Opleiding Bioinformatica

Impact of GO evolution on covid-19 results

Benthe Brouwer

Supervisors:
K. Wolstencroft & L. Cao

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

20/06/2023

Abstract

The Gene Ontology is a structured knowledge base often used in the final stages of Omics research [Gen]. It reflects the current state of our knowledge regarding gene products and their functions. During the COVID-19 pandemic, the GO evolved along as our knowledge of COVID-19 evolved. By reproducing the existing enrichment analyses from a subset of papers relating to COVID-19 using the current version of the Gene Ontology, the effect of the evolution of the Gene Ontology can be seen. This can lead to new insights compared to the ones reported by the original papers. This not only highlights the evolution of the Gene Ontology and specific areas that evolved rapidly during the COVID-19 pandemic but also illustrates the importance of using enrichment analysis tools that use the most recent version of the Gene Ontology.

Contents

1 Introduction

1.1 The Gene Ontology

The key reason for the creation of the Gene Ontology (GO) in 1998 came from the observation that similar genes often have similar functions in different organisms. This created a demand for a shared vocabulary to encompass this knowledge and allow for clearer comparison of orthologous genes across species. Often, gene products are involved in different and many complex functions and processes. The GO aimed to allow researchers to classify these gene products and their functions to help researchers find patterns in large datasets, such as gene expression data or protein interaction data. At all times, the GO provides a snapshot of the current shared knowledge researchers have of gene products and their functions. Additionally, the increasing speed at which more molecular sequences were determined, led to the consequence that gene annotation was not keeping up with this progress. These circumstances indicated a growing need for some sort of way to bring together all of the findings and find a way to standardize nomenclature across the field of Biology. This was the goal that led to the creation of the GO [ABB+00].

The GO consists of two parts; the GO itself, which is divided into three different ontologies, and the annotations. These three different ontologies are all used to describe gene products and their functions. These ontologies are biological process, molecular function and cellular component. The biological process indicates that a particular gene or gene product has some function or plays a role towards achieving a biological objective. The molecular function of a gene or gene product indicates a biochemical activity, without specification of where or when this activity occurs. The cellular component is an indication of the location within the cell where a specific gene or gene product is active.

Annotations are added to GO to link genes and gene products to specific GO-terms. Each and every annotation also has a source and a database entry connected to it, to ensure that the annotation that is added can be backed up with evidence. Annotations are one way that causes the GO to evolve and change; with every new edition of the GO new connections are added, obsolete terms are removed and connections are edited. Another way is through new research being performed. As researchers complete experiments and contribute to the GO, the GO terms change over time and adapt according to these new discoveries.

1.2 Using GO

The structure that GO uses can be described as a hierarchical directed acyclic graph. Where each term can be described as a node within this hierarchy. Since the purpose of GO is to reflect the current state of our knowledge in biology, both the hierarchy as well as the terms itself are updated monthly according to new annotations that are added.

The GO can be used for many different purposes, one of which is as a structured knowledge base. It can help narrow down the search to other genes that are involved in a biological process due to connections within GO. GO can also be used as a tool to analyse results in high through-put experiments. It can be used to help infer functions of genes that have not been annotated yet and it can also help compare two sets of genes to see which GO-terms might be represented differently between the two sets [dPSD11]. This in particular, is what the focus of this thesis will be on; to investigate the effect that GO evolution has on interpreting the results of high throughput analyses.

1.3 Enrichment Analysis

During the final stages of many omics analyses, the researchers often end up with a list of differentially expressed genes (DEGs); genes that are expressed different in one gene set as compared to another gene set. Researchers then turn to enrichment analysis to infer further connections between these genes and whether the genes are associated in the same biological processes, molecular function and/or cellular components. Enrichment analysis aims to systemically map genes to their biological annotation, as can be found in GO. Enrichment analysis can identify whether a GO-term is over- or underrepresented in the list of DEGs, this can give an indication of which processes might be affected by the conditions that are being tested. To perform an enrichment analysis you need the list of DEGs, you also need to provide most tools with an input species as well as an analysis species. The input species is the species where the DEGs were taken from and the analysis species is the species which will be used to find enriched terms from [TH10].

1.4 COVID-19

For the purpose of this thesis, the choice was made to examine a period of time where GO was changing rapidly, due to a large amount of annotations being contributed. A logical choice that came to mind was the very recent COVID-19 pandemic, a large amount of research was done on a specific topic that led to many significant changes within GO, as can be viewed in this thesis.

A total of 90 different COVID-19 related papers were collected that also involved some sort of enrichment analysis. Out of these 90 different papers, a sample of 10 papers were selected, the reasoning for why these 10 papers were chosen is elaborated on in the methods section of this thesis.

The goal of this thesis was to examine the evolution of the GO using the COVID-19 pandemic. The motivation for this specific period of time was that a large amount of research was done on a singular topic which led to quick evolution of very specific areas of GO. The aim is to try and use enrichment analysis to pinpoint these areas of GO that have undergone changes during the COVID-19 pandemic and illustrate how this may affect the reusability of data and whether we can identify new insights using a newer version of GO than was available at the time of the research.

This leads to the research question: what was the impact of GO evolution on COVID-19 results?

2 Related Works

2.1 The creation of GO

As mentioned above, the GO was created out of an increasingly important demand to describe and classify gene products and functions, there was a need for a shared vocabulary. But for the purpose of this thesis, and to further examine the way the GO evolves, it is important to understand how GO was designed and created [Gen].

The GO consortium was formed to create some kind of shared vocabulary that was sufficient enough for the annotation of gene products and functions across multiple organisms, the shared vocabulary was meant to help support cross-database queries but it soon became clear that the set of annotations from a model organism in its entirety was an invaluable source for the scientific community. The GO consortium was established in 1998 as a collaboration between 3 different model organism databases; Flybase, the Saccharomyces Genome Database and the Mouse Genome Database.

