Briefings in Bioinformatics, 22(2), 2021, 701–713

doi: 10.1093/bib/bbaa328
Advance Access Publication Date: 7 December 2020
Articles

A review of COVID-19 biomarkers and drug targets: resources and tools

Francesca P. Caruso[†], Giovanni Scala[†], Luigi Cerulo*, Michele Ceccarelli

Corresponding author: Luigi Cerulo. Tel: +39 0824 305154; Fax: +39 0824 305143. E-mail: lcerulo@unisannio.it;
Michele Ceccarelli. Tel: +39 0817683787; Fax: +390817683787. E-mail: michele.ceccarelli@unina.it
[†]These authors contributed equally to this work.

Abstract

The stratification of patients at risk of progression of COVID-19 and their molecular characterization is of extreme importance to optimize treatment and to identify therapeutic options. The bioinformatics community has responded to the outbreak emergency with a set of tools and resource to identify biomarkers and drug targets that we review here. Starting from a consolidated corpus of 27 570 papers, we adopt latent Dirichlet analysis to extract relevant topics and select those associated with computational methods for biomarker identification and drug repurposing. The selected topics span from machine learning and artificial intelligence for disease characterization to vaccine development and to therapeutic target identification. Although the way to go for the ultimate defeat of the pandemic is still long, the amount of knowledge, data and tools generated so far constitutes an unprecedented example of global cooperation to this threat.

Introduction

The crisis generated by the spread of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the corresponding COVID-19 disease was declared a pandemic by the World Health Organization on 11 March 2020. The origin of SARS-CoV-2 was traced to the Huanan Seafood Wholesale Market in the city of Wuhan, China. The causative pathogen was identified as a beta-coronavirus with high sequence homology to bat coronaviruses (CoVs) using angiotensin-converting enzyme 2 (ACE2) receptor as the dominant mechanism of cell entry [1]. Human-to-human transmission events were confirmed with clinical presentations ranging from no symptoms to mild fever, cough and dyspnea to cytokine storm, respiratory failure and death. The scientific

Biogem, Istituto di Biologia e Genetica Molecolare, Via Camporeale, Ariano Irpino, Italy
Department of Electrical Engineering and Information Technology, University of Naples ”Federico II”, Via Claudio,
82100, Naples, Italy
Department of Biology, University of Naples ”Federico II”, Via Monte Sant’Angelo, 82100, Naples, Italy
Department of Science and Technology, University of Sannio,Benevento, Italy

Francesca Pia Caruso is a postdoctoral researcher at the University of Naples ‘Federico II’. Her research interests focus on cancer genomic and its interactions with host immune system.

Giovanni Scala is a bioinformatics researcher at the Department of Biology of the University of Naples Federico II. His interests are in the definition of systems biology models for the response of biological systems to external stimuli by using integrative statistical methods and machine learning based analyses of omics data.

Luigi Cerulo is an associate professor of computational biology at University of Sannio in Benevento. His research focuses on machine learning in sequence analysis and cancer bioinformatics. He contributed recently in the development of supervised approaches for the classification of non coding RNA sequences and the identification of biomarkers in cancer contexts.

Michele Ceccarelli is a full professor of Computer Science and Engineering at University of Naples ‘Federico II’ where he serves as chair of PhD program in Computational and Quantitative Biology. Michele has given several contributions in Cancer Genomics and Computational Systems Biology.

Submitted: 31 July 2020; **Received (in revised form):** 5 October 2020

community responded to the crisis with an extraordinary effort involving thousands of scientists and hundreds of laboratories worldwide. This produced a vast amount of biological data allowing the computational biology community to characterize the molecular bases of the diseases, the spread and evolution of the virus and the identification of potential drugs.

The identification of biomarkers for stratification of patients at risk of progression of COVID-19 and their molecular characterization is of extreme importance to optimize treatment and to identify therapeutic options.

We refer to a biomarker as a measurable characteristic—e.g. expression level of a group of genes—used as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention [2, 3]. Depending on the context of use, a biomarker can be categorized as susceptibility/risk, diagnostic, monitoring, prognostic, predictive, pharmacodynamic/response and safety biomarker. It is important to distinguish between the prognostic biomarkers that are useful to identify patients more likely to have a particular outcome independently from treatment and predictive biomarkers that involve a comparison of a treatment to a control in patients with and without the biomarker.

Several prognostic COVID-19 biomarkers predicting disease severity have been already validated in clinical settings [4]. Among biomakers that segregate severe from non-severe patients, obtained by retrospective analysis of large cohorts, of particular interest are those associated to dysregulation of immune response. Infection-related biomarkers, such as inflammatory cytokines TNFα, interleukines IL-2R and IL-6 and other blood cell counts, are seen in much higher dosage in severe groups with respect to the non-severe group [5], whereas the platelet count tends to be significantly decreased in severe cases [6]. Genomewide association studies have also identified a gene cluster on chromosome 3 as a the major genetic risk factor for severe SARS-CoV-2 infection and hospitalization [7, 8]. This genomic segment of 50 kb is inherited from Neanderthals and is carried by about 50% of people in South Asia and about 16% of people in Europe today [9]. Other prognostic biomarkers of disease progression and mortality are related to cardiovascular damages involved in COVID-19 and make use of the cardiac troponin [10] or to the occurrence of chronic kidney diseases where an increase of creatinine levels is observed in severe patients [11]. Other than these clinical biomarkers, there is already a vast literature of molecular biomarkers that characterize the disease associated with SARS-CoV-2 viral infection and that can be exploited to identify therapeutic targets.

In this paper, we focus on Bioinformatics resources, tools and approaches connected to molecular COVID-19 biomarkers. To this aim, we needed to address the vast amount of information produced by the recent explosion of COVID-19-related scientific literature.

The paper is organized following the induced set of biomarker-related topics as follows: the next section describes the methods adopted to mine COVID-19-related scientific literature and to extract relevant topics; in Section 5, we report machine learning tools developed to characterize COVID-19 disease, especially from the image scans; Section 6 describes the relevant molecular datasets available for the characterization of COVID-19 biomarkers from genomics and proteomics profiling; Section 7 focuses on immune repertoire sequencing and antibody isolation; Section 8 collects methods and tools related to vaccine development; and finally Section 9 reports approaches and tools for the discovery of therapeutic targets.

Methods adopted to mine COVID-19 literature

Topic modeling

We adopted latent Dirichlet analysis to extract relevant topics from over 27 000 research papers, appeared in the past 10 months and indexed in PubMed or uploaded on preprint servers, such as bio and med rxiv [12]. The overall procedures implemented in Python and R, including details about the adopted analysis, are reported at https://github.com/bioinformatics-sannio/covidLiterature. We started with a set of 27 894 articles downloaded on 20 June 2020 from LitCovid, a curated open-resource literature of PubMed research papers related to COVID-19 [13] and from the COVID-19/SARS-CoV-2 collection of medRxiv and bioRxiv preprint servers. The document text content, composed by joining article's title and abstract, was tokenized, stemmed and filtered by stopwords. Duplicates, due to PubMed edited paper available also on preprint servers, have been removed by comparing vectors of term frequencies with cosine distance obtaining a consolidated corpus of 27 570 papers. In this corpus, we discovered an optimal set of 36 topics showing the lowest perplexity (Supplemental Figure 1). Among them, we selected five topics, topic #0, strongly related to computational models, and four topics related to biomarker research (Figure 1). From the consolidated corpus, we selected papers with a content associated with such a set of topics. Specifically, we considered papers not distributed on many topics (i.e. Shannon entropy less than half of its maximum $\frac{1}{2}log2(\frac{1}{36})$) and having one of the biomarker-related topics shown in (Figure 1, Table S1) as the top most probable. The final set of 3032 papers, which we made available as Table S2, was manually evaluated and the most relevant discussed in this work.

Attention of studied elements

COVID-19 literature can also be mined to extract valuable information regarding molecular elements (i.e. gene, proteins, etc.) that received more attention in this particular subset of scientific literature. Here we considered gene attention and reported two different analyses: the first showing genes that received more attention in the selection of manuscripts reported in this review compared to all manuscript published in the same time-frame and the second showing genes that received more attention in the first half of 2020 (the pandemic time-window) compared to the whole 2019. The concept of attention can be formulated in different ways; here we choose the number of manuscripts citing a gene as a proxy for the attention received by the gene.

The association between genes and citing manuscripts can be obtained by the NCBI NIH gene2pubmed table [14] while the temporal information associated to manuscript was obtained using the NCBI NIH PMC-ids table [15] and the RISmed R package [16]. For the analyses presented in this review, we only considered human, mouse, rat and SARS-CoV-2 genes by filtering the gene2pubmed table for accession IDs of genes annotated for these species and mapping the corresponding ENTREZ gene IDs to gene symbols. Given a set of manuscripts, we computed the attention score of a given gene in that set, by summing up the number of times it was cited the manuscripts from the set.

For the 1st analysis, we select 15 652 gene/manuscript associations from the filtered gene/manuscript table, covering 2904 manuscripts published from the 1 January 2020 to 17 June 2020. This latter set of articles was intersected with the selection of COVID-19 manuscripts provided here, generating a partition of 182 COVID-19 gene citing manuscripts and 2722 non-COVID-19 gene-citing manuscripts. For each symbol, we than compared the number of times COVID-19 manuscripts cited the gene with the corresponding number of citations in non-COVID-19 manuscripts and computed its statistical enrichment by using a Fisher's exact test and finally correcting all *P*-values with False Discovery Rate (FDR) correction.

