Briefings in Bioinformatics, 22(2), 2021, 616–630

doi: 10.1093/bib/bbaa297
Advance Access Publication Date: 7 December 2020
Method Review

Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities

Matteo Chiara, Anna Maria D'Erchia, Carmela Gissi, Caterina Manzari, Antonio Parisi, Nicoletta Resta, Federico Zambelli, Ernesto Picardi, Giulio Pavesi, David S. Horner and Graziano Pesole

Corresponding author: Graziano Pesole, Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari "A. Moro" and Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Via Orabona 4, Bari 70126, Italy. E-mail: graziano.pesole@uniba.it

Abstract

Various next generation sequencing (NGS) based strategies have been successfully used in the recent past for tracing origins and understanding the evolution of infectious agents, investigating the spread and transmission chains of outbreaks, as well as facilitating the development of effective and rapid molecular diagnostic tests and contributing to the hunt for treatments and vaccines. The ongoing COVID-19 pandemic poses one of the greatest global threats in modern history and has already caused severe social and economic costs. The development of efficient and rapid sequencing methods to reconstruct the genomic sequence of SARS-CoV-2, the etiological agent of COVID-19, has been fundamental for the design of diagnostic molecular tests and to devise effective measures and strategies to mitigate the diffusion of the pandemic. Diverse approaches and sequencing methods can, as testified by the number of available sequences, be applied to SARS-CoV-2

Matteo Chiara is Assistant Professor in molecular biology and bioinformatics at the University of Milan. His research interests include comparative genomics and the development of bioinformatics methods for the analysis of NGS data.

Anna Maria D'Erchia is Associate Professor in molecular biology at the University of Bari and research associate at the Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies of the National Research Council in Bari. Her research focuses on comparative and functional genomics for the study of gene expression in normal and pathological tissues.

Carmela Gissi is Associate Professor in molecular biology at the University of Bari and research associate at the Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies of the National Research Council in Bari. Her research focuses on comparative and evolutionary genomics.

Caterina Manzari is Research Assistant at the Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies of the National Research Council in Bari. Her research focuses on the development of NGS experimental workflows for the study of molecular biodiversity.

Antonio Parisi heads the Genetic and Molecular Epidemiology Laboratory at the Experimental Zooprophylactic Institute of Apulia and Basilicata. His research focuses on the epidemiology of animal diseases and food-borne zoonosis.

Nicoletta Resta is Associate Professor of Medical Genetics at the University of Bari. She heads the Laboratory Unit of Medical Genetics and the School of Specialization in Medical Genetics. Her research focuses on the pathogenetic mechanisms responsible of hereditary cancer predisposition syndromes, and on the congenital disorders caused by mutations in mTOR/Akt/Pten/PI3KCA pathway genes

Federico Zambelli is Assistant Professor in molecular biology and bioinformatics at the University of Milan. His research interests are mainly focused on analysis of sequencing data for the characterization of gene expression and the underlying regulatory mechanisms.

Ernesto Picardi is Associate Professor in molecular biology and bioinformatics at the University of Bari and research associate at the Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies of the National Research Council in Bari. His research focuses on bioinformatics methods for the analysis of NGS data with emphasis on computational strategies to detect RNA editing events in eukaryotic genomes.

Giulio Pavesi is Associate Professor of bioinformatics at the University of Milan (Italy). His research interests are mainly focused on bioinformatics in general, and the development and application of bioinformatic analysis methods and workflows for the characterization of gene expression and its regulation at the genetic and epigenetic level.

David S. Horner is Associate Professor in molecular biology and bioinformatics at the University of Milan. His research interests include comparative genomics and development of bioinformatics methods for the analysis of NGS data.

Graziano Pesole is Full Professor in molecular biology at the University of Bari and Research Associate at the Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies of the National Research Council in Bari. His research focuses on comparative genomics and bioinformatics methodologies for the analysis of NGS data.

Submitted: 8 August 2020; **Received (in revised form):** 27 September 2020

genomes. However, each technology and sequencing approach has its own advantages and limitations. In the current review, we will provide a brief, but hopefully comprehensive, account of currently available platforms and methodological approaches for the sequencing of SARS-CoV-2 genomes. We also present an outline of current repositories and databases that provide access to SARS-CoV-2 genomic data and associated metadata. Finally, we offer general advice and guidelines for the appropriate sharing and deposition of SARS-CoV-2 data and metadata, and suggest that more efficient and standardized integration of current and future SARS-CoV-2-related data would greatly facilitate the struggle against this new pathogen. We hope that our '*vademecum*' for the production and handling of SARS-CoV-2-related sequencing data, will contribute to this objective.

Key words: COVID-19; SARS-CoV-2; omics data; sequencing technologies; data integration; data deposition

Introduction

In January 2020, a novel betacoronavirus, subsequently designated SARS-CoV-2, was identified as the etiological agent of a cluster of pneumonia cases in Wuhan City, Hubei Province, China [1–4]. COVID-19 (coronavirus disease 2019), the disease caused by the infection of this novel pathogen, spread rapidly and on the 11 March 2020, with 118 000 cases reported from 110 countries, the World Health Organization (WHO) declared a pandemic [5]. At the time of writing (25 September 2020), COVID-19 has affected more than 200 countries worldwide, with more than 33 Million confirmed individual infections and a death toll of about 1 Million, posing the greatest global health and socioeconomic threat since World War II [6]. SARS-CoV-2 is primarily transmitted between humans through respiratory droplets and physical contact [7], although some airborne transmission seems probable [8]. The incubation period ranges between 2 and 14 days, but longer intervals have been reported [9]. Fever, dry-cough and general fatigue are the most common symptoms. Less common symptoms include muscle pain, nasal congestion, runny nose, sore throat and diarrhea [10, 11]. A minority of patients develop pneumonia, severe acute respiratory syndrome and/or kidney failure [12, 13]. Estimated fatality rates vary greatly between countries [14], probably due to differences in testing strategies, demographic factors [15, 16], background comorbidities and other factors. While the pandemic has prompted an unprecedented global effort to find therapeutic targets and develop treatments and vaccines [17, 18], to date, decisive remedies are lacking.

The first complete genomic sequences of the novel betacoronavirus were obtained in late December 2019 through metatranscriptomics approaches, supplemented by PCR and Sanger sequencing [2–4]. The availability of a reference genome assembly facilitated the development of diagnostic tests based on real time PCR [19].

SARS-CoV-2 falls into the severe acute respiratory syndrome-related coronavirus (SARSr-CoV) group defined by the International Committee on Taxonomy of Viruses (ICTV) [20]. Along with numerous isolates from bats and other mammals, the SARSr-CoV group contains SARS-CoV-1, the causal agent of a large epidemic of viral pneumonia (Severe Acute Respiratory Syndrome, SARS) that affected China and 25 other countries in 2003 and 2004 [21]. Phylogenetic analyses demonstrate that SARS-CoV-1 and SARS-CoV-2 are relatively distantly related, and that their spill-over into humans were distinct events [22]. The positive sense RNA genome of SARS-CoV-2 is approximately 30 000 nt long, and shows the highest levels of genome identity (96%) with a SARSr-CoV (denoted RaTG13) isolated from a bat in the Yunnan province of China [2]. The recent isolation of SARSr-CoVs closely related to SARS-CoV-2 (genome identity 91%) from Malayan pangolins, illegally imported into China, indicates that many similar coronaviruses circulate among mammals [23, 24]. Indeed, various studies have suggested 'intermediate' hosts in the zoonotic process [25], although the exact chain of events that allowed SARS-CoV-2 to acquire the molecular features required for human to human transmission remains unclear [26]. Further environmental sampling and meta-transcriptomic sequencing will be required to conclusively resolve these issues.

The arrangement of the SARS-CoV-2 genome is not atypical. The replicase gene, which consists of two long, overlapping open reading frames, ORF1a and ORF1b [27, 28] occupies the two thirds of the genome at the 5′ end. ORF1a is translated to polyprotein 1a (pp1a), while the polyprotein 1b (pp1ab) is generated by −1 ribosomal frameshifting [29]. These polyproteins are subsequently processed into 16 nonstructural proteins (nsps), required for viral genome replication and transcription (Figure 1A). The 3′ terminal end of the genome encodes four structural proteins required for the assembly of the viral capsid, and six other accessory proteins which are less well characterized and are not universally conserved among coronaviruses (Figure 1B).

These genes are transcribed through a complex mechanism of discontinuous transcription that generates a set of nested sub-genomic transcripts, called sub-genomic mRNAs (sgmRNAs). Antisense RNAs whose synthesis is prematurely terminated at specific transcription regulatory sequences (TRSs) upstream of each of the accessory genes are directed to continue synthesis of the complement of the 67–72 nt 'leader' at the extreme 5′ end of the positive sense genomic RNA. Transcription of these negative sense sgmRNAs results in positive sense sgmRNAs which are 5′ and 3′ coterminal with the genome sequence. Discontinuous transcription is mediated by sequence identity between a donor RNA (body, TRS-B) and hairpin structures present in the acceptor RNA (leader, Transcription Regulatory Leader Sequence [TRS-L]) and is probably modulated by long-distance RNA–RNA interactions (see also Figure 1B). For a complete review of coronavirus replication and transcription mechanisms, we refer readers to [27, 28].

Recent experience with emerging infectious diseases, such as SARS, MERS, Zika and Ebola has demonstrated that NGS technologies represent powerful tools for tracing origins, spread and transmission chains of outbreaks, as well as for monitoring the evolution of the etiological agents [30–34]. Accordingly, the COVID-19 pandemic has triggered unprecedented efforts for the development of effective real-time surveillance strategies based on sequencing of the genome of its causative agent [35–40] with more than 100 000 complete or near complete SARS-CoV-2 having been deposited in dedicated repositories such as EpiCov [41] and others [35, 42]. These data have already fostered several studies on the evolutionary dynamics of the virus, and the identification of variants of potential clinical relevance [43–45].