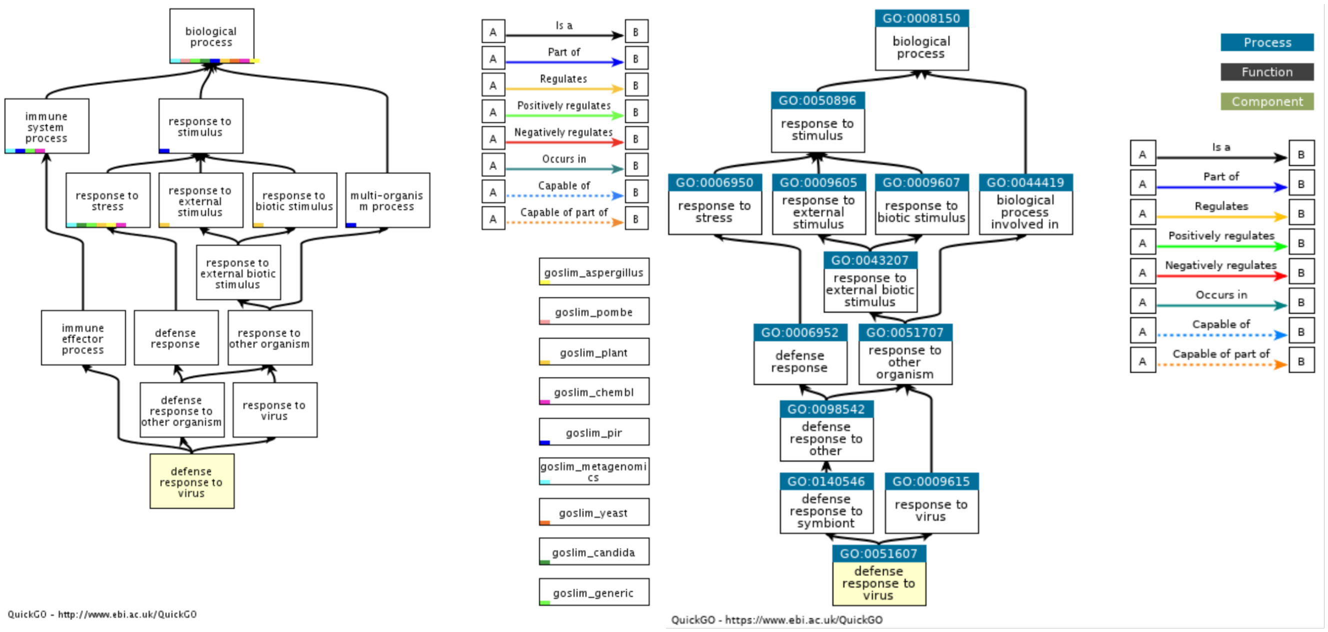
To accomplish this, three different ontologies were formed. An ontology for biological processes, another for molecular functions and a last one for cellular components. This was necessary because they represent sets of information that can be found in all organisms. The ontologies are structured as directed acyclic graphs wherein a ”child” term can be connected to at least one ”parent” term. Each term in the ontology is its own accessible object within the GO structured knowledge base. The term has a unique identifier, also known as a GO accession number and the term is also defined, with a clear source for this definition.

In its essence, there are a few standards that were defined when GO was created that the GO and any of its annotations or terms had to adhere to. All paths within the GO hierarchy had to be true, meaning a clear source for an addition or deletion of a path had to be included when contributing to GO. Terms should not be specific to a single species but should be able to be used as broadly as possible for as many species as feasible. Any and all new or edited attributes of GO had to be accompanied with citations that were deemed appropriate. Finally, all annotations contributed to gene products and gene functions had to incorporate statements that supported the relationship to which the annotation was added.

2.2 The evolution of GO

It is also vital to understand how GO could change over a period of time and how such changes and evolution can affect research done in the future or even, past research. This should also highlight the importance of the research question this thesis aims to answer.

An illustration of the evolution of GO can easily be visualised as below. In fig 1a and fig 1b we see the hierarchical directed acyclic graph of the term ”Defense response to virus”. Fig 1a is based on the May 2016 release of GO, whereas fig 1b is based on the April 2023 release of GO. We can clearly see that in these 7 years terms have been added that led to these changes in the hierarchy chart.


biological process
immune system process
response to stimulus
response to stress
response to external stimulus
response to biotic stimulus
multi-organism process
response to external biotic stimulus
immune effector process
defense response
response to other organism
defense response to other organism
response to virus
defense response to virus
QuickGO - http://www.ebi.ac.uk/QuickGO
Is a
Part of
Regulates
Positively regulates
Negatively regulates
Occurs in
Capable of
Capable of part of
A
B
goslim_aspergillus
goslim_pombe
goslim_plant
goslim_chembl
goslim_pir
goslim_metagenomics
goslim_yeast
goslim_candida
goslim_generic
GO:0008150 biological process
GO:0050896 response to stimulus
GO:0006950 response to stress
GO:0009605 response to external stimulus
GO:0009607 response to biotic stimulus
GO:0044419 biological process involved in
GO:0043207 response to external biotic stimulus
GO:0006952 defense response
GO:0051707 response to other organism
GO:0098542 defense response to other
GO:0140546 defense response to symbiont
GO:0009615 response to virus
GO:0051607 defense response to virus
Process
Function
Component
QuickGO - https://www.ebi.ac.uk/QuickGO

(a) Defense response to virus in May, 2016

(b) Defense response to virus in April, 2023

Figure 1: Comparison of the hierarchy charts in 2016 and 2023

Another thing to note about the evolution of GO, is that it may suffer from annotation bias. Although research is done on many different gene functions and gene products, not all gene products and functions are annotated at the same pace and frequency as others. This can affect the results of analyses that are performed using GO such as enrichment analysis and protein function prediction, we can often find low consistency between enrichment analysis performed in earlier and later versions of GO [TMW+18].

2.3 Enrichment analysis methods

As stated above, many researchers use enrichment analysis in the final stages of Omics research to help infer further connections between found DEGs. It is important to understand that there are many different methods and tools that are used for enrichment analysis, though these tools usually use one of three different classes of algorithms, all of which will be elaborated on below [TH10]. An important note is that there is currently no perfect enrichment analysis method, this is why new tools are still developed every year. Knowing the difference between these algorithm classes and the way the different tools use these can help researchers understand which tool is most suitable for their omics research.