For the 2nd analysis, we selected 76 658 gene/manuscript associations from the gene/manuscript table, with 16 480 associations covering 3324 manuscripts published during the year 2020 and 60 178 gene/manuscript associations from the same table, covering 20 208 manuscripts published in the year 2019. We then selected the genes being cited in at least five manuscripts during 2020 and ranked them based on their attention score in each considered year (2019 and 2020). We defined Δ–*rank* as the difference (positive or negative) in rank of each paper between the 2 years as a measure of gain or loss of attention for each gene between 2019 and 2020 was. We assigned an empiric *P*-value to the Δ–*rank* of each gene using a bootstrap procedure (1000 iteration) where the same procedure describe above was applied to a random selection of 16 480 gene/manuscript associations and shuffling of gene IDs with respect to manuscripts in each realization.

Genes/proteins with high attention score in the COVID-19 infection process

Genes that received significant attention in the subset of scientific literature considered are shown in Figure 1A–B, both in terms of enrichment of attention score and significant Δ–*rank*, include important genes involved in SARS-CoV-2 parthenogenesis (Figure 1C). The main mechanism of adhesion and viral entry into the cell involves the viral protein Spike (S), which binds the human ACE2 receptor through its receptor-binding domain (RBD) with a binding affinity 10 times higher than that of the spike protein of the SARS virus. The very efficient cellular entry of SARS-CoV-2 is also due to the action of the Furin enzyme that is expressed in significant concentrations in the lung and activate the spike protein [17, 18]. Some recent evidence suggests that many other genes may contribute to virus entry and are being studied as potential therapeutic targets in the treatment of coronavirus infections. For example, the host cell protease TMPRSS2 acts as a primer for the spike protein [19, 20]; the membrane protein DPP4 acts as a co-receptor of SARS-CoV-2 and is a key factor for the hijacking and virulence in the respiratory tract [21]; the AAK1 gene is a known regulator of the clathrin-mediated endocytosis [22]. The uncontrolled and excessive release of pro-inflammatory cytokines and chemokines (like IL-1β, IL-6, IL-12, CXCL8, CXCL9, CXCL10, IFNs, TNF, etc.) is the most damaging and potentially fatal effect related to the COVID-19 and therefore it is the subject of several studies. The IL-6 gene is the main prognostic biomarkers since it plays a key role in cytokine storm, and high levels of this cytokine are associated with respiratory failure and mortality risk [23]. Unfortunately, the efficacy of cell-mediated immunity against SARS-CoV-2 is still unclear and many studies are aimed at clarifying the role of T cells in the resolution of COVID-19 [24]. Some recent evidence has shown an increase in the expression of the CD8 T cell marker (CD8A) in COVID-19 patients to support hyper-activation of cytotoxic T lymphocytes [25].

Main topics in COVID-19 biomarker research

We adopted a semiautomatic approach, based on topic analysis, to select from over 27 000 research papers, appeared from September 2019 and indexed by PubMed or uploaded on preprint servers such as bio- and med- *rxiv*, a manageable set of resources that can be manually revised. From the overall corpus of documents, we induced 36 relevant topics, 5 of which are associated to biomarkers and are depicted in Figure 2, whereas the breakdown of the papers per topic is summarized in Table S1.

Topic #0 refers to the use of artificial intelligence (AI), in particular deep learning approaches for the analysis of biomedical images, such as computed tomography (CT) scans or ultrasonography (LUS) images, to diagnose and predict the prognosis of COVID-19 patients. Topic #1 is related to the study of neutralizing antibodies and cellular immune response to SARS-CoV-2 and focuses on the design of serological tests to identify seroconversion prognostic biomarkers. Topic #20 is about drug discovery and is specific to structural and functional analysis of SARS-CoV-2 to identify therapeutic targets. Topic #27 is related to the discovery of biomarkers that trigger an immune response and could be adopted for vaccine development. Topic #33 encloses genome- and proteome-wide studies with publicly available datasets, a valuable source of information for biomarker discovery.

Machine learning and AI for image-based disease characterization

Imaging is the main tool for the identification of patients with higher risks of developing acute respiratory failure due to SARS-CoV-2 virus pneumonia [26]. Lesion characteristics such as number, size, density and bilateral and multi-lobar glass ground opacifications (mainly posteriorly and/or peripherally distributed) are indicators of lung damage and remaining lung reserve [27]. They are effectively used as biomarkers to train an automatic diagnostic system or to assist the accurate diagnosis of disease severity and to distinguish between normal and SARS-CoV-2 virus pneumonia. In [28] the authors collected a dataset of 532 506 CT scans from 3777 patients for the purpose of training a diagnostic system (Table 1) and showed that a convolutional neural network, adapted from 3D ResNet-18, trained on lung-lesion maps, obtained by different automatic segmentation algorithms, achieves 92.49% accuracy, 94.93% sensitivity, 91.13% specificity and an area under the curve (AUC) of 0.9797 [29]. The use of multiple features, such as texture, surface, volume histogram and intensity, has also been shown to improve the diagnostic accuracy [30] of chest CT scans up to 93.9%. As an alternative to CT scans, lung ultrasound (LUS) has been shown to be a more widely available, cost-effective, safe and real-time imaging technique [31].

Virus and host genomics, transcriptomics and proteomics profiling

The genomic sequence of SARS-Cov-2 has 29 903 nucleotides [1] and is available with accession number NC_045512.2. It has 89.1% similarity with a bat SARS-like coronavirus (CoV) isolate-bat SL-CoVZC45 (accession number MG772933) and is organized in *replicase ORF1ab* (21,291 nt), *spike* (3,822 nt), *ORF3a* (828 nt), *envelope* (228 nt), *membrane* (669 nt) and *nucleocapsid* (1260 nt). As of 21 June 21 2020, a total number of 49 239 sequences have been deposited on GISAID EpiFlu Database (www.gisaid.org), which is the main source of genomic data associated with SARS-Cov-2 [32]. To get insight into the complex pathogenesis caused by novel coronavirus, sequencing of single cells (scRNA-seq), RNA (RNA-seq), adaptive immune receptor repertoire (AIRR-seq), image datasets and proteomic assay have been massively adopted to unveil the