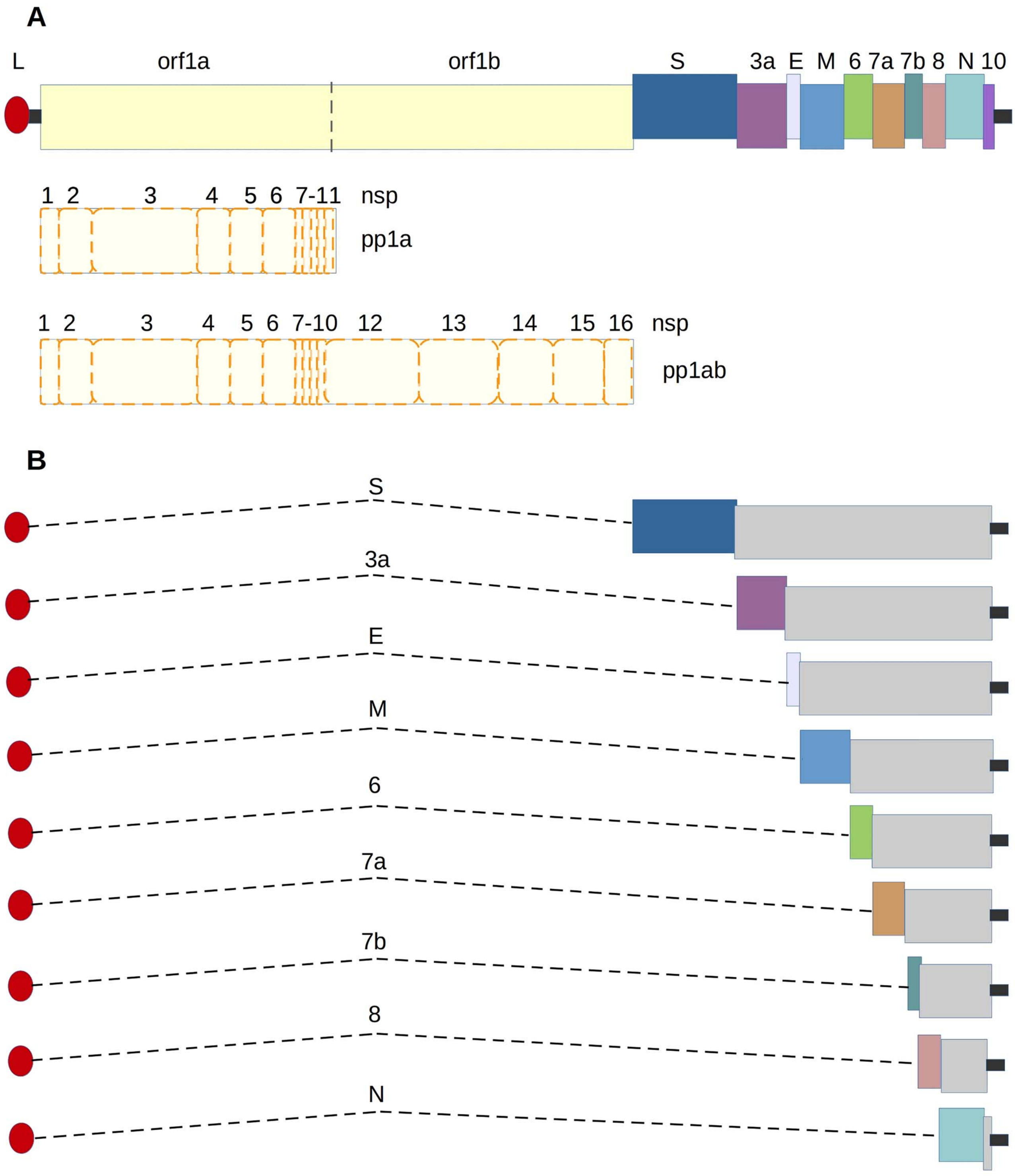

Figure 1. Architecture of the genome of SARS-CoV-2. (**A**) SARS-CoV-2 genome structure. Labels indicate gene names. The red circle indicates the TRS-L. The lower panel depicts the nsps derived from processing of the pp1a and pp1ab polyproteins. (**B**) sgmRNAs. Dotted lines are used to link the TRS-L with the body of each individual sgmRNA. The specific gene product, obtained from each individual sgmRNA is indicated by the colored boxes and the corresponding labels.

A critical need for consistent handling, labeling and deposition of sequence data has become apparent, given our incomplete understanding of the complexity of virus replication and gene expression, the possibility of RNA modifications of either RNA strand during replication or transcription, and, not least, to facilitate access to coherent and relevant metadata. These challenges can only be addressed through shared and coordinated efforts [46]. While data standards represent a recurring theme in the 'omics' era [47–49], in the case of SARS-CoV-2 the need to guarantee straightforward, unrestricted and rapid access to large volumes of processed and, in many cases, raw molecular data are unprecedented.

This review provides a brief, but hopefully comprehensive summary of state of the art for NGS applications in SARS-CoV-2 genomics. Along with detailed descriptions of currently available sequencing approaches, we present an overview of the repositories and databases that provide access to SARS-CoV-2 genomic data and metadata, together with general advice for their correct sharing and deposition. By offering a clear and detailed *vademecum* for the production and handling of COVID-19-related sequencing data, and a detailed picture of the state of the art, we hope to contribute to more efficient and informative curation, integration and exploitation of SARS-CoV-2 sequencing data and metadata.

High-throughput sequencing for COVID-19 pandemic

Sample collection

Available SARS-CoV-2 sequence data derive mainly from clinical diagnostic samples, with high viral loads that permit the extraction of enough RNA for the sequencing and reconstruction of complete or nearly complete viral genomes. The WHO (Interim guidance; [50]) lists several types of clinical specimens that can be collected for laboratory diagnosis of COVID-19 [51], mostly deriving from the upper or lower respiratory tract. Some studies report that specimens from the lower respiratory tract may contain a higher viral load than those from the upper respiratory tract (see [51] and references therein). However, during the course of infection, the viral load changes dynamically between different respiratory districts as well as between respiratory and non-respiratory tissues [52–57].

SARS-CoV-2 genome assemblies have also been obtained from non-respiratory clinical specimens including urine and feces (see Supplementary Table S1 available online at https://academic.oup.com/bib). However, to our knowledge, they have not, until now, been generated from blood or serum, probably due to the low viral loads associated with these samples [58]. Viral genetic material can also be isolated from the supernatant of infected cell lines, but viral populations grown in cell lines often accumulate novel genetic variants during laboratory passage [59], and show relevant differences in the composition of viral quasi-species with respect to matched clinical samples for both SARS-CoV-2 [60] and SARS-CoV-1 [61]. These factors have profound implications for the study of viral evolution and the suitability of laboratory-adapted viruses in downstream applications.

A very limited number of complete/nearly complete SARS-CoV-2 genomes have been obtained from environmental specimens, such as wastewater, air samples and undefined 'environmental swabs.' In these cases, the choice of the sequencing strategy and technology is greatly influenced by the low viral load and the consequent scarcity and poor quality of viral RNA [62–64]. Specific protocols for the sequencing of SARS-CoV-2 from wastewater are of emerging importance for epidemiological studies [65, 66] and can be used not only as a proxy to monitor viral prevalence in a population but also for genotyping the predominant genomic variant circulating in a specific geographical area [63].

While not exhaustive, Supplementary Table S1, available online at https://academic.oup.com/bib, lists the isolation source of the 23 791 SARS-CoV-2 genome sequences available in the NCBI virus database [67] (on 25 September 2020). It is evident that clinical respiratory specimens predominate but, for many entries, the isolation source is not mentioned or insufficiently/unclearly described, underlining the widespread incompleteness of metadata associated with viral genomes (see also Data Deposition and Access).

RNA extraction

A schematic of the common wet-lab workflow used for SARS-CoV-2 RNA extraction is represented in Supplementary Figure S1, available online at https://academic.oup.com/bib. Viral RNA extraction requires biosafety level (BSL) 2 laboratories. RNA can be extracted and purified from clinical specimens, cultured isolates or environmental samples, using any of a large variety of commercially available kits for total RNA extraction or enrichment of viral RNA (see Supplementary Table S2 available online at https://academic.oup.com/bib). Standard methodologies include the usage of Guanidine salt, which inhibits nucleases, ensuring viral RNA is not degraded, and of phenol, to denature and dissolve protein, effectively inactivating the virus. Viral RNA extraction protocols usually recommend the addition of carrier RNA, such as poly-A RNA, to increase RNA recovery. While the presence of carrier RNA does not affect SARS-CoV-2 genome sequencing methods based on amplicon or hybrid-capture, it may notably bias metatranscriptomic methods (as described below). Its use should thus be carefully evaluated. Alternatively, addition of linear polyacrylamide to the lysis buffer has been proposed for viral RNA extractions [68]. During or after RNA extraction, a DNase treatment is also recommended, especially for metatranscriptomic library preparations. RNA can be qualitatively analyzed with the Agilent 2100 Bioanalyzer, using a high sensitivity RNA assay (RNA 6000 Pico Kit), quantified by NanoDrop spectrophotometers (ThermoFisher) or Qubit Fluorometer (ThermoFisher) and stored at −80°C until use. Before sequencing, the presence and quantity of SARS-CoV-2 RNA can be evaluated using qRT-PCR targeting one or more viral genes (i.e. RdRp, orf1ab, E and N [69]) providing Ct (threshold cycle) values for each target. Ct values are inversely correlated with the viral load in the sample (i.e. the lower the Ct value, the higher the viral title) and their interpretation is specific to each amplicon.

Sequencing strategies

NGS sequencing technologies have rapidly become the method of choice for various applications in virology, including the identification of novel viruses from metagenomic samples [70], the reconstruction of complete or nearly complete viral genome sequences [71], and the analysis of viral evolution and quasispecies [72] (see [73] for a recent review). One of the most relevant advantages of NGS-based approaches is that full-length viral genomes can be reconstructed even for unknown or poorly characterized viruses, starting either from culture-enriched viral preparations, or directly from clinical samples. In the case of SARS-CoV-2, both second and third generation of NGS technologies have been successfully applied, and several specific library preparation protocols have developed independently by different manufacturers [74–78].

The final objectives of the project and the type of biological sample at hand are key considerations informing the choice of the most appropriate sequencing strategy. The type of sample (e.g. clinical specimens, environmental samples, infected cultured cells), viral load (often related to the sample source),

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])

RNA extraction procedure, RNA quality, requirements for parallelization/automation and other considerations must all be reconciled with the experimental objectives (investigation of inter- or intra-sample variations of the viral genome, study of the viral and host transcriptome and epitranscriptome, single cell studies, etc.). To date, four conceptually different approaches have been applied: (i) shotgun metatranscriptomics, (ii) hybrid capture-enrichment, (iii) amplicon sequencing and (iv) direct RNA sequencing (Table 1). In the following sections, we will discuss the merits and limitations of each of these strategies and their application using different sequencing platforms.