The first class of algorithms is the singular enrichment analysis (SEA). SEA is the most traditional method for enrichment analysis. SEA iteratively compares each GO-term to the DEGs. To calculate the enriched value of a term, the observed frequency of a term when compared to the DEGs is compared to the frequency that would be expected by pure chance. An important downside of these algorithms is the fact that it does not take the relationships between terms or the hierarchical nature of the GO into consideration while calculating whether a term is enriched or not. The result of SEA is usually a large list of terms that includes redundant terms due to SEA not taking into account that some terms are very closely related. A tool that uses this method that is also performed by a paper looked at within this thesis is GOstats, an R-based package.

The second class of algorithms is the gene set enrichment analysis (GSEA). As the name implies, this class of algorithms looks at all genes during an analysis, not just DEGs. GSEA is used in cases where clear phenotypes are selected or samples were taken at different points of time since this method requires a quantative value that is used to rank each gene. After this is accomplished, a maximized enrichment score is calculated from the ranked list compared to possible enriched terms. Like SEA, GSEA does not take the relationships between terms or the hierarchical nature into consideration. A tool that uses this method that is also performed by a paper looked at within this thesis is the GSEA tool itself and tools such as Metascape and Enrichr.

The third class of algorithms is the modular enrichment analysis (MEA). This class is the only class to take the relationships between terms and the hierarchical nature of the GO into account when calculating whether a term is considered enriched or not. This can reduce the redundancy in the results generated by an enrichment analysis. A commonly used MEA tool is DAVID.

2.4 Semantic similarity

The semantic similarity between two terms measures the relation between terms using the background knowledge that can be found within an ontology [KSGH20]. This can take into account attributes such as the processes in which a term is involved or the distance between two terms within the hierarchy. In the context of this thesis, the semantic similarity can be used to determine whether a GO term is enriched or not.

A few different tools have been created to specifically calculate the semantic similarity between GO terms. One of these tools is called GOSemSim, which is an R-based package. This tool calculates the semantic similarity between terms by implementing five different methods, four methods that look at the information content of a term and one that looks at the graph structure within the ontologies [Yu20]. Another example of a tool that calculates the semantic similarity is ViSEAGO which is another R-based package. ViSEAGO stands for visualisation, semantic similarity and enrichment analysis of GO. As the name indicates this tool is multifunctional and performs the three main components for the analyses used during the final stage of omics research. This tool uses the same five methods for calculating the semantic similarity between terms [BJHA19].

For the purpose of this thesis, the semantic similarity of terms were calculated using code provided by Yi Chen [CVW]. This code uses an information content method, specifically it looks at the information content of the most informative common ancestor. The more generalised, and higher up in the ancestry, a term is, the lower the information content.

2.5 Data reusability, reproducability and FAIR

In this thesis, the aim was to reproduce the enrichment analyses performed within each paper from the sample as accurately and as close to the original as possible. To accomplish this, an important aspect that had to be examined was the data reusability of each paper. Could we reuse the data provided within each paper to reproduce the analysis and would this result in similar results? To gain more insight on judging whether a paper's analysis was reproducible, the FAIR principles was examined.

In 2016, a paper was published detailing the guidelines of the FAIR principles for scientific data with the intent to improve the reusability of research [Wil]. FAIR stands for Findability, Accessibility, Interoperability and Reusability. Findability means that any data included within the paper is easy to find, by both computers and humans. Accessibility indicates that anyone that should be able to access the data, is able to and has the proper authorization. Interoperability says that the data needs to be well integrated with other data and that it should be able to interoperate with applications and workflows. Reusability indicates that the data should be reusable [GOF22].

3 Methods

3.1 Performing enrichment analysis

To perform the enrichment analyses that were performed in each of the papers using Metascape, the same steps were taken for each paper. The first step was determining a list of differentially expressed genes (DEGs) from the input data that was used in each paper. Sometimes, the paper included a supplement file that had all of the found DEGs. Other times, however, the DEGs had to be found from the raw gene reads using EdgeR. EdgeR is an R package that can be used to process raw reads data and additionally perform different analyses such as to find the DEGs. When neither was possible in some cases the DEG list was taken from a graphic, this could mean that the complete DEG list was not used for the re-analysis.

The next step was to perform the enrichment analysis using the Metascape web-tool. As an input in Metascape, a txt file with the found DEGs was found. Next, the user has to select the input species as well as the analysis species. For the purpose of this thesis, the setting H. sapiens was always used.

3.2 Calculating the semantic similarity

As noted above, Yi Chen's code for calculating the semantic similarity between terms was used. To use his code, a small script had to be written that gave Yi's code two lists featuring the terms that had to be compared. In the case of this thesis this was a list of the found GO terms in the original analysis compared to the list found during the reproduction of the analysis.

3.3 Reporting on the availability of papers data

Since the aim is the reproduce the enrichment analyses performed in a subset of papers about COVID-19 to show the effects of the evolution of the GO, it was important to examine the availability of the data and to look at whether the datasets that were used originally had been updated. To this extent, a review was done on each of the chosen papers to see which papers had usable DEG lists and/or had datasets that had been updated after the publishing date of the papers themselves, which can be viewed in the table below. The papers were chosen due to the relative simplicity of their enrichment analyses.

Table 1: Comparison of data availability of papers

Paper	Dataset updated	DEG list available
PMC 8342995	No	Yes
PMC 7321036	No	No
PMC 8245040	No	Yes
PMC 7252184	Yes	No
PMC 8476510	No	No
PMC 7805430	Yes	Yes
PMC 8069812	Yes	No
PMC 8356054	Yes	No
PMC 8438179	Yes	Yes

4 Results

To examine the evolution of GO during the pandemic, the choice was made to perform the enrichment analyses of 10 different COVID-19 related papers using the Metascape web-tool according to the protocol as described above. The subset of papers chosen was part of a larger analysis that examined all enrichment analyses found in COVID-19 related papers. The choice to use Metascape was made based on the fact that Metascape is kept up to date with the latest version of GO, it is updated monthly. All the papers selected for this purpose were published in 2020-2022 (See figure: 2, papers using the same GO-version have been coloured the same) and all performed at least one enrichment analysis that specifically resulted in a list of GO-terms, for the purpose of this paper the results looking at KEGG pathways or other sources have not been included, though they were often also performed.

