OXFORD

Briefings in Bioinformatics, 00(0), 2020, 1–14

doi: 10.1093/bib/bbaa377
Advance Access Publication Date: 00 Month 20xx
Method Review

Comprehensive pathway enrichment analysis workflows: COVID-19 case study

Giuseppe Agapito, Chiara Pastrello and Igor Jurisica

Corresponding author: Giuseppe Agapito, Department of Legal, Economic and Social Sciences, Magna Graecia University of Catanzaro, Catanzaro, Italy. E-mail: agapito@unicz.it

Abstract

The coronavirus disease 2019 (COVID-19) outbreak due to the novel coronavirus named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been classified as a pandemic disease by the World Health Organization on the 12th March 2020. This world-wide crisis created an urgent need to identify effective countermeasures against SARS-CoV-2. *In silico* methods, artificial intelligence and bioinformatics analysis pipelines provide effective and useful infrastructure for comprehensive interrogation and interpretation of available data, helping to find biomarkers, explainable models and eventually cures. One class of such tools, pathway enrichment analysis (PEA) methods, helps researchers to find possible key targets present in biological pathways of host cells that are targeted by SARS-CoV-2. Since many software tools are available, it is not easy for non-computational users to choose the best one for their needs. In this paper, we highlight how to choose the most suitable PEA method based on the type of COVID-19 data to analyze. We aim to provide a comprehensive overview of PEA techniques and the tools that implement them.

Key words: biological pathways; pathway enrichment analysis; COVID-19; statistical analysis; network analysis.

Introduction

Coronaviruses (CoV) are a broad family of respiratory viruses that can cause mild to moderate illnesses, including the common cold, to severe respiratory syndromes such as Middle East respiratory syndrome (MERS) and severe acute respiratory syndrome (SARS) [1]. CoV owe their name to the spikes on their surface, forming a structure similar to a crown (*corona* in Latin). CoV are prevalent in many animal species, such as camels, pangolins and bats [2]. Rarely, they can evolve and infect humans. The novel coronavirus, named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [3], was reported for the first time in Wuhan, China, in December 2019, and subsequently spread across the world.

While there are over 1900 registered trials as of August 2020 (https://www.covid-trials.org/), there are neither approved treatments nor effective vaccines against coronavirus disease 2019 (COVID-19) yet [4]. Albeit SARS-CoV-2 has a lower mutation rate than other coronaviruses [5], genomic diversity (although limited) is manifested both among singular patients and within the same virus class [6, 7]. Genetic diversity provides viral adaptation to different hosts and different environments within the hosts and it is often associated with disease progression, drug resistance and treatment results. Consequently, even a minimal but continuous mutations of the virus would reduce the efficacy of a vaccine to clearly contrast the advance of COVID-19. Therefore, it will be essential to obtain information on the evolution and pathogenesis of the virus to control this pandemic.

Giuseppe Agapito is an Assistant Professor at the Magna Grœcia University, Catanzaro, Italy. His research interests include biological networks, genomics data analysis and parallel computing.
Chiara Pastrello is a Research Associate at the Krembil Research Institute (KRI), Toronto, Canada. Her research interests include network biology and cancer genetics.
Igor Jurisica: is a senior scientist at KRI and a professor at UofT. His interests include integrative computational biology, and explainable modeling and analysis of chronic diseases.
Submitted: 8 August 2020; **Received (in revised form):** 2 November 2020

Researchers are sharing their results to obtain insight into the genome and the evolution of SARS-CoV-2 globally. Many studies are focusing on success and failures of prevention and treatment, aiming at faster re-positioning of existing drugs [8]. Drug re-positioning may offer a faster and cheaper treatment compared to de novo drug development.

In the arsenal of *in silico* tools, useful for drug re-positioning and understanding individual differences in response to the virus, pathway enrichment analysis (PEA), can improve interpretation of SARS-CoV-2 data by identifying biological pathways of host cells affected by the virus, help characterize possible pharmacological targets and drug mechanism of action. Thus, aiding not only drug repurposing, but enabling *de novo* drug design as well. The large amount of data released so far is quite useful, but knowing the connections among events and context of specific effects is key to understand and address the disease. Thus, extracting knowledge from the data that characterize SARS-CoV-2 host infection is of paramount importance to succeed in subsequent steps linked to identification of an efficacious treatment.

Biological pathways are human representations of coordinated molecular actions within a cell that can include genes, proteins, small molecules, tissues and organs. Pathways are classified into three main broad categories: signaling pathways, metabolic pathways and regulatory pathways. Some pathway databases focus on only one category of pathways—for example, STKE collects only signaling pathways [9] and HumanCyc focuses on metabolic pathways [10]; as a result, both are very specific and their genome coverage is limited. Generic pathway databases (including pathways from all three categories) are larger and more frequently used. Several generic pathway databases are available, some stemming from large curation efforts (primary databases, the oldest being KEGG [11] and the largest being Reactome [12]), some integrating data from primary databases (integrated databases, the oldest being PantherDB [13] and the largest being PathwayCommons [14]) and some using a hybrid integration-curation approach (hybrid databases, the oldest and largest being WikiPathways [15, 16]). Nonetheless, each database includes a largely different number of pathways, annotates only part of the genome and has poor overlap with other databases. Moreover, even upon integration of all major pathway databases, only about 13 000 genes are annotated with any pathway [17]. Numbers are even scanter for organisms other than humans, as the majority of pathway databases focus on *Homo sapiens*-related pathways. Researchers performing pathway enrichment analysis should consider the overlap of their genes of interest with the database they are planning on using. Alternatively, one could take advantage of and select pathway databases with broader coverage provided by the inclusion of both curated and predicted pathway associations (the oldest being PathExpand [18], and the largest being pathDIP [17] that annotates more than 18 000 human genes, and covers pathways for 17 organisms). It is important to highlight that good quality results depend on choosing the appropriate tool for the data to be analyzed, but also the quality of the databases used for enrichment analysis. PEA tools can help researchers to characterize the role of host genes in affecting the biological response to SARS-CoV-2 and different outcomes of COVID-19. Considering that many PEA tools are available, each one suitable to analyze a specific type of input data, it is not straightforward to select the most appropriate framework to investigate a particular kind of COVID-19 related data. Thus, knowing the main characteristics of the available PEA methods, it not only helps researchers to improve the quality and speed of the analysis by using the most appropriate tool for the input data to analyze, but it also contributes to the production of more accurate and meaningful results.

PEA methods can be clustered into three principal categories: i) over-representation analysis (ORA), ii) gene set enrichment analysis (GSEA) and iii) topological enrichment analysis (TEA).

Each PEA approach is available as a software tool, and we briefly describe features for the most frequently used tools per category. In particular, we want to highlight how to choose the most suitable PEA method based on the type of COVID-19 data to analyze, and to present an overview of PEA techniques and their most frequently used software tools. We provide a comparison for all the frameworks reviewed based on the supported input types and formats, the provided output results, along with the supported pathway databases used to define the biological context of the investigated proteins/genes, the practicality, ease of use, programming language and type of interface. This way, even users not well-accustomed to pathway analysis can easily select the best software tool to use with their COVID-19 data.

Approaches

In this Section, we briefly describe the most used pathway enrichment methods in the scientific community.

Pathway enrichment analysis

PEA can be broadly divided into three categories, as listed above.

ORA: one of the most simple PEA methods, ORA methods perform statistical evaluation of the fraction of pathway components found among a user-selected list of biological components. The enrichment is accomplished through an iterative methodology. It calculates, for each pathway, the amount of input genes that belong to the current pathway, repeating this process for each pathway in the database. The most used tests are based on the hypergeometric, chi-square, Fisher's test, Jaccard Index or binomial distribution as reported in [19]. The final results from an ORA method generally consist of a list of relevant pathways, ordered according to a *P*-value or a multiple hypothesis tests corrected *P*-value.

Any list of genes or proteins collected from a COVID-19 paper or database can be used with ORA, that is the most popular adopted pathway analysis method as it is easy to perform. However, a limitation of ORA is that it considers each gene with equal importance, which is often biologically inaccurate.

GSEA: it exploits the hypothesis that few major gene expression changes have a considerable effect on pathways function, and the sum of several weaker and concurrent changes in pathways' genes impact the general functioning as well. GSEA methods compute pathway enrichment analysis using a three steps methodology. i) The first step ranks the genes according to their differential expression across groups of samples, and calculates the over-representation of these genes among the highest and lowest ranking positions. Statistics are calculated using the Kolmogorov–Smirnov-like test. ii) The second step consists of calculating a *P*-value comparing the score obtained in the first step to a null distribution obtained permuting the phenotype labels. iii) The last steps performs multiple hypothesis testing adjustment.

Only ranked list of genes can be used in GSEA, such as those from differential expression analysis (e.g. infected with SARS-CoV-2 vs non-infected). GSEA may consider genes with different importance, making it possible to use more information

than ORA for analyzing pathways and yielding more biologically meaningful models. A limitation of GSEA, though, is that it assumes that genes at the top of the ranking (those more differentially expressed) are more crucial, even though this is not always biologically true (signaling pathways, for example, are very sensitive to minor expression changes). Moreover, the need for a ranking value (e.g. fold change or *P*-value) limits the type of data that can be used.