Shotgun metatranscriptomics

Shotgun metagenomics sequencing is a culture-independent technique that can interrogate all of the DNA in a sample, allowing the characterization of complex communities of microorganisms, without any prior knowledge of their genome sequences [79]. Metagenomic sequencing is an extremely powerful tool for the identification of previously uncharacterized pathogens, see [80, 81] for a recent review. By offering detailed and quantitative information on the composition of microbial communities, this approach also provides added value in clinical microbiology where it can be used to inform therapeutic strategies.

Shotgun metatranscriptomics—saturation RNA sequencing—has been successfully applied to obtain complete or nearly complete assemblies of the genome of SARS-CoV-2 from several types of clinical samples. Since metagenomics/metatranscriptomics can also identify other viral and bacterial DNA/RNAs, these methods can also provide information regarding secondary infections, potentially informing treatment decisions and predicting patient outcomes. Moreover, since metatranscriptomics can recover host transcripts from infected epithelial and activated immune cells, this approach can provide an accurate snapshot of immune system reaction in patients, potentially informing studies of virus–host interactions [82, 83] and even facilitate limited genotyping of patients.

Most RNA sequencing protocols were originally developed to monitor host gene expression and employ either enrichment of the poly(A) + RNA fraction, or depletion of host rRNA. Full length SARS-CoV-2 genomes and mature transcripts are polyadenylated [27, 28] and can thus be enriched using poly(T) oligonucleotides. However, such approaches may be less appropriate if the characterization and (potentially), the quantification of negative-strand intermediates in coronavirus transcription and genome replication are experimental objectives. In such cases, the adoption of strand-specific RNA-seq libraries should be considered.

A typical workflow consists of RNA fragmentation, first- and second-strand cDNA synthesis, and library preparation according to the NGS technology of choice. Supplementary Table S3, available online at https://academic.oup.com/bib, reports a selection of protocols and NGS platforms that have been used for metatranscriptomic SARS-CoV-2 sequencing.

While most studies have employed the Illumina platform, the Oxford Nanopore Technology (ONT) has been also exploited for shotgun metatranscriptomics [84], through modification of a protocol designed for influenza viruses from clinical samples [85]. A sequence-independent single-primer amplification (SISPA) step [68, 86] is employed, to meet the requirement for ≥1 µg of cDNA for ONT library preparation. Notwithstanding possible biases introduced by SISPA, this approach allows rapid generation of complete SARS-CoV-2 genome assemblies, even from low amounts of RNA [84]. The Pacific Bioscience (PacBio) technology is also suitable for shotgun metatranscriptomics of SARS-Cov-2, although its use has limited to date (e.g. [23]).

The shotgun metatranscriptomics approach was employed in the discovery of SARS-CoV-2 [2–4] and is, in many senses, the method of choice for sequencing emerging SARS-CoV-2 strains. It requires no prior knowledge of the viral sequence, and avoids potential effects of divergent regions on capture and amplicon approaches. In principle, other than the viral genome, viral subgenomic RNAs (derived from discontinuous transcription), possible post-transcriptional modifications and, depending on the library preparation workflow, negative-strand intermediates can all be studied with shotgun metagenomics. Moreover, when adequate levels of genome coverage are obtained, in addition to some insight into host gene expression, this approach can provide an accurate evaluation of intra-sample virus variants, from quasispecies or coinfections, and, as previously mentioned, allow insights into host gene expression patterns during infection. The major limitation of shotgun metatranscriptomics is the requirement for a high viral load to obtain complete virus assemblies. Moreover, compared to targeted enrichment based approaches, a substantially higher sequencing depth (>2 G bases) is required. Viral load shows enormous variation in clinical specimens due to variation in sampling technique as well as from inherent differences in load between patients. The proportion of reads derived from SARS-CoV-2 can vary greatly between samples, even where viral loads (as measured by Ct values) are similar [82, 83, 87]. High coverage can be easily obtained from viral cell cultures, prepared by infecting cell cultures with viruses derived from clinical samples. However, this last approach is time consuming, labor-intensive, requires access to a BSL3 laboratory environment (https://www.cdc.gov/coronavirus/2019-ncov/lab/lab-biosafety-guidelines.html) and carries the risk of identifying variants of questionable physiological origin (see previous section on sample collection).

Amplicon-based sequencing

Amplicon sequencing enables researchers to restrict the scope of their analysis only to a limited number/type of sequences of choice. This approach is highly specific, but requires significant a priori knowledge of the sequence that is to be 'targeted.' Diagnostic RT-PCR tests for the detection of SARS-CoV-2 nucleic acids from clinical specimens, which are based on very specific primers for the amplification of discrete regions of the genome of the virus, could be considered a specialized form of amplicon sequencing. Amplicon-based approaches for the sequencing of SARS-CoV-2 adopt an enrichment workflow consisting of first-strand cDNA synthesis followed by genome amplification with multiplex PCRs. The objective is to produce pools of amplicons that cover either the entire length or the discrete portions of the viral genome (see Supplementary Table S3 available online at https://academic.oup.com/bib). Several different multiplex PCR designs, differing in the number and size of amplicons, have been proposed for SARS-CoV-2.

Amplicon sequencing is highly specific and robust to low amounts of RNA and degraded samples, and less sequencing is required with respect to the metatranscriptomic approach since non-viral reads are rare. While amplicon sequencing is theoretically convenient and cheap, it presents some limitations which should be considered. Firstly, because of differences in primer efficiency, or possible variants in the primer annealing regions, amplification across the genome can be biased, with decreased coverage in specific genomic regions (see V1 version of ARTIC protocol [88, 89]) and/or 3′ and 5′ UTRs regions missed altogether (see Supplementary Table S3 available online at https://academic.oup.com/bib) leading to an incomplete assembly. Moreover, since the primers are designed on the reference SARS-CoV-2 genome sequence, this approach may not identify large structural variants and can present systematic limitations in the presence of high levels of genomic divergence.

While the amplicon-based approach is highly dependable for the reconstruction of the most prevalent genome variant in a viral population, a recent study suggests that it provides highly biased representation of minor allele frequencies with respect to that derived from metatranscriptomics experiments performed on the same samples [87].

Several commercial kits and non-commercial protocols are available for SARS-CoV-2 amplicon preparation, some of which are tailored to particular NGS platforms (see Supplementary Table S3, available online at https://academic.oup.com/bib, in the Additional Supporting File). Since sequencing depth is a marginal consideration, libraries can be sequenced on benchtop platforms with a mid-throughput (i.e. Illumina NextSeq and Miseq; Ion torrent platforms, etc.). Additionally, when combined with the short turn-around times of Single Molecule Sequencing (SMS) technologies such as ONT and PacBio, amplicon sequencing of SARS-CoV-2 can be used for rapid surveillance of transmission chains, as exemplified by the approach adopted by the ARTIC network for real-time monitoring of the COVID-19 outbreak in the United Kingdom [35], where a fast, amplicon-based protocol successfully applied to previous viral outbreaks (see https://artic.network/ncov-2019 for a complete list of the protocols and methods) has been adapted to SARS-CoV-2. Wang *et al.* [90] established a rapid in house tiling multiplex PCR protocol for the simultaneous detection and sequencing of several respiratory viruses which includes a large part of the SARS-CoV-2 genome. The Wang protocol has also been suggested for diagnostic usage as it shows higher sensitivity than approved RT-qPCR tests [90].

While several SARS-Cov-2 genome sequencing protocols using tiled amplicons are available for the PacBio platform (see https://www.pacb.com/research-focus/microbiology/COVID-19-sequencing-tools-and-resources/), to our knowledge they have been scarcely used until now, although a major study of the introduction and spread of SARS-CoV-2 in the New York City area used both the PacBio and the Illumina technologies [37].

The robustness of amplicon-sequencing to degraded and low concentrations of RNA is evident from studies of environmental specimens, where this approach is followed by sequencing with Ion torrent [62] or ONT [63] for wastewater samples, and by Sanger sequencing for a patient breathing air sample and for a door handle swab ([64] and J.A.Lednicky, personal communication).

Hybrid capture-enrichment sequencing

Similar to amplicon-based sequencing, hybrid capture is a sequencing strategy that enables researchers to target only predefined sequences or regions of a genome that are relevant to their specific interests. Target-enrichment strategies using hybrid capture were originally developed for human genomic studies, to enable the rapid and cost-effective sequencing of the exons of protein coding genes (exome sequencing) [91]. Exome sequencing is still considered the method of choice for the study of genetic variation in protein coding loci in humans [92], as it achieves a good trade-off between the specificity of amplicon based enrichment, and the sensitivity (to different types of genetic variants) of shotgun sequencing at significantly lower costs.

Hybrid capture enriches targeted genetic material through hybridization to specific biotinylated probes, allowing a considerably reduced sequencing depth compared with shotgun metatranscriptomics. Libraries can be sequenced on benchtop platforms (Illumina NextSeq and Miseq, Ion torrent, etc.). In general hybrid capture-enrichment methods are based on a larger number of fragments/probes than amplicon-based methods (see Supplementary Table S3 available online at https://academic.oup.com/bib), and provide more complete profiling of the target sequences. Moreover, since the capture of target regions is less dependent on perfect complementarity than PCR-amplicon generation, capture by hybridization is generally more robust to genomic variability. While one hand, Xiao *et al.* [87] found that hybrid capture sequencing is less sensitive than amplicon-based methods for the sequencing of SARS-CoV-2 genomes, and did not recommend its application for challenging samples with low viral loads, in other studies enrichment by hybridization has been successful even for samples with very low viral loads [93]. Capture-based methods may also offer unbiased representation of intra-sample variants. Xiao *et al.* [87] reported high levels of concordance between allele frequency distributions estimated by shotgun metatranscriptomics and/or hybrid capture on the same sample. The SARS-CoV-2 genome capture enrichment workflow developed by Illumina is noteworthy as it includes probes for the simultaneous detection of SARS-CoV-2 and other respiratory viruses (see Supplementary Table S3, available online at https://academic.oup.com/bib, of Additional Supporting Material).