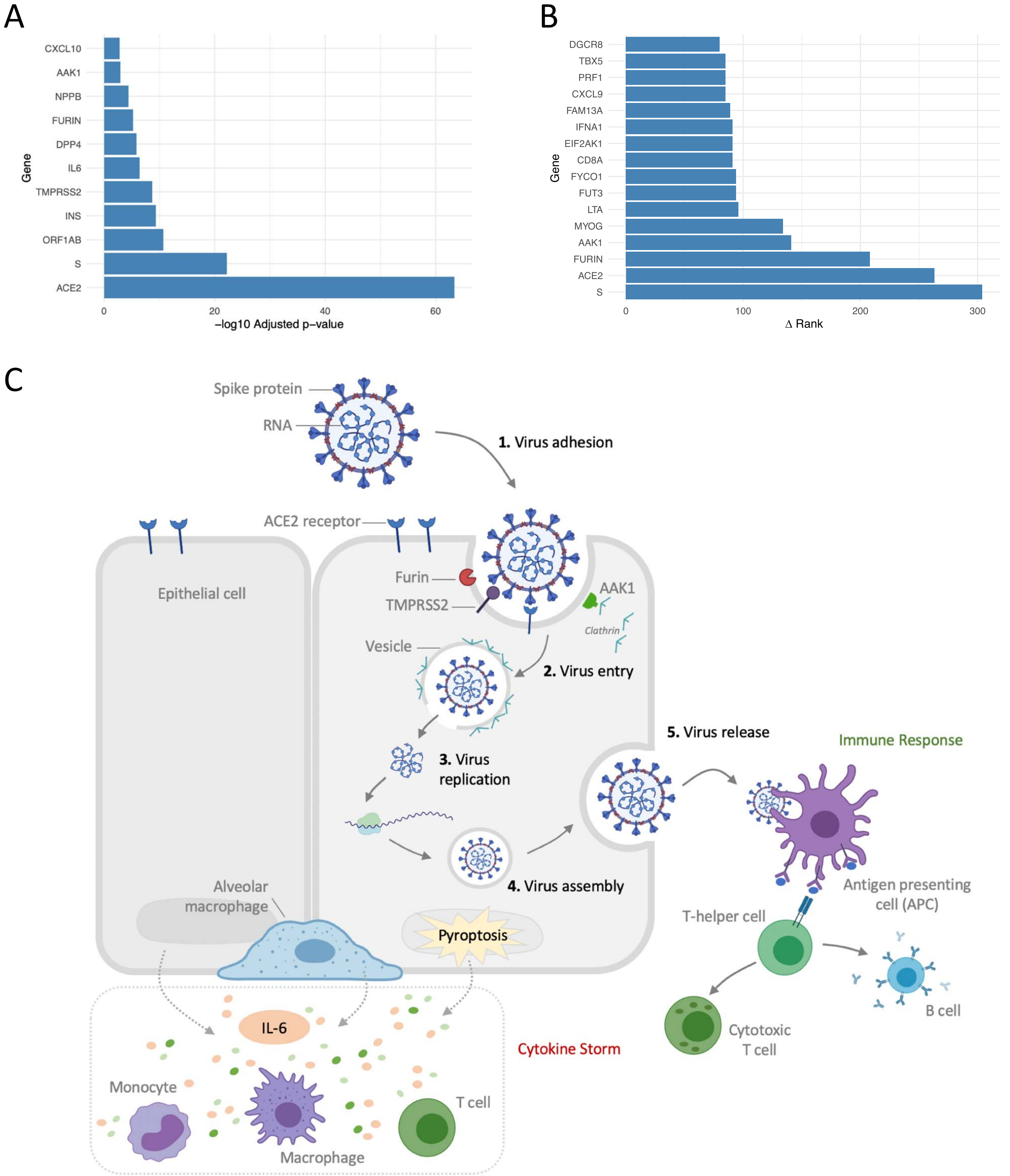

Figure 1. (A) the P-value of the gene with significant attention score enrichment, (B) right panel shows the genes with a significant (P-value ¡ 0.1) Δ−*rank*. (C) SARS-CoV-2 infection. The Sars-Cov-2 virus adheres to and enters human cells through the interaction of the viral Spike protein and the human ACE2 receptor. The virus entry mechanism is favored by the presence of some cleaving enzymes, such as TMPRSS2 and Furin, which activate the Spike protein. Virus endocytosis within clathrin-coated vesicles is regulated by the AAK1 gene. The release of the viral RNA, the subsequent replication and assembly of new particles cause pyroptosis of the host cell and the release of damage-associated molecular patterns. These molecules are recognized by adjacent cells that secrete pro-inflammatory cytokines and chemokines. The pro-inflammatory stimulus attracts monocytes, macrophages and T cells to the site of infection, which contribute to the inflammatory process with a positive feedback loop. In the physiological immune response, antigen-presenting cells engulf the viral particles and stimulate the activation of T-helper cells. The latter trigger the adaptive immune response by stimulating B cells to produce antibodies against the virus and T cytotoxic cells that recognize and destroy other virus-infected cells. On the other hand, the accumulation of immune cells at the site of infection due to excessive pro-inflammatory stimulus, such as the release of IL-6, causes the cytokine storm, damage to lung tissues and increases the risk of death.

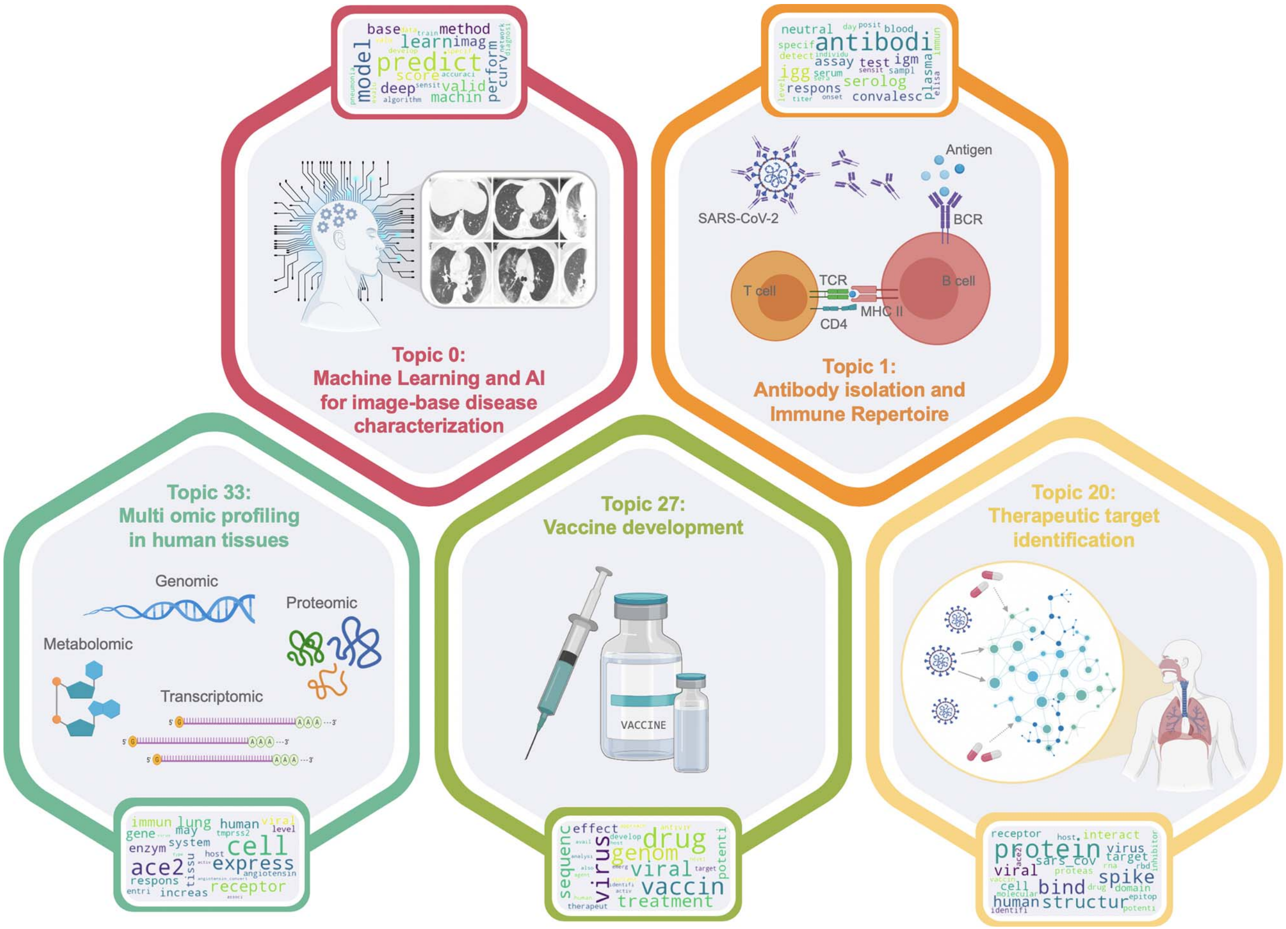

Figure 2. Main topics associated to biomarkers research activity automatically extracted from a corpus of 27 894 scientific papers.

characteristics of the immune response triggered in patients affected by COVID-19. Single cell sequencing is often combined with RNA or AIRR sequencing as the pulmonary microenvironment and peripheral immune response allow to reveal potential mechanisms underlying the pathogenesis of COVID-19 and the identification of diagnostic and therapeutic biomarkers. Most of data are available on public databases, such as Gene Expression Omnibus (GEO) [33], Sequence Read Archive (SRA) [34], European Nucleotide Archive (ENA) [35], European Genome-phenome Archive (EGA) [36] and Genome Sequence Archive (GSA) [37]. Single cell transcriptomic data can be interactively explored through the Single Cell Portal [38]. Table 1 provides a curated list of 28 transcriptomic, 2 image datasets and 6 proteomic studies, publicly available datasets. The list contains 14 scRNA-seq studies derived from peripheral blood mononuclear cells (PBMCs) (*n* = 8), nasopharyngeal swabs and bronchial branches (*n* = 1), bronchoalveolar lavage fluid (BALF) (*n* = 1) and lung tissue (*n* = 1) in COVID-19 patients. There are also scRNA-seq datasets of lung organoids (*n* = 2) and human cell lines infected with SARS-CoV-2 (*n* = 3). Similarly, there are 9 RNA-seq studies that include datasets of infected human cell lines (*n* = 4) and organoids (*n* = 3), nasopharyngeal swabs (n= 1), BALF and PBMC (*n* = 1) and several tissue (i.e. lung, heart, liver, kidney, bowel, skin, fat, marrow) (*n* = 2) from COVID-19 patients. The AIRR-seq datasets, including data of BCR, TCR, IGH and antibody sequencing, were derived from PBMCs (n = 10) and BALF (n = 1) in COVID-19 patients.

Proteomics datasets are created in this context to characterize the set of SARS-CoV-2 encoded proteins and to investigate their interaction with the human proteome during the different phases of the infection. Gordon *et al.* [71] recently developed a protein interaction map by expressing all of the 29 SARS-CoV-2 proteins in human cells and then assessing their affinity with human proteins by means of affinity-purification mass spectrometry, obtaining a list of 332 SARS-CoV-2-human protein–protein interactions that is available as a supplementary file at [72]. Bojkova *et al.* analyzed human cell lines infected with SARS-CoV-2 [73] and characterized their translatome and the proteome at different time points after the infection and made this dataset available at [74]. Li *et al.* [75] by using genome wide yeast-two hybrid and co-immunoprecipitation approach, identified 58 distinct intra-viral protein–protein interactions. In the same study, the authors studied the viral–host interactome by over-expressing all the SARS-CoV-2 genes into HEK293 cells and defined a list of 631 viral–host protein–protein. Interaction data from this work are available in the IntAct database (imex:IM-27901).