TEA: in this set of methods, pathways are represented as graphs, where nodes represent pathway's components (e.g. genes, proteins and small molecules), and edges provide information about the interactions among such components (e.g. activation, deactivation, conditions where the interaction happens). The main difference compared with other methods is that TEA uses topology information as additional information to compute pathway enrichment values. Researchers could use condition specific interaction networks to shed light on different pathways and related different responses activated by SARS-CoV-2. However, one limitation of current TEA methods is that they only handle the static properties of the network topology; thus, they are not well suited for dynamic systems modelling.

Tools

In this Section, we describe the characteristics of the principal software tools to perform pathway analysis, along with some databases focusing on COVID-19 data.

COVID-19 databases

To improve our response to COVID-19 and to speed up development of better tests, guidelines, treatments and vaccines, many databases have been updated to include virus relevant data deposited at a rapid pace. Below is a list of some of such databases.

- **CORona Drug InTEractions database (CORDITE)** collects and aggregates data for SARS-CoV-2 available in the literature from PubMed, MedRxiv, BioRxiv, ChemRxiv and ClinicalTrials.gov [20]. Its main focus is set on drug interactions either addressing viral proteins or human proteins that could be used to treat COVID and it is available at https://cordite.mathematik.uni-marburg.de/#/. CORDITE provides up to date information on computational predictions, *in vitro*, *in vivo* studies data and clinical trials.
- **Comparative Toxicogenomics database (CTD)** [21] is a public database that aims at advancing the understanding about how environmental exposures affect human health. CTD is freely available at http://ctdbase.org/. The database provides manually curated information about chemical–gene/protein interactions, chemical–disease and gene–disease relationships. Data are integrated with functional and pathway data to aid in development of hypotheses about the mechanisms underlying environmentally influenced diseases. Moreover, CTD provides the set of genes associated with COVID-19 at http://ctdbase.org/detail.go?type=disease&acc=MESH∖%3aC000657245&view=gene. A gene has either a curated association to the disease (marker/mechanism and/or therapeutic) or an inferred association through a curated chemical interaction.
- **CoronaVirus Explorer (CoVex)** is an interactive web-based platform that collects data for SARS-CoV-2 - host interactions, human protein-protein interactions and drug-target interactions [22]. It includes a visual tool that provides the ability to explore the collected interactome, and a systems medicine algorithms for network-based drug re-positioning. It is available at https://exbio.wzw.tum.de/covex/.
- **DisGeNET** is a comprensive database containing collections of genes and variants associated to human diseases [23]. DisGeNET integrates data from expert curated repositories, GWAS databases, animal models and the scientific literature. DisGeNET data are homogeneously annotated with controlled vocabularies and community-driven ontologies. In addition, DisGeNET provides access to COVID-19 specific data at https://www.disgenet.org/covid/diseases/summary/.
- **IntAct Molecular Interaction database** is a freely available database for molecular interaction data, available at https://www.ebi.ac.uk/intact/ [24]. All interactions are derived from literature curation in a coordinated effort by multiple databases belonging to the IMEx consortium [25] . Recently, IntAct introduced an update including interaction data from the high-throughput (HT) multi-level proteomics studies on SARS-CoV-2, SARS and other Coronaviridae [26]. The IntAct update provides information about molecular interactions extracted from publications concerning viral proteins from the Coronaviridae family and human proteins and other organisms. The data includes protein-protein and RNA–protein interactions. The COVID-19 data are available at https://www.ebi.ac.uk/intact/query/annot:∖%22dataset:coronavirus∖%22.
- **SIGnaling Network Open Resource (SIGNOR)** stores and organizes in a structured format signaling pathways available in the scientific literature [27]. The collected information is represented as a directed graph. Each interaction in the graph is associated with an effect (up/down-regulation) and a mechanism (e.g. binding, phosphorylation, inhibition, etc.). The current version of SIGNOR stores almost 23 000 manually annotated causal relationships between proteins and other biologically relevant entities (e.g. chemicals, phenotypes and complexes). SIGNOR data can be freely downloaded at https://signor.uniroma2.it/downloads.php. In addition, the SIGNOR team has added available evidence that is likely to be relevant for the COVID-19 pathology. Evidence obtained using related human coronaviruses diseases such as SARS and MERS is also mapped to the networks. COVID-19 data are freely available at https://signor.uniroma2.it/covid/.
- **VirHostNet** is a database for the management and the investigation of proteome-wide virus–host interaction networks linked to functional annotations [28]. VirHostNet integrates a comprehensive and original literature-curated dataset of the virus–virus and virus–host interactions from several distinct viral species and one of the largest human interactome reconstructed from publicly available data. Public access to the VirHostNet database is available at http://pbildb1.univ-lyon1.fr/virhostnet. Recently, VirHostNet added a manual curation of Coronaviridae–host protein–protein interactions that is freely accessible at http://virhostnet.prabi.fr:9090/psicquic/webservices/ current/search/query/pubid:https*.

Pathway enrichment tools

PEA are commonly implemented as stand-alone software, web-based applications or program libraries. The first two categories are usually more straightforward to use, as they do not require

analytical skills or programming abilities. The last class is mostly coded in R and Python languages and shared openly in the Bio-Conductor [29] and GitHub [30] repositories. The main benefits of using PEA programming packages is the potential customization of every step of the analysis and the feasibility to automate the process through a scripting analysis pipelines. Deciding between software platforms and program libraries may be influenced by user skills and the cost-benefit ratio of time invested in orchestrating everything necessary to run the analysis.

ORA tools

The main advantages of using ORA methodologies is that they provide a biological context for COVID-19 data without the need for further annotations. Following is a list of commonly used software tools using ORA method:

BioPAX-Parser (BiP)

BiP [31] performs PEA using pathways encoded in Biological Pathway Exchange (BioPAX) format [32]. BioPAX is a meta language defined in OWL (Web Ontology Language) and represented in the RDF/XML (Resource Description Framework / eXtensible Meta Language) format, and is the language of choice to store and exchange pathway data.

- **Availability**: BiP can be freely downloaded as a stand alone application at https://gitlab.com/giuseppeagapito/bip.
- **OS platform**: it is fully developed in Java making it platform independent—it can be executed on all OS compatible with Java.
- **Input data format**: BiP requires as input a plain text file containing the list of genes/proteins of interest to be investigated.
- **Output data format**: it allows users to export ranked PEA results in tabular format, along with further information available in the selected pathway database for the analysis.
- **Analysis**: PEA in BiP is obtained using the hypergeometric test, along with multiple statistical corrector such as false discovery rate (FDR) and Bonferroni.
- **Supported database**: PEA can be performed using information from all the available pathway databases compliant with the BioPAX format, e.g. KEGG, Humancyc [10], NetPath [33], Panther [13], PID [34], Reactome [12], PathwayCommons [14] and WikiPathways [15].

Enrichr

Enrichr [35] is an easy to use enrichment analysis web-based tool. It provides various types of visualization summaries of different categories of biological functions.

- **Availability**: Enrichr is available as an HTML5 web-based application and also as a mobile app at: http://amp.pharm.mssm.edu/Enrichr. It is also available as an R package at https://cran.r-project.org/web/packages/enrichR/vignettes/enrichR.html.
- **OS platform**: Enrichr is delivered as web-based tool, making it compatible with all the available OS. In addition, Enrichr can also be accessed via Android, iOS, and BlackBerry phone apps. The R implementation is intended for experienced computational users and is compatible with any OS running R.
- **Input data format**: To start PEA in Enrichr, users must upload a file containing a list of genes in text plain format.
- **Output data format**: Enrichr provides various ways to visualize the results from the enrichment analysis. Enriched terms can be visualized on a grid of squares, or as a network of terms. All the available visualization in Enrichr can be downloaded as scalable vector graphics (SVG), portable network graphics (PNG) or joint photographic experts group (JPG) files. Results can also be presented in an hypertext markup language (HTML) sortable table with various columns showing the enriched terms with the various scores. Table results can be exported as tab-delimited files.
- **Analysis**: The functional enrichment of the input gene list is evaluated using a customized implementation of the *Fisher's Test*. To correct results for multiple testing Enrichr uses *FDR* and *Bonferroni* correctors.
- **Supported database**: it can obtain information from KEGG [11], WikiPathways [15], BioCarta [36], HumanCyc [10], Panther [13], BioPlanet [37], Elsevier Pathway Collection [38], PID [34] and Reactome [12] pathway databases.

g:Profiler

g:Profiler [39] maps genes to known functional annotations and detects statistically significantly enriched pathways.

