrid-capture genome sequencing, it can skew metatranscriptomic methods significantly.As a result, due thought should be given to its implementation.For viral RNA extractions, it has been suggested that linear polyacrylamide be added to the lysis buffer.

DNase treatment is also recommended during or after RNA extraction, particularly for metatranscriptomic library preparation.RNA can be qualitatively analysed with the Agilent 2100 Bioanalyzer, quantified with NanoDrop spectrophotometers (ThermoFisher), and processed at 80°C until use with a high sensitivity RNA assay(RNA 6000 Pico Kit).qRT-PCR, which targets one or more viral genes(e.g., RdRp, orf1ab, E, and N)and provides Ct (threshold cycle) values for each target, can be used to assess the presence and quantity of SARS-CoV-2 RNA prior to sequencing.The viral load in the sample is inversely related to Ct values(the lower the Ct value, the higher the viral title),and each amplicon has its own .meaning

Sequencing strategies*

NGS sequencing technologies have rapidly become the go-to method in virology for a range of applications, including detecting novel viruses from metagenomic samples, reconstructing complete or nearly complete viral genome sequences, and studying viral evolution and quasispecies. Except for unknown or poorly identified viruses, one of the most significant advantages of NGS-based approaches is that full-length viral genomes can be reconstructed from culture-enriched viral preparations or directly from clinical samples. In the case of SARS-CoV-2, both second and third generation NGS technologies were successfully introduced,Several companies have developed their own library planning procedures.When choosing a sequencing method, keep in mind the project's overall objectives as well as the type of biological sample at hand. The viral load (which is often related to the sample source), the RNA extraction process, RNA accuracy, parallelization/automation criteria, and other variables must all be considered in relation to the experimental goals (investigation of inter- or intra-sample variations of the viral genome, study of the viral and host transcriptome and epitranscriptome, single cell studies, etc.).Shotgun metatranscriptomics, hybrid capture-enrichment, amplicon sequencing, and direct RNA sequencing are the four conceptually different methods that have been used so far (Table 1). In the parts that follow, we'll go through the benefits and drawbacks of both of these techniques, as well as how to apply them to various .sequencing platforms

Shotgun metatranscriptomics*

Shotgun metagenomics sequencing is a culture-independent technique that analyses all of the DNA in a sample to identify complex populations of microorganisms without knowing their genome sequences. Metagenomic sequencing is a valuable tool for detecting pathogens that were previously unknown. This approach advances clinical microbiology by presenting quantitative and accurate data on the presence of microbial species that can be used to guide therapeutic decisions.To control host gene expression, most RNA sequencing protocols use either poly(A) + RNA fraction enrichment or host rRNA depletion. SARS-CoV-2 genomes and mature transcripts may be enriched with polyadenylated poly(T) oligonucleotides. However, if the characterization and (potentially) quantification of negative-strand intermediates in coronavirus transcription and genome replication are experimental goals, such methods may be less suitable. Strand-specific RNA-seq libraries .should be considered in such cases

Amplicon-based sequencing*

Since non-viral reads are uncommon, amplicon sequencing is more specific, immune to low levels of RNA, and requires less sequencing than metatranscriptomic sequencing. Amplicon sequencing has some drawbacks due to its technical simplicity and low cost. Amplification across the genome can be biassed due to variations in primer efficiency or potential variants in the primer annealing regions, resulting in reduced coverage in particular genomic regions and incomplete assembly. Furthermore, since the primers are based on the SARS-CoV-2 virus's genome sequence as a guide,this approach may not identify large structural variants and can present systematic limitations in the presence of .high levels of genomic divergence

Hybrid capture-enrichment sequencing*

Hybrid capture enriches genetic material by fusing it to biotinylated probes, allowing for much lower sequencing depth than shotgun metatranscriptomics. On benchtop platforms, libraries can be sequenced (Illumina NextSeq and Miseq, Ion torrent, etc.). Hybrid capture-enrichment methods use more fragments/probes and provide more comprehensive profiling of target sequences than amplicon-based approaches. Furthermore, since it is less dependent on perfect complementarity, target region capture by hybridization is more resistant to genomic heterogeneity than PCR-amplicon generation. Although Xiao et al. discovered that hybrid capture sequencing is less sensitive than amplicon-based methods for sequencing SARS-CoV-2 genomes, and did not suggest its use for challenging samples with low viral loads, enrichment by hybridization has been effective even for samples with very low viral loads in other studies. Intra-sample variants can also be represented unbiasedly using capture-based methods. On the same study, Xiao et al. found strong concordance between allele frequency distributions estimated by shotgun metatranscriptomics and/or hybrid capture.The Illumina SARS-CoV-2 genome capture enrichment workflow is noteworthy because it contains probes for simultaneously .detecting SARS-CoV-2 and other respiratory viruses