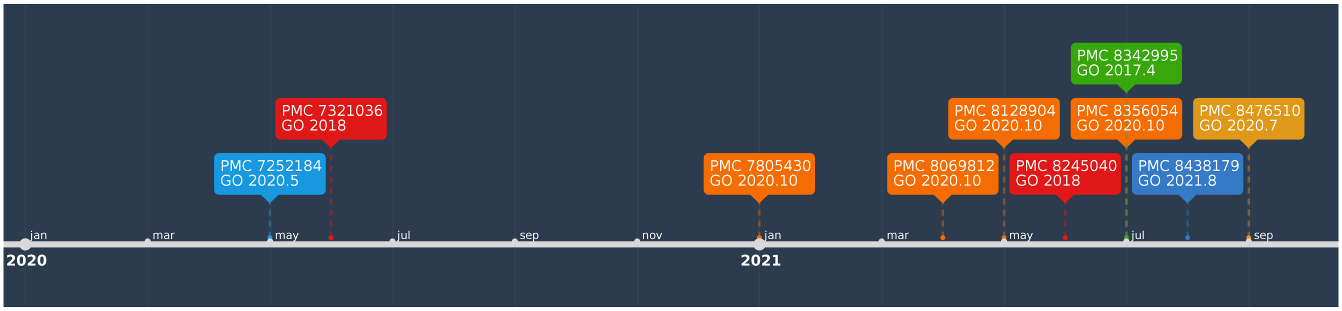


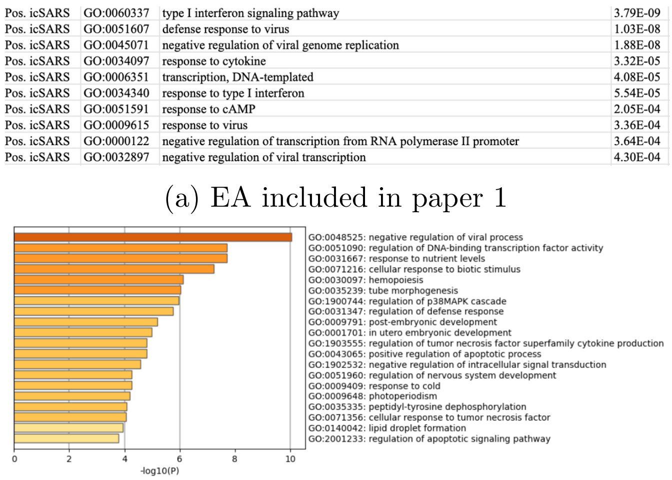
Figure 2: Timeline of all papers with the GO-version date

In this results section we will look at each of these papers in more detail, ordered from the most outdated GO-version to the most recent. The GO-version was determined by looking at which tool was used to perform the enrichment analyses and then identifying which version of GO each specific tool used.

4.1 PMC 8342995

The first paper was titled *Gene Expression Meta-Analysis Reveals Interferon-Induced Genes Associated With SARS Infection in Lungs* and was published July 23rd, 2021 [PH21]. The goal of this paper was to identify therapeutic targets that could possibly be associated with COVID-19 infections, they accomplished this by analyzing mRNA expression data from COVID-19 infections by meta-analysing gene signatures. This paper used GSEA 3.0 to perform their enrichment analysis, this means they used GO-version 2017.04.

The paper looked at both a positive and negative icSARS panel, for the purpose of this thesis only the positive icSARS panel enrichment analysis (fig: 3a) was re-analysed (fig: 3b).
As we can tell, in the original the most enriched terms were related to the type I interferon as well as the defense response to viruses. The re-analysed enrichment analysis also included terms relating to the defense response to viruses but also included terms that were not included in the original analysis. Terms such as *response to nutrient levels* and *cellular response to biotic stimulus.*



Pos. icSARS	GO:0060337	type I interferon signaling pathway	3.79E-09
Pos. icSARS	GO:0051607	defense response to virus	1.03E-08
Pos. icSARS	GO:0045071	negative regulation of viral genome replication	1.88E-08
Pos. icSARS	GO:0034097	response to cytokine	3.32E-05
Pos. icSARS	GO:0006351	transcription, DNA-templated	4.08E-05
Pos. icSARS	GO:0034340	response to type I interferon	5.54E-05
Pos. icSARS	GO:0051591	response to cAMP	2.05E-04
Pos. icSARS	GO:0009615	response to virus	3.36E-04
Pos. icSARS	GO:0000122	negative regulation of transcription from RNA polymerase II promoter	3.64E-04
Pos. icSARS	GO:0032897	negative regulation of viral transcription	4.30E-04

(a) EA included in paper 1

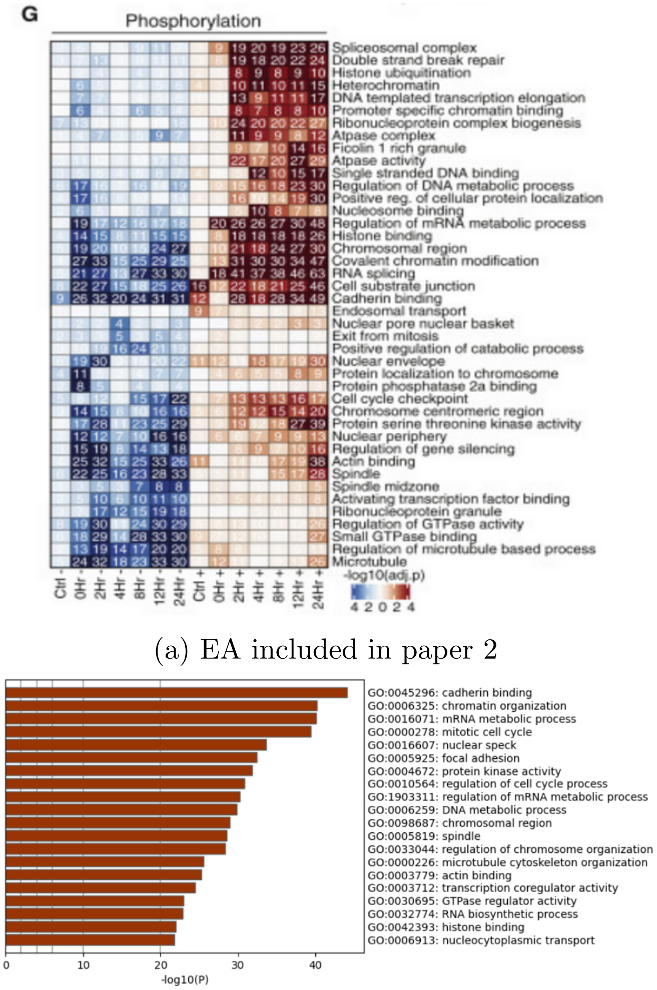
GO:0048525: negative regulation of viral process
GO:0051090: regulation of DNA-binding transcription factor activity
GO:0031667: response to nutrient levels
GO:0071216: cellular response to biotic stimulus
GO:0030097: hemopoiesis
GO:0035239: tube morphogenesis
GO:1900744: regulation of p38MAPK cascade
GO:0031347: regulation of defense response
GO:0009791: post-embryonic development
GO:0001701: in utero embryonic development
GO:1903555: regulation of tumor necrosis factor superfamily cytokine production
GO:0043065: positive regulation of apoptotic process
GO:1902532: negative regulation of intracellular signal transduction
GO:0051960: regulation of nervous system development
GO:0009409: response to cold
GO:0009648: photoperiodism
GO:0035335: peptidyl-tyrosine dephosphorylation
GO:0071356: cellular response to tumor necrosis factor
GO:0140042: lipid droplet formation
GO:2001233: regulation of apoptotic signaling pathway
0
2
4
6
8
10
-log10(P)