- **Availability**: g:Profiler is available freely as a web application at https://biit.cs.ut.ee/gprofiler/. In addition, g:Profiler is available as both Python and R client libraries, and as API. The backend of g:Profiler is implemeted using *Python 3.6*.
- **OS platform**: g:Profiler is a web application that can be used from every available OS with active internet connection.
- **Input data format**: The default input data format of g:Profiler is a list of genes/proteins. The input gene list can be either unordered or ordered (as default option list is considered in the order of decreasing importance). The ordered query option is useful when the genes can be placed in some biologically meaningful order.
- **Output data format**: g:Profiler provides the computation results in three separate tabs—Results, Detailed Results and Query Info. g:Profiler can show the users enrichment analysis results through an interactive Manhattan plot, or as an interactive result table containing the information about the enriched terms. Both data results can be exported as images (i.e. in SVG and PNG formats), or as tables in comma separated values (CSV) format.
- **Analysis**: The functional enrichment of the input gene list is evaluated using the well-known cumulative hypergeometric test. To correct results obtained employing multiple testing, g:Profiler uses the FDR and Bonferroni correctors.
- **Supported database**: g:Profiler can obtain information from KEGG [11], WikiPathways [15] and Reactome [12] pathway databases.

pathDIP

pathDIP [17] is an integrated database of pathways in human, model organisms and domesticated animals, comprising core pathways from major curated pathway databases, and gene pathway associations predicted using physical protein interactions.

- **Availability**: it is publicly available at http://ophid.utoronto.ca/pathDIP.
- **OS platform**: pathDIP is a web application making it platform independent and compatible with all available OS.

In addition, pathDIP is available as application program interface (API) in Java, R, or Python.

- **Input data format**: pathDIP uses an input list of proteins/-genes, and requires the selection of the appropriate organism. Users can also select the databases of choice for the analysis.
- **Output data format**: enrichment results and detailed annotations for input list are exported in tab separated format. In addtion pathDIP can visualize results as diagrams or interactive tables.
- **Analysis**: to calculate the enrichment score, pathDIP uses the Fisher's exact test, followed by correction for multiple hypothesis testing by two different methods, Bonferroni and FDR.
- **Supported database**: pathDIP integrates ASCN2 [40], BioCarta [41], EHMN [42], HumanCyc [10], INOH [43], IPAVS [44], KEGG [11], NetPath [33], OntoCancro [45], Panther[13], PharmGKB [46], PID [34], RB-pathways [47], Reactome [12], Signalink2.0 [48], SIGNOR2.0 [49], SMPDB [50], SPIKE [51], STKE [9], System-biology.org [52], UniProt Pathways (https://www.uniprot.org/help/pathway) and WikiPathways [15].

ORA tools can be used with lists of genes obtained from all the databases listed in 3.1. When using data from host-pathogen interaction studies, lists will be composed of host genes only.

GSEA tools

GSEA has been developed for gene expression data obtained through microarrays, but it is a frequently used method for pathway enrichment analysis in any set of genes for which differential gene expression is available. Below are some tools performing GSEA:

clusterProfiler

clusterProfiler [53] focuses on the enrichment and comparison of gene clusters and the classification and visualization of biological terms.

- **Availability**: it can be downloaded freely as an R library/-package directly from https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html.
- **OS platform**: clusterProfiler can be executed on each OS compatible with the R language.
- **Input data format**: clusterProfiler requires as input a list of gene or protein identifiers of interest for ORA, and a ranked gene list (with fold change or other numeric variable) for GSEA.
- **Output data format**: clusterProfiler allows users to export results as images (i.e. in SVG, JPG and PNG formats), or as a ranked list of enriched pathways.
- **Analysis**: clusterProfiler support ORA, GSEA and biological theme comparison. Enrichment is obtained by using the hypergeometric test or enrichment score. To correct the possible errors due to the multiple hypothesis tests, clusterProfiler uses the *FDR* corrector.
- **Supported database**: clusterProfiler can use pathways information coming from KEGG [11] and WikiPathways [15]. A separate package is required to use Reactome database.

GSEA

GSEA [54] uses gene expression data coming from samples belonging to two classes, e.g. responder or non-responder, treated or non-treated. GSEA is particularly suitable when ranks are available for all the genes under investigation.

- **Availability**: GSEA is freely available as a stand alone application. To download the GSEA software, users have to register at https://www.gsea-msigdb.org/gsea/login.jsp. GSEA is available as Java desktop application with an easy to use graphical interface (i.e. recommended for all users with little programming skills), as well as a Java jar file command line interface, useful for expert programmers to analyze large datasets or running analyses on high performance machines. Finally, GSEA R implementation is intended for experienced computational biologists.
- **OS platform**: Both Java and R versions of GSEA are compatible with all the OS that support Java and R programming languages.
- **Input data format**: to run PEA, GSEA needs four types of data files: an expression dataset in any of the formats RES, Gene Cluster Text (GCT), PCL or txt file formats; phenotype labels in categorical class (CLS) format; gene sets defined using the Gene MatriX (GMXe) or gene matrix transposed (GMT) file format, and CHIP: Chip file file format with microarray annotations. All four file types have to be tab-delimited.
- **Output data format**: GSEA produces different types of reports: enrichment in phenotype, dataset details, gene set details, gene markers, global statistics and plots, detailed enrichment results and gene set details report. All the reports can be exported as HTML, or xlsx (Excel) format.
- **Analysis**: The functional enrichment of the input gene list is evaluated using *Enrichment Score*. To correct results obtained through multiple testing, GSEA uses the FDR and Bonferroni correctors.
- **Supported database**: GSEA can retrieve information from KEGG [11], BioCarta [41], PID [34] and Reactome [12] pathway databases. Moreover, users can upload any other pathway database data in GMT format.

WEB-based Gene SeT AnaLysis Toolkit (WebGestalt)

WebGestalt is a collection of tools for functional enrichment analysis in several biological contexts. The current version of WebGestalt [55] supports 12 organisms, 342 gene identifiers from different databases, and 155,175 functional categories.

- **Availability**: WebGestalt is available as web application at http://www.webgestalt.org. WebGestalt is also available as an *R* package at the CRAN archive at https://cran.r-project.org/web/packages/WebGestaltR/index.html.
- **OS platform**: WebGestalt web application is OS independent, and can be used in all the OS supporting *R* language.
- **Input data format**: WebGestalt receives as input a list of genes, along with functional categories with their own gene identifiers.
- **Output data format**: Enrichment results can be explored and analyzed interactively through a graphical user friendly interface, in the form of tab-based and interactive report. In addition, the GUI allows users to export enrichment results as reports and figures (i.e. in SVG and PNG formats) that can be used in publications.
- **Analysis**: WebGestalt ORA enrichment method is based on the the Fisher's test. From the WebGestalt-2019 version, it supports TEA, ORA and GSEA.
- **Supported database**: WebGestalt can obtain information from KEGG [11], WikiPathways [15], Reactome [12] and Panther [13] pathway databases.

GSEA tools can be used with list of genes paired with fold changes or p-values. This type of data can usually be found in tables linked to COVID-19 specific publications or it can be retrieved (but needs to be pre-processed) from databases such as GEO (https://www.ncbi.nlm.nih.gov/geo/), SRA (https://www.ncbi.nlm.nih.gov/sra/), ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) or PRIDE (https://www.ebi.ac.uk/pride/archive/).

TEA tools

Pathways are representations of biological events describing the interactions of genes, proteins or metabolites within cells, tissues or organisms, not simple lists of genes. Genes and proteins are not independent, they perform a variety of functions and tasks through their interactions and connections. To take advantage of the pathway topology information for assessing statistical relevance of the pathways a third category of methods called TEA has been proposed. TEA includes the following tools (as well as WebGestalt, already mentioned above):

EnrichNet

EnrichNet [56] uses topological information to identify, prioritize and analyze functional associations between user-defined gene or protein sets and cellular pathways using information from molecular interaction networks.

- **Availability**: EnrichNet is a web application for pathway analysis using topological information available at http://www.enrichnet.org. It is available as Python package as well as RESTful API.
- **OS platform**: EnrichNet is platform independent and can be used through a web browser.
- **Input data format**: To perform TEA it requires as input a list of gene or protein identifiers of interest. The list of proteins/genes of interest cannot exceed the 5 000 genes/proteins identifier per single analysis.
- **Output data format**: The main output produced by EnrichNet is a ranking of pathways or a rank of gene ontology (GO) in terms of their predicted functional association with the provided gene/protein list.
- **Analysis**: To perform TEA, EnrichNet implements GSEA by a new association measure that integrates information from the known network structure of interactions between proteins. The interactions are computed using the random walk with restart (RWR) algorithm, and the statistical relevance is computed using the Fisher's test and the FDR method for multiple testing adjustment.
- **Supported database**: EnrichNet can use pathways information coming from the following pathway and process databases: KEGG [11], BioCarta [41], Reactome [12], WikiPathways [15], GO [57] and NCI pathway database [58].

Pathway analysis using Network information (PathNet)

PathNet [59] uses topological information present in pathways and differential expression levels of genes, to identify pathways significantly enriched associated in the context of gene expression data.