Knowledge of the SARS-CoV-2-encoded proteins structure can be exploited to search for molecules showing structural affinity and hence acting as potential inhibitors of these latter. Protein Data Bank is a public resource collecting user deposited structures of all of the 29 COVID-19-related PDB structures [76]. Using used model-validation metrics Wlodawer *et al.* [77] defined a refined version of COVID-19-related PDB structures present in

Table 1. Transcriptomics and proteomics datasets

			# samples		
Ref	Data availability	Biotype	COVID-19	Control	Data type
[39]	GSE150728	Peripheral blood mononuclear cells (PBMCs)	7	6	scRNAseq
[40]	GSE148697—pending	hPSC-derived lung organoids	na	na	scRNAseq
[41]	Pending	SARS-CoV-2 infected human bronchial epithelial cells	na	na	scRNAseq
[42]	CRA002509—pending	PBMCs	2	na	scRNAseq
[43]	EGAS00001004481	Nasopharyngeal and bronchial	19	5	scRNAseq
[44]	Under request	PBMCs	10	na	scRNAseq
[45]	Pending	PBMCs	4	na	scRNAseq
[46]	CNP0001102	PBMCs	16	3	scRNAseq
[47]	GSE147507	SARS-CoV-2 infected human cell lines and Lung	23	2	RNAseq
[48]	CRA002390	Bronchoalveolar lavage fluid (BALF) and PBMCs	7	3	RNAseq
[49]	GSE152075	Nasopharyngeal	430	54	RNAseq
[50]	GSE150392	SARS-CoV-2 infected iPSC-cardiomyocyte cells	3	3	RNAseq
[51]	GSE150819	SARS-CoV-2 infected human bronchial organoids	6	9	RNAseq
[52]	GSE150316	Various	83	5	RNAseq
[53]	GSE149312	SARS-CoV-2 infected intestinal organoids	8	10	RNAseq
[54]	PRJNA628125	PBMCs	14	na	AIRRseq
[55]	PRJNA630455	PBMCs	42	na	AIRRseq
[56]	PRJNA633317	PBMCs	120	na	AIRRseq
[57]	Web page at [58]	PBMCs	149	na	AIRRseq
[59]	Web page at [60]	PBMCs	na	na	AIRRseq
[61]	Pending	PBMCs	na	na	AIRRseq
[62]	PRJEB38339	PBMCs	215	na	AIRRseq
[63]	GSE148729	SARS-CoV-1/2 infected human cell lines	167	na	RNAseq + scRNAseq
[64]	GSE151803	SARS-CoV-2 infected human cell lines, organoids and lung	12	9	RNAseq + scRNAseq
[65]	Pending	PBMCs	na	na	AIRRseq + scRNAseq
[66]	Pending	PBMCs	na	na	AIRRseq + scRNAseq
[67]	EGAS00001004412	PBMCs	na	na	AIRRseq + scRNAseq
[68]	GSE145926	BALF	12	9	AIRRseq + scRNAseq
[28]	Web page at [69]	Chest scan of COVID-19 patients and normal controls	1386	1105	CT images
[31]	Web page at [70]	LUS images of COVID-19 patients and normal controls	na	na	LUS images
[71]	Web page at [72]	Infected human kidney derived cell lines	na	na	interactome
[73]	Web page at [74]	Infected human colon derived cell lines	na	na	Translatome + Proteome
[75]	Intact imex:IM-27901	PBMCs	na	na	interactome
[76]	Web page at [76]	Various	na	na	Protein structures
[77]	Web page at [78]	Various	na	na	Protein structures
[79]	IPX0002106000 and IPX0002171000	Blood serum	46	53	Proteomic and metabolomic

Protein Data Bank and made them available at [78]. As of July 2020, this repository hosts 285 SARS-CoV-2 protein structures and 23 additional structures of other coronaviruses. A recent study, based on the analysis of proteomic and metabolomic profiles from COVID-19 patients, identified possible biomarkers related to the severity of the pathology [79]. The machine learning-based approach has highlighted important changes in the serum of COVID-19 patients involving the deregulation of complement system processes, macrophage and platelet activity and metabolic suppression. All data are deposited in ProteomeXchange Consortium (Table 1).

Immune repertoire sequencing and antibody isolation

SARS-CoV-2 infection affects adaptive immunity, immune cell architecture and function [80]. Exposure to viral antigens stimulates the cellular immune response of T cells and the humoral immune response of B cells, which can be studied in detail through the immune repertoire high-throughput sequencing. The analysis of the sequences of T and B cells repertoires for different cohorts of patients, from non-hospitalized infected patients to patients with severe symptoms, may reveal the nature of protective versus detrimental B and T cell responses and can be be used as a prognostic biomarker. For example, significant highly clonal T cell repertoires in active COVID-19 patients versus patients recovered from COVID-19 without medical intervention has been recently reported [62]. The Adaptive Immune Receptor Repertoire Community (AIRR-C) has defined standards for sharing and interoperability of B-cell and T-cell receptor repertories [81], and sequences of are being deposited in multiple repositories such as [82] which (at the date of writing this paper) contains 178 190 149 sequences from 285 patients.

T and B cell sequencing is important for the development of monoclonal antibodies against SARS-CoV-2 but also to determine the optimal T cell engagement strategy for vaccine development. SARS-CoV-2-reactive and neutralizing antibodies have now been isolated from COVID-19 survivors. Neutralizing antibodies could block viral entry by preventing the S protein from binding to host cell receptors, such as ACE2. Neutralizing antibodies could also mimic receptor binding and prematurely trigger fusogenic conformational changes in the S protein before it

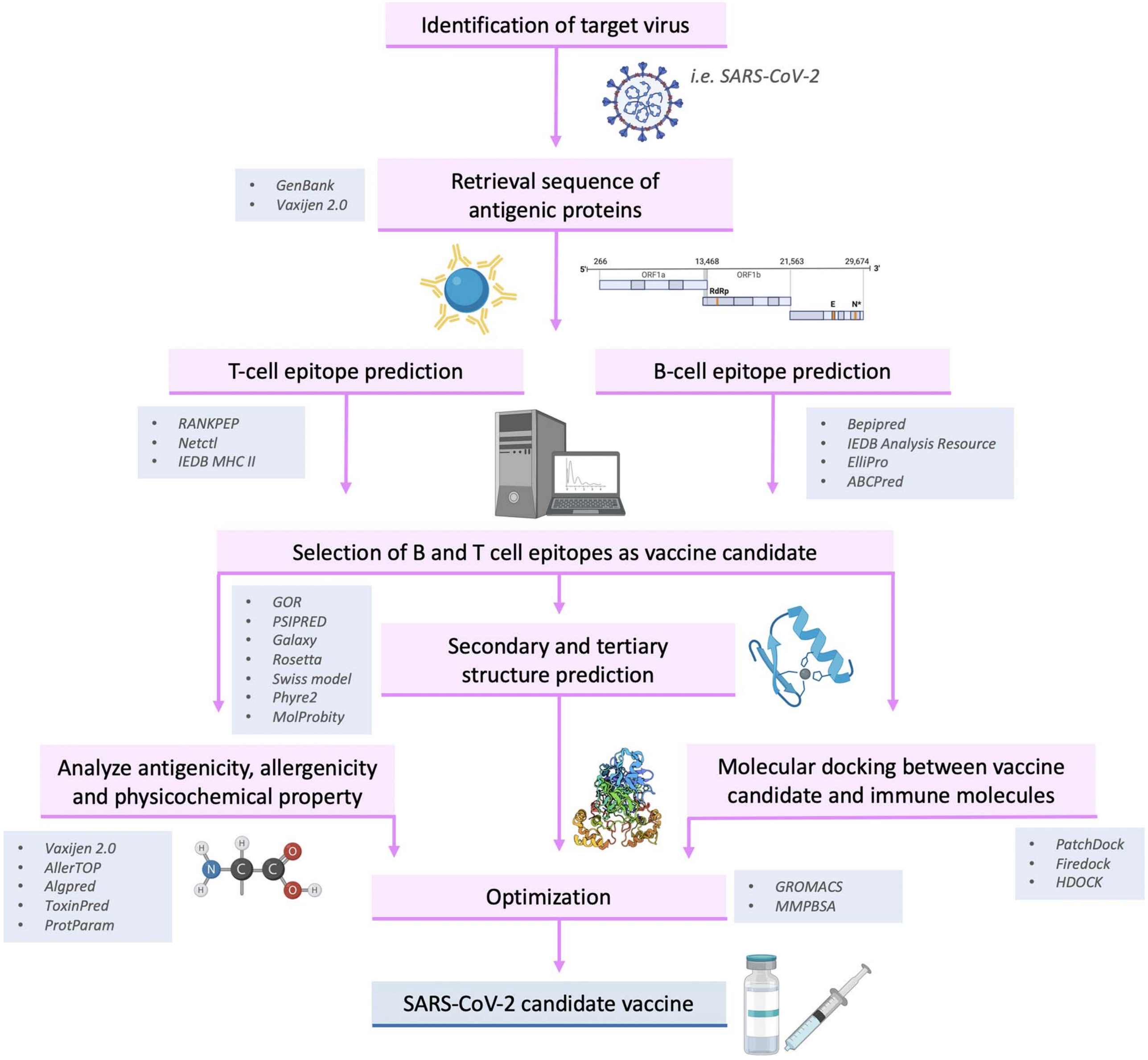
Identification of target virus
i.e. SARS-CoV-2
GenBank
Vaxijen 2.0
Retrieval sequence of antigenic proteins
266
13,468
21,563
29,674
5'
3'
ORF1a
ORF1b
RdRp
E
N*
T-cell epitope prediction
B-cell epitope prediction
RANKPEP
Netctl
IEDB MHC II
Bepipred
IEDB Analysis Resource
ElliPro
ABCPred
Selection of B and T cell epitopes as vaccine candidate
GOR
PSIPRED
Galaxy
Rosetta
Swiss model
Phyre2
MolProbity
Secondary and tertiary structure prediction
Analyze antigenicity, allergenicity and physicochemical property
Molecular docking between vaccine candidate and immune molecules
Vaxijen 2.0
AllerTOP
Algpred
ToxinPred
ProtParam
H
N
C
O
PatchDock
Firedock
HDOCK
Optimization
GROMACS
MMPBSA
SARS-CoV-2 candidate vaccine

Figure 3. A general in silico vaccine development workflow.

engages ACE2. The Coronavirus Antibody Database, CoV-AbDab [83], is a publicly available resource to query and download coronavirus-binding antibody sequences and structures. It actually contains 460 records. A recent study isolated 19 antibodies with high neutralizing power from infected SARs-Cov-2 patients [84]. This collection includes antibodies directed towards the spike protein RBD domain, which compete strongly with the ACE protein and are promising candidates for vaccine development, and non-RBD antibodies, which are instead mainly directed towards the NTD domain. The sequences of these 19 antibodies are deposited on Genbank.