(b) EA re-performed in Metascape

Figure 3: Side-by-side of both EAs performed on the DEG-list from paper 1

4.2 PMC 7321036

The second paper was titled *The Global Phosphorylation Landscape of SARS-CoV-2 Infection* and was published June 28th, 2020 [bou]. The goal of this paper was to perform a quantitative mass spectrometry-based phosphoproteomics survey of COVID-10. This paper used Enrichr to perform their enrichment analysis, this means they used GO-version 2018.



(a) EA included in paper 2

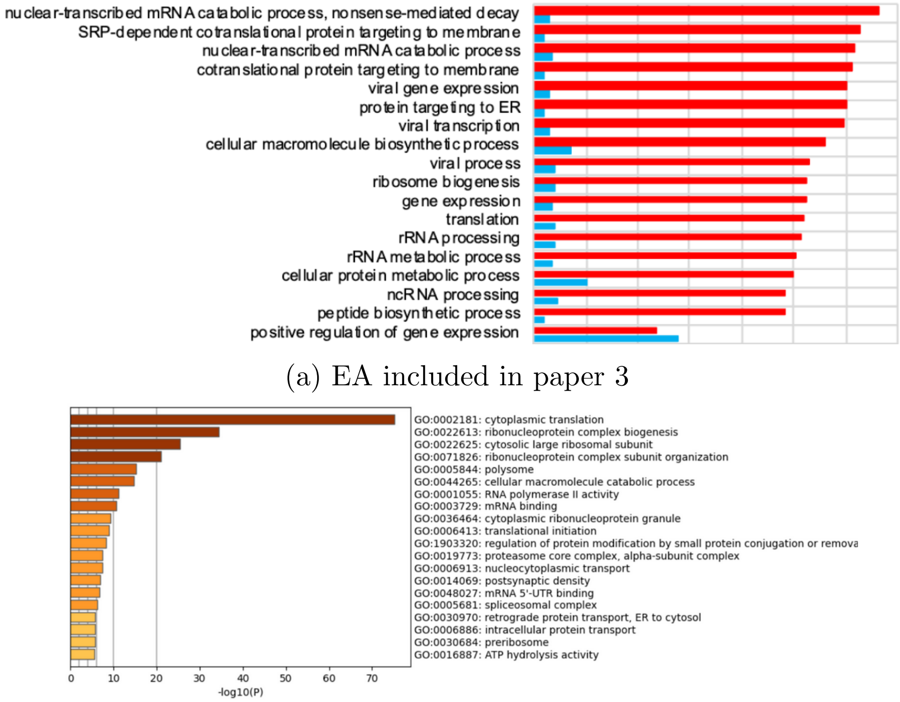
(b) EA re-performed in Metascape

Figure 4: Side-by-side of both EAs performed on the DEG-list from paper 2

As we can tell, in the original the most enriched terms were related to protein binding and RNA processing. The re-analysed enrichment analysis also included many similar terms as well as a few terms regarding organelle organization such as *chromatin organization* and *microtubule cytoskeleton organization*.

4.3 PMC 8245040

The third paper was titled *Identification of SARS-CoV-2–induced pathways reveals drug repurposing strategies* and was published June 30th, 2021 [HHT+21]. The goal of this paper was to construct a COVID-19 induced protein network to identify drugs that are predicted to target COVID-19 induced pathways. This paper used Enrichr to perform their enrichment analysis, this means they used GO-version 2018.



(a) EA included in paper 3

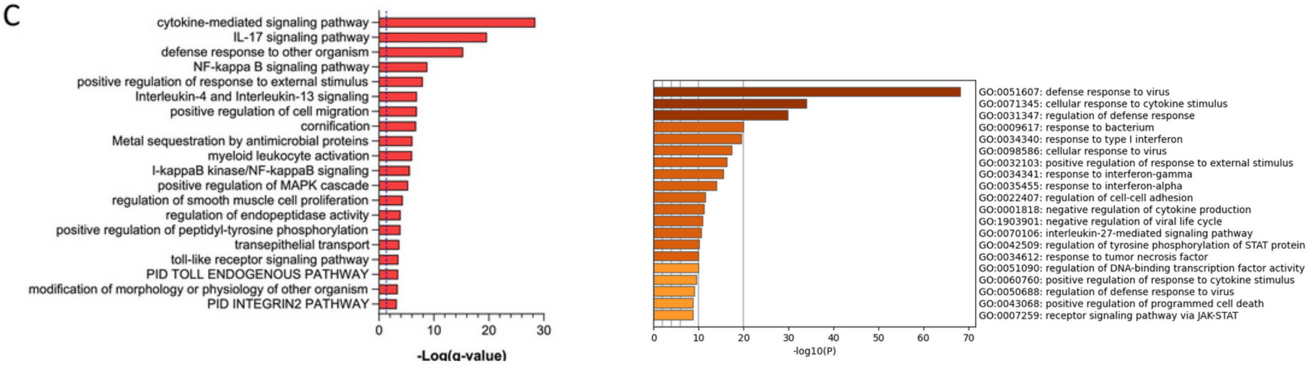
(b) EA re-performed in Metascape

Figure 5: Side-by-side of both EAs performed on the DEG-list from paper 3

As we can tell, in the original the most enriched terms were related to RNA metabolic processes as well as viral processes. The re-analysed enrichment analysis also included terms related to RNA metabolic processes but also included terms such as *cytoplasmic translation* and *cytosolic large ribosomal subunit*.