- **Availability**: PathNet is a set of R functions for pathway analysis using topological information. PathNet is available as an R workspace image from http://www.bhsai.org/downloads/pathnet/.
- **OS platform**: PathNet is Platform independent and can be executed on all OS supporting the *R* language.
- **Input data format**: PathNet requires differential expression levels, an adjacency matrix file containing the connectivity information among genes in the list of interest, and pathway information. PathNet is distributed with example text files to use as a reference when creating new datasets for analysis.
- **Output data format**: PathNet enrichment analysis results are displayed on the screen and stored in two plain-text files.
- **Analysis**: To perform TEA, PathNet combines all pathways under consideration into a pooled pathway. The interactions among genes in the pooled pathway are represented by an adjacency matrix, and given the network, PathNet computes the molecular relevance using Fisher's test.
- **Supported database**: PathNet uses topology information collected from KEGG [11] pathway database only. A researcher willing to use a different pathway database needs to format it as the KEGG pathways file provided with the package.

Topology-based pathway enrichment analysis (TPEA)

TPEA [60] method, computes the relevance of nodes based on its upstream/downstream positions and the degrees in pathways.

- **Availability**: TPEA is available at the Comprehensive R Archive Network (CRAN) repository (https://cran.r-project.org/web/packages/TPEA/).
- **OS platform**: TPEA is compatible with all the OS supporting *R* programming language.
- **Input data format**: The gene set of interest must be a list in *"Entrez ID"* format.
- **Output data format**: TPEA allows users to save ranked PEA results in a tabular format.
- **Analysis**: TPEA computes the area under the enrichment curve (AUEC), which was obtained based on the cumulative weighted node score of the relevant nodes, to evaluate the enrichment significance of pathways.
- **Supported database**: TPEA uses only KEGG database [11] to compute PEA.

TEA tools can be used with gene lists similar to the ones used for ORA or GSEA, but they also require interaction information that, for COVID-19, can be retrieved in CoVex, IntAct and SIGNOR (the latter being focused on COVID-19 related signaling pathways). Virus - host and drug - target networks cannot be used for TEA, as the tools require interactions among the genes present in the pathways for topological analysis.

Table 1 lists scientific papers focused on COVID-19 that used HT assays and PEA methodologies. The HT experimental assays mostly used to produce SARS-CoV-2-related genes/proteins of interest include mass spectrometry and RNA-Seq. Before performing PEA, each dataset has been properly pre-processed to be suitable to perform the appropriate PEA. Most of the PEA tools listed in Table 1 used ORA or GSEA methodologies; specifically, 13 papers used ORA and 10 GSEA. ORA is the methodology easiest to apply, as it needs only a list of proteins/genes, and hence it is the one used more frequently. GSEA requires additional data and processing, but the method has been increasingly used since its publication in 2005. While TEA methods are more complete they are also more complex to use. They require interaction and topology data to be performed, and they have not been used in COVID-19 papers yet.

Table 1. List of COVID-19 research papers using high throughput method to generate raw data analyzed using PEA methodologies

Paper	HT method	PEA method	Tool	Reference
Proteomic and metabolomic characterization of COVID-19 Ppatient Ssera	RNA-Seq	ORA	IPA[a]	[61]
Multi-level proteomics reveals host-perturbation strategies of SARS-CoV-2 and SARS-CoV	MS and RNA-seq	ORA	IPA	[62]
Bulk and single-cell gene expression profiling of SARS-CoV-2 infected human cell lines identifies molecular targets for therapeutic intervention	sc and bulk RNA-seq	GSEA	clusterProfiler	[63]
Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients	RNA-seq	GSEA	clusterProfiler	[64]
A SARS-CoV-2 – host proximity interactome	BioID	ORA	g:Profiler	[65]
Transcriptomic profiling of human corona virus (HCoV)-229E -infected human cells and genomic mutational analysis of HCoV-229E and SARS-CoV-2	RNA-Seq	ORA	IPA	[66]
Host metabolic reprogramming in response to SARS-Cov-2 infection	RNA-Seq	ORA	EnrichR	[67]
Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19	sc RNA-seq	GSEA	clusterProfiler	[68]
Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19	RNA-Seq	ORA	enricher	[69]
A single-cell atlas of the peripheral immune response in patients with severe COVID-19	scRNA-seq	ORA	IPA	[70]
Generation of human bronchial organoids for SARS-CoV-2 research	RNA-seq	GSEA	PGSEA	[71]
In vivo antiviral host response to SARS-CoV-2 by viral load, sex, and age [dataset I]	shotgun RNA sequencing	GSEA	GSEA	[72]
Severely ill COVID-19 patients display a defective exhaustion program in SARS-CoV-2 reactive CD8+ T cells	scRNA-seq	GSEA	GSEA	[73]
Type I and Type III IFN Restrict SARS-CoV-2 Infection of Human Airway Epithelial Cultures	RNA-seq	GSEA	GSEA[b]	[74]
Modulating the transcriptional landscape of SARS-CoV-2 as an effective method for developing antiviral compounds	RNA-Seq	ORA	Enricher	[75]
In vivo antiviral host response to SARS-CoV-2 by viral load, sex, and age [dataset II]	shotgun RNA-seq	GSEA	GSEA	[76]

[a] Ingenuity pathway analysis [b]MSigDB molecular signature database

Choosing a PEA method

This section provides basic guidelines on how to choose the most suitable PEA method considering intended data and research goal.

All three PEA methodologies use genes/proteins of interest and a pathway database to identify critical pathways that may be affected in a condition by correlating information in a pathway with genes/proteins for the disease. Thus, the type of the genes/proteins list dictates the choice of the PEA method and suggests the database to use to calculate the enrichment. Another aspect to consider when performing PEA is that databases contain different representations of the same biological pathway, which may lead to varying PEA results. Thus, before performing the PEA, a researcher should carefully select the most suitable pathway database for his/her research purpose. For example, to investigate signal transduction a signaling specific database should be used (e.g. SIGNOR), or to explore metabolic aspects a metabolism specific database should be used (e.g. MetaCyc). Thus, the selection of a suitable pathway database depends on the biological context that is under investigation.

To compute the enrichment, **ORA** methods require a simple input list of genes/proteins, along with a pathway database (or more databases if supported by the tool). Among the reviewed ORA software tools, BiP allows users to retrieve pathway information from each pathway database compatible with the BioPAX format, while Enrichr, g:Profiler, and pathDIP perform PEA with a pre-selected set of pathway databases.

GSEA methods need to use the gene annotations to compute pathway enrichment. **GSEA** methods use the entire list of genes/proteins obtained from gene expression studies, along with additional annotations obtained from fold changes or *P*-values. Moreover, GSEA methods requires that the set of samples used to produce the data to analyze contain at least 3–5 samples per group. Thus, to obtain consistent pathway enrichment results using GSEA methods, the choice of the specific software should be data-driven. In particular, if expression dataset, phenotype labels, and gene sets with microarray annotations are available, GSEA software tools should be chosen to perform PEA. On the other hand, if microarray data sets are incomplete or only a list of proteins/genes is available, PEA can be computed

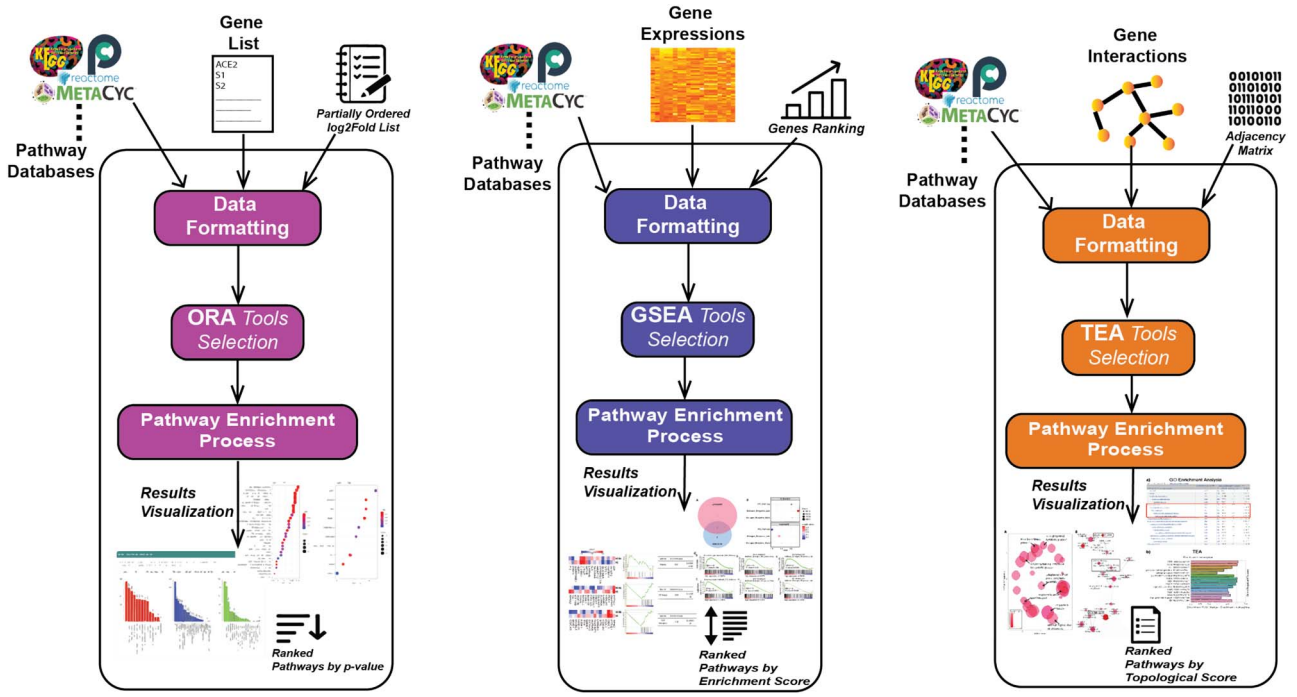

Figure 1. The main steps needed to perform PEA, by using a ORA, GSEA or TEA software tool.