Vaccine development

Although several research groups around the world are engaged in the development of a vaccine against SARS-CoV-2, currently there are no approved treatments for humans.

Reverse vaccinology is a methodology that uses bioinformatics tools and genomic data for the identification of pathogen antigens [85]. In silico vaccine development improves the potential for successful vaccine design reducing time and cost to identify the effective epitopes that could trigger the immune response without causing disease [86]. Figure 3 shows a general workflow of in silico vaccine development, including the main resources used in COVID-19 vaccine discovery so far. Initially, amino acid sequences of proteins that are potentially antigenic or essential for virus replication must be retrieved from sequence databases, such as GenBank [87]. The nucleocapsid (N) protein of SARS-CoV-2 is a suitable vaccine candidate because it is a crucial structural protein, highly conserved with antigenic properties [88]. Also, other structural and non-structural proteins, such as the membrane (M) protein, spike glycoprotein (S), open reading frame 3a (ORF3a), etc., are putative antigenic targets in vaccine design [89]. The identification of antigenic proteins and prediction of T-cell and B-cell epitopes are major steps in developing in silico vaccine. Supplemental Table S3 provides a list of the main bioinformatics resources useful for the prediction of MHC Class-I and II epitopes. Prediction tools for continuous B cell epitopes and T cell epitope are very similar and include algorithms based on (i) machine learning and artificial neural network (ANN) approaches (i.e. NetMHC, NetMHCII, NetCTL, nHLAPred, BepiPred, MHC2Pred, SVMHC, etc.); (ii) the amino acid properties and secondary structure (i.e. VaxiJen, MHCPred, Bcepred, SEPPA, etc.); (iii) position-specific scoring matrix (PSSM) matrix (i.e. RANKPEP). Instead, discontinuous B cell epitope prediction employs resources based on 3D structure resolution of the antigen (i.e. Discotope, ElliPro, etc.) Many other on-line tools are also available to analyze the physiochemical properties and allergenicity and to predict secondary and tertiary structure of vaccine candidate (Figure 3). The EPV-CoV19, a candidate vaccine in the clinical trial phase, was entirely designed using the iVax Toolkit [90], a web-based work environment including several computational immunology tools to develop epitope-driven in silico vaccine.

Therapeutic target identification

Biomarkers for drug repurposing (or drug targets) are molecular elements that are part of the pathophysiologic mechanism of action of a disease. In the context of viral infection, such elements are represented by (i) viral targets, proteins encoded by the viral genome that are essential to the infection process; (ii) viral/host interactors, host proteins that directly interact with viral proteins acting as entry-points for the infection process; and (iii) host response targets, host proteins not directly interacting with the viral proteins but whose inhibition/activation is able to block the signaling pathways that are essential for the infection process to succeed.

Table 2 shows a list of bioinformatics tools developed for therapeutic target identification that have been applied in the context of COVID-19 disease. Most of them have developed in different contexts (e.g. cancer) and can be virtually applied to the targets categories described above. Each of the proposed approach/tool is based on different input structures that can be classified in the following categories: (i) protein–protein networks along with a selection of subsets of proteins of interest (e.g. COVID-19 direct interactors and drug targets); (ii) transcriptomic networks inferred from infected samples; and (iii) proteins/ligands structure and composition.

Tools based on protein–protein networks

In the case of viral infections, at least three pieces of information should be modeled within the network structure: (i) virus–host protein interactions, (ii) host protein–protein interactions and (iii) drug–protein interactions. Pure (unimodal) protein–protein network based approaches consider only proteins as nodes of the network and protein–protein interactions as edges. In multi-modal networks, nodes can be proteins, drugs and diseases, while edges represent interactions among them (protein–protein, drug–protein, drug–diseases, drug–drug, disease–protein, disease–disease). The basic idea is that the closer are the drug targets to disease-related components (such as viral-host interactors), the higher are the odds for the drug to affect the adverse phenotype. A commonly used distance measure between nodes over a graph is the length of the shortest path connecting them. By extending this notion to a set of nodes (e.g. candidate biomarker nodes and COVID-19 nodes),

Table 2. Tools for therapeutic target identification

						Reference	
Tool	First release	Approach	Main input	Available as	Link	Tool	COVID-19
GPSnet	Dec 2018	Network distance	Protein–protein network	Matlab lib	https://github.com/ChengF-Lab/GPSnet	[91]	[92, 93]
TrustRank/Steiner tree	Jul 2020	Network distance	Protein–protein network	Web tool	https://exbio.wzw.tum.de/covex	[94, 95]	[95]
MONET-DSD	Jun 2020	Network diffusion	Protein–protein network	Command line	https://github.com/BergmannLab/MONET	[96]	na
Decagon/AI-Net	Sep 2018	Graph conv. network	Protein–protein network	Python script	https://github.com/mims-harvard/decagon	[97]	[93]
VIPER	Jun 2016	Master regulator analysis	Transcriptome network	R package	http://califano.c2b2.columbia.edu/viper	[98]	[99]
corto	Jun 2020	Master regulator analysis	Transcriptome network	R package	https://github.com/federicogiorgi/corto	[100]	[101]
MANTRA	Aug 2010	Gene set enrichment	Transcriptional signature	Web tool	https://mantra.tigem.it	[102]	[103]
AutoDock Vina	Feb 2011	Docking	Proteins/ligand structure	C++ lib, Command line	http://vina.scripps.edu	[104]	[105]
Deep Docking	Aug 2010	Docking	Proteins/ligand structure	Command line	https://github.com/vibudh2209/D2	[106]	[107]
MT-DTI	Aug 2010	Affinity prediction	Proteins/ligand structure	Web tool	https://mt-dti.deargendev.me	[108]	[109]

the length of the minimum connecting shortest path (MSP) is a proxy for the biomarker or target relevance [110]. The MSP approach has been proved to be an effective metric for ranking and re-purposing drugs against COVID-19 infection [92, 93]. This approach has been originally developed to repurposing drugs in cancer-derived networks with the GPSnet tool [91] and can be easily adapted to COVID-19 networks as shown in [92] and [93] where the authors integrated virus/host interactome data from [72] (Table 1) with a human protein–protein interaction network to rank repurposable drugs based on the distance of their molecular targets from COVID-19 nodes.

Shortest path methods ultimately rely their estimate on the length of a single path (i.e. the minimal one); other methods, such as the TrustRank method [94], try to define the relevance of a node-set (the candidate bio-marker) with respect to another (COVID-19 targets) based on global characteristics like the connectivity level between the two. It is a variant of the Google's PageRank algorithm and is implemented in the CoVex tool [95]. This method can be used to rank a set of protein nodes based on how well they are connected to a set of trusted seed proteins (e.g. SARS-CoV-2 target proteins from [72] (Table 1). In particular, the algorithm propagates such a trustiness information from seed nodes to other non-seed nodes and, based on these propagated values, ranks the all other protein nodes based on their connectivity with the seeds.

The Steiner tree problem aims at finding the minimum cost subgraph connecting a given set of seed nodes. In the case of COVID-19 derived networks, it can be mapped to the problem of finding the minimal subgraph connecting a selection of COVID-19 interactors (acting as seeds), in order to have a representation of the mechanism of action related to such interactors and consequently identify potential drug targets and drug candidates. The Steiner tree problem belongs to the class of NP-hard problems, but different efficient approximation algorithms exist for this problem. An implementation based on finding and merging multiple 2-approximate solutions to the Steiner tree over a protein–protein network and seed nodes selected from [72] (Table 1) is presented in the CoVex tool [95].

