Scenarios for the integration of microarray gene expression profiles in COVID-19-related studies

Anna Bernasconi and Silvia Cascianelli

Abstract The COVID-19 pandemic has hit heavily many aspects of our lives. At this time, genomic research is concerned with exploiting available datasets and knowledge to fuel discovery on this novel disease. Studies that can precisely characterize the gene expression profiles of human hosts infected by SARS-CoV-2 are of significant relevance. However, not many such experiments have yet been produced to date, nor made publicly available online. Thus, it is of paramount importance that data analysts explore all possibilities to integrate information coming from similar viruses and related diseases; interestingly, microarray gene profile experiments become extremely valuable for this purpose. This chapter reviews the aspects that should be considered when integrating transcriptomics data, considering mainly samples infected by different viruses and combining together various data types and also the extracted knowledge. It describes a series of scenarios from studies performed in literature and it suggests possible other directions of noteworthy integration.

Keywords. COVID-19; Microarray experiments; gene expression; viral infection; data integration; genomics.

1 Introduction

The coronavirus emerged at the end of 2019 has caused many outbreaks at global level and profound societal disruption. Importantly, the COVID-19 disease, which is caused by the SARS-CoV-2 virus, exhibits infection characteristics that have been only partially understood. As they resemble SARS-like complications, it has been argued that comparative analysis of genomic and transcriptomic data from similar viruses can lead to valuable insights [1].

This review chapter addresses data analysts who are familiar with microarray gene expression datasets. Due to the unavailability of many new experiments specifically targeting tissues/cell lines infected by the new virus, microarray datasets related to other (genetically close) viruses' infections – available from public databases and validated by well-recognized literature – can be re-purposed for the objective of investigating the novel COVID-19.

We thus propose a series of guidelines that can be used by data analysts to infer knowledge on COVID-19 gene expression profiles by performing *integrative studies*, i.e., using jointly data and results from microarray and next-generation sequencing (NGS) experiments concerning SARS-CoV-2, similar viruses (i.e., meta-analysis), and/or associated diseases. The presented typical scenarios include experiments based on expression data of human hosts. Our base assumption is that microarray platforms have been used more widely in the past, while recently NGS approaches are strongly preferred. Accordingly, microarray technologies have been adopted for performing experiments focused on infections caused by previous viruses that are considered genetically close to SARS-CoV-2. Such studies can be conveniently employed for integration with studies specifically concerned with SARS-CoV-2 infections, to derive enriched resulting knowledge.

The possibility of integrating datasets on infected host genetics profiles with datasets that analyze the viral genomics is not discussed in this chapter. Indeed, from the perspective of a data analyst who can only access publicly available data, such advanced integration is still not practically feasible; to the best of our knowledge, no human COVID-19 gene expression dataset (on Gene Expression Omnibus [2], ArrayExpress [3], or similar repositories) has yet been linked to the viral sequences responsible for the infection. We refer the interested reader to [4] for a discussion on this topic. Data integration efforts regarding sequences of SARS-CoV-2 have been produced [5], but so far viral sequences and human profiles are kept separate [6, 7].

The chapter is organized as follows. In Section 2 we provide an overview of the differences between gene expression profile experiments performed with microarray technology [8] and next-generation sequencing technology [9], also known as high-throughput sequencing. Then, Section 3 discusses why data and knowledge derived from infections caused by viruses similar to SARS-CoV-2 or from other associated diseases associated could be transferred to COVID-19-related studies. Section 4 describes common procedures to acquire datasets of public data to conveniently analyze the domain at hand. Section 5 overviews the integration activities from two perspectives, i.e., the data one and the knowledge one, outlining several possible levels and instantiating them on example studies from literature. Section 6 explains our general view of COVID-19 integrative studies, decomposing them in materials, methods and results blocks; seven different scenarios, existing in literature or proposed for future applications are hence presented. Finally, Section 7 concludes the chapter.

2 Microarray and Next-Generation Sequencing technologies for human host expression profiling

Many transcriptomics platforms are commonly used to generate expression data of human tissues or cell lines affected by viral infections, such as those of our interest: yet, to date, the two main technological options are represented by microarrays and by NGS methods.

Microarrays are platforms that can host thousands of DNA fragments, called probes, orderly and firmly arranged in well-known locations on a solid holder, as to identify each gene of interest. A hybridization process occurs when these probes are placed in presence of the complementary ones, which are obtained from messenger-RNA (mRNA) of the sample(s) under exam and marked with fluorochrome. In case of strong binding interactions of complementary strands, corresponding gene activity can be evaluated by measuring local fluorescence intensity: the amount of fluorescent signal arising from a given position is indeed directly proportional to the amount of the corresponding gene transcript.

Microarrays are first characterized by the number of samples hybridized within the same experiment: a single-channel array provides gene expression estimates of one sample at a time, while a dual-channel (or two-color) microarray is meant to quantify the expression levels from competitive hybridization of two samples, representing two biological conditions under analysis. In both cases, microarray technologies are intrinsically designed to provide a relative quantification measure, due to the relative intensity of each fluorescent dye during the hybridization process. Therefore, despite their differences, both types of microarray are primarily used in comparative settings (e.g., to compare an infected sample with a mock control): reliable values of relative expression (called gene expression ratio) are commonly obtained evaluating two alternative conditions either on the same dual-channel array or on two single-channel

ones. Accordingly, microarray experiments often aim at identifying differentially expressed genes (DEGs), i.e., genes experiencing statistically significant changes in their expression levels between the two biological conditions, usually assessed across a population of such alternative cases.