using clusterProfiler and WebGestalt software tools. Both clusterProfiler and WebGestalt software tools require as an input the list of ranked genes/proteins to perform the enrichment and a pathway database. A possible drawback related to clusterProfiler is that it includes only KEGG and wikiPathways databases. GSEA software tool can obtain pathway information from BioCarta, KEGG, PID and Reactome, while WebGestalt can get pathway information from KEGG, WikiPathways, Reactome and Panther databases.

TEA methods utilize the number and type of interactions between gene products, which may or may not be a part of a pathway database, in addition to functional annotations. TEA methods have been developed to use additional information regarding pathway topology to compute gene-level statistics. One obvious problem is that proper pathway topology is dependent on the type of cell used to produce the data, due to cell-specific gene expression profiles and conditions being studied, and this type of data is not always available. The assessed **TEA** software tools use genes/proteins lists and interaction information available in specific interaction databases such as IntAct. PathNet requires differential expression levels, an adjacency matrix file containing the connectivity information among genes in the list of interest, and pathway information. On the other hand, both EnrichNet and TPEA software tools require a list of genes/proteins; for EnrichNet the list of interest cannot exceed 5000 genes/proteins, whereas, for TPEA the genes/proteins identifiers have to be in 'EnterezID' format. Finally, the choice of a TEA method to use is related to the biological context that will be investigated. To that end, TPEA and PathNet can get pathway information only from KEGG, whereas EnrichNet can obtain pathway information from BioCarta, KEGG, Reactome, NCI Pathway and WikiPAthways databases, providing broader context and less bias.

Discussion

The description of each tool provides the necessary information for a user to decide what type of tool and what type of data are compatible and best suited for an accurate and appropriate analysis.

PEA methods help researchers to identify the pathways significantly impacted from a collection of proteins/genes of interest i.e. the list of COVID-19 related genes. To approach this problem, PEA methods need at least *i*) a collection of pathways of an organism (usually collected from a single or multiple pathway databases), and *ii*) experimental data such as gene expressions or proteins/genes list. Figure 1 graphically highlights the main steps required for the available PEA approaches.

In Figure 2 we compare all frameworks reviewed based on the supported input types and formats, the provided output results, along with the supported pathway databases used to define the biological context of the investigated proteins/genes.

Subset of frameworks analyzed in Figure 2 accept as input the whole list of genes/proteins considered in the experiment together with their expression values. Other frameworks use only the list of differential expressed genes, without the corresponding expression values, and other frameworks need additional input data. Among the surveyed frameworks, BiP, clusterProfiler, EnrichNet, Enrichr, g:Profiler, pathDIP, TPEA and WebGestalt use all genes/proteins with or without their expression values as input, whereas, GSEA and PathNet use all genes/proteins with their expression values along with further additional files as input.

Moreover, for input datasets to be suitable for PEA they must contain a minimum number of samples. In particular, sample size affects the specificity and sensitivity of the enrichment results. The sample size heavily affects the reproducibility of the PEA results for both ORA and GSEA methodologies. A methodological procedure able to assist users in defining the opportune

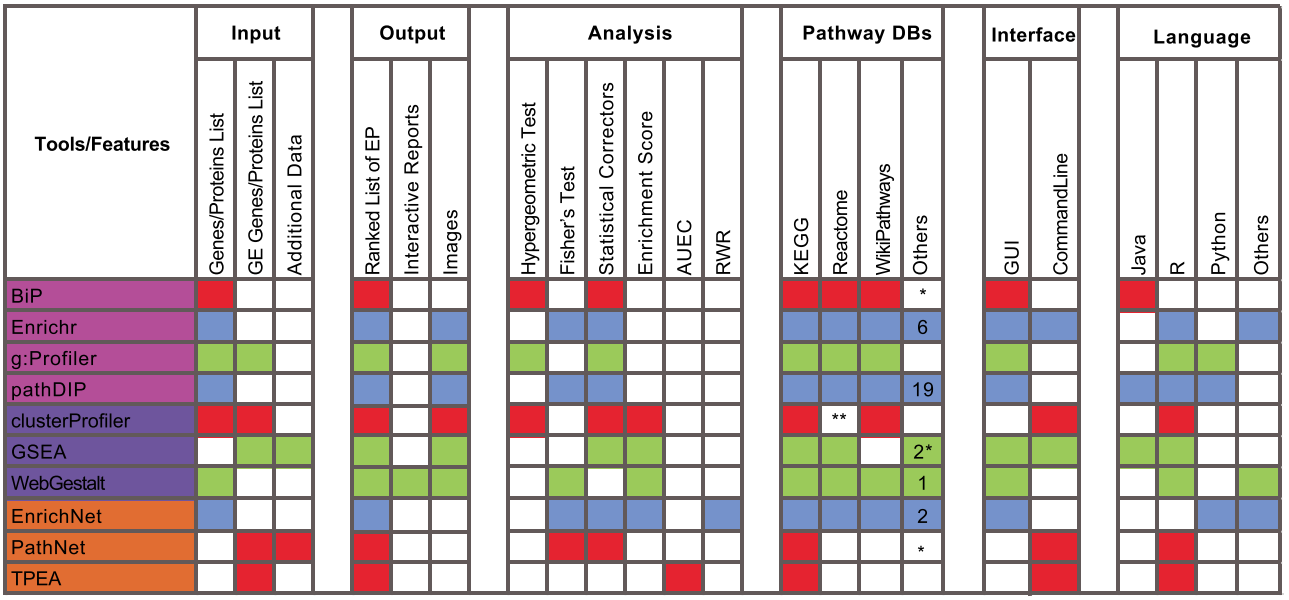

Figure 2. Heatmap representation of the comparison among the surveyed PEA software tools. The violet color indicates ORA tools, the purple color indicates GSEA tools, orange color indicates TEA tool. The red color indicates the features of the standalone tools, blue color represent the features of the web-tools, and green color refers to the features of the tools available as both standalone and web applications. '*' Indicates the tools can import pathway from further databases other than KEGG, Reactome and WikiPathways. The values in the *Other* column, indicates the number of additional supported pathway database for each tool. '**' Indicates that data can be obtained from the database through additional modules.

sample size of the datasets to use with ORA and GSEA software tools, is provided in [77], where it is highlighted that enrichment results reproducibility is unlikely for both methods by using small sample sizes (3-5 samples per group). In the GSEA user's guide it is specified as well that an input dataset must contain at least three samples for each group to be suitable for enrichment analysis. Even enrichment results obtained by using TEA tools are affected by the sample size, with more samples per group providing better results, as reported in [78] and [79].

Pathway data are the second type of input for the surveyed PEA frameworks. Pathway data generally are collected from a single or multiple sources (pathway databases). Most of the surveyed frameworks in Figure 2 use multiple pathway databases; for example, *pathDIP* integrates 22 pathway databases, BiP can use all databases encoded in BioPAX format and *GSEA* can run the analysis on any database in GMT format. The majority of tools, though, by default can run PEA using KEGG (all the tools mentioned in this paper), Reactome or WikiPathways (7 tools each). Of these, Reactome is the largest curated primary database, including 2423 pathways for *Homo sapiens* and annotating 10 923 proteins. Reactome includes curated pathways for 15 additional organisms, is freely available and can be downloaded in the most frequently used pathway exchange formats. KEGG is the oldest primary curated pathway database, with its first version having being released in 1996. The database is now available for download only through subscription, but can be queried through API. KEGG is also the second largest database in terms of genome coverage, with 7217 annotated proteins, and WikiPathways is the third with 6233. Although the pathway analysis results should be a ranked list of pathways, not all tools reviewed here provide this. Some return interactive data while other provide plots that can be useful for presentations or publications. Other tools enable further analysis, such as identifying the most common terms in pathway names as in pathDIP. Among the inspected frameworks, BiP, clusterProfiler, Enrichr, pathDIP, PathNet and TPEA provide a ranked list of enriched pathways, whereas g:Profiler, Enrichr, WebGestalt and GSEA provide interactive reports that can be easily exported as images. Tools like clusterProfiler, Enrichr, g:Profiler and WebGestald, provide static images in SVG, JPG and PNG formats that can be used for publication.