Direct RNA sequencing

The aforementioned strategies all require retrotranscription of RNA, and a greater or lesser degree of manipulation of nucleic acids prior to library construction, and can result in the loss of information, including post-transcriptional modifications and accurate representation of the stoichiometry of the transcripts. SMS is a relatively recent development in sequencing technologies, allowing the direct determination of the sequence of single nucleic acid molecules, without amplification and, in some cases (e.g. direct RNA sequencing by ONT), retrotranscription. SMS technologies usually provide longer reads than 'classic' NGS methods, but with reportedly higher error rates [94]. A direct RNA sequencing protocol setup by ONT potentially permits the detection of post transcriptional modifications (see the following section). Additionally, by virtue of the long reads, these technologies are able to provide very accurate reconstructions of single mature and precursor transcripts, and of complex transcriptional patterns, such as those taking place during coronavirus infection (recombination, alternative transcript maturation, rare transcriptional isoforms, etc.). In a recent study, Kim *et al.* [95] applied ONT direct RNA sequencing with DNA nanoball sequencing, to obtain a complete representation of the SARS-CoV-2 transcriptome and epitranscriptome (see section below), using RNA from SARS-CoV-2-infected cultures and from SARS-CoV-2 RNA fragments produced by *in vitro* transcription.

SARS-CoV-2 transcriptome and epitranscriptome

Current large-scale SARS-CoV-2 transcriptome investigations, mostly based on ONT direct RNA sequencing and DNA nanoball sequencing, have confirmed that transcription in SARS-CoV-2 is a discontinuous and highly controlled process (Figure 1B), in which a template switch during the synthesis of subgenomic negative-strand RNA adds a copy of the leader sequence [27, 28, 95]. Counting RNA-seq reads spanning template switch sites allows quantification of individual sgmRNAs [96]. Bulk and single cell RNA-seq data from infected human cell lines have revealed hierarchies of viral and host gene expression through time that appear to be linked to innate antiviral responses [96].

Epitranscriptome modifications, including transient changes such as N6-methyladenosine (m6A) and 5-methylcytosine (5mC) or non-transient changes such as RNA editing, may play relevant roles in host–virus interactions [97, 98]. ONT direct RNA sequencing of SARS-CoV-2 infected Vero cells revealed an 'AAGAA-like' motif enriched the 3′ region of the viral genome, which is strongly associated with probable post transcriptional modifications [95]. Putative post transcriptional modifications are more frequent in longer viral transcripts and are associated with shorter poly(A) tails, indicating an involvement in the control of viral RNA stability [95]. Consistent patterns of 5mC have been detected in HCoV-229E infected cells by ONT direct RNA sequencing [99]. RNA-seq and metatranscriptome sequencing of SARS-CoV-2 infected cell lines and clinical samples have shown strong signatures of A-to-I and C-to-U RNA editing, likely mediated by ADAR and APOBEC enzymes, respectively [100, 101]. Interestingly, computational analyses of RNAseq data from infected human cell lines detected A-to-I hyper-edited regions, distributed along the entire viral genome and responsible for multiple nonsynonymous changes [101].

Data analysis, deposition and access

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

Since the genome of SARS-CoV-2 is relatively compact in size, and does not contain any large repetitive sequence, the assembly of the viral genome is *per se* a relatively straightforward process. Provided that the results of the sequencing reaction offer a complete and accurate representation of the genome, any state of the art method for the assembly of NGS data—based on Overlap Layout Consensus, de Bruijn graphs or, in general on reference based assembly—see [102, 103] for an up-to date review—should be capable of producing highly contiguous and accurate assemblies. Since 30x theoretical coverage of the genome is generally considered sufficient to generate high-quality assembly, SARS-CoV-2 genomes should be tractable with as little as a Megabase of sequencing data. However, depending on the sequencing platform and most importantly on the sequencing strategy, different considerations may apply.

In principle, data obtained from targeted-enrichment-based library preparations methods, such as hybrid capture and amplicon sequencing, should be highly enriched for viral genomic reads. This notwithstanding, variable levels of 'contaminant' sequences, have been reported [104]. Moreover (see below) these strategies often generate dishomogeneous genome coverage—which can confound several assemblers. Data derived from metagenomics sequencing protocols tend to provide more uniform coverage, but variable proportions of viral reads can be obtained depending (although not linearly) on the viral load of the sample. Moreover, (see above) these data might also contain reads derived from viral subgenomic RNAs and replication intermediates.

Although highly efficient software tools for the assembly of metagenomics reads are currently available [105], in general, this process is considerably more complex and computationally intensive than the assembly of a single genome and can be confounded by several factors, including the relative abundance of different species/transcripts in the sample. For these reasons, we strongly suggest that filtering of 'non-viral' reads should be performed prior to assembly, a process that can also be beneficial in the assembly of reads derived from targeted sequencing approaches. Simple similarity filters can be applied by mapping the complete collection of reads against the reference genome assembly of SARS-CoV-2 and retaining only SARS-CoV-2-like reads. However to avoid the systematic loss of reads at polymorphic loci, relatively relaxed similarity filters should be implemented. For shotgun metagenomic libraries, prior alignment of SARS-CoV-2 like reads to the reference genome can be useful also for the identification and filtering of reads or pairs of reads derived from subgenomic mRNA, by excluding reads with discontiguous mapping or read pairs mapping at an aberrant (with respect to the insert size) distance on the genome. In a similar vein, filtration of PCR duplicates can be a useful approach to obtain a more uniform coverage profile of the genome, particularly for libraries derived from targeted enrichment. If the aim of the study is to obtain an accurate representation of the genomic sequence of a novel strain of SARS-CoV-2, *de-novo* assembly should always be preferred to reference-guided assembly methods, as this type of approach is in general more sensitive to possible (although unlikely) large-scale rearrangements events [106, 107]. However, reference-guided approaches, or approaches based on variant calling may provide a clear advantage if the objective is of the study is to obtain a fine grained representation of the viral population in a sample, including rare variants, or the study of viral quasi-species. In such cases, a vcf file reporting the occurrence and the frequency of all the genetic variants observed in a sample is probably the most relevant type of output file that should be provided/generated. In this respect, it should be noted that in the presence of co-infection by more than one viral strain, *de-novo* assembly of viral genomes based on short—second generation—NGS reads cannot provide an accurate reconstruction of the different viral haplotypes. While, by virtue of a longer read size, in principle this should be possible when long SMS sequencing reads are available.

Currently available resources and guidelines for data deposition

Particularly during the current pandemic, timely deposition of available information and straightforward access to open data are essential and enabling elements for implementing effective mitigation strategies, supporting pharmaceutical and vaccine development, and understanding the disease and its effects [46, 108]. Careful curation and deposition of SARS-CoV-2 sequencing data and associated metadata has profound implications for both epidemiological studies and in enabling extensive association studies and, in future, follow-up studies [109].

While the first half of 2020 has seen a boom in the release of COVID-19 related scientific manuscripts, questions have been raised concerning the quantity and quality of data sharing [110, 111]. However, the pandemic has also seen a renewed effort by open-data-aware scientific entities and communities towards the dissemination of best practices and recommendations for COVID-19 data sharing (e.g. [112]), analysis (e.g. [113]), and for the effective coordination of national scientific infrastructures (e.g. [46]).

At present, the GISAID [41] EpiCov portal represents the most widely used repository of SARS-CoV-2 genomic data. It provides a collection of over 100 000 complete SARS-CoV-2 genomes, isolated from over 80 countries (data collected on 25 September 2020). Limited metadata, including the type of sample, the sequencing technology and sequencing protocols are associated with each viral genome, and basic clinical annotations, i.e. the patient status (e.g. hospitalized or released), are available for a subset of ∼5000 genomes. Other potentially important patient information (e.g. gender, age) are not collected systematically. Although data in EpiCov are publicly accessible, users must register and agree not to redistribute data to third parties, data use is limited to research purposes, raw sequencing data cannot be deposited, and programmatic access is not available. For these reasons, we welcome recommendations, such as those from the Research Data Alliance [112], that, in addition to GISAID, SARS-CoV-2 genomes and sequencing data should be submitted to repositories more compliant with FAIR principles [47]. In particular, raw and processed viral sequence data should be made available in one of the International Nucleotide Sequence Database Collaboration (INSDC) [114] repositories. Gene expression data should be deposited to ArrayExpress [115] or Gene Expression Omnibus [116], while the EGA [117] and GWAS Catalog [118] should be the choice for genome association data. We underline the fact that human genetic data must always be managed in compliance with applicable laws and regulations and, where possible, made available through dedicated secure repositories such as EGA and dbGAP [119]. It should also be noted that, for all omics data types, careful adherence to relevant metadata standards is essential for maximizing the utility and future reusability of datasets [120].

Development and reporting of computational methods

As for the reporting and availability of raw data and metadata, the reproducibility of bioinformatics analyses and workflows constitutes a crucial issue in modern biology [121]. For this reason, we highly recommend that all the tools and workflows used in the analysis of COVID-19 data should be made readily available through dedicated infrastructures and repositories. In this respect, the set of best practices and principles outlined in [122] represents an excellent guideline for software developers and bioinformaticians working in the development and application of software tools for COVID-19 data. However, these considerations extend to the analyses of clinical microbiology data in general. Highly curated catalogs of bioinformatics software and applications, such as https://bio.tools/ [123], represent important resources for the discovery and advertising of novel bioinformatics methods. Moreover, the usage of well-established workflow managers, as for example those provided by the Galaxy platform [124] or the Microreact [125] portal can foster collaborative analysis of data and the development of standard operative protocols and pipelines. Finally, deposition of software tools and methods in specialized repositories, specifically developed for the COVID-19 community, for example the OpenAIRE COVID-19 gateway [126], link relevant expertise and know-how and can greatly improve the discussion within the COVID-19 bioinformatics community, further facilitating the development of new software and methods.

All in all, analogously to the situation for sharing and integration of data and metadata, a wealth of repositories and platforms are already available for sharing and integrating software tools and methods. We strongly believe that the promotion of best practice in software development and usage will be critical in the fight against COVID-19.