Direct RNA sequencing*

Both of the above techniques include RNA retrotranslation and nucleic acid manipulation prior to library construction, which may result in the loss of details such as post-transcriptional modifications and accurate representation of the transcripts' stoichiometry. SMS is a recent advancement in sequencing technology that allows for the direct sequencing of single nucleic acid molecules without the need for amplification or, in certain cases (such as direct RNA sequencing by ONT), retrotranslation. SMS technologies, according to the report, produce longer reads but have a higher error rate than "traditional" NGS methods.ONT has established a direct RNA sequencing protocol that could help researchers detect post-transcriptional modifications. Because of the long reads, these technologies can also provide very precise reconstructions of single mature and precursor transcripts, as well as complex transcriptional patterns like those seen during coronavirus infection (recombination, alternative transcript maturation, rare transcriptional isoforms, etc.). Kim et al.used RNA from SARS-CoV-2-infected cultures and

SARS-CoV-2 RNA fragments formed by in vitro transcription to obtain a complete representation of the SARS-CoV-2 transcriptome and epitranscriptome.

SARS-CoV-2 transcriptome and epitranscriptome

SARS-CoV-2 transcription is a discontinuous and highly regulated mechanism in which a template switch during the synthesis of subgenomic negative-strand RNA adds a copy of the leader sequence, according to current large-scale SARS-CoV-2 transcriptome investigations, which are mainly focused on ONT direct RNA sequencing and DNA nanoball sequencing. Individual sgmRNAs can be quantified by counting RNA-seq reads spanning template switch sites. Infected human cell lines have revealed hierarchies of viral and host gene expression over time that appear to be linked to innate antiviral responses using bulk and single cell RNA-seq results.

Data analysis, deposition and access**

Guidelines for the generation of SARS-CoV-2 genome assemblies*

Since the SARS-CoV-2 genome is small and lacks long repetitive sequences, it is relatively simple to put together. If the sequencing reaction results provide a complete and accurate representation of the genome, any state-of-the-art method for NGS data assembly—based on Overlap Layout Consensus, de Bruijn graphs, or, in general, reference-based assembly for an up-to-date review—should be capable of producing highly contiguous and accurate assemblies. Since 30x theoretical coverage of the genome is widely considered adequate for high-quality assembly, SARS-CoV-2 genomes should be tractable with as little as a Megabase of sequencing data. Depending on the sequencing platform and, more importantly, the sequencing strategy, different factors can apply.

Currently available resources and guidelines for data deposition*

The GISAID EpiCov portal is currently the most widely used source of SARS-CoV-2 genomic data. In the database, there are over 100 000 SARS-CoV-2 genomes from over 80 countries (data collected on 25 September 2020). Any viral genome is associated with limited metadata such as sample form, sequencing technology, and sequencing protocols, and a subset of 5000 genomes has comprehensive clinical annotations such as patient status (e.g. hospitalised or released). Sex and age are two examples of patient data that are not regularly obtained.Users must register and consent not to redistribute EpiCov data to third parties; data use is restricted to research purposes; raw sequencing data cannot be deposited; and programmatic access is not accessible.As a result, we applaud the Research Data Alliance's recommendation to send SARS-CoV-2 genomes and sequencing data to FAIR-compliant repositories rather than GISAID. Virus sequence data, both raw and compiled, should be archived in one of the INSDC servers. Gene expression data should be analysed with ArrayExpress or Gene Expression Omnibus, while genome association data should be analysed with EGA and GWAS Catalog.We stress the importance of handling human genetic data in compliance with applicable laws and regulations, and making it available where necessary through dedicated secure repositories such as EGA and dbGAP.

It's also worth noting that strict adherence to acceptable metadata specifications is crucial .for maximising dataset usefulness and future reusability for all forms of omics data

Development and reporting of computational methods*

Bioinformatics analyses and workflows are needed in modern biology, as is the documentation and availability of raw data and metadata. As a result, we strongly urge the public release of all COVID-19 data analysis methods and workflows through dedicated infrastructures and repositories. In this regard, the collection of best practises and principles outlined in is an excellent guideline for software developers and bioinformaticians working on COVID-19 data software creation and implementation. On the other hand, these considerations apply to all clinical microbiology outcome studies.Carefully curated bioinformatics tools and application catalogues are useful resources for learning about and promoting emerging bioinformatics approaches. Workflow administrators like the Galaxy system and the Microreact portal will aid collaborative data analysis and the creation of standard operating protocols and pipelines.Finally, depositing software tools and methods in specialised repositories, such as the OpenAIRE COVID-19 gateway, which are explicitly designed for the COVID-19 community, would greatly improve debate within the COVID-19 bioinformatics .community, promoting the creation of new software and methods