Although the main strength of a framework lies in its computational algorithm, its implementation plays a fundamental role in the user experience. Practicality, ease of use, and type of interface play a central role in choosing one framework over another. The surveyed frameworks can be classified as web-based or standalone frameworks.

Web-based frameworks run on a remote server, providing computational power and graphical interface. They can be used through a web browser by uploading the data to be analyzed and the results can be collected in the format provided by the web framework. Moreover, the users are not requested to configure or maintain high-performance hardware and do not need computational expertise. Among the web-based frameworks there are EnrichNet, Enrichr, g:Profiler, pathDIP and WebGestalt.

Standalone frameworks need to be installed on local machines, a task that requires some programming skills. Advantages include data security and privacy, as well as the possibility to perform data analysis locally without any limitation and without depending on network connection. However, standalone tools have to be installed on a high-performance machine to provide good scalability when analyzing massive datasets. Standalone frameworks category encompass BiP, clusterProfiler, g:Profiler, GSEA, PathNet and TPEA.

The programming language and style used for software implementation play an essential role in the future adoption of a software tool. Software tools that are skillfully implemented through an effective graphical user interface (GUI) are more appealing than those that do not have a GUI requiring some programming skills to use the full functionalities.

Table 2. Number of pathways enriched per tool per database, considering a *P*-value lower than 0.05.

Database	BiP	Enrichr	g:Profiler	pathDIP	clusterProfiler	GSEA	WebGestalt	EnrichNet	PathNet	TPEA
KEGG	63	228	5	2,140	NE	1	10	228	NE	109
Reactome	1,300	805	26	3,512	NA	33	10	805	NA	NA
WikiPathways	NE	300	6	1,162	NE	NA	10	300	NA	NA

NE = not enriched; NA = not available.

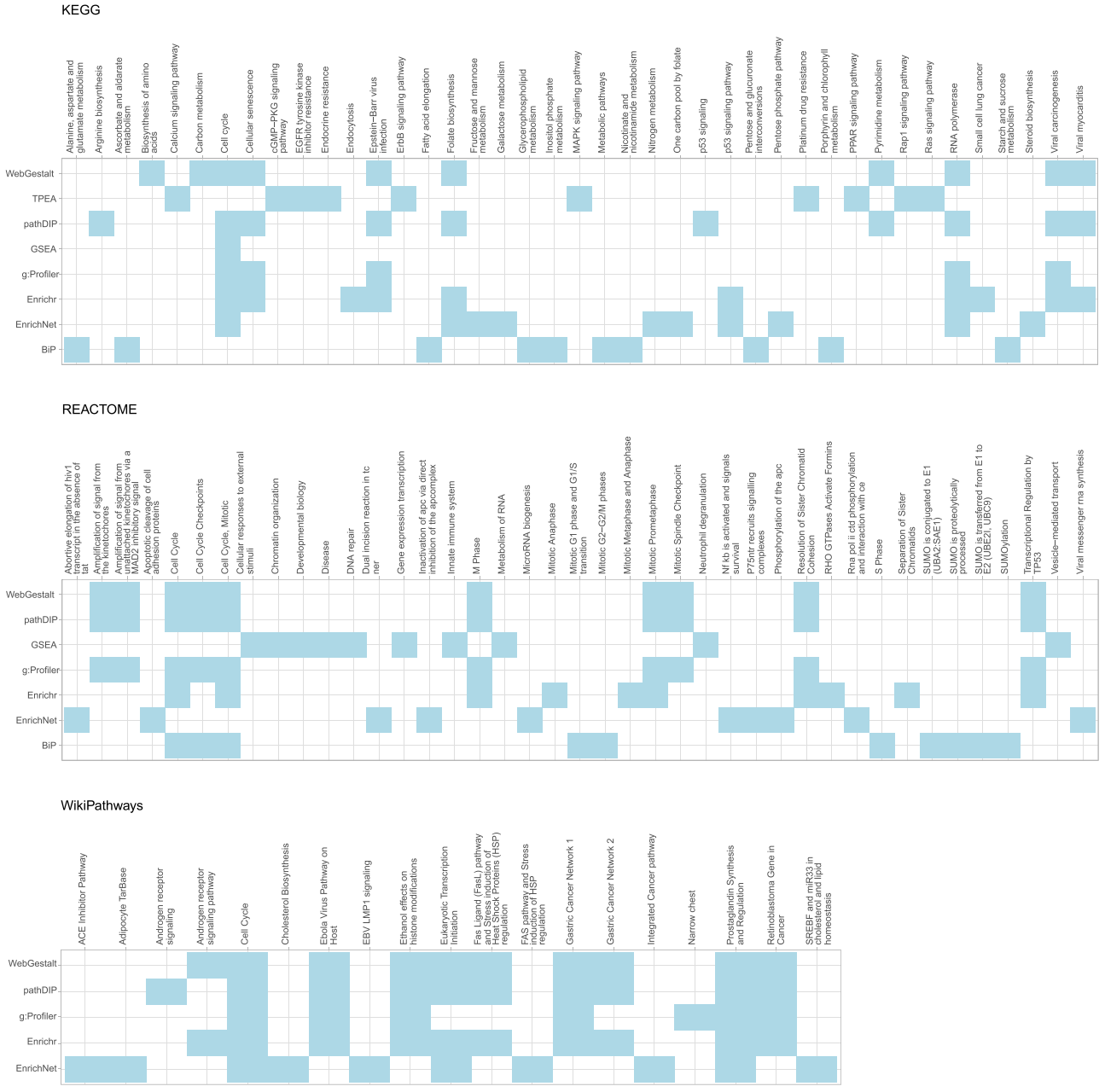

Figure 3. Top 10 pathways enriched per database. Rows represent tools listed in this paper and columns pathways enriched.

clusterProfiler, g:Profiler, GSEA, PathNet and TPEA are developed using the R programming language and are available as software packages either from Bioconductor and CRAN repositories. Their demand among biologists and bioinformaticians is due to the availability of many R bioinformatics packages.

BiP, GSEA and Enrichr have been implemented in Java, providing access to the full functionalities through a GUI for users with lower computational expertise. In addition, g:Profiler, GSEA and WebGestalt provide both web and standalone implementations.

A summary of the analysis methods, framework type, implementation, and availability details for the surveyed methods is presented in Figure 2.

Our PEA guide is appropriate to analyze lists of genes or biomolecules from any organism and is not limited to COVID-19 only, but we are focusing this section on the data available at the time of writing for the current pandemic.

A gene list can be generated using HT assays (such as RNA-seq, genome-wide association studies and proteomics) that generate a high amount of raw data that must be processed to obtain input information suitable for PEA. After processing, the data available can be stored as differential expression gene lists—with or without associated expression—or as gene, proteins and microRNAs lists as well, ranked by other features. A user can also collect a gene list of interest from the databases listed in COVID-19 databases.

A differential gene expression list can be obtained by a user after the analysis of their HT assays or from databases like GEO (https://www.ncbi.nlm.nih.gov/geo/), SRA (https://www.ncbi.nlm.nih.gov/sra/) and ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) that focus on expression related datasets. This type of list can be used as input in GO-Elite, GSEA and *WebGestalt*.

Partially ranked or unranked gene lists can be obtained from the literature (as of August 7th, PubMed lists 38 521 COVID-19 and SARS-CoV-2 related papers, while medRxiv/bioRxiv list 7465) as well as from databases like DisGeNet, CORDITE and CTD. This type of list can be analyzed using BiP, clusterProfiler, Enrichr, EnrichNet, g:Profiler or pathDIP. CTD can also provide ranked gene lists based on the effect of a compound on the expression of a set of genes of interest. Such a list can be used with GSEA and WebGestalt. Databases like IntAct store host–host interactions, providing the user with data useful for TEA that can be performed using EnrichNet, PathNet, TPEA and WebGestalt.

To showcase possible results obtained by analysis of COVID-19 data, we collected significant COVID-19 genes along with the related log-fold change values from Supplementary Table 5 of Stukalov *et al.* [62]. The extracted data have been properly formatted to be analyzed using each tool reviewed in this paper. As first step we collected all the significant COVID-19 genes available in the Supplementary Table 5 of Stukalov *et al.* [62], along with the detected log-fold change values. As next step, we produced two genes' list ordered and unordered, both suitable to perform PEA with all the following framework tools: BiP, clusterProfiler, Enrichr, EnrichNet, g:Profiler, pathDIP and TPEA. To perform PEA, using both GSEA and PathNet, it is necessary to provide the data used for ordering the list (in our example, *P*-value). To perform PEA with PathNet it is also necessary to provide an adjacency matrix file containing the connectivity information among genes in the list of interest.