Secondary analysis of the data and specialized repositories

Notwithstanding relevant limitations of the type and extent of data that are shared at different levels by the SARS-CoV-2 research community [127], and the requirement for a more thorough and systematic sharing of primary data, many dedicated computational infrastructures have been established to facilitate access and retrieval of COVID-19 omics data. By allowing a

Table 2. Summary statistics of methods applied in the sequencing of SARS-CoV-2

Library preparation	Sequencing technology	Records	Notes
Amplicon	Illumina	24 311	21 142 from COG-UK (ARTIC)
	Oxford Nanopore	16 811	16 137 from COG-UK (ARTIC)
Hybrid capture	Illumina	468	
Metatranscriptomics	Illumina	1987	

Data are related from records in INSDC public databases, for which an associated genome assembly is available

smooth integration of different types of data, these platforms have greatly facilitated the execution of complex meta-analyses including the monitoring of adaptive evolution in the genome of SARS-CoV-2 and a fine grained control of the prevalence of different viral strains in different geographic regions. In this respect, the system for the identification of emerging mutations in the S protein of SARS-CoV-2 developed Korber *et al*. [44] is probably one of the most remarkable examples. In brief, by monitoring the prevalence of different missense substitutions in the S protein of SARS-CoV-2, the authors have observed a systematic increase in the prevalence of a specific amino acid substitution, D614G, at the regional level in distinct geographic locations. Retrospective analyses of viral loads, as measured by Ct values, indicated a relatively modest, but statically significant increase of viral loads (decrease in Ct) in patients infected by viruses carrying the D614G haplotype. This suggests a likely association of this variant with an increased infectivity. However, no relevant differences were observed in the severity of symptoms manifested by the patients. While a detailed discussion of the functional relevance of the D614G substitution lies outside the scope of this review (we refer readers to [45] for a more detailed discussion), we would like to underline the importance of this and similar approaches for the generation of testable biological hypothesis and the monitoring of the evolution of SARS-CoV-2. The Nextstrain [128] and the Hyphy COVID-19 [129] portals are further notable examples of highly flexible and interactive systems for the real time monitoring of the evolution SARS-CoV-2 strains. By providing real time information of the worldwide distribution of different clades and lineages of SARS-CoV-2 (Nexstrain), and detailed phylogenetic analyses of SARS-CoV-2 protein coding genes (Hyphy), these systems provide, respectively, a one shop stop for the monitoring the prevalence of SARS-CoV-2 strains worldwide and the identification of amino acid residues that are possibly under selection. In this respect, we underscore that any initiative aiming to apply well established standards and protocols for the sharing of SARS-CoV-2 genetic/genomic data, like for example the application or modification of the Beacon [130] protocol, as available from [131] should be fully supported by the SARS-CoV-2 research community. Finally, we stress the importance of developing highly curated resources and databases to allow the seamless integration of different types of data/and or the execution of complex queries, which could represent an important added value for data mining and meta-analyses, as exemplified by [132]. By allowing the seamless and rapid integration of different types of data and metadata, these and similar resources can—at least in part—mitigate some of the most important limitations for a rapid and widespread access to the COVID-19 data.

The EBI COVID-19 Data Portal (https://www.covid19dataportal.org/) and the equivalent SARS-CoV-2 resource portal at the NCBI (https://www.ncbi.nlm.nih.gov/sars-cov-2/) probably, provide the most complete catalog of resources to navigate, access and retrieve SARS-CoV-2 data from open access repositories, including bioinformatics tools and online resources. The Vipr portal [133], an integrated system that facilitates the retrieval of SARS-CoV-2 genomic sequence data and provides access to a set of sophisticated tools for the execution of detailed comparative genomic analyses. COV3D [134] is a centralized resource for spike and other coronavirus protein structures, which provides effective and yet simple tools for the visualization of protein structures, along with the annotation of relevant functional elements or genomic variants. The Galaxy Europe server [135] incorporates a highly curated collection of tools and expert-made workflows for the analysis of COVID-19 data, along with pointers to many relevant datasets.

Portals and resources for the sharing COVID-19-related knowledge are not limited to bioinformatics methods and applications, but also include sites that disseminate wet lab and sequencing protocols. The open source protocols.io portal (https://www.protocols.io/) provides access to a collection of more than 150 wet-lab and in-silico protocols, for the generation, handling and deposition of SARS-CoV-2 data in public repositories. Similar initiatives at the national level, e.g. the COVID-19 Genomics UK Consortium page (https://www.cogconsortium.uk), and COVID-19 Data Portal Sweden (https://www.covid19dataportal.se/), or made available by Research Infrastructures, such as the ELIXIR COVID-19 support page (https://elixir-europe.org/services/covid-19), provide pointers to a wealth of resources including guidelines, protocols, best practices, data analysis tools and computational platforms. Similarly, a detailed list of lab protocols, bioinformatics methods and primary repositories of SARS-CoV-2 sequencing data is also provided through publicly accessible Github repository by the US Centers for Disease Control and Prevention (https://github.com/CDCgov/SARS-CoV-2_Sequencing).

Data integration and exploratory analyses of currently available data

Although the aforementioned resources provide access to a wealth of sequencing data and metadata for SARS-CoV-2, their integration is not straightforward.

Exploratory analyses of currently available genomic sequences, as obtained from three of the most popular resources for SARS-CoV-2 genome data: COG-UK [35], GISAID EpiCoV [41] and the NCBI virus portal [67], highlight apparent inconsistencies between databases. For example, analyses of strain identifiers and available metadata suggests that, of the more than 100 000 genomes currently available in GISAID EpiCoV, 22 599 are derived from the COG-UK database. However, these assemblies do not represent the entirety of COG-UK, which currently contains over 48 000 sequences. Similarly, only about 10% (1695 out of 17 106) of the genomic assemblies contained in the NCBI virus database can be linked directly or indirectly (through strain identifiers, or BioSample metadata) to sequences also deposited at

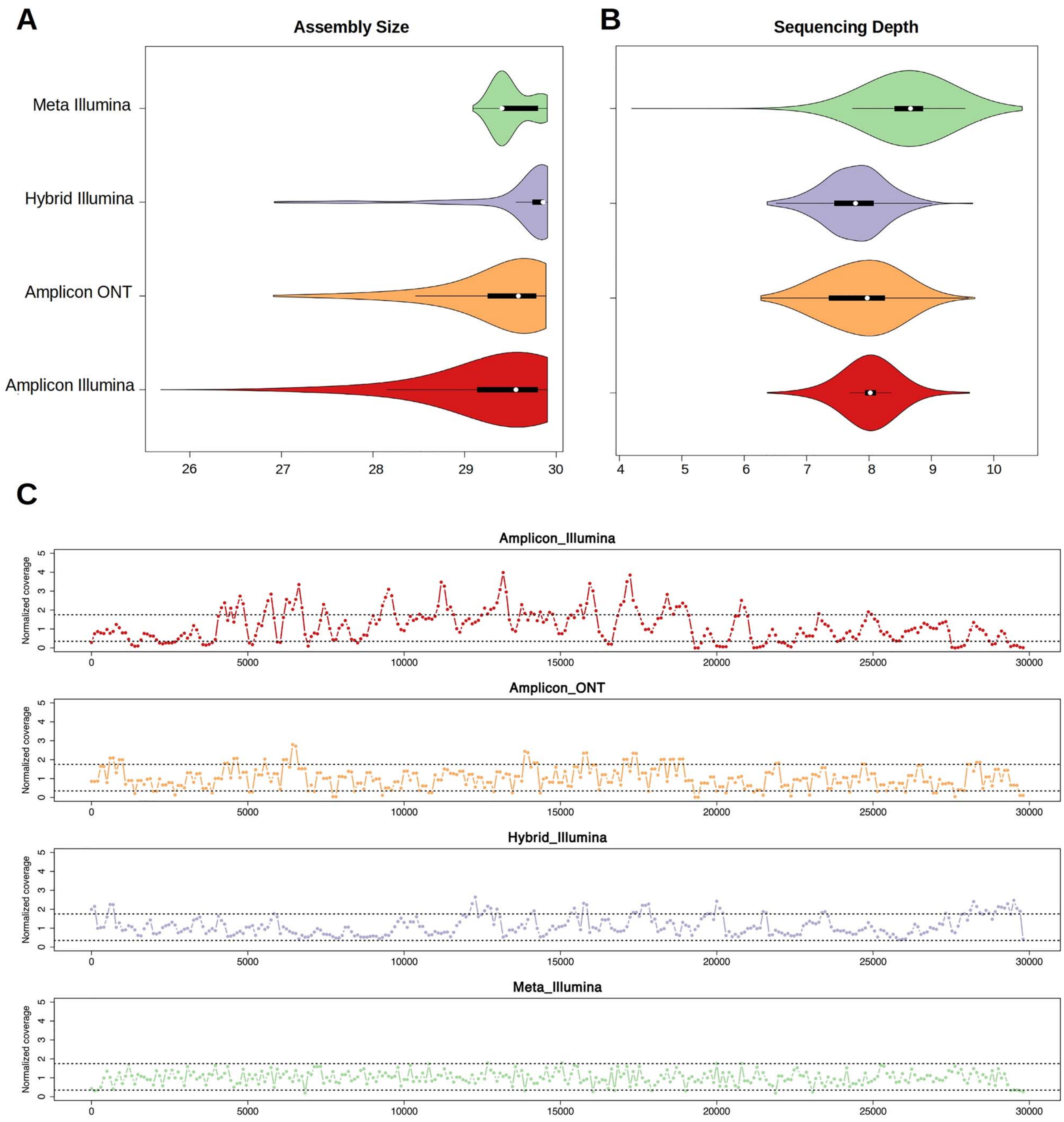

Figure 2. Overview of the properties of different approaches for SARS-CoV-2 genome sequencing. (**A**) Violin plot of the size of SARS-CoV-2 genome assemblies obtained through different sequencing approaches. Assembly size in Knt (Kilonucleotides), is reported on the x-axis. (**B**) Violin plot of the sequencing depth (log10 of the total number of sequenced bases) obtained by different sequencing approaches. (**C**) Profile of normalized coverage levels of the genome of SARS-CoV-2 as obtained from different sequencing approaches. Coverage profiles were calculated on 300 non-overlapping genomic windows of 100 nt in size. A subset of 100 distinct records as available from public repositories of raw sequencing data has been considered to estimate the coverage profile of every sequencing approach. Coverage values were normalized by using the upper quartile normalization, and averaged for every data point (genomic window).

GISAID EpiCoV. At present, even establishing the levels of overlap between data stored at different repositories is challenging.