Another dichotomy within microarray technology concerns the probe generation method: spotted microarrays are two-color arrays having probes of 1-2 kilo bases length obtained using cDNA-libraries; oligo-chips are instead arrays synthesized in situ and typically including multiple oligonucleotide probes of fixed and short length. Even if the former can rely on hybridization on longer fragments and are globally cheaper and more flexible, they can be affected by colorimetric problems as well as possible differences in the amounts of deposited probes and of mRNA coming from the two samples: this makes it more difficult to compare results obtained from different arrays. Conversely, oligo-chips can assess a higher number of genes while showing less variability between one chip and another. Though this improves the comparability of the results obtained from different arrays, it limits the investigation to sets of more generic predetermined sequences.

All microarrays inspect a set of previously identified target genes to provide their expression levels; more recent and complete arrays support analysis within the overall known transcriptome with fast profiling of coding and non-coding genes, exons, and transcript isoforms, i.e., all possible known RNA sequences for a gene. Nonetheless, shortcomings and issues with microarrays include the need to know a priori the sequences to be investigated, cross-hybridization artifacts, poor quantification of lowly and highly expressed genes as well as complexity of the required normalization pipelines [10]. Due to these technical limitations and the progressively decreasing cost of NGS experiments, transcriptomics is transitioning mostly to sequencing-based methods.

NGS technologies provide useful information on the sequences in addition to more accurate estimations of corresponding gene expression levels; particularly, RNA-Sequencing (RNA-Seq) [11] is a high-throughput NGS technology that has undoubtedly become the most popular method for transcriptome analysis. Briefly, RNA-Seq works by sequencing every RNA molecule and profiling the expression of a given gene by counting the number of times its transcripts have been sequenced. Beyond being able to offer a qualitative and quantitative screening of multiple types of RNA in biological samples, RNA-Seq profiling has allowed to identify thousands of novel isoforms and show the complexity of protein-coding transcriptome [12], in a much wider way than before. Additionally, for the discovery of differentially expressed genes, whole-transcriptome sequencing grants the interrogation of any gene in the sample, with a wide dynamic range, allowing also DEG analysis of low-expressing genes.

Although microarrays show some weaknesses compared with NGS technologies (e.g., they can only investigate known sequences), they have been largely employed and their protocols have been enhanced in the last two decades; also, robust statistical methods exist to comprehensively analyse them. Thus, microarrays have left researchers with a wide and rich legacy of data and analyses that can be leveraged to improve future studies focused on gene expression, even if primarily based on sequencing methods. At the same time, huge amounts of NGS data are being progressively collected due to the decreasing cost of sequencing experiments. These are particularly valuable to study viral infections, where host samples can be sequenced and aligned to the human genome as well as to viral genomes; this allows indeed to investigate virus integration in host genomes and perform comparative analyses of viral transcripts and host gene responses [13]. Hence, it is essential to exploit the benefits of integrating data and knowledge available from these transcriptomics technologies in relevant scopes such as the investigation of the novel COVID-19

infectious disease: when taking care of all the aspects involved in their mutual integration [14], their joint use becomes a precious resource for ongoing researches.

3 COVID-19 and its responsible virus

SARS-CoV-2 is the seventh coronavirus known as a human infection driver. The first reports of the disease associated to this virus, the novel pneumonia denominated COVID-19, were sent in Wuhan, Hubei province, China, at the end of 2019. In the following we discuss the possibility of using experiments regarding infections caused by different viruses or other diseases.

3.1 Knowledge transfer from other viruses' infections to SARS-CoV-2 ones

SARS-CoV-2 is an enveloped, positive sense, single stranded RNA virus. Even if hypotheses on laboratory origins of the virus have been raised, there have been proofs of its proximal origin [15], backed by comparative analysis of genomic data with similar viruses. Based on phylogenetic analyses [1], SARS-CoV-2 emerged as genetically close to **SARS-CoV** (or SARS-CoV-1) – responsible for the SARS pandemic of started in China in 2003. Both SARS-CoV-2 and SARS-CoV share a common ancestral origin with **MERS-CoV** – causing the MERS Saudi Arabia outbreak in 2012. These three viruses show high percentages of sequence similarity [16], which has been accordingly exploited also for vaccine design [17]. While a core feature as the polybasic cleavage site at the junction of S1 and S2 in the Spike protein was not previously observed in other beta coronaviruses (as claimed in [15]) many specific similarities at the level of the sequence are proven. These are, for example, in the

Receptor Binding Domain (RBD) [18], in the codon usage of RNA-dependent RNA polymerase (RdRp) and Main protease (Mpro) genes [19] (these are key proteins in the virus' life cycle), as well as at the level of viral growth and of gene expression patterns [20].

Note that the idea of comparing gene expression profiles in case of infections by different but similar viruses is not new; for example, Bosworth *et al.* [21] had proposed to relate human respiratory syncytial virus (HRSV) profiles with Ebola virus ones, based on their similar genome organisation and replication strategy. Overall, the potentiality of integrating datasets from different viruses is exploited in so-called meta-analysis studies: these kinds of studies are crucial to face a new viral infectious disease, especially for assessing the comparability of gene responses and for the applications of drug repurposing, which is the identification of new opportunities to reuse already existing safe drugs.

In addition, comparative studies and meta-analyses may also include other pathogenic viruses that primarily attack the respiratory system and may cause death. For brevity, in the following discussion, among these viruses, we only consider the influenza A virus (**IAV**), which was responsible for several popular pandemics, such as the Spanish influenza in 1918 and the Avian influenza in 2009. In literature this virus has been compared to SARS-CoV-2 in terms of excess years of life lost. Indeed, the mean age (~80 years) of COVID-19 fatalities is considered similar to the 1957 and 1968 influenza pandemics [22]. Note that other comparisons, especially with other coronaviruses, may be meaningful [23].

Datasets and studies available for these different viruses become particularly useful when considering the current scarcity of expression profiling experiments on patients/cell lines infected by SARS-CoV-2, especially with microarray technologies [1]. In this way, it is possible to maximize the use of previously available

expression data while not many experiments have been performed (or have not been made public) for COVID-19.