Figure 3 shows top 10 pathway enrichment results for each database obtained by the PEA software tools listed in this paper. The number of enriched pathways are obtained considering a *P*-value lower than 0.05. Table 2 shows which tools did not support or did not obtain any enrichment for a specific database. Interestingly, all three databases and the majority of tools show enrichment for pathways related to *cell cycle*, a process well known to be disrupted after viral infection [80].

Conclusion

Considering the veracity and volume of COVID-19 information, we have an urgent need to identify effective ways of extracting knowledge from quickly produced data in the fastest and most accurate way. PEA approaches are a core strategy to obtain knowledge and annotate SARS-CoV-2 data. PEA applications allow users to find and characterize key components crucial to identify efficacious treatments and highlight individual differences. In this manuscript, we presented the main software tools for pathways enrichment analysis. We described the features to use in order to choose the most suitable pathway enrichment tool for the specific type of COVID-19 data to be investigated. This way, researchers can find information about various tools and methods in a single place, and can make an informed decision about the more appropriate tool to use.

Conflict of interest

All the authors declare that the research was conducted in the absence of any potential conflict of interest.

Key Points

- COVID19 research generated an unprecedented volume of papers and data, which makes selecting the best data, tools and analysis challenging.
- This paper provides a comprehensive list of bioinformatics methods and resources used to analyze available COVID-19 data with pathway enrichment analysis.
- We describe a simple guide of the main steps of a general pathway enrichment analysis procedure to quickly gain insight into the genes/proteins list of interest derived from COVID-19 data.

References

1. Habibzadeh P, Stoneman EK. The novel coronavirus: a bird's eye view. *Int J Occupational Environ Med* 2020; **11**(2): 65.
2. Zhou P, Yang X-L, Wang X-G, *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020; **579**(7798): 270–3.
3. Yu C, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. *J Med Virol* 2020; **92**(4): 418–23.
4. Thorlund K, Dron L, Park J, *et al.* A real-time dashboard of clinical trials for COVID-19. *Lancet Digital Health* 2020; **2**(6): e286–7. doi: 10.1016/S2589-7500(20)30086-8.
5. Jia Y, Shen G, Zhang Y, *et al.* Analysis of the mutation dynamics of SARS-COV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity. *BioRxiv* 2020.
6. Shen Z, Xiao Y, Lu K, *et al.* Genomic diversity of SARS-COV-2 in coronavirus disease 2019 patients. *Clin Infect Dis* 2020a.
7. Phan T. Genetic diversity and evolution of SARS-COV-2. *Infect Genet Evol* 2020; **81**:104260.
8. Kumar S. COVID-19: a drug repurposing and biomarker identification by using comprehensive gene-disease associations through protein-protein interaction network analysis. 2020.
9. Gough NR. Science's signal transduction knowledge environment: the connections maps database. *Ann N Y Acad Sci* 2002; **971**(1): 585–7.
10. Trupp M, Altman T, Fulcher CA, *et al.* Beyond the genome (BTG) is a (PGDB) pathway genome database: Humancyc. *Genome Biol* 2010; **11**(1): 1–1.
11. Ogata H, Goto S, Fujibuchi W, *et al.* Computation with the KEGG pathway database. *Biosystems* 1998; **47**(1–2): 119–28.

12. Joshi-Tope G, Gillespie M, Vastrik I, *et al*. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005; **33**(suppl_1): D428–32.
13. Mi H, Lazareva-Ulitsky B, Loo R, *et al*. The panther database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* 2005; **33**(suppl_1): D284–8.
14. Cerami EG, Gross BE, Demir E, *et al*. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res* 2010; **39**(suppl_1): D685–90.
15. Pico AR, Kelder T, Van Iersel, *et al*. Wikipathways: pathway editing for the people. *PLoS Biol* 2008; **6**(7):e184.
16. Sara Rahmati, Chiara Pastrello, Andrea E.M. Rossos, *et al*. Two decades of biological pathway databases: Results and challenges. In Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach, editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 1071–84. Academic Press, Oxford, 2019a. ISBN 978-0-12-811432-2. doi:10.1016/S2589-7500(20)30086-810.1016/B978-0-12-809633-8.20496-2. http://www.sciencedirect.com/science/article/pii/B9780128096338204962.
17. Sara Rahmati, Mark Abovsky, Chiara Pastrello, *et al*. pathDIP 4: an extended pathway annotations and enrichment analysis resource for human, model organisms and domesticated species. *Nucleic Acids Res*, **48** (D1): D479–88, 2019b. ISSN 0305-1048. doi: 10.1016/S2589-7500(20)30086-810.1093/nar/gkz989.
18. Glaab E, Baudot A, Krasnogor N, *et al*. Extending pathways and processes using molecular interaction networks to analyse cancer genome data. *BMC Bioinformatics* 2010; **11**(1): 597. doi: 10.1186/1471-2105-11-597.
19. Marco-Ramell A, Palau-Rodriguez M, *et al*. Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. *BMC Bioinformatics* **19**(1, 2018): 1. doi: 10.1016/S2589-7500(20)30086-810.1186/s12859-017-2006-0 ISSN 1471-2105.
20. Roman Martin, Hannah F Löchel, Marius Welzel, *et al*. CORDITE: the curated CORona drug InTERactions database for SARS-CoV-2. *iScience*, **23** (7), jul 2020. doi: 10.1016/j.isci.2020.101297.
21. Davis AP, Grondin CJ, Johnson RJ, *et al*. The comparative Toxicogenomics database: update 2019. *Nucleic Acids Res* jan 2019. ISSN 1362-4962; **47**(D1): D948–54. doi: 10.1093/nar/gky868. https://pubmed.ncbi.nlm.nih.gov/30247620https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323936/.
22. 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): 3518. doi: 10.1038/s41467-020-17189-2.
23. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, *et al*. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 11 2019; **48**(D1): D845–55. doi: 10.1093/nar/gkz1021.
24. Orchard S, Ammari M, Aranda B, *et al*. The mintact project–intact as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* January 2014. ISSN 0305-1048; **42**(Database issue): D358–63. doi: 10.1093/nar/gkt1115. https://europepmc.org/articles/PMC3965093.
25. Orchard S, Kerrien S, Abbani S, *et al*. Protein interaction data curation: the international molecular exchange (IMEx) consortium. *Nat Methods* 2012; **9**(4): 345–50. doi: 10.1038/nmeth.1931.
26. Perfetto L, Pastrello C, Del-Toro N, *et al*. The IMEX coronavirus interactome: an evolving map of coronaviridae-host molecular interactions. *BioRxiv* 2020.
27. Licata L, Surdo PL, Iannuccelli M, *et al*. SIGNOR 2.0, the SIGnaling network open resource 2.0: 2019 update. *Nucleic Acids Res* 10 2019; **48**(D1): D504–10. doi: 10.1093/nar/gkz949.
28. Navratil V, de Chassey, Meyniel L, *et al*. VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic Acids Res* 11 2008; **37**(suppl 1): D661–8. doi: 10.1093/nar/gkn794.
29. Gentleman RC, Carey VJ, Bates DM, *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004; **5**(10): R80.
30. Dabbish L, Stuart C, Tsay J, *et al*. Social coding in github: transparency and collaboration in an open software repository. In: *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 2012, 1277–86.
31. Agapito G, Pastrello C, Guzzi PH, *et al*. BioPAX-parser: parsing and enrichment analysis of BioPAX pathways. *Bioinformatics* 05 2020. doi: 10.1093/bioinformatics/btaa529.
32. Demir E, Cary MP, Paley S, *et al*. The BioPAX community standard for pathway data sharing. *Nat Biotechnol* 2010; **28**(9): 935–42. doi: 10.1038/nbt.1666.
33. Kandasamy K, Mohan SS, Raju R, *et al*. Netpath: a public resource of curated signal transduction pathways. *Genome Biol* 2010; **11**(1): 1–9.
34. Schaefer CF, Anthony K, Krupa S, *et al*. Pid: the pathway interaction database. *Nucleic Acids Res* 2009; **37**(suppl_1): D674–9.
35. Chen EY, Tan CM, Kou Y, *et al*. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013; **14**(1): 128. doi: 10.1186/1471-2105-14-128.
36. Rouillard AD, Gundersen GW, Fernandez NF, *et al*. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* 2016, 07 2016. doi: 10.1093/database/baw100.
37. Huang R, Grishagin I, Wang Y, *et al*. The ncats bioplanet – an integrated platform for exploring the universe of cellular signaling pathways for toxicology, systems biology, and chemical genomics. *Front Pharmacol* 2019. ISSN 1663-9812; **10**:445. doi: 10.3389/fphar.2019.00445. https://www.frontiersin.org/article/10.3389/fphar.2019.00445.
38. Nesterova AP, Yuryev A, Klimov EA, *et al*. *Disease Pathways: An Atlas of Human Disease Signaling Pathways*. Elsevier, 2019.
39. Raudvere U, Kolberg L, Kuzmin I, *et al*. G:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* 05 2019; **47**(W1): W191–8. doi: 10.1093/nar/gkz369.
40. Kuperstein I, Bonnet E, Nguyen H-A, *et al*. Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with google maps. *Oncogene* 2015; **4**(7): e160–0.
41. Nishimura D. Biocarta. *Biotech Software Internet Report: Comput Software J Scient* 2001; **2**(3): 117–20.
42. Ma H, Sorokin A, Mazein A, *et al*. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* 2007; **3**(1): 135.
43. Yamamoto S, Sakai N, Nakamura H, *et al*. INOH: ontology-based highly structured database of signal transduction pathways. *Database* 11 2011; **2011**. doi: 10.1093/database/bar052.
44. Sreenivasaiah PK, Rani S, Cayetano J, *et al*. Ipavs: integrated pathway resources, analysis and visualization system. *Nucleic Acids Res* 2012; **40**(D1): D803–8.