Currently, INSDC repositories collectively provide access to more than 65 000 distinct depositions of raw sequencing data for SARS-CoV-2. Of these, 43 577 can be/are associated with a genome assembly. As outlined in Table 2, the majority of the raw sequencing data records (37 279) have been deposited by the COG-Consortium, and are the result of the application of the ARTIC amplicon protocol, combined with either Illumina (21 142) or Nanopore (16 137) sequencing. The remaining data offer a more unbiased representation of the approaches to the sequencing of SARS-CoV-2, and include metatranscriptomics libraries (1987 distinct depositions), amplicon libraries (3843) and a small number (468) of libraries based on hybrid capture protocols.

Although these data provide an incomplete representation of sequencing protocols and strategies, visualization of their respective outputs and the completeness of associated genomic assemblies offers some relevant observations. As outlined in Figure 2A, metatranscriptomics and hybrid capture approaches seem to provide—on average—more complete representations of the SARS-CoV-2 genome. For amplicon-based sequencing, ONT assemblies tend to be slightly more complete than those obtained from Illumina sequencing technologies. As shown in Figure 2B, the quantity of data generated by each sequencing approach for which raw data depositions are available, is in line with expectations, and—metatranscriptomics sequencing datasets typically contain in the order of 10x more reads than those from targeted sequencing approaches. Interestingly, metatranscriptomics libraries show a highly uniform profile of genome coverage (Figure 2C), although a considerable reduction in coverage is observed at both ends of the genome, and in particular at the 3′ UTR where 53% of assemblies are incomplete. Hybrid-capture based methods also provide relatively uniform and reproducible coverage. Finally, amplicon-based approaches provide a generally more skewed coverage of the genome, with spikes in coverage corresponding with the overlaps between different amplicons.

Conclusions

In the last decades, significant policy attention has focused on the need to identify and limit emerging outbreaks that might lead to pandemics and to expand and sustain investment to build preparedness and health capacity [136–139]. In this context, ultra-rapid and cost-effective methods for the reconstruction of the genomic sequences of emerging pathogens represent important tools for monitoring and countering the spread of novel human infectious diseases, as exemplified by recent experience with SARS, MERS, Zika and Ebola [30–34].

NGS methods have been rapidly adapted to the SARS-CoV-2 paradigm and shown to be applicable to a wide variety of associated biological questions [35, 69, 82, 88, 90, 93, 99]. The rate of data production and analysis has been unprecedented and would have been inconceivable only a few years ago.

In just a few months, genome sequence data have allowed reconstruction of the probable time of spillover of SARS-CoV-2 into the human population [140–142], the development of systems for the classification of viral strains which have been fundamental for monitoring the spread of the virus [41, 140, 142], and for the identification of sites in the genome of SARS-CoV-2 that might be under the influence of various selective pressures [129, 143]. High-throughput transcriptomics has provided novel mechanistic insights into SARS-CoV-2 gene expression, the stoichiometry of their gene products, and possible molecular mechanisms—including post transcriptional modifications—of regulation of gene viral gene expression [95, 100, 101]. Several authors have already highlighted genetic variants in the genome of SARS-CoV-2 that could possibly be linked with increased/decreased virulence or possible adaptation to human hosts [43–45].

The integration of host and virus genome-wide variant information, ideally with other clinical, demographic and social parameters might provide both mechanistic hints and add predictive value for clinical outcomes. However, substantial numbers of individuals need to be incorporated in association studies to obtain the required statistical power. Indeed, notwithstanding, some remarkable initiatives [144, 145] at present few large scale association studies on COVID-19 have been presented.

Here, we have attempted to provide a concise summary of the relative merits and applications of different sequencing strategies and platforms for SARS-CoV-2-related applications, emphasizing the considerations that should be borne in mind when establishing an experimental pipeline.

A wealth of databases and resources providing access to SARS-CoV-2 sequence data are already available. However, to maximize their utility, associated raw data and metadata (which must be as extensive as possible, presented in standard formats and ideally available through FAIR compliant databases) are critical elements. Importantly, highly curated resources for the secondary analysis of the data and the integration of different types of metadata are already available, which can greatly facilitate the execution of complex meta-analyses, and/or retrospective cross-sectional studies.

The challenge of fully exploiting this ongoing deluge of COVID-19-related sequence data lies ahead. It is clear that an equally unprecedented widespread acceptance of data standards will be required to fully capitalize on the productivity that has already been attained in data production. Availability and integration of (in many cases) publicly funded data are fundamental for open science and the progress of humanity at the best of times, but time is currently short. Winter is coming.

Key Points

- The application of ‘omics technologies’ to SARS-CoV-2 have been fundamental in epidemiological and other aspects of the fight against COVID-19.
- Different approaches, with different advantages and limitations, can be applied to the sequencing of SARS-CoV-2 genomes. Various considerations should influence the choice of approach in different clinical and research contexts.
- While more than 100 thousand complete SARS-CoV-2 genomes are currently available in public repositories, the integration of these data and of associated metadata is, at present, problematic.
- Coordinated efforts are required to promote the principles of open science and data sharing in order to facilitate more efficient and comprehensive analyses of SARS-CoV-2 data.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgments

We thank ELIXIR Italy for providing the computing and bioinformatics facilities and Laura Marra for technical assistance. We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID’s EpiFlu™ Database on which this research is based.

Funding

Italian Ministero dell’Università e Ricerca (PRIN 2017 to G.P. and A.M.D.), Consiglio Nazionale delle Ricerche, H2020 projects EOSC-Life (GA 824087), EOSC-Pillar (GA 857650) and ELIXIR Converge (GA 871075), and Elixir-IIB.

References

1. Poon LLM, Peiris M. Emergence of a novel human coronavirus threatening human health. *Nat Med* 2020;**26**:317–9.
2. Zhou P, Yang X-L, Wang X-G, *et al*. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;**579**:270–3.
3. Wu F, Zhao S, Yu B, *et al*. A new coronavirus associated with human respiratory disease in China. *Nature* 2020;**579**: 265–9.
4. Lu R, Zhao X, Li J, *et al*. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 2020;**395**:565–74.
5. WHO Director, World Health Organization. *WHO Director-General's opening remarks at the media briefing on COVID-19*. 11 March 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Accessed 07 August 2020.
6. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020;**20**:533–4.
7. Riou J, Althaus CL. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Euro Surveill* 2020;**25**:2000058. doi: 10.2807/1560-7917.ES.2020.25.4.2000058.
8. van Doremalen N, Bushmaker T, Morris DH, *et al*. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *N Engl J Med* 2020;**382**:1564–7.
9. Lauer SA, Grantz KH, Bi Q, *et al*. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann Intern Med* 2020;**172**:577–82.
10. Deng Y, Liu W, Liu K, *et al*. Clinical characteristics of fatal and recovered cases of coronavirus disease 2019 in Wuhan, China: a retrospective study. *Chin Med J (Engl)* 2020;**133**: 1261–7.
11. Singhal T. A review of coronavirus Disease-2019 (COVID-19). *Indian J Pediatr* 2020;**87**:281–6.
12. Cevik M, Bamford CGG, Ho A. COVID-19 pandemic-a focused review for clinicians. *Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis* 2020;**26**:842–7.
13. Tay MZ, Poh CM, Rénia L, *et al*. The trinity of COVID-19: immunity, inflammation and intervention. *Nat Rev Immunol* 2020;**20**:363–74.
14. World Health Organization. Estimating mortality from COVID-19. 4 August 2020. https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19. (07 August 2020, Accessed).
15. Dowd JB, Andriano L, Brazel DM, *et al*. *Demographic science aids in understanding the spread and fatality rates of COVID-19*. Proc Natl Acad Sci U S A. 2020; **117**(18):9696–9698. doi: 10.1073/pnas.2004911117.
16. Niedzwiedz CL, O'Donnell CA, Jani BD, *et al*. Ethnic and socioeconomic differences in SARS-CoV-2 infection: prospective cohort study using UK biobank. *BMC Med* 2020;**18**:160.
17. Sanders JM, Monogue ML, Jodlowski TZ, *et al*. Pharmacologic treatments for coronavirus disease 2019 (COVID-19): a review. *JAMA* 2020; 12; **323**(18):1824–1836. doi: 10.1001/jama.2020.6019.
18. Amanat F, Krammer F. SARS-CoV-2 vaccines: status report. *Immunity* 2020;**52**:583–9.
19. Corman VM, Landt O, Kaiser M, *et al*. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill* 2020; **25**(3):2000045. doi: 10.2807/1560-7917.ES.2020.25.3.2000045.
20. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* 2020;**5**:536–44.
21. Ksiazek TG, Erdman D, Goldsmith CS, *et al*. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med* 2003;**348**:1953–66.
22. Andersen KG, Rambaut A, Lipkin WI, *et al*. The proximal origin of SARS-CoV-2. *Nat Med* 2020;**26**:450–2.
23. Lam TT-Y, Jia N, Zhang Y-W, *et al*. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* 2020;**583**:282–5.
24. Wong MC, Javornik Cregeen SJ, Ajami NJ, *et al*. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *BioRxiv Prepr Serv Biol* 2020; 2020.02.07.939207. Published 13 February. 2020, doi: 10.1101/2020.02.07.939207.
25. Zhang T, Wu Q, Zhang Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Curr Biol* 2020;**30**:1346–1351.e2.
26. Yz Z, Ec HA. Genomic perspective on the origin and emergence of SARS-CoV-2. *Cell* 2020;**181**:223–7.
27. Sawicki SG, Sawicki DL, Siddell SG. A contemporary view of coronavirus transcription. *J Virol* 2007;**81**:20–9.
28. Sola I, Almazán F, Zúñiga S, *et al*. Continuous and discontinuous RNA synthesis in coronaviruses. *Annu Rev Virol* 2015;**2**:265–88.
29. Plant EP, Dinman JD. The role of programmed-1 ribosomal frameshifting in coronavirus propagation. *Front Biosci J Virtual Libr* 2008;**13**:4873–81.
30. de Wit E, van Doremalen N, Falzarano D, *et al*. SARS and MERS: recent insights into emerging coronaviruses. *Nat Rev Microbiol* 2016;**14**:523–34.
31. Kamelian K, Montoya V, Olmstead A, *et al*. Phylogenetic surveillance of travel-related Zika virus infections through whole-genome sequencing methods. *Sci Rep* 2019;**9**: 16433.
32. Shrivastava S, Puri V, Dilley KA, *et al*. Whole genome sequencing, variant analysis, phylogenetics, and deep sequencing of Zika virus strains. *Sci Rep* 2018;**8**:15843.
33. Kugelman JR, Wiley MR, Mate S, *et al*. Monitoring of Ebola virus Makona evolution through establishment of advanced genomic capability in Liberia. *Emerg Infect Dis* 2015;**21**:1135–43.
34. Quick J, Loman NJ, Duraffour S, *et al*. Real-time, portable genome sequencing for Ebola surveillance. *Nature* 2016;**530**:228–32.
35. Meredith LW, Hamilton WL, Warne B, *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* 2020; **20**(11):1263–1272. doi: 10.1016/S1473-3099(20)30562-4.
36. Rockett RJ, Arnott A, Lam C, *et al*. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nat Med* 2020; **26**(9):1398-1404. doi: 10.1038/s41591-020-1000-7.
37. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, *et al*. Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* 2020;**369**:297–301.