Diffusion based methods can be used to rank candidate drug-related biomarkers based on a graph diffusion state similarity measure. A diffusion state can be obtained for a node x by computing for all the other nodes y the expected number of random walks originating in x and passing through y. This approach has been employed by [93] to score a set of drugs based on a the similarity of diffusion states between each drug target node-set and COVID-19 target nodes. This methods can be easily implemented using the diffusion state distance (DSD) tool available in the MONET toolbox [96]

Another interesting approach to drug-related biomarker definition is the possibility to numerically encode all of the semantic contained in the network under study in a low-dimensional space and look for similarities between encoded entities in this new space using vector-based distance measures. Graph embedding methods are based on neural networks implementing an encoder-decoder architecture; this latter able to translate network entities in numeric vectors. It is possible to represent the knowledge network containing interactions between proteins, drugs and diseases in a low-dimensional space (an hyperplane) where each node of the graph can be represented as a scalar vector and distances between points in the encoded feature space are representative of (i) the association between drugs and diseases, (ii) the similarity between diseases and (iii) similarities between drugs' mechanism of action. Gysi *et al.* [93] report an example of drug repositioning based on the embedding of a multi-modal graph containing information on three distinct types of biomedical entities (i.e. drugs, proteins, diseases) and edges representing four types of relationships between the entities (i.e. protein–protein interactions, drug–target associations, disease–protein associations and drug–disease treatments). This approach can be implemented by using an adaptation of the Decagon tool [97] that implements a graph convolutional neural network model for detecting polypharmacy side effects.

Tools based on transcriptomic regulatory networks

While the previous strategies can be more suited to target viral/host interactors, functional annotation-based approaches can be used to identify biomarkers related to the host response to the infection. These approaches can exploit omics data generated from infected samples to infer activated protein modules and/or biochemical pathways that in turn can be used to produce biomarker-targets for drug repositioning.

Li *et al.* [111] followed this kind of approach using transcriptomic data of infected NHBE, A549_ACE2 and Calu3 human lung epithelial cells from [47] (Table 1) and their normal counterparts to identify differentially expressed genes and dysfunctional signaling KEGG pathways activated by these latter. Drug bank data were then exploited in order to find drugs potentially inhibiting one or more of discovered pathways.

Master regulator analysis (MRA) exploits network models derived from omics assays [112]. In the context of viral infection, a master regulator (MR) can be identified as a regulatory protein whose activity is sufficient to determine the success of the infection process. In this setting, also the concept of tumor checkpoints [112] (i.e. a hyperconnected and autoregulated module built around MR proteins) can be translated in the concept of infection checkpoint and thus regarded as a biomarker. In particular, it is possible to extrapolate a set of crucial biomarkers of the infection process, constituted by modules (subnetworks) linked to an MR, i.e. a key-responsive transcriptional regulator along with its direct targets. The VIPER tool [98] can be used to identify transcription factors controlling the infection process given a regulatory network built over infection transcriptome data. This approach has been implemented in Laise *et al.* [99] where the authors used transcriptome data from Calu-3 lung adenocarcinoma cells infected with SARS-CoV to identify master regulator proteins related to SARS-CoV infection process.

These models can be further enhanced by integrating omics derived regulators with functional networks (e.g. known protein–protein networks, pathway-based networks, etc.) thus obtaining functional modules linked to MRs. Such an approach has been successfully adopted in [101] where the authors used their corto algorithm [100] to identify disease sub-modules related to SARS-CoV infection derived co-expression network.

The availability of omics data from infected samples makes it possible to derive biomarkers based on omics signatures (i.e. omics profiles that are characteristics of the particular infection). In this case, the biomarker is represented by the set of molecular features (e.g. genes, proteins, miRNAs) differentiating COVID-19 infected tissues from the normal counterparts. Complex biomarker such as gene signatures can also be used to discover potential drugs that can inhibit its components' activity. In particular, it can be compared to known drug signatures (e.g. drug gene expression signatures from the Connectivity Map dataset [113]) by using a Gene Set Enrichment-based Analysis against transcriptional signatures associated to known drugs. This approach is implemented in the MANTRA tool [102] and has been applied to COVID-19 in Napolitano *et al.* [103] where the authors exploited the transcriptomics data from primary human bronchial epithelial cell line (NHBE) [47] (Table 1).

Tools based on protein/ligand affinity structures

Structure-based approaches rely on the study of the structural affinity between proteins and drug molecules that in turn drives the interaction potential between these two. This approach is particularly suited in targeting viral proteins and, in particular, in the discovery of drugs with potential inhibitory effects against these latter. Structure-based approaches for targeting viral proteins goes through three main steps: (i) identifications of target viral proteins; (ii) modelling of the 3D target proteins structures; (iii) search for potentially interacting ligands. Several data sources reporting the protein sequences of all known COVID-19 proteins along with models of their 3D structures are listed in Section 6 and Table 1.

The search for drug repurposing biomarkers in this case is reduced to the identification of (viral and/or host) proteins involved in the infection process and showing structural affinity for known compounds.

The task of determining the structural affinity can be addressed following a rule-based approach, using molecular docking screens, or by indirect approaches, inferring possible protein/drug interacting pairs from molecular derived features of these latter, given a statistical model trained on known and validated molecule/protein interaction.

Docking simulations work by generating different poses between a ligand and a protein given their 3D structure, obtained by testing different orientations and conformations, and scoring all these poses to determine the ligand affinity between the two structures. This approach can be implemented by using protein models from [76] (Table 1) and through different tools like AutoDock Vina [104], applied by Yu *et al.* [105] to SARS-CoV-2 structural and non-structural proteins, and the Deep Docking tool [106] applied by Ton *et al.* [107] to SARS-CoV-2 main protease.

Indirect approaches for structural affinity screening can be implemented by means of machine learning tools.

These methods are capable of learning the high-dimensional structure of a molecule starting from its raw sequence and encode (embed) it in a low dimensional space, where the relationships between interacting proteins/ligands can be learnt by means of (deep) neural networks or other machine learning approaches. This approach has been implemented in the MT-DTI tool [108] that is based on the natural language processing based Bidirectional Encoder Representations from Transformers (BERT) framework [114] and has been applied to SARS-CoV-2 protein sequences extracted from the SARS-CoV-2 genome [1] (accession NC_045512.2, see Section 6) to discover six coronavirus-related targeted by FDA approved antivirals in [109].

Conclusion

The Bioinformatics community responded to the SARS-CoV-2 emergency with an unprecedented amount of work and research outputs. We have shown that the vast amount of scientific literature related to the computational approaches for the identification of biomarkers can be classified in five main categories. Some categories are more focused on the data generation and sharing such as transcriptomics profiling to identify the markers of the viral infection in host tissues and to characterize the T cell repertoire. A vast amount of work has also been performed to develop AI-based automatic diagnostic tools to characterize the severity of the disease image scans. However, the area where the computational biology community has exploited all the arsenal of approaches that were also developed in other fields such as cancer and neuroscience is the identification of therapeutic targets of existing molecules. However, we want also to mention some potential limitations and opportunities for improvements in some areas. In order to make significant inroads in terms of diagnostic development, it would be necessary for profiles of hundreds, if not thousands, of patients to be available. And it seems that 9 months into this pandemic, we are still very far from the mark. For example, regarding AIRR-seq, while sequencing performed on bulk samples can be informative it will be at some point necessary to determine repertoires among subpopulations separately. For TCR-seq, for instance, it would be quite important to consider separately the repertoire of T helper cells, effectors, memory or regulatory populations.

Overall, we have briefly described the most advanced approaches, mainly based on the inhibition of the signaling cascades activated by viral infection using the knowledge encoded in gene regulatory networks and/or protein–protein interaction networks. Indeed, a plethora of algorithms developed in the area of systems biology has been successfully exploited to prioritize existing drug and molecules; some of the predicted drug are already in clinical trials. Finally, we have also reported the main bioinformatics tools needed in the process of vaccine development that is the ultimate way to combat the emerging COVID-19 pandemic.

Key Points

- A vast amount of literature about COVID-19 biomarkers has been already published so far; automatic text categorization methods are useful to identify key topics
- The analysis of a corpus of 27 000 papers resulted in 36 topics, five of them related to biomarker discovery and drug target identification
- Selected topics span from machine learning and AI for disease characterization to vaccine development and to systems biology for therapeutic target identification.
- We include an up to date catalog of public transcriptomics and proteomics dataset available to the computational biology community for discovery of biomarkers and disease characterization.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

Associazione Italiana per la Ricerca sul Cancro (IG 2018 - ID. 21846 project, P.I. MC) and Italian Ministry of Research Grant PRIN 2017XJ38A4_004.

References

1. Wu F, Zhao S, Yu B, *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* 2020; **579**(7798): 265–9.
2. Downing G. Biomarkers definitions working group. Biomarkers and surrogate endpoints. *Clin Pharmacol Ther* 2001; **69**:89–95.