45. Simão ÉM, Cabral HB, Castro MAA, *et al*. Modeling the human genome maintenance network. *Physica A: Stat Mechanics Its Appl* 2010; **389**(19): 4188–94.
46. Whirl-Carrillo M, McDonagh EM, Hebert JM, *et al*. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Therapeutics* 2012; **92**(4): 414–7.
47. Calzone L, Gelay A, Zinovyev A, *et al*. A comprehensive modular map of molecular interactions in RB/E2F pathway. *Mol Syst Biol* 2008; **4**(1): 0174.
48. Fazekas D, Koltai M, Türei D, *et al*. Signalink 2–a signaling pathway resource with multi-layered regulatory networks. *BMC Syst Biol* 2013; **7**(1): 1–15.
49. Perfetto L, Briganti L, Calderone A. Andrea Cerquone Perpetuini, Marta Iannuccelli, Francesca Langone, Luana Licata, Milica Marinkovic, Anna Mattioni, Theodora Pavlidou, et al. Signor: a database of causal relationships between biological entities. *Nucleic Acids Res* 2016; **44**(D1): D548–54.
50. Jewison T, Yilu S, Disfany FM, *et al*. *Nucleic Acids Res* 2014; **42**(D1): D478–84.
51. Paz A, Brownstein Z, Ber Y, *et al*. Spike: a database of highly curated human signaling pathways. *Nucleic Acids Res* 2011; **39**(suppl_1): D793–9.
52. Kitano H, Funahashi A, Matsuoka Y, *et al*. Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol* 2005; **23**(8): 961–6.
53. Yu G, Wang L-G, Han Y, *et al*. Clusterprofiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012; **16**(5): 284–7. doi: 10.1089/omi.2011.0118.
54. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 2005. ISSN 0027-8424; **102**(43): 15545–50. doi: 10.1073/pnas.0506580102. https://www.pnas.org/content/102/43/15545.
55. Wang J, Vasaikar S, Shi Z, *et al*. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res* 05 2017; **45**(W1): W130–7. doi: 10.1093/nar/gkx356.
56. Glaab E, Baudot A, Krasnogor N, *et al*. Enrichnet: network-based gene set enrichment analysis. *Bioinformatics (Oxford, England)* 09 2012; **28**(18): i451–7. doi: 10.1093/bioinformatics/bts389. https://pubmed.ncbi.nlm.nih.gov/22962466.
57. Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res* 2019; **47**(D1): D330–8.
58. Krupa S, Anthony K, Buchoff J, *et al*. The NCI-nature pathway interaction database: a cell signaling resource. *Nat Prec* 2007;1–1.
59. Dutta B, Wallqvist A, Reifman J. PathNet: a tool for pathway analysis using topological information. *Source Code Biol Med* 2012; **7**(1): 10. doi: 10.1186/1751-0473-7-10.
60. Yang Q, Wang S, Dai E, *et al*. Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. *Brief Bioinform* 2019; **20**(1): 168–77.
61. Shen B, Yi X, Sun Y, *et al*. Proteomic and Metabolomic characterization of COVID-19 patient sera. *Cell* jul 2020b; **182**(1): 59–72.e15. doi: 10.1016/j.cell.2020.05.032.
62. Stukalov A, Girault V, Grass V, *et al*. Multi-level proteomics reveals host-perturbation strategies of SARS-COV-2 and SARS-COV. *bioRxiv* 2020. doi: 10.1101/2020.06.17.156455. https://www.biorxiv.org/content/early/2020/06/17/2020.06.17.156455.
63. Emanuel W, Kirstin M, Vedran F, *et al*. Bulk and single-cell gene expression profiling of SARS-COV-2 infected human cell lines identifies molecular targets for therapeutic intervention. *bioRxiv* 2020. doi: 10.1101/2020.05.05.079194. https://www.biorxiv.org/content/early/2020/05/05/2020.05.05.079194.
64. 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 Infections* 2020; **9**(1): 761–70. doi: 10.1080/22221751.2020.1747363.
65. Samavarchi-Tehrani P, Abdouni H, Knight JDR, *et al*. A SARS-COV-2 – host proximity interactome. *bioRxiv* 2020. doi: 10.1101/2020.09.03.282103. https://www.biorxiv.org/content/early/2020/09/04/2020.09.03.282103.
66. Friedman N, Jacob-Hirsch J, Drori Y, *et al*. Transcriptomic profiling of human corona virus (HCOV)-229E -infected human cells and genomic mutational analysis of HCOV-229E and SARS-COV-2. *bioRxiv* 2020. doi: 10.1101/2020.08.17.253682. https://www.biorxiv.org/content/early/2020/08/17/2020.08.17.253682.
67. Moolamalla STR, Chauhan R, Priyakumar UD, *et al*. Host metabolic reprogramming in response to SARS-COV-2 infection. *bioRxiv* 2020. doi: 10.1101/2020.08.02.232645. https://www.biorxiv.org/content/early/2020/08/05/2020.08.02.232645.
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**(6): 842–4. doi: 10.1038/s41591-020-0901-9.
69. Lee JS, Park S, Jeong HW, *et al*. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci Immunol* jul 2020; **5**(49):eabd1554. doi: 10.1126/sciimmunol.abd1554. https://pubmed.ncbi.nlm.nih.gov/32651212https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7402635/.
70. 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* jul 2020. ISSN 1546-170X; **26**(7): 1070–6. doi: 10.1038/s41591-020-0944-y. https://pubmed.ncbi.nlm.nih.gov/32514174ttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7382903/.
71. Suzuki T, Itoh Y, Sakai Y, *et al*. Generation of human bronchial organoids for SARS-COV-2 research. *bioRxiv* 2020. doi: 10.1101/2020.05.25.115600. https://www.biorxiv.org/content/early/2020/06/01/2020.05.25.115600.
72. Lieberman N A P AUID-ORCID. In vivo antiviral host transcriptional response to SARS-CoV-2 by viral load, sex, and age. doi: 10.1371/journal.pbio.3000849.
73. Kusnadi A, Ramírez-Suástegui C, Fajardo V, *et al*. Severely ill COVID-19 patients display augmented functional properties in SARS-CoV-2-reactive CD8 (+) T cells. *bioRxiv*. doi: 10.1101/2020.07.09.194027. https://pubmed.ncbi.nlm.nih.gov/32676602https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7359524/.
74. Vanderheiden A, Ralfs P, Chirkova T, *et al*. Type I and type III interferons restrict SARS-COV-2 infection of human airway epithelial cultures. *J Virol* 2020. ISSN 0022-538X; **94**(19). doi: 10.1128/JVI.00985-20. https://jvi.asm.org/content/94/19/e00985-20.
75. Hoagland DA, Clarke DJB, Møller R, *et al*. Modulating the transcriptional landscape of SARS-COV-2 as an effective method for developing antiviral compounds. *bioRxiv* 2020. doi: 10.1101/2020.07.12.199687. https://www.biorxiv.org/content/early/2020/07/13/2020.07.12.199687.

76. Lieberman NAP, Peddu V, Xie H, *et al*. In vivo antiviral host transcriptional response to SARS-COV-2 by viral load, sex, and age. *PLoS Biol* 09 2020; **18**(9): 1–17 doi: 10.1371/journal.pbio.3000849.
77. Maleki F, Ovens K, McQuillan I, *et al*. Size matters: how sample size affects the reproducibility and specificity of gene set analysis. *Hum Genomics* oct 2019. ISSN 1479-7364; **13**(Suppl 1): 42. doi: 10.1186/s40246-019-0226-2. https://pubmed.ncbi.nlm.nih.gov/31639047https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6805317/.
78. Ma J, Shojaie A, Michailidis G. A comparative study of topology-based pathway enrichment analysis methods. *BMC Bioinformatics* 2019; **20**(1): 546. doi: 10.1186/s12859-019-3146-1.
79. Ihnatova I, Popovici V, Budinska E. A critical comparison of topology-based pathway analysis methods. *PLoS One* jan 2018; **13**(1):e0191154. doi: 10.1371/journal.pone.0191154.
80. Sumedha Bagga and Michael J Bouchard. Cell cycle regulation during viral infection. In *Cell Cycle Control*, pages 165–227. Springer, 2014.