38. Center of Disease Control and Prevention (CDC). Coronavirus disease 2019 (COVID-19). *Cent Dis Control Prev* 2020. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/spheres.html. Accessed 7 August 2020
39. Gudbjartsson DF, Helgason A, Jonsson H, *et al*. Spread of SARS-CoV-2 in the Icelandic Population. *N Engl J Med*. 2020; **382**(24):2302–2315. doi: 10.1056/NEJMoa2006100.
40. SeqCOVID. *Genomic epidemiology of SARS-CoV-2 in Spain*. 2020. http://seqcovid.csic.es/. (07 August 2020, Accessed).
41. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance* 2017;**22**:30494.
42. Goodacre N, Aljanahi A, Nandakumar S, *et al*. A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. *mSphere* 2018; 14; **3**(2):e00069-18. doi: 10.1128/mSphereDirect.00069-18.
43. Pachetti M, Marini B, Benedetti F, *et al*. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med* 2020;**18**:179.
44. Korber B, Fischer WM, Gnanakaran S, *et al*. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 2020; **182**(4):812-827.e19. doi: 10.1016/j.cell.2020.06.043.
45. Grubaugh ND, Hanage WP, Rasmussen AL. Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. *Cell* 2020; **182**(4):794-795. doi: 10.1016/j.cell.2020.06.040.
46. Blomberg N, Lauer KB. Connecting data, tools and people across Europe: ELIXIR's response to the COVID-19 pandemic. *Eur J Hum Genet* 2020;**28**:719–23.
47. Wilkinson MD, Dumontier M, IjJ A, *et al*. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 2016;**3**:160018.
48. Conesa A, Beck S. Making multi-omics data accessible to researchers. *Sci Data* 2019;**6**:251.
49. Chervitz SA, Deutsch EW, Field D, *et al*. Data standards for Omics data: the basis of data sharing and reuse. *Methods Mol Biol* 2011;**719**:31–69.
50. World Health Organization. *Laboratory Testing for Coronavirus Disease (COVID-19) in Suspected Human Cases: Interim Guidance, 19 March 2020*. https://apps.who.int/iris/handle/10665/331501. Accessed 24 July 2020.
51. Li C, Zhao C, Bao J, *et al*. Laboratory diagnosis of coronavirus disease-2019 (COVID-19). *Clin Chim Acta* 2020;**510**:35–46.
52. Pan Y, Zhang D, Yang P, *et al*. Viral load of SARS-CoV-2 in clinical samples. *Lancet Infect Dis* 2020;**20**:411–2.
53. Zhang W, Du R-H, Li B, *et al*. Molecular and serological investigation of 2019-nCoV infected patients: implication of multiple shedding routes. *Emerg Microbes Infect* 2020; **9**:386–9.
54. Yu F, Yan L, Wang N, *et al*. Quantitative detection and viral load analysis of SARS-CoV-2 in infected patients. *Clin Infect Dis*. 2020 Jul 28; **71**(15):793–798. doi: 10.1093/cid/ciaa345.
55. Walsh KA, Jordan K, Clyne B, *et al*. SARS-CoV-2 detection, viral load and infectivity over the course of an infection. *J Infect* 2020; **81**(3):357–371. doi:10.1016/j.jinf.2020.06.067.
56. Yan Y, Chang L, Wang L. Laboratory testing of SARS-CoV, MERS-CoV, and SARS-CoV-2 (2019-nCoV): current status, challenges, and countermeasures. *Rev Med Virol* 2020;**30**:e2106.
57. Wang Y, Zhang L, Sang L, *et al*. Kinetics of viral load and antibody response in relation to COVID-19 severity. *J Clin Invest* 2020; **130**(10):5235–5244. doi:10.1172/JCI138759.
58. Chen W, Lan Y, Yuan X, *et al*. Detectable 2019-nCoV viral RNA in blood is a strong indicator for the further clinical severity. *Emerg Microbes Infect* 2020;**9**:469–73.
59. Riojas MA, Frank AM, Puthuveetil NP, *et al*. A rare deletion in SARS-CoV-2 ORF6 dramatically alters the predicted three-dimensional structure of the resultant protein. *BioRxiv Prepr Serv Biol*; 2020. 2020.06.09.134460. Published 2020 Jun 10. doi:10.1101/2020.06.09.134460.
60. Capobianchi MR, Rueca M, Messina F, *et al*. Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. *Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis* 2020;**26**:954–6.
61. Poon LLM, Leung CSW, Chan KH, *et al*. Recurrent mutations associated with isolation and passage of SARS coronavirus in cells from non-human primates. *J Med Virol* 2005;**76**:435–40.
62. Rimoldi SG, Stefani F, Gigantiello A, *et al*. Presence and infectivity of SARS-CoV-2 virus in wastewaters and rivers. *Sci Total Environ* 2020;**744**:140911.
63. Nemudryi A, Nemudraia A, Wiegand T, *et al*. Temporal Detection and Phylogenetic Assessment of SARS-CoV-2 in Municipal Wastewater. *Cell Rep Med*. 2020; **1**(6):100098. doi: 10.1016/j.xcrm.2020.100098.
64. Lednicky JA, Shankar SN, Elbadry MA, *et al*. Collection of SARS-CoV-2 virus from the air of a clinic within a university student health care Center and analyses of the viral genomic sequence. *Aerosol Air Qual Res* 2020;**20**: 1167–71.
65. La Rosa G, Iaconelli M, Mancini P, *et al*. First detection of SARS-CoV-2 in untreated wastewaters in Italy. *Sci Total Environ* 2020;**736**:139652.
66. Ahmed W, Angel N, Edson J, *et al*. First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: a proof of concept for the wastewater surveillance of COVID-19 in the community. *Sci Total Environ* 2020;**728**: 138764.
67. Brister JR, Ako-adjei D, Bao Y, *et al*. NCBI viral genomes resource. *Nucleic Acids Res* 2015;**43**:D571–7.
68. Greninger AL, Chen EC, Sittler T, *et al*. A metagenomic analysis of pandemic influenza a (2009 H1N1) infection in patients from North America. *PLOS One* 2010;**5**:e13381.
69. Carter LJ, Garner LV, Smoot JW, *et al*. Assay techniques and test development for COVID-19 diagnosis. *ACS Cent Sci* 2020;**6**:591–605.
70. Kohl C, Brinkmann A, Dabrowski PW, *et al*. Protocol for metagenomic virus detection in clinical specimens. *Emerg Infect Dis* 2015;**21**:48–57.
71. Smits SL, Bodewes R, Ruiz-Gonzalez A, *et al*. Assembly of viral genomes from metagenomes. *Front Microbiol* 2014; 18;**5**:714. doi: 10.3389/fmicb.2014.00714. PMID: 25566226.
72. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. *Microbiol Mol Biol Rev MMBR* 2012;**76**:159–216.
73. Chiu CY, Miller SA. Clinical metagenomics. *Nat Rev Genet* 2019;**20**:341–55.
74. Pillay S, Giandhari J, Tegally H, *et al*. Whole genome sequencing of SARS-CoV-2: adapting Illumina protocols for quick and accurate outbreak investigation during a pandemic. *Genes* 2020;**11**:949.