3. Kraus VB. Biomarkers as drug development tools: discovery, validation, qualification and use. *Nat Rev Rheumatol* 2018; **14**(6): 354–62.
4. Kermali M, Khalsa RK, Pillai K, *et al*. The role of biomarkers in diagnosis of covid-19—a systematic review. *Life Sci* 2020; **254**: 117788.
5. Qin C, Zhou L, Hu Z, *et al*. Dysregulation of immune response in patients with covid-19 in Wuhan, China. *Clin Infect Dis* 2020; **71**(15): 762–8. doi: 10.1093/cid/ciaa248.
6. Lippi G, Plebani M, Henry BM. Thrombocytopenia is associated with severe coronavirus disease 2019 (covid-19) infections: a meta-analysis. *Clin Chim Acta* 2020; **506**: 145–8.
7. Ellinghaus D, Degenhardt F, Bujanda L, *et al*. Genomewide association study of severe covid-19 with respiratory failure. *New Engl J Med* 2020; **383**(16): 1522–34. PMID: 32558485.
8. 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.
9. Zeberg H, Pääbo S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* 2020. doi: 10.1038/s41586-020-2818-3.
10. Zhou F, Yu T, Ronghui D, *et al*. Clinical course and risk factors for mortality of adult inpatients with covid-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020; **395**(10229): 1054–62.
11. Cheng Y, Luo R, Wang K, *et al*. Kidney disease is associated with in-hospital death of patients with covid-19. *Kidney Int* 2020; **97**(5): 829–38.
12. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res* 2003; **3**:993–1022.
13. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. *Nature* 2020; **579**(7798): 193.
14. Web page. https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz. Retrieved June 20, 2020.
15. Web page. https://ftp.ncbi.nlm.nih.gov/pub/pmc/PMC-ids.csv.gz. Retrieved June 20, 2020.
16. Kovalchik S. RISmed: Download Content from NCBI Databases, 2017, R package version 2.1.7.
17. Shang J, Wan Y, Luo C, *et al*. Cell entry mechanisms of sars-cov-2. *Proc Natl Acad Sci* 2020; **117**(21): 11727–34.
18. Mallapaty S. Why does the coronavirus spread so easily between people?, 2020.
19. Hoffmann M, Kleine-Weber H, Schroeder S, *et al*. Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell* 2020; **181**(2): 271–280.e8.
20. Shen LW, Mao HJ, Wu YL, *et al*. Tmprss2: a potential target for treatment of influenza virus and coronavirus infections. *Biochimie* 2017; **142**:1–10.
21. Solerte SB, Di Sabatino A, Galli M, *et al*. Dipeptidyl peptidase-4 (dpp4) inhibition in covid-19. *Acta Diabetol* 2020; **57**(7): 779–83.
22. Stebbing J, Phelan A, Griffin I, *et al*. Covid-19: combining antiviral and anti-inflammatory treatments. *Lancet Infect Dis* 2020; **20**(4): 400–2.
23. Grifoni E, Valoriani A, Cei F, *et al*. Interleukin-6 as prognosticator in patients with covid-19: Il-6 and covid-19. *J Infect* 2020; **81**(3): 452–82.
24. Chen Z, Wherry EJ. T cell responses in patients with covid-19. *Nat Rev Immunol* 2020; **20**(9): 529–36.
25. Ganji A, Farahani I, Khansarinejad B, *et al*. Increased expression of cd8 marker on t-cells in covid-19 patients. *Blood Cells Mol Dis* 2020; **83**: 102437.
26. Xu X, Yu C, Jing Q, *et al*. Imaging and clinical features of patients with 2019 novel coronavirus sars-cov-2. *Eur J Nucl Med Mol Imaging* 2020; **92**(9): 1449–59.
27. Salehi S, Abedi A, Balakrishnan S, *et al*. Coronavirus disease 2019 (covid-19): a systematic review of imaging findings in 919 patients. *Am J Roentgenol* 2020; **215**(1): 87–93.
28. Zhang K, Liu X, Shen J, *et al*. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell* 2020; **181**(6): 1423–33.
29. Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3d residual networks for action recognition. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops* 2017;3154–60.
30. Kang H, Xia L, Yan F, *et al*. Diagnosis of coronavirus disease 2019 (covid-19) with structured latent multi-view representation learning. *IEEE Trans Med Imaging* 2020; **39**(8): 2606–14. doi: 10.1109/TMI.2020.2992546.
31. Roy S, Menapace W, Oei S, *et al*. Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound. *IEEE Trans Med Imaging* 2020; **39**(8): 2676–87. doi: 10.1109/TMI.2020.2994459.
32. Chiara M, Horner DS, Gissi C, *et al*. Comparative genomics suggests limited variability and similar evolutionary patterns between major clades of sars-cov-2. *bioRxiv* 2020. doi: 10.1101/2020.03.30.016790.
33. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002; **30**(1): 207–10.
34. Leinonen R, Sugawara H, Shumway M, *et al*. The sequence read archive. *Nucleic Acids Res* 2010; **39**(suppl_1): D19–21.
35. Leinonen R, Akhtar R, Birney E, *et al*. The european nucleotide archive. *Nucleic Acids Res* 2010; **39**(suppl_1): D28–31.
36. 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**(7): 692–5.
37. Wang Y, Song F, Zhu J, *et al*. GSA: genome sequence archive. *Genomics Proteomics Bioinformatics* 2017; **15**(1): 14–8.
38. Web page. https://singlecell.broadinstitute.org/single_cell. Retrieved June 20, 2020.
39. Wilk AJ, Rustagi A, Zhao NQ, *et al*. A single-cell atlas of the peripheral immune response in patients with severe covid-19. *Nat Med* 2020; **26**:1070–6.
40. Han Y, Yang L, Duan X, *et al*. Identification of candidate covid-19 therapeutics using hpsc-derived lung organoids. *bioRxiv* 2020. doi: 10.1101/2020.05.05.079095.
41. Ravindra NG, Alfajaro MM, Gasque V, *et al*. Single-cell longitudinal analysis of sars-cov-2 infection in human bronchial epithelial cells. *bioRxiv* 2020. doi: 10.1101/2020.05.06.081695.
42. Guo C, Li B, Ma H, Wang X, Cai P, Yu Q, *et al*. Single-cell analysis of two severe COVID-19 patients reveals a monocyteassociated and tocilizumab-responding cytokine storm. *Nat Commun* 2020; **11**(1): 3924.
43. Chua RL, Lukassen S, Trump S, *et al*. Covid-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nat Biotechnol* 2020; **38**(8): 970–9.
44. Patterson BK, Seethamraju H, Dhody K, *et al*. Disruption of the ccl5/rantes-ccr5 pathway restores immune homeostasis and reduces plasma viral load in critical covid-19. *medRxiv* 2020. doi: 10.1101/2020.05.02.20084673.

45. Wei L, Ming S, Zou B, *et al*. Viral Invasion and Type I Interferon Response Characterize the Immunophenotypes During Covid-19 Infection. *SSRN Electron J*., 2020. doi: 10.2139/ssrn.3564998.
46. Zhu L, Yang P, Zhao Y, *et al*. Single-cell sequencing of peripheral blood mononuclear cells reveals distinct immune response landscapes of covid-19 and influenza patients. *Immunity* 2020; **53**(3): 685–696.e3. doi: 10.1016/j.immuni.2020.07.009.
47. Blanco-Melo D, Nilsson-Payant BE, Liu W-C, *et al*. Imbalanced host response to sars-cov-2 drives development of covid-19. *Cell* 2020; **181**(5): 1036–1045.e9.
48. Xiong Y, Liu Y, Cao L, *et al*. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in covid-19 patients. *Emerging Microbes Infect* 2020; **9**(1): 761–70.
49. Lieberman NAP, Peddu V, Xie H, *et al*. In vivo antiviral host transcriptional response to SARSCoV-2 by viral load, sex, and age. *PLoS Biol* 2020; **18**(9). doi: 10.1371/JOURNAL.PBIO.3000849.
50. Sharma A, Garcia G, Wang Y, *et al*. Human iPSC-Derived Cardiomyocytes Are Susceptible to SARS-CoV-2 Infection. *Cell Reports Med* 2020; **1**(4): 100052.
51. Suzuki T, Ito Y, Sakai Y, *et al*. Generation of human bronchial organoids for sars-cov-2 research. *bioRxiv* 2020. doi: 10.1101/2020.05.25.115600.
52. Web page. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150316. Retrieved June 20, 2020.
53. Lamers MM, Beumer J, van der Vaart J, *et al*. Sars-cov-2 productively infects human gut enterocytes. *Science* 2020; **369**(6499): 50–4.
54. Nielsen SCA, Yang F, Hoh RA, *et al*. B cell clonal expansion and convergent antibody responses to sars-cov-2. *Cell Host Microbe* 2020; **28**(4): 516–525.e5. doi: 10.21203/rs.3.rs-27220/v1.
55. Kuri-Cervantes L, Pampena MB, Meng W, *et al*. Immunologic perturbations in severe covid-19/sars-cov-2 infection. *bioRxiv* 2020. doi: 10.1101/2020.05.18.101717.
56. Minervina AA, Komech EA, Titov A, *et al*. Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T cell memory formation after mild COVID-19 infection. arXiv Prepr arXiv200508290. 2020. doi: 10.1101/2020.05.18.100545.
57. Robbiani DF, Gaebler C, Muecksch F, *et al*. Convergent antibody responses to sars-cov-2 in convalescent individuals. *Nature* 2020; **28**(4): 516–525.e5. doi: 10.1038/s41586-020-2456-9.
58. Web page. https://github.com/stratust/igpipeline. Retrieved June 20, 2020.
59. Shomuradova AS, Vagida MS, Sheetikov SA, *et al*. Sars-cov-2 epitopes are recognized by a public and diverse repertoire of human t-cell receptors. *medRxiv* 2020. doi: 10.1101/2020.05.20.20107813.
60. Web page. https://vdjdb.cdr3.net. Retrieved June 20, 2020.
61. Galson JD, Schaetzle S, Bashford-Rogers RJM, *et al*. Deep sequencing of b cell receptor repertoires from covid-19 patients reveals strong convergent immune signatures. *bioRxiv* 2020. doi: 10.1101/2020.05.20.106294.
62. Schultheiß C, Paschold L, Simnica D, *et al*. Next generation sequencing of t and b cell receptor repertoires from covid-19 patients showed signatures associated with severity of disease. *Immunity* 2020; **53**(2): 442–455.e4. j.immuni.2020.06.024.
63. Wyler E, Mösbauer K, Franke V, *et al*. Bulk and single-cell gene expression profiling of sars-cov-2 infected human cell lines identifies molecular targets for therapeutic intervention. *bioRxiv* 2020. doi: 10.1101/2020.05.05.079194.
64. Yang L, Han Y, Nilsson-Payant BE, *et al*. A human pluripotent stem cell-based platform to study sars-cov-2 tropism and model virus infection in human cells and organoids. *Cell Stem Cell* 2020; **27**(1): 125–36.
65. Huang L, Shi Y, Gong B, *et al*. Blood single cell immune profiling reveals the interferon-mapk pathway mediated adaptive immune response for covid-19. *medRxiv* 2020. doi: 10.1101/2020.03.15.20033472.
66. Wen W, Wenru S, Tang H, *et al*. Immune cell profiling of covid-19 patients in the recovery stage by single-cell sequencing. *Cell Discov* 2020; **6**(1): 1–18.
67. Cao Y, Bin S, Guo X, *et al*. Potent neutralizing antibodies against sars-cov-2 identified by high-throughput single-cell sequencing of convalescent patients' b cells. *Cell* 2020; **182**(1): 73–84.
68. Liao M, Yang L, Yuan J, *et al*. Single-cell landscape of bronchoalveolar immune cells in patients with covid-19. *Nat Med* 2020; **26**:842–4.
69. Web page. http://ncov-ai.big.ac.cn/download. Retrieved June 20, 2020.
70. Web page. https://www.disi.unitn.it/iclus. Retrieved June 20, 2020.
71. Gordon DE, Jang GM, Bouhaddou M, *et al*. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 2020; **583**:459–68.
72. Web page. BioChemPantry, 2020, https://github.com/momeara/BioChemPantry/. Retrieved June 20, 2020.
73. Bojkova D, Klann K, Koch B, *et al*. Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* 2020; **583**:469–72.
74. Web page. http://corona.papers.biochem2.com. Retrieved June 20, 2020.
75. Li J, Guo M, Tian X, *et al*. Virus-host interactome and proteomic survey of PMBCs from COVID-19 patients reveal potential virulence factors influencing SARS-CoV-2 pathogenesis. *bioRxiv* 2020. 2020.03.31.019216.
76. Web page. 2020. http://www.rcsb.org. Retrieved June 20, 2020.
77. Wlodawer A, Dauter Z, Shabalin IG, *et al*. Ligand-centered assessment of sars-cov-2 drug target models in the protein data bank. *FEBS J* 2020; **287**(17): 3703–18.
78. Web page. https://covid-19.bioreproducibility.org. Retrieved June 20, 2020.
79. Shen B, Yi X, Sun Y, *et al*. Proteomic and metabolomic characterization of covid-19 patient sera. *Cell* 2020; **182**(1): 59–72.
80. Vabret N, Britton GJ, Gruber C, *et al*. Immunology of covid-19: current state of the science. *Immunity* 2020; **52**(6): 910–41.
81. Rubelt F, Busse CE, Bukhari SAC, *et al*. Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. *Nat Immunol* 2017; **18**(12): 1274–8.
82. Web page. https://gateway.ireceptor.org.
83. Raybould MIJ, Kovaltsuk A, Marks C, Deane CM CoV-AbDab: the coronavirus antibody database. *Bioinformatics* 2020. doi: 10.1093/bioinformatics/btaa739.
84. Liu L, Wang P, Nair MS, *et al*. Potent neutralizing antibodies directed to multiple epitopes on sars-cov-2 spike.