Secondary analysis of the data and specialized repositories*

Despite significant limitations in the type and scale of data shared at various levels by the SARS-CoV-2 research community, as well as the need for more robust and systematic sharing of primary data, a number of dedicated computational infrastructures have been built to make COVID-19 omics data more accessible and retrievable. By allowing the smooth integration of different types of data, these platforms have greatly facilitated the execution of complex meta-analyses, such as the monitoring of adaptive evolution in the genome of SARS-CoV-2 and fine-grained control of the prevalence of different viral strains in different geographic regions.One of the most notable examples is the Korber et almethod for detecting emerging mutations in the S protein of SARS-CoV-2. By tracking the prevalence of different missense substitutions in the S protein of SARS-CoV-2, the authors discovered a systematic increase in the prevalence of a specific amino acid substitution, .D614G, at the regional level in different geographic locations

Data integration and exploratory analyses of currently available data*

Exploratory analyses of currently accessible genomic sequences from three of the most commonly used methods for SARS-CoV-2 genome data: COG-UK, GISAID EpiCoV, and the NCBI virus portal, reveal database differences. The COG-UK database is responsible for 22 599 of the over 100 000 genomes currently available in GISAID EpiCoV, according to strain identifiers and open metadata. These assemblies do not cover the COG-UK database, which currently comprises over 48 000 sequences.. Similarly, only about 10% (1695 out of 17 106) of the genomic assemblies in the NCBI virus database can be linked to sequences deposited at GISAID EpiCoV directly or indirectly (via strain identifiers or BioSample

metadata). Determining the levels of overlap between data held in various repositories is .also difficult at the moment

Conclusions**

The need to identify and track emerging diseases that could lead to pandemics, as well as increase and sustain investment in preparedness and health capacity, has received a lot of attention in recent decades. As the recent outbreaks of SARS, MERS, Zika, and Ebola have shown, ultra-rapid and cost-effective methods for reconstructing the genomic sequences of emerging pathogens are useful tools for monitoring and combating the spread of novel human infectious diseases.NGS techniques were easily adapted to the SARS-CoV-2 model and shown to be applicable to a wide range of biological questions. The rate at which data is produced and analysed has never been seen before, and just a few years ago, it would have been unimaginable.In just a few months, genome sequence data enabled researchers to reconstruct the likely time of SARS-CoV-2 spillover into the human population, develop viral strain classification systems that were critical for monitoring the virus's spread, and identify sites in the SARS-CoV-2 genome that could be influenced by various selective toxins.High-throughput transcriptomics has revealed new mechanistic insights into SARS-CoV-2 gene expression, the stoichiometry of their gene products, and potential molecular mechanisms of gene viral gene expression regulation, including post-transcriptional modifications. Several researchers have found genetic variants in the SARS-CoV-2 genome that could be related to virulence or human host adaptation.Integration of host and virus genome-wide variant data, ideally with other clinical, demographic, and social parameters, may provide mechanistic hints as well as improve clinical outcome prediction. However, association studies must involve a large number of people in order to obtain the necessary statistical power. Despite a few notable campaigns, few large-scale interaction studies on COVID-19 have been presented to date.In this paper, we've tried to provide a quick overview of the relative merits and implementations of various sequencing strategies and platforms for SARS-CoV-2-related applications, with a focus on things to think about when setting up an experimental pipeline.SARS-CoV-2 sequence data is now available in a number of databases and resources. However, for full utility, relevant raw data and metadata are needed (which must be as comprehensive as possible, given in standard formats, and preferably accessible via FAIR compliant databases). Importantly, highly curated resources for secondary data analysis and the integration of various types of metadata are already available, making complex meta-analyses and/or retrospective cross-sectional studies far easier to carry out.It's difficult to completely exploit the recent avalanche of COVID-19-related sequence data. To fully realise the gains in data processing efficiency already made, a similarly unparalleled standardised implementation of data standards would be needed. In the best of circumstances, open science and humanity's progress involve the availability and inclusion of (in many cases) publicly funded data, but .time is running out. Winter is approaching

Table 1
Characteristics of SARS-CoV-2 sequencing approaches

	Shotgun metatranscriptomics	Amplicon-based	Hybrid capture-enrichment	Direct RNA sequencing[a]
Goals	SARS-CoV-2, host microbiota, and host response to infection	SARS-CoV-2 genome	SARS-CoV-2 genome	SARS-CoV-2 and host transcriptome and epitranscriptome
Co-infection detection	Yes	No	No/yes (depending on gene panel)	Yes
Minimum number of reads	20–50 M	5–20 M	5–20 M	0.5 M
Genome Coverage	≥99%	≥95–99%	≥95–99%	≥99%
Accuracy in SNV identification	High	High	Moderate	Low
Sample viral load (Ct) requested (ref Xiao)	<24–28	≥24–28	≥24–28	<24–28
Sample RNA input (ng)	10–200	1–50	10–50	≥1000
Sample type	Patient specimens	Patient specimens, environmental samples	Patient specimens, environmental samples	Viral cell cultures
Cost	High	Low	Moderate	High
NGS sequencing platforms	High- or ultra high-throughput platforms	Mid-throughput platforms	Mid- or high-throughput platforms	ONT

[a]Only 1 dataset from direct RNA sequencing is currently available in public repositories (Kim *et al.* [95])