3.2 Knowledge transfer from related diseases to COVID-19

Another perspective for improving the current knowledge about a new viral infectious disease such as COVID-19 is to enrich the gene-level analyses with data and insights coming from linked diseases; such related diseases can emerge from the *diseasome*, a complete conceptual model of human diseases, including also their responsible or involved genes as well as the ways in which these genes are expressed [24, 25].

Expression profiling has indeed become a powerful and increasingly affordable way to provide quantitative estimates of gene level activity. It is crucial to examine, explore and stratify different diseases. Particularly, a deeper knowledge of the genes underlying a disease has made it possible – when a given gene appears significantly dysregulated/mutated in the associated disease – to represent and explore the diseasome as a bipartite graph, consisting of disease nodes connected to gene nodes.

The links at the transcriptional level can also highlight another model of relevant associations, i.e., the network of homogeneous disease-disease interactions obtained through the connection of pairs of diseases that share the same genes in gene-disease links. Additionally, in a homogeneous disease network, further disease-disease interactions can reflect proved similarities and relationships between diseases at phenotypic level (e.g., comorbidities, side-effects, ...) [25]. Accordingly, within the scope of the integrative studies presented in this work, different kind of similarities between diseases are exploited in two types of relevant contexts, discussed as follows.

On the one hand, an integrative study can compensate for the often-limited gene expression data of infected hosts with additional gene expression profiles of patients affected by diseases that are commonly recognized as complications of the viral

infectious disease of interest [1]. This kind of data integration implies a knowledge transfer, based on the assumption that causally related diseases are linked through disease-disease associations that may, in turn, hide similar underlying genes at the transcriptional level. Therefore, causal relationships involving the infectious disease of interest and some of its complications are worth investigating with comparative analyses, which can be directly applied to genes emerged as differentially expressed in such diverse but associated diseases.

On the other hand, a work focused on an infectious disease can be enhanced through knowledge about protein-drug-disease associations, where proteins of interest usually come from dysregulated genes, often examined also with enrichment analyses. This kind of knowledge integration can lead to study homogeneous disease network [1] or heterogeneous networks such as protein-drug interaction networks [26, 1]. In both cases, the so-obtained associations are rooted in the genes emerged as significantly involved in the hosts' infection onset/progression; they are also able to guide the search of potential therapeutic targets as well as the use of computational methods for drug-repurposing (e.g., [27, 28, 29]), based on the assumption that similar diseases can be treated with similar drugs [28].

4 Data acquisition

From the point of view of a data analyst, the well-known Gene Expression Omnibus [2] is the most common resource to access publicly available datasets for integrative studies on COVID-19. On GEO DataSets page (https://www.ncbi.nlm.nih.gov/gds), a user can perform various search sessions, e.g., using the query `("Homo sapiens"[Organism] AND ("gse"[Filter] AND "Expression profiling by array"[Filter])) AND "Sars"[Title]`.
Similar ones can be designed by using strings for matching titles, e.g., re-

garding "mers" or "coronavirus". Alternatively, users may employ the Browser endpoint, looking for Series; to date (February 8th, 2021) the API request https://www.ncbi.nlm.nih.gov/geo/browse/?view=series&search=sars allows to retrieve 158 results, which are reduced to 11 when the filters on homo sapiens organism and microarray series type are applied.

Other recommended sources for finding relevant datasets include ArrayExpress [3], the Genomic Expression Archive maintained by the DNA DataBank of Japan [30], and All Of gene Expression [31], an integrator of publicly available gene expression data. As metadata are often not structured, especially the ones regarding the characteristics of analyzed samples, semi-automated methods of metadata extraction may be needed for achieving more tailored search processes (see, e.g., [32] for a proposal to extract information from GEO Samples descriptions, or [33] for a structured instance of metadata).

[Table 1 near here]

A number of very recent papers [34, 1, 26, 35, 36] provide a set of relevant GEO Series with microarray data of interest in this chapter. In Table 1, we report a list of such datasets, plus additional ones; these are all produced with microarray technology, which refer to human gene expression analysis on tissues or cell lines infected by the most interesting viruses for our purpose (i.e., SARS-CoV-2, SARS-CoV, MERS-CoV, and Influenza A). Almost all these datasets are used in the five mentioned studies, which are thoroughly explained in Section 6, while a few others have been retrieved performing classical search sessions on GEO. The last three listed datasets involve tissues with diseases that are not viral, i.e., pneumonia (PNA), shortness of breath (SHOB), and diarrhea (DRA), commonly regarded as COVID-19 complications. Note that here we do not report relevant SARS-CoV-2 experiments used in the five studies when they are performed with NGS technology (i.e., GSE147507 and GSE162835).

For each experiment from GEO, we report the Series ID, the virus/disease causing the infection event, the type of analyzed samples (either extracted from human individuals or from cell lines), the number of infected vs. control (healthy) samples; the name of the employed platform; the number of genes profiled in the array; the year of first public appearance of the Series on GEO; and the PubMed ID of the linked academic publication.

5 Integration levels

Focusing on the scope presented in the preliminaries, we observe that, both methodologies already adopted in COVID-19-related studies (recently published in literature) and other data investigation techniques (that could be soon applied in the context of ongoing research), require one or more integration levels. It seems appropriate to categorize these as *data integration* and *knowledge integration*. Both levels of integration are, in turn, characterized by a plethora of possible cases: *data integration* captures the situations in which different kinds of data – experimental transcriptional data (especially from different technologies), experimental metadata, database annotations and/or associations – must be considered together as the main actors of one or more analytical steps. Conversely, *knowledge integration* concerns the problem of combining and exploiting together pre-existing knowledge and information that has been derived by previous steps of analysis, may it regard a set of differentially expressed genes, interesting pathways, or drug targets.