Nature 2020; **182**(1): 73–84.e16. doi: 10.1038/s41586-020-2571-7.

85. María RR, Arturo CJ, Alicia J-A, *et al.* The impact of bioinformatics on vaccine design and development. In: *Vaccines*, InTech, Rijeka, Croatia, 2017.
86. Tahir ul Qamar M, Rehman A, Ashfaq UA, *et al.* Designing of a next generation multiepitope based vaccine (mev) against sars-cov-2: Immunoinformatics and in silico approaches. *BioRxiv* 2020. doi: 10.1101/2020.02.28.970343.
87. Web page. https://www.ncbi.nlm.nih.gov/genbank. Retrieved June 20, 2020.
88. Zeng W, Liu G, Ma H, *et al.* Biochemical characterization of sars-cov-2 nucleocapsid protein. *Biochem Biophys Res Commun* 2020; **527**(3): 618–23.
89. Enayatkhani M, Hasaniazad M, Faezi S, *et al.* Reverse vaccinology approach to design a novel multi-epitope vaccine candidate against covid-19: an in silico study. *J Biomol Struct Dyn* 2020;1–16.
90. Moise L, Gutierrez A, Kibria F, *et al.* Ivax: an integrated toolkit for the selection and optimization of antigens and the design of epitope-driven vaccines. *Hum Vaccin Immunother* 2015; **11**(9): 2312–21.
91. Cheng F, Lu W, Liu C, *et al.* A genome-wide positioning systems network algorithm for in silico drug repurposing. *Nat Commun* 2019; **10**(1): 1–14.
92. Zhou Y, Hou Y, Shen J, *et al.* Network-based drug repurposing for novel coronavirus 2019-ncov/sars-cov-2. *Cell Discov* 2020; **6**(1): 14.
93. Gysi DM, Do Valle Í, Zitnik M, *et al.* Network medicine framework for identifying drug repurposing opportunities for COVID-19. arXiv:2004.07229. 2020.
94. Gyöngyi Z, Garcia-Molina H, Pedersen J. Combating web spam with trustrank. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases—Volume 30, VLDB '04*. VLDB Endowment, 2004, 576–87.
95. Sadegh S, Matschinske J, Blumenthal DB, *et al.* Exploring the sars-cov-2 virus-host-drug interactome for drug repurposing. *Nat Commun* 2020; **11**(1).
96. Tomasoni M, Gómez S, Crawford J, *et al.* MONET: a toolbox integrating top-performing methods for network modularization. *Bioinformatics* 2020; **36**(12): 3920–1.
97. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 2018; **34**(13): i457–66.
98. Alvarez MJ, Shen Y, Giorgi FM, *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* 2016; **48**(8): 838–47.
99. Laise P, Bosker G, Sun X, *et al.* The host cell virocheckpoint: identification and pharmacologic targeting of novel mechanistic determinants of coronavirus-mediated hijacked cell states. *bioRxiv* 2020. doi: 10.1101/2020.05.12.091256.
100. Mercatelli D, Lopez-Garcia G, Giorgi FM. Corto: a lightweight R package for gene network inference and master regulator analysis. *Bioinformatics* 2020; **36**(12): 3916–7.
101. Guzzi PH, Mercatelli D, Ceraolo C, *et al.* Master regulator analysis of the sars-cov-2/human interactome. *J Clin Med* 2020; **9**(4): 982.
102. Iorio F, Bosotti R, Scacheri E, *et al.* Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci* 2010; **107**(33): 14621–6.
103. Napolitano F, Gambardella G, Carrella D, *et al.* Computational drug repositioning and elucidation of mechanism of action of compounds against sars-cov-2, arXiv 2020; Prepr arXiv200407697.
104. Trott O, Olson AJ. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 2009; **31**(2): 455–61.
105. Yu R, Liang C, Lan R, *et al.* Computational screening of antagonists against the sars-cov-2 (covid-19) coronavirus by molecular docking. *Int J Antimicrob Agents* 2020; **56**(2): 106012.
106. Gentile F, Agrawal V, Hsing M, *et al.* Deep docking: a deep learning platform for augmentation of structure based drug discovery. *ACS Cent Sci* 2020; **6**(6): 939–49.
107. Ton A-T, Gentile F, Hsing M, *et al.* Rapid identification of potential inhibitors of sars-cov-2 main protease by deep docking of 1.3 billion compounds. *Mol Inf* 2020; **39**(8). doi: 10.1002/minf.202000028.
108. Shin B, Park S, Kang K, *et al.* Self-attention based molecule representation for predicting drug-target interaction. In: Doshi-Velez F, Fackler J, Jung K, Kale D, Ranganath R, Wallace B, Wiens J, (eds), *Proceedings of the 4th Machine Learning for Healthcare Conference*, vol. 106. *Proceedings of Machine Learning Research*, 230–248, Ann Arbor, Michigan, 09–10 Aug 2019. PMLR.
109. Beck BR, Shin B, Choi Y, *et al.* Predicting commercially available antiviral drugs that may act on the novel coronavirus (sars-cov-2) through a drug-target interaction deep learning model. *Comput Struct Biotechnol J* 2020; **18**: 784–90.
110. Guney E, Menche J, Vidal M, *et al.* Network-based in silico drug efficacy screening. *Nat Commun* 2016; **7**(1).
111. Li F, Michelson AP, Foraker R, *et al.* Repurposing drugs for covid-19 based on transcriptional response of host cells to sars-cov-2. 2020; arXiv 2006.01226.
112. Califano A, Alvarez MJ. The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nat Rev Cancer* 2016; **17**(2): 116–30.
113. Subramanian A, Narayan R, Corsello SM, *et al.* A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017; **171**(6): 1437–1452.e17.
114. Devlin J, Chang M-W, Lee K, *et al.* Bert: pre-training of deep bidirectional transformers for language understanding. In: *NAACL-HLT*, 2019.