4.4 PMC 7252184

The fourth paper was titled *Transcriptional landscape of SARS-CoV-2 infection dismantles pathogenic pathways activated by the virus, proposes unique sex-specific differences and predicts tailored therapeutic strategies* and was published May 3rd, 2020 [FCL+20]. The goal of this paper was to identify a specific set of biological pathways that were altered in primary human lung epithelium when infected with COVID-19. This paper used Metascape to perform their enrichment analysis, Metascape is kept up to date monthly with the most recent GO-version.



(a) EA included in paper 4

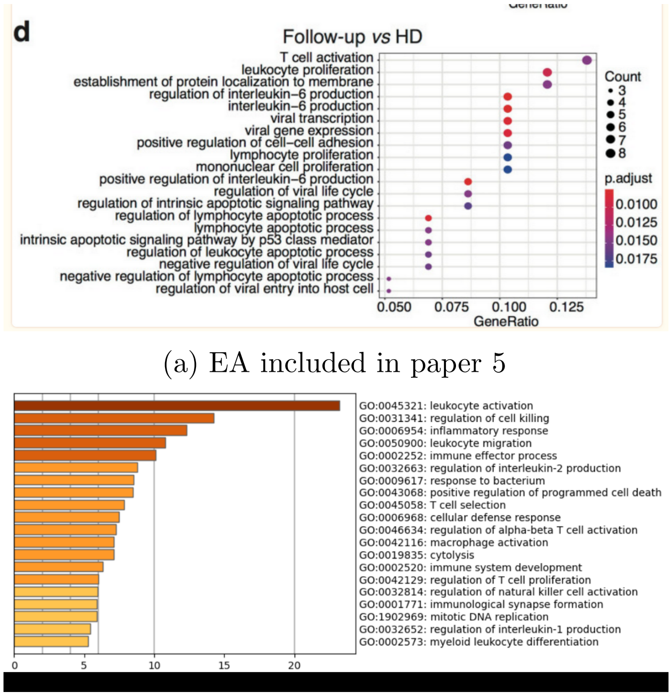
(b) EA re-performed in Metascape

Figure 6: Side-by-side of both EAs performed on the DEG-list from paper 4

As we can tell, in the original the most enriched terms were related to signaling pathways. The re-analysed enrichment analysis included more terms related to the response to viruses and other external simuli such as *defense response to virus* and *response to bacterium*.

4.5 PMC 8476510

The fifth paper was titled *Dynamics of TCR repertoire and T cell function in COVID-19 convalescent individuals* and was published September 28th, 2021 [LLP+21]. The goal of this paper was to analyse the dynamics of TCR repertoire and immune metabolic functions found in blood T cells that were taken from patients that had recently recovered from COVID-19. This paper used GSEA v4.1.0, which means this paper made use of GO-version 2020.07.



(a) EA included in paper 5

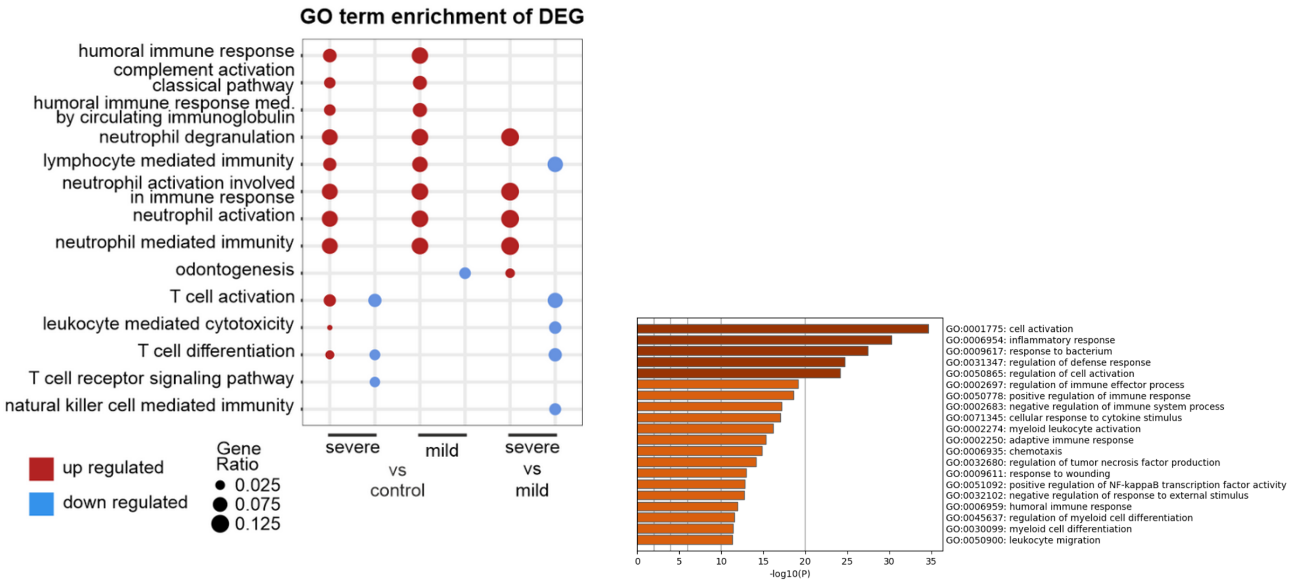
(b) EA re-performed in Metascape

Figure 7: Side-by-side of both EAs performed on the DEG-list from paper 5

As we can tell, in the original the most enriched terms were related to viral processes and interleukin-6. The re-analysed enrichment analysis included more terms related to immune system processes such as *leukocyte activation* and *leukocyte migration*.

4.6 PMC 7805430

The sixth paper was titled *Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients* and was published January 13th, 2021 [AMK+21]. The goal of this paper was to provide novel insights in the different molecular phenotypes found in COVID-19 patients. This paper used ClusterProfiler which is an R-based tool, this means this paper made use of GO-version 2020.10. This paper performed several EAs, the choice was made to re-perform the analysis that compared a patient with a severe case of COVID-19 to a patient with a mild case of COVID-19.



(a) EA included in paper 6

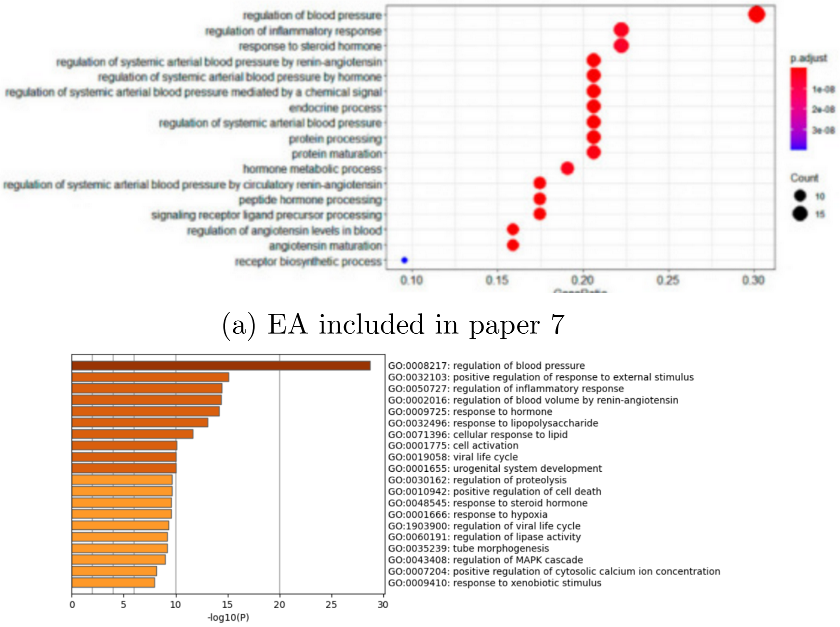
(b) EA re-performed in Metascape

Figure 8: Side-by-side of both EAs performed on the DEG-list from paper 6

As we can tell, in the original the most enriched terms were related to immune responses, lymphocytes and T cells. The re-analysed enrichment also had terms relating to immune responses but also included unique terms such as *chemotaxis*.

4.7 PMC 8069812

The seventh paper was titled *Drug Repurposing for COVID-19 Treatment by Integrating Network Pharmacology and Transcriptomics* and was published April 14th, 2021[LLL+21]. The goal of this paper was to screen drug candidates for new COVID-19 treatments by using integrated network-based pharmalogic and transcriptomic approaches. This paper used ClusterProfiler which is an R-based tool, this means this paper made use of GO-version 2020.10.



(a) EA included in paper 7

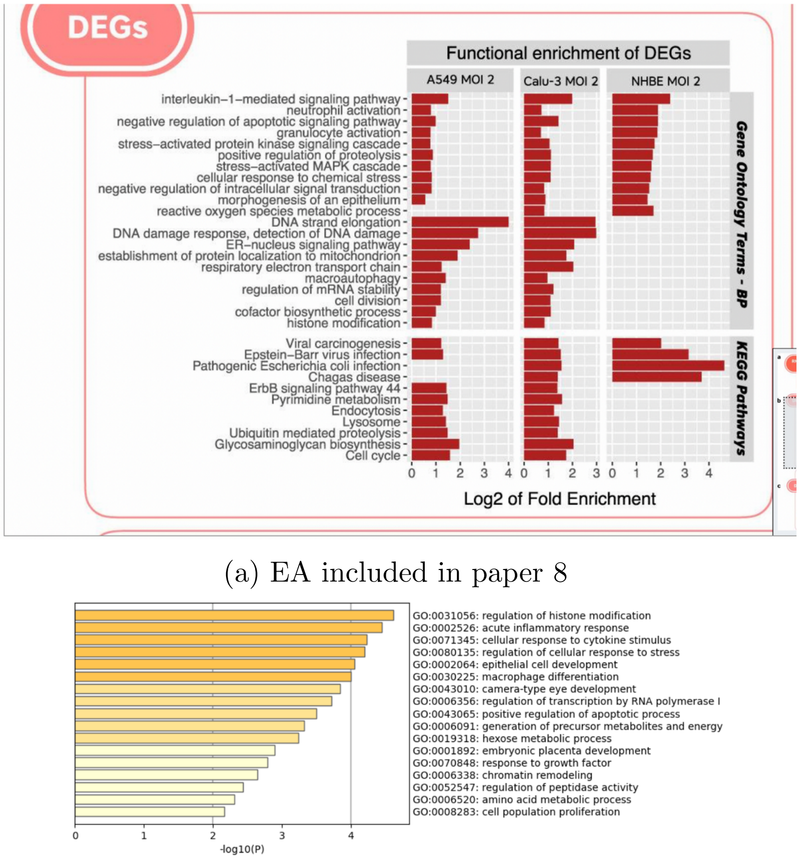
(b) EA re-performed in Metascape

Figure 9: Side-by-side of both EAs performed on the DEG-list from paper 7

As we can tell, in the original the most enriched terms were related to blood pressure and hormone regulation. The re-analysed enrichment had a lot of similar terms but also included unique terms such as *cell activation* and *viral life cycle*.

4.8 PMC 8128904

The eighth paper was titled *Genome-wide bioinformatic analyses predict key host and viral factors in SARS-CoV-2 pathogenesis* and was published May 17th, 2021[FLR+21]. The goal of this paper was to identify genes, isoforms and transposable element families that are specifically altered in COVID-19 infected respiratory cells. This paper used the GOstats package which is an R-based package, this means this paper made use of GO-version 2020.10.



(a) EA included in paper 8

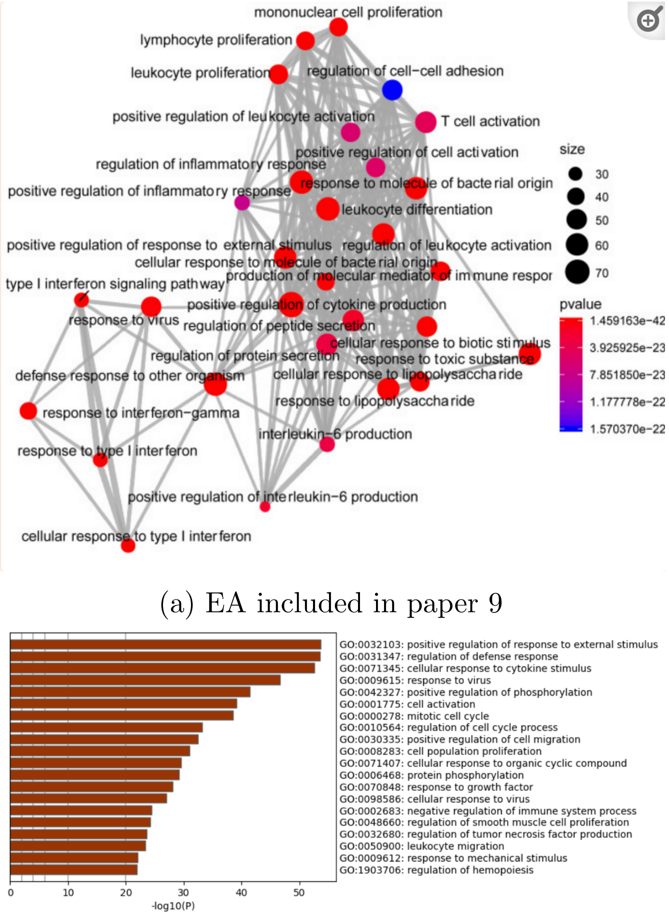
(b) EA re-performed in Metascape

Figure 10: Side-by-side of both EAs performed on the DEG-list from paper 8

As we can tell, in the original the most enriched terms were related to DNA metabolic processes. The re-analysed enrichment included many different terms that were not similar to those included in the original, terms such as *epithelial cell development*, *camera-type eye development* and *embryonic placenta development*.

4.9 PMC 8356054

The ninth paper was titled *Identification of COVID-19 and Dengue Host Factor Interaction Networks Based on Integrative Bioinformatics Analyses* and was published July 28th, 2021[ZWL+21]. The goal of this paper was to identify the host factor interaction network for COVID-19 and dengue fever to identify possible drugs that could be used for clinical practice. This paper used the ShinyGO package which is an R-based package, this means this paper made use of GO-version 2020.10.



(a) EA included in paper 9

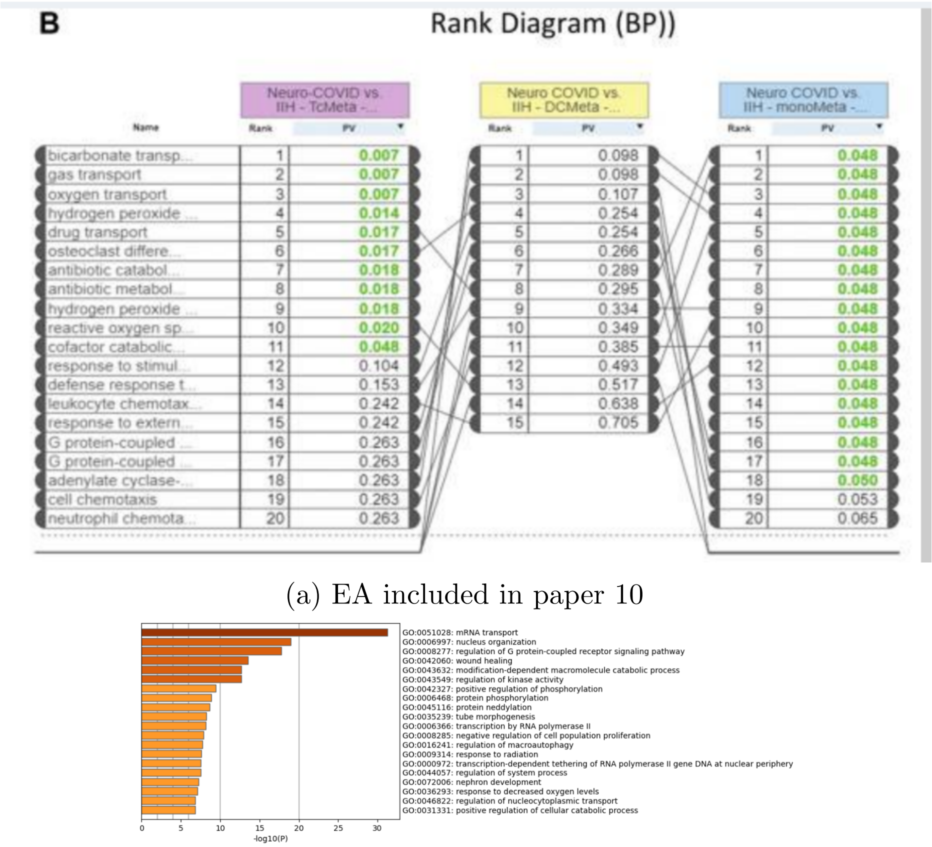
(b) EA re-performed in Metascape

Figure 11: Side-by-side of both EAs performed on the DEG-list from paper 9

As we can tell, in the original the most enriched terms were related to immune system processes and responses to external stimuli. The re-analysed enrichment analysis included many similar terms but also included unique terms such as *regulation of tumor necrosis factor production*.

4.10 PMC 8438179

The tenth paper was titled *Identification of Novel Gene Signatures using Next-Generation Sequencing Data from COVID-19 Infection Models: Focus on Neuro-COVID and Potential Therapeutics* and was published August 31st, 2021[PAN21] . The goal of this paper was to use the gene expression profiles of COVID-19 infection models to identify possible therapeutics for COVID-19. This paper used Metascape to perform their enrichment analysis. Metascape is updated monthly to reflect the most current version of the GO. The enrichment analysis performed was based on the DEG's found from the first column in 12a.



	Neuro-COVID vs. IIH - TcMeta -...		Neuro COVID vs. IIH - DCMeta -...		