In the following, we report the most common cases of data integration and knowledge integration observed in several COVID-19-related studies. These levels were first identified in the five integrative studies [34, 36, 1, 26, 35] on which this chapter is focused, as deemed particularly explanatory of our scope of interest; relevant details are provided in Table 2.

Data integration levels include:

D1: Microarray experiments from viruses that are different from SARS-CoV-2, integrated with RNA-Seq complete gene profiles of infected patients.

D1b: Partial microarray experiments (on selected genes) from viruses that are different from SARS-CoV-2, integrated with RNA-Seq complete gene profiles of infected patients. Note that this level is separated from D1, as partial microarrays impose an additional constraint to integrative studies that already deal with different viruses and aggregate NGS with microarray technologies.

D2: Microarray experiments from viruses that are different from SARS-CoV-2, integrated with microarray experiments on patients with diseases that are typically considered as comorbidities or complications of COVID-19.

D3: Microarray experiments from SARS-CoV-2, integrated with RNA-Seq complete gene profiles of infected patients.

Knowledge integration levels include:

K1: Meta-analysis of different viruses, extracting differentially expressed genes (sets intersections and differences, possibly based on the up/down regulation). This can be done among SARS-CoV-2 and its similar viruses or only among similar viruses.

K2: Integration of gene/proteins dysregulated and/or involved in processes/pathways enriched in diseases that are associated to COVID-19, e.g., its common complications.

K3: Identification of potential therapeutic targets and/or drugs through the integration of relevant gene/proteins (selected via network or gene set enrichment analyses) with i) external data banks for drugs/chemical agents, ii) external ontologies (the Coronavirus Infectious Disease Ontology [37] can define relations between drugs and roles or mechanisms of action).

K4: Identification of gene signatures related to the expected prognosis through the integration of predictive models able to provide clinically relevant predictions, e.g., survival models.

The described levels of integration are interlinked in many analytical steps that intrinsically involve data integration and knowledge integration together. In turn, a given level could be addressed multiple times in the same study. For instance, consider the data integration effort that is needed to aggregate experimental data of infectious diseases caused by similar viruses: such experimental data are commonly derived from different platforms and transcriptomics technologies. COVID-19 datasets are being progressively produced using the more widespread NGS technologies; on the contrary, other data of interest – such as the one from comparable infectious diseases (e.g., MERS, SARS, Type A influenza) – are mainly obtained with microarrays, as they were generated prior to the revolution and spread of NGS technologies in the last five years.

[Table 2 near here]

Table 2 reports an overview of integrative studies existing in literature, aiming to derive knowledge on COVID-19. For each of them we specify the first author of the publication, the publication (studies are listed in chronological order), the list of employed datasets – divided by technology group, either NGS or microarray – and the used integration levels identified by the data (D) and knowledge (K) points listed previously in this section. When appropriate, we specified the objects of such integration. For example, Loganathan *et al.* [36] performs knowledge integration between SARS-CoV-2 information with other similar viruses (K1); Nain *et al.* [1] integrates datasets from diseases typically recognized as COVID-19 complications (D2) and performs a knowledge-level integration of genes involved in processes of such diseases and of SARS infection (K2); Moni *et al.* [26] integrate knowledge acquired from a lung-related disease – using SARS-CoV-2 infection response genes –

to perform survival analysis on lung adenocarcinoma patients (K4). Note that we did not find relevant works including the D3 level or the K4 level (with a predictive model applied directly on COVID-19 patients); however these are crucial contributions that future works should address (as suggested in Section 6).

6 Integrative studies: possible scenarios

The integrative studies proposed in this chapter can be described comprehensively by considering a set of different materials to be integrated at the data level, then processed using methods allowing to derive information that can be integrated at the knowledge level, with the final purpose of achieving results that are meaningful for COVID-19 research.

The specific framework is described in Figure 1. On the left we outline a set of *materials*, focused on gene expression; the use of datasets from different viruses and COVID-19 related diseases is usually supported by, respectively, a phylogenetic analysis and a diseasome analysis. The datasets are input into a collection of *methods* that include the differentially expressed genes (DEG) analysis, the gene set enrichment analysis (GSEA), as well as survival analysis or classification with feature importance analysis. In addition, many network analyses (NA) can be performed by building node-relationship models of co-expressed genes, of genes associated with diseases or with regulatory biomarkers, and of protein interactions with proteins and with drugs. The application of these methods leads to the achievement of a series of important *results* such as the identification of relevant genes, of targets for novel therapies and of drugs that may be repurposed from other uses. Machine learning goals in this scope are focused on prediction of different classes (e.g., classes of infectious disease severity) and their comparison; for data analysts, computational methods for classification and feature importance analysis are indeed valid data-

driven alternatives to more biologically oriented ways of selecting valuable gene signatures.

[Figure 1 near here]

Fig. 1 Materials, methods and results modules typical of gene expression-based integrative studies for COVID-19 knowledge discovery. Note that the 'classification' method and the 'class prediction' result are rendered in yellow as they are tightly linked to each other, co-existing in the scenarios that include them. Instead, all the other modules can be combined and repeated in different orders.

The different types of data and the possible analyses that involve them open up a series of scenarios where knowledge, either already available or acquired through the workflow, is integrated with such data to progressively generate new knowledge about COVID-19. Figure 1 reports a general view for understanding the single scenarios reported in the following; other analytical workflows could be designed in addition.We do not aim at covering all the possible examples of analytical workflows; instead, we provide illustrative scenarios to show how analysts can make use of already available transcriptional data (primarily from microarrays) to improve studies and current knowledge about COVID-19.

Our scenarios are mostly focused on integrative studies from already published COVID-19-related works; we summarize their workflows in Figure 2, while a discussion of each of them is provided below. Specifically, the first five retrace the previously mentioned integrative studies described in [34, 36, 1, 26, 35]. Despite using different analytical strategies, these studies target similar results for: extraction of relevant genes, drug repurposing, and therapeutic targets identification.

In addition, we propose two novel scenarios, as study options for data analysts: soon, both of them will become completely feasible thanks to the increasing number of data that is being progressively collected, as a consequence of the global COVID-19 pandemic.

[Figure 2 near here]

Fig. 2 Seven scenarios of integrative studies: five extracted from literature [36, 1, 26, 35, 34] and two proposed as prompts for future studies. Note that the acronym NA stands for Network Analysis.

As we can see from Figure 2, each of the seven scenarios includes from two up to fifteen different gene expression datasets: these expression data come mainly from microarrays of patients and/or cell lines infected some years ago with viruses different from SARS-CoV-2. Conversely, among the few exceptions, we find several RNA-Seq datasets of SARS-CoV-2 infection, already available on GEO repository (GSE156544, GSE162835, GSE147507), which clearly reflect the recent transition of transcriptomics to sequencing-based methods.

Scenario 1 [36]. This meta-analysis integrates 13 different datasets of SARS-CoV, MERS-CoV, and SARS-CoV-2. Data come mostly from microarrays except for one RNA-Seq dataset of MERS-CoV and one of SARS-CoV-2. Loganathan *et al.* present a workflow aiming to propose candidate repurposable drugs against COVID-19. Specifically, DEG analysis is applied separately on SARS-CoV-2 and together on SARS-CoV and MERS-CoV data. Notice that, when not differently specified, DEG analysis compares infected/diseased samples versus mock controls and an adjusted P-value threshold is used to identify the significantly DEGs. Selected differentially expressed genes are then used to build a protein interaction network while GSEA is applied to trace significantly annotated pathways and Gene Ontology (GO) terms. Eventually, protein-drug associations are extracted for potential therapeutic targets from DrugBank (http://www.drugbank.ca/) and used in a computational method of drug repurposing.

Scenario 2 [34]. This scenario includes the highest number of datasets (15), of which almost one-third have data of SARS-CoV-2. Further, ten IAV and one SARS-CoV microarray expression datasets are integrated into the meta-analysis. Gardinassi *et al.* use DEG analysis and GSEA to assess COVID-19 samples and then to com-

pare them with cases infected with SARS-CoV-1 or IAV. In this way, they trace a core transcriptional signature that is comparable between infections caused by SARS-CoV-2, SARS-CoV, and IAV, and is enriched in cell-cycle and proliferation. Additionally, for SARS-CoV-2, they find multiple relevant immune and metabolic signatures. These, together with the core signature, could be used in future studies for a feature selection step prior to the classification of COVID-19 cases into valuable clinical classes (e.g., severity classes) or before tracing relevant clusters (i.e., groups of samples characterized by even non-trivial relationships of similarity at the transcriptional level).

Scenario 3 [1]. This scenario is analytically complex. In their integrative study, Nain *et al.* collect and compare gene expression data of the viral infection that is phylogenetically closest to COVID-19, i.e., SARS-CoV, and of three diseases that are identified as COVID-19 complications (i.e., pneumonia, severe acute respiratory syndrome, and shortness of breath and diarrhea). With such rich meta-analysis framework they can compensate for the lack of samples infected with SARS-CoV-2. A cross-comparative analysis of DEGs found in SARS-CoV and in each complication's dataset is performed. Using significantly dysregulated genes during a SARS-CoV infection, a gene-disease associations network is built and explored to obtain further useful disease-disease links; enrichment analysis of signaling pathways is then used to shed light on the molecular mechanisms underlying such disease-disease interactions. Starting from the shared DEGs, instead, the study uncovers their relationships with regulatory biomarkers (transcription factors and microRNAs). Finally, a protein interaction network is built to find highly connected (hub) genes as potential biomarkers or therapeutic targets for COVID-19, while a protein-drug/chemical agents network is built to retrieve therapeutic options worthy of further investigation.

Scenario 4 [35]. This scenario is a meta-analysis completely focused on SARS-CoV microarray data to extract relevant genes and potential therapeutic targets for

COVID-19. Particularly, Ramesh *et al.* aim to provide a deep understanding of the mechanisms behind the immune dysregulation and the cytokine storm development of the severe cases, which are phenomena shared between COVID-19 and the SARS infectious disease. DEG analysis – together with a protein interaction network – are used to identify the hub genes: among them, GSEA based on GO annotations allows to find two genes which could induce higher levels of pro-inflammatory cytokines in the lungs and could be potential therapeutic targets to avoid cytokine storm.

Scenario 5 [26]. In a meta-analysis context, this scenario integrates four expression datasets: two are of RNA-Seq (one for SARS-CoV-2 and one for IAV) and two are of microarray (for SARS-CoV and MERS-CoV). Even if authors confirm that SARS-CoV species is the phylogenetically closest one to SARS-CoV-2, each dataset is used separately for DEG analyses. Most significantly dysregulated genes are examined with GSEA of pathways and GO terms, finding enriched inflammatory and infection responses to SARS-CoV-2 infection. Additionally, to confirm that there are noteworthy disease associations of the genes involved in the SARS-CoV-2 infection response, Moni *et al.* use them to perform a survival analysis (including a Cox regression model and Kaplan Meyer survival curves) in a different – but also lung-related disease – context. This is the case of expression data of lung cancer (LC) adenocarcinoma patients. SARS-CoV-2 infection response genes succeed to stratify LC patients according to prognosis: differential expressions of the responding genes appear associated with significantly reduced survival, supporting the notion of giving critical care to classes of patients with lung-related co-morbidities, such as the LC patients.

Scenario 6. We propose to perform a meta-analysis where SARS-CoV and MERS-CoV microarray data of cell lines are integrated with already available RNA-Seq and microarray datasets for SARS-CoV-2 with primary goals of therapeutic targeting and drug repurposing. After DEG analysis for each infectious disease dataset alone,

we compare and merge (e.g., intersection) the most significantly dysregulated genes. A gene co-expression network can also be built on SARS-CoV-2 expression data and used to identify all the co-expressed genes (above a given similarity threshold) starting from the significant DEGs. The genes obtained in this way are likely to be involved in the same biological processes and may be useful in providing a broader view of the mechanisms underlying the infectious disease. Hence, we suppose to use these genes as features for a basic classifier, which is meant to recognize infected samples from mock samples, regardless of the virus. From feature importance analysis, which ranks the features/genes according to their role in the classification task, we can then extract the most characterising infectious state and use them for further GSEA of pathways and GO terms. The most interesting ones can be thus analysed building a gene-disease and a protein-drug/chemical agent interaction network, so as to find possible therapeutic targets as well as potential drugs to be repurposed.

Scenario 7. The main focus of this last proposed scenario is to develop a classifier able to distinguish COVID-19 severity classes and provide accordingly a more tailored drug repurposing. In a meta-analysis setting, we collect MERS-CoV, SARS-CoV and IAV microarray expression data of patients, featuring some annotations about severity. Then, we integrate them with an RNA-Seq dataset of SARS-CoV2 (GSE162835) that, to the best of our knowledge, is the only one already available having also metadata of disease severity [38]. We conduct the usual DEG analysis on each infectious disease separately; we repeat DEG analyses of each pairwise comparison among the three classes of COVID-19 severity (severe, moderate and mild). Using the most significant DEGs as features we can train a classifier to recognize only severe cases or, possibly, all the three classes of severity: training data may include patients infected by SARS-CoV-2 or the other viruses, but testing data should be COVID-19 cases, in order to assess the reliability of classification for this specific infectious disease. Also, the samples infected with other viruses can

be used always together with a label indicating their severity annotations, or even without any label, if they are not available. In the former case, we have a completely supervised setting, where the number of samples could be lower but each sample is annotated to a severity class; in the latter case, a semi-supervised setting is chosen, in which also samples without severity annotation contribute to acquire knowledge about the input data, despite not contributing to learning the classification task. Once the classifier is built, feature importance analysis extracts signatures consisting of the most useful genes/features to distinguish each class of severity. These signatures are examined with GSEA of pathways and GO terms. The most interesting genes, which are the ones that allow to distinguish severe cases, are considered to build gene-regulatory biomarkers and protein-drug/chemical agent interaction networks. Potential therapeutic targets and repurposable drugs can be found as a consequence.

7 Conclusions

COVID-19 is a very complex disease, and its mechanisms are far from being completely understood. For this reason, it is important to exploit all the available resources, including datasets produced with microarray technologies and referring to viruses similar to SARS-CoV-2 or diseases related to COVID-19. Many integration scenarios are feasible, as richly described in existing literature and overviewed in this chapter. Accordingly, here we proposed possible future directions for integrative studies on COVID-19, combining NGS and microarray datasets of SARS-CoV-2 but also of other infections and diseases, including machine learning techniques that can investigate in depth the hidden mechanisms of gene expression in human hosts infected by a virus. In increasingly complex scenarios, it is paramount that integration efforts are devoted to both data and knowledge aspects, with the ideal goal of reusing existing datasets and information (even if not collected ad-hoc for this research) to

increase the amount of exploitable material and improve the variety and reliability of methods useful to provide meaningful results in COVID-19 related studies.

Acknowledgements A.B. is supported by the European Research Council Executive Agency under the EU Framework Programme Horizon 2020, ERC Advanced Grant number 693174 GeCo (data-driven Genomic Computing).

References

[1] Nain Z, Rana HK, Liò P, Islam SMS, Summers MA, Moni MA (2020) Pathogenetic profiling of COVID-19 and SARS-like viruses. Briefings in Bioinformatics

[2] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. (2012) NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Research 41(D1):D991–D995

[3] Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca NA, Petryszak R, Papatheodorou I, et al. (2018) ArrayExpress update–from bulk to single-cell expression data. Nucleic Acids Research 47(D1):D711–D715

[4] Bernasconi A, Canakoglu A, Masseroli M, Pinoli P, Ceri S (2020) A review on viral data sources and search systems for perspective mitigation of COVID-19. Briefings in Bioinformatics

[5] Bernasconi A, Canakoglu A, Pinoli P, Ceri S (2020) Empowering Virus Sequence Research Through Conceptual Modeling. In: Dobbie G, Frank U, Kappel G, Liddle SW, Mayr HC (eds) Conceptual Modeling, Springer International Publishing, Cham, pp 388–402

[6] Canakoglu A, Bernasconi A, Colombo A, Masseroli M, Ceri S (2019) GenoSurf: metadata driven semantic search system for integrated genomic datasets. Database 2019, baz132

[7] Canakoglu A, Pinoli P, Bernasconi A, Alfonsi T, Melidis DP, Ceri S (2020) ViruSurf: an integrated database to investigate viral sequences. Nucleic Acids Research 49(D1):D817–D824

[8] eds2002microarray (2002) Microarray standards at last. Nature 419(323)

[9] Schuster SC (2008) Next-generation sequencing transforms today's biology. Nature methods 5(1):16–18

[10] Quackenbush J (2002) Microarray data normalization and transformation. Nature genetics 32(4):496–501

[11] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by rna-seq. Nature methods 5(7):621–628

[12] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology 28(5):511–515

[13] Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA (2012) Rapid identification of non-human sequences in high-throughput sequencing datasets. Bioinformatics 28(8):1174–1175

[14] Ma T, Liang F, Oesterreich S, Tseng GC (2017) A joint bayesian model for integrating microarray and rna sequencing transcriptomic data. Journal of Computational Biology 24(7):647–662

[15] Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020) The proximal origin of SARS-CoV-2. Nature medicine 26(4):450–452

[16] Abdelrahman Z, Li M, Wang X (2020) Comparative review of SARS-CoV-2, SARS-CoV, MERS-CoV, and influenza a respiratory viruses. Frontiers in immunology 11:2309

[17] Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A (2020) A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host & Microbe 27(4):671–680.e2

[18] Wan Y, Shang J, Graham R, Baric RS, Li F (2020) Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. Journal of virology 94(7)

[19] Kandeel M, Ibrahim A, Fayez M, Al-Nazawi M (2020) From SARS and MERS CoVs to SARS-CoV-2: Moving toward more biased codon usage in viral structural and nonstructural genes. Journal of medical virology 92(6):660–666

[20] Jang Y, Seo SH (2020) Gene expression pattern differences in primary human pulmonary epithelial cells infected with MERS-CoV or SARS-CoV-2. Archives of virology 165(10):2205–2211

[21] Bosworth A, Dowall SD, Garcia-Dorival I, Rickett NY, Bruce CB, Matthews DA, Fang Y, Aljabr W, Kenny J, Nelson C, et al. (2017) A comparison of host gene expression signatures associated with infection in vitro by the Makona and Ecran (Mayinga) variants of Ebola virus. Scientific reports 7(1):1–15

[22] Petersen E, Koopmans M, Go U, Hamer DH, Petrosillo N, Castelli F, Storgaard M, Al Khalili S, Simonsen L (2020) Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. The Lancet infectious diseases

[23] Alsamman AM, Zayed H (2020) The transcriptomic profiling of SARS-CoV-2 compared to SARS, MERS, EBOV, and H1N1. PloS one 15(12):e0243270

[24] Goh KI, Choi IG (2012) Exploring the human diseasome: the human disease network. Briefings in functional genomics 11(6):533–542

[25] Del Valle EPG, García GL, Santamaría LP, Zanin M, Ruiz EM, Rodriguez-Gonzalez A (2019) Disease networks and their contribution to disease understanding: a review of their evolution, techniques and data sources. Journal of biomedical informatics 94:103206

[26] Moni MA, Quinn JM, Sinmaz N, Summers MA (2020) Gene expression profiling of SARS-CoV-2 infections reveal distinct primary lung cell and systemic immune infection responses that identify pathways relevant in COVID-19 disease. Briefings in Bioinformatics

[27] Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. science 313(5795):1929–1935

[28] Li J, Lu Z (2012) A new method for computational drug repositioning using drug pairwise similarity. In: 2012 IEEE International Conference on Bioinformatics and Biomedicine, IEEE, pp 1–4

[29] Ceddia G, Pinoli P, Ceri S, Masseroli M (2020) Matrix Factorization-based Technique for Drug Repurposing Predictions. IEEE journal of biomedical and health informatics 24(11):3162–3172

[30] Kodama Y, Mashima J, Kosuge T, Kaminuma E, Ogasawara O, Okubo K, Nakamura Y, Takagi T (2017) DNA Data Bank of Japan: 30th Anniversary. Nucleic Acids Research 46(D1):D30–D35

[31] Bono H (2020) All of gene expression (AOE): an integrated index for public gene expression databases. PloS one 15(1):e0227076

[32] Cannizzaro G, Leone M, Bernasconi A, Canakoglu A, Carman MJ (2020) Automated Integration of Genomic Metadata with Sequence-to-Sequence Models. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, (*in print*)

[33] Chen G, Ramírez JC, Deng N, Qiu X, Wu C, Zheng WJ, Wu H (2019) Restructured GEO: restructuring Gene Expression Omnibus metadata for genome dynamics analysis. Database 2019, bay145

[34] Gardinassi LG, Souza COS, Sales-Campos H, Fonseca SG (2020) Immune and Metabolic Signatures of COVID-19 Revealed by Transcriptomics Data Reuse. Frontiers in Immunology 11:1636

[35] Ramesh P, Veerappapillai S, Karuppasamy R (2020) Gene expression profiling of corona virus microarray datasets to identify crucial targets in COVID-19 patients. Gene reports 22:100980

[36] Loganathan T, Ramachandran S, Shankaran P, Nagarajan D, et al. (2020) Host transcriptome-guided drug repurposing for COVID-19 treatment: a meta-analysis based approach. PeerJ 8:e9357

[37] He Y, Yu H, Ong E, Wang Y, Liu Y, Huffman A, Huang Hh, Beverley J, Hur J, Yang X, et al. (2020) CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Scientific Data 7(1):1–5

[38] Jain R, Ramaswamy S, Harilal D, Uddin M, Loney T, Nowotny N, Alsuwaidi H, Varghese R, Deesi Z, Alkhajeh A, et al. (2021) Host transcriptomic profiling of covid-19 patients with mild, moderate, and severe clinical outcomes. Computational and structural biotechnology journal 19:153–160

[39] Zoubarev A, Hamer KM, Keshav KD, McCarthy EL, Santos JRC, Van Rossum T, McDonald C, Hall A, Wan X, Lim R, et al. (2012) Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. Bioinformatics 28(17):2272–2273

Table 1 A collection of datasets that can be suitably employed by data analysts to perform COVID-19-related cross-species gene expression profiles analyses. Datasets are sorted by virus or disease; #inf and #cont indicate respectively the number of infected and control samples used in the experiment; almost all the listed platforms are used in their single channel mode (this information was not available for all models); #genes indicates the number of unique genes represented in the platform. For Agilent and Illumina models we retrieve the information from the Gemma database [39], for Affymetrix from the documentation of the manufacturer. GSE5972 is an exception, as we alternatively provide the number of cDNA clones from IMAGE.

GSE ID	Virus/disease	Sample type (cell line/patient)	#inf	#cont	Platform	#genes	Year	PubMed
GSE156544	SARS-CoV2	Primary epithelial organoids from colon	4	4	Agilent-026652 Whole Human Genome Microarray 4x44K v2	21278	2021	-
GSE30589	SARS-CoV	Primate cells using human platform	24	9	Affymetrix Human Genome U133 Plus 2.0 Array	38,500	2011	22028656
GSE5972	SARS-CoV	Patients	60	10	UHNMAC Homo sapiens 19K Hu19Kv8	19,200	2007	17537853
GSE1739	SARS-CoV	Patients	10	4	Affymetrix Human HG-Focus Target Array	8,500	2005	15655079
GSE47963	SARS-CoV	HAE cultures	89	102	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	18498	2013	23935999
GSE33267	SARS-CoV	2B4 cells (clonal deriv. of Calu-3 cells)	33	33	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	18498	2011	23365422
GSE17400	SARS-CoV	2B4 cells (clonal deriv. of Calu-3 cells)	9	9	Affymetrix Human Genome U133 Plus 2.0 Array	38,500	2010	20090954
GSE48142	SARS-CoV	2B4 cells (clonal deriv. of Calu-3 cells)	24	22	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	18498	2014	-
GSE37827	SARS-CoV	2B4 cells (clonal deriv. of Calu-3 cells)	30	30	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	18498	2012	-
GSE56677	MERS-CoV	2B4 cells (clonal deriv. of Calu-3 cells)	18	15	Agilent-039494 SurePrint G3 Human GE v2 8x60K Microarray 039381	21087	2014	25534508
GSE45042	MERS-CoV	2B4 cells (clonal deriv. of Calu-3 cells)	17	15	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	18498	2013	23631916
GSE100496	MERS-CoV	Human fibroblasts	25	25	Agilent-026652 Whole Human Genome Microarray 4x44K v2	21278	2017	-
GSE86528	MERS-CoV	Human fibroblasts	25	25	Agilent-026652 Whole Human Genome Microarray 4x44K v2	21278	2016	-
GSE100509	MERS-CoV	Human fibroblasts	25	25	Agilent-026652 Whole Human Genome Microarray 4x44K v2	21278	2017	-
GSE79172	MERS-CoV	Human fibroblasts	15	14	Agilent-026652 Whole Human Genome Microarray 4x44K v2	21278	2016	-
GSE81909	MERS-CoV	Human fibroblasts	25	25	Agilent-026652 Whole Human Genome Microarray 4x44K v2	21278	2016	-
GSE100504	MERS-CoV	Primary human airway epithelial cells	5	5	Agilent-026652 Whole Human Genome Microarray 4x44K v2	21278	2017	-
GSE65574	MERS-CoV	2B4 cells (clonal deriv. of Calu-3 cells)	15	3	Agilent-026652 Whole Human Genome Microarray 4x44K v2	21278	2015	28830941
GSE34205	IAV	Patients	28	12	Affymetrix Human Genome U133 Plus 2.0 Array	38,500	2012	22398282
GSE6269	IAV	Patients	18	7	Affymetrix Human Genome U133A Array	14,500	2007	17105821
GSE20346	IAV	Patients	19	18	Illumina HumanHT-12 V3.0 expression beadchip	16643	2011	21408152
GSE29366	IAV	Patients	19	12	Illumina HumanWG-6 v3.0 expression beadchip	16643	2015	-
GSE40012	IAV	Patients	39	18	Illumina HumanHT-12 V3.0 expression beadchip	16643	2012	22898401
GSE38900	IAV	Patients	16	39	Illumina HumanWG-6 v3.0 expression beadchip	16643	2013	24265599
GSE52428	IAV	Patients	41	0	Affymetrix Human Genome U133A 2.0 Array	14,500	2013	23326326
GSE61754	IAV	Patients	66	22	Illumina HumanHT-12 V4.0 expression beadchip	18366	2014	25345603
GSE68310	IAV	Patients	747	133	Illumina HumanHT-12 V4.0 expression beadchip	18366	2015	26070066
GSE90732	IAV	Patients	86	22	Illumina HumanHT-12 V4.0 expression beadchip	18366	2017	28595644
GSE14841	DRA	Patients	5	4	Affymetrix Human Genome U133 Plus 2.0 Array	38,500	2009	33244004
GSE103119	PNA	Patients	152	20	Illumina HumanHT-12 V4.0 expression beadchip	18366	2017	30425971
GSE137268	SHOB	Patients	54	15	Illumina humanRef-8 v2.0 bead chip	11950	2019	-

Table 2 Example integrative studies described in recent papers, employing several datasets from Table 1, Note that, among the microarray GSE IDs, GSE47962 was not previously inserted in Table 1 as it is subseries of an already included series. In addition, here we include also NGS experiments not previously mentioned.

Publication/date	Datasets (by GEO Series ID)	Integration levels
Loganathan *et al.* [36] 10 June 2020	**NGS**: GSE147507, GSE122876 **Microarray**: GSE17400, GSE45042, GSE47962, GSE33267, GSE37827, GSE100496, GSE86528, GSE100509, GSE79172, GSE81909, GSE48142	D1 K1 (SARS-CoV-2 + similar viruses) K3
Gardinassi *et al.* [34] 26 June 2020	**NGS**: GSE147507 **Microarray**: GSE1739, GSE34205, GSE6269, GSE20346, GSE29366, GSE40012, GSE38900, GSE52428, GSE61754, GSE68310, GSE90732	D1 D1b K1 (SARS-CoV-2 + similar viruses) K3
Nain *et al.* [1] 11 August 2020	**Microarray**: GSE30589, GSE103119, GSE14841, GSE137268, GSE1739	D2 (SARS, DRA, PNA, SHOB) K1 (only similar viruses) K2 (SARS + other diseases) K3
Ramesh *et al.* [35] 27 November 2020	**Microarray**: GSE33267, GSE1739	K1 (only similar viruses) K3
Moni *et al.* [26] 18 December 2020	**NGS**: GSE147507, GSE89008 **Microarray**: GSE47963, GSE100504	D1 K1 (SARS-CoV-2 + similar viruses) K3 K4 (survival in lung adenocarcinoma)