75. Paden CR, Tao Y, Queen K, *et al*. Rapid, sensitive, full-genome sequencing of severe acute respiratory syndrome coronavirus 2. *Emerg Infect Dis* 2020;**26**(10):2401–5.
76. Oxford Nanopore Technologies. Nanopore sequencing the SARS-CoV-2 genome: introduction to protocol. March 8 2020. http://nanoporetech.com/resource-centre/nanopore-sequencing-sars-cov-2-genome-introduction-protocol. Accessed 26 September 2020.
77. Campos GS, Sardi SI, Falcao MB, *et al*. Ion torrent-based nasopharyngeal swab metatranscriptomics in COVID-19. *J Virol Methods* 2020;**282**:113888.
78. Pacific Biosciencies. COVID-19 sequencing tools and resources. 2020. https://www.pacb.com/research-focus/microbiology/covid-19-sequencing-tools-and-resources/. (27 September 2020, Accessed).
79. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* 2004;**68**:669–85.
80. Quince C, Walker AW, Simpson JT, *et al*. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 2017;**35**:833–44.
81. Gilbert JA, Dupont CL. Microbial metagenomics: beyond the genome. *Ann Rev Mar Sci* 2011;**3**:347–71.
82. Zhang H, Ai J-W, Yang W, *et al*. Metatranscriptomic characterization of COVID-19 identified a host transcriptional classifier associated with immune signaling. *Clin Infect Dis*. 2020; ciaa663. doi: 10.1093/cid/ciaa663.
83. Butler DJ, Mozsary C, Meydan C, *et al*. Shotgun transcriptome and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions. *BioRxiv Prepr Serv Biol* 2020; 2020.04.20.048066. Published 2020 May 1. doi: 10.1101/2020.04.20.048066.
84. Chan JF-W, Yuan S, Kok K-H, *et al*. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* 2020;**395**:514–23.
85. Lewandowski K, Xu Y, Pullan ST, *et al*. Metagenomic Nanopore sequencing of influenza virus direct from clinical respiratory samples. *J Clin Microbiol* 2019;**58**(1):e00963–19. doi: 10.1128/JCM.00963-19.
86. Kafetzopoulou LE, Efthymiadis K, Lewandowski K, *et al*. Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. *Euro Surveill*. 2018; **23**(50):1800228. doi: 10.2807/1560-7917.ES.2018.23.50.1800228.
87. Xiao M, Liu X, Ji J, *et al*. Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. *Genome Med* 2020;**12**(1):57. doi: 10.1186/s13073-020-00751-4.
88. ARTICnetwork. artic-network/artic-ncov2019. 2020. https://github.com/artic-network/artic-ncov2019. Accessed 7 August 2020.
89. Itokawa K, Sekizuka T, Hashino M, Tanaka R, Kuroda M. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. *PLoS One* 2020; Sep 18; **15**(9):e0239403. doi: 10.1371/journal.pone.0239403.
90. Wang M, Fu A, Hu B, *et al*. Nanopore targeted sequencing for the accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. *Small* 2020; **16**(32):e2002169. doi: 10.1002/smll.202002169.
91. Albert TJ, Molla MN, Muzny DM, *et al*. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 2007;**4**:903–5.
92. Warr A, Robert C, Hume D, *et al*. Exome sequencing: current and future perspectives. *G3 (Bethesda)* 2015;**5**:1543–50.
93. Maurano MT, Ramaswami S, Zappile P, *et al*. Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City region Genome Res. 2020. doi: 10.1101/gr.266676.120.
94. Amarasinghe SL, Su S, Dong X, *et al*. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 2020;**21**:30.
95. Kim D, Lee J-Y, Yang J-S, *et al*. The architecture of SARS-CoV-2 transcriptome. *Cell* 2020;**181**:914–921.e10.
96. Emanuel W, Kirstin M, Vedran F, *et al*. Bulk and single-cell gene expression profiling of SARS-CoV-2 infected human cell lines identifies molecular targets for therapeutic intervention. *BioRxiv Prepr Serv Biol* 2020; 2020.05.05.079194. Published 5 May 2020. doi: 10.1101/2020.05.05.079194.
97. Tan B, Gao S-J. RNA epitranscriptomics: regulation of infection of RNA and DNA viruses by N6-methyladenosine (m6A). *Rev Med Virol* 2018;**28**:e1983.
98. O'Connell MA, Mannion NM, Keegan LP. The epitranscriptome and innate immunity. *PLoS Genet* 2015;**11**:e1005687.
99. Viehweger A, Krautwurst S, Lamkiewicz K, *et al*. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Res* 2019;**29**:1545–54.
100. Giorgio SD, Martignano F, Torcia MG, *et al*. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci Adv* 2020;**6**:eabb5813.
101. Picardi E, Mansi L, Pesole G. A-to-I RNA editing in SARS-COV-2: real or artifact? *BioRxiv Prepr Serv Biol* 2020; 2020.07.27.223172. Published 27 July 2020. doi: 10.1101/2020.07.27.223172.
102. Wee Y, Bhyan SB, Liu Y, *et al*. The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. *Brief Funct Genomics* 2019;**18**:1–12.
103. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics* 2010;**95**:315–27.
104. Doddapaneni H, Cregeen SJ, Sucgang R, *et al*. Oligonucleotide capture sequencing of the SARS-CoV-2 genome and subgenomic fragments from COVID-19 individuals. *BioRxiv Prepr Serv Biol* 2020; 2020.07.27.223495. Published Jul 27. 2020. doi: 10.1101/2020.07.27.223495.
105. Vollmers J, Wiegand S, Kaster A-K. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective—not only size matters! *PLOS One* 2017;**12**:e0169662.
106. Simpson JT, Pop M. The theory and practice of genome sequence assembly. *Annu Rev Genomics Hum Genet* 2015;**16**:153–72.
107. Dominguez Del Angel V, Hjerde E, Sterck L, *et al*. Ten steps to get started in genome assembly and annotation. *F1000Res* 2018;**7**:ELIXIR-148. doi: 10.12688/f1000research.13598.1.
108. Molloy JC. The open knowledge foundation: open data means better science. *PLoS Biol* 2011;**9**:e1001195.
109. Hayes B. Overview of statistical methods for genome-wide association studies (GWAS). *Methods Mol Biol* 2013;**1019**:149–69.
110. Gkiouras K, Nigdelis MP, Grammatikopoulou MG, *et al*. Tracing open data in emergencies: the case of the COVID-

19 pandemic. *Eur J Clin Invest*. **50**(9):e13323. doi: 10.1111/eci.13323.

111. Homolak J, Kodvanj I, Virag D. Preliminary analysis of COVID-19 academic information patterns: a call for open science in the times of closed borders. *Scientometrics* 2020; 1–15. doi: 10.1007/s11192-020-03587-2.
112. Research Data Alliance (RDA). The final version of the RDA COVID-19 Recommendations and Guidelines for Data Sharing. 30 June 2020. https://www.rd-alliance.org/group/rda-covid19-rda-covid19-omics-rda-covid19-epidemiology-rda-covid19-clinical-rda-covid19-1. (6 August 2020, Accessed).
113. Baker D, van den Beek M, Blankenberg D *et al*. No more business as usual: Agile and effective responses to emerging pathogen threats require open data and open analytics. *PLoS Pathog*. 2020; **16**(8):e1008643. Published 2020 Aug 13. doi: 10.1371/journal.ppat.1008643.
114. Cochrane G, Karsch-Mizrachi I, Takagi T, *et al*. The international nucleotide sequence database collaboration. *Nucleic Acids Res* 2016;**44**:D48–50.
115. Athar A, Füllgrabe A, George N, *et al*. ArrayExpress update—from bulk to single-cell expression data. *Nucleic Acids Res* 2019;**47**:D711–5.
116. Edgar R. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;**30**:207–10.
117. Lappalainen I, Almeida-King J, Kumanduri V, *et al*. The European genome-phenome archive of human data consented for biomedical research. *Nat Genet* 2015;**47**:692–5.
118. MacArthur J, Bowler E, Cerezo M, *et al*. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 2017;**45**:D896–901.
119. Mailman MD, Feolo M, Jin Y, *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007;**39**: 1181–6.
120. Schriml LM, Chuvochina M, Davies N, *et al*. COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci Data* 2020;**7**:188.
121. Bauer DC, Tay AP, Wilson LOW, *et al*. Supporting pandemic response using genomics and bioinformatics: a case study on the emergent SARS-CoV-2 outbreak. *Transbound Emerg Dis* 2020;**67**:1453–62.
122. Black A, MacCannell DR, Sibley TR, *et al*. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nat Med* 2020;**26**:832–41.
123. Ison J, Rapacki K, Ménager H, *et al*. Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Res* 2016;**44**:D38–47.
124. Afgan E, Baker D, Batut B, *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 2018;**46**:W537–44.
125. Argimón S, Abudahab K, Goater RJE, *et al*. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* 2016;**2**(11):e000093. doi: 10.1099/mgen.0.000093.
126. OpenAIRE. OpenAIRE COVID-19 gateway. *OpenAIRE*. 2020. https://www.openaire.eu/openaire-covid-19-gateway. Accessed 27th September 2020.
127. Cosgriff CV, Ebner DK, Celi LA. Data sharing in the era of COVID-19. Lancet Digit Health. 2020; **2**(5):e224. doi: 10.1016/S2589-7500(20)30082-0.
128. Hadfield J, Megill C, Bell SM, *et al*. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 2018;**34**: 4121–3.
129. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 2005;**21**:676–9.
130. ELIXIR. A beacon in the ocean of SARS-CoV-2 data. Mon 7 September 2020. https://elixir-europe.org/news/beacon-ocean-sars-cov-2-data. Accessed: 26 September 2020.
131. CSIRO and CSIR-IGIB. The COVID-19 Beacon is a collaborative endeavour between Transformational Bioinformatics group at Covid19 beacon. 2020. http://covid19.genomes.in/. Accessed 26 September 2020.
132. Canakoglu A, Pinoli P, Bernasconi A, *et al*. ViruSurf: an integrated database to investigate viral sequences. *Nucleic Acids Res*. 2020; gkaa846. doi: 10.1093/nar/gkaa846.
133. Pickett BE, Sadat EL, Zhang Y, *et al*. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res* 2012;**40**:D593–8.
134. Gowthaman R, Guest JD, Yin R, *et al*. CoV3D: a database of high resolution coronavirus protein structures [published online ahead of print, 2020 Sep 5]. *Nucleic Acids Res*. 2020; gkaa731. doi: 10.1093/nar/gkaa731.
135. Baker D, van den Beek M, Blankenberg D, *et al*. No more business as usual: agile and effective responses to emerging pathogen threats require open data and open analytics. *PLoS Pathog* 2020;**16**:e1008643.
136. Katz R. Use of revised international health regulations during influenza A (H1N1) epidemic, 2009. *Emerg Infect Dis* 2009;**15**:1165–70.
137. Wolicki SB, Nuzzo JB, Blazes DL, *et al*. Public health surveillance: at the Core of the Global Health Security Agenda. *Health Secur* 2016;**14**:185–8.
138. Schuchat A, Tappero J, Blandford J. Global health and the US Centers for Disease Control and Prevention. *Lancet Lond Engl* 2014;**384**:98–101.
139. Ghsa Preparation Task Force Team Null. Global Health Security: the lessons from the West African Ebola virus disease epidemic and MERS outbreak in the Republic of Korea. *Osong Public Health Res Perspect* 2015;**6**:S25–7.
140. Chiara M, Horner DS, Gissi C, *et al*. Comparative genomics provides an operational classification system and reveals early emergence and biased spatio-temporal distribution of SARS-CoV-2. 2020; 2020.06.26.172924. Published June 30 2020. doi: 10.1101/2020.06.26.172924.
141. Zehender G, Lai A, Bergna A, *et al*. Genomic characterization and phylogenetic analysis of SARS-COV-2 in Italy. *J Med Virol* 2020;10.1002/jmv.25794. doi: 10.1002/jmv.25794.
142. Boni MF, Lemey P, Jiang X, *et al*. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol* 2020; **5**(11):1408–1417. doi: 10.1038/s41564-020-0771-4.
143. HyPhy team. HyPhy COVID-19. 2020. http://hyphy.org/covid/. Accessed 27 September 2020.
144. The COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur J Hum Genet* 2020;**28**: 715–8.
145. Casanova J-L, Su HC, Abel L, *et al*. A global effort to define the human genetics of protective immunity to SARS-CoV-2 infection. *Cell* 2020;**181**:1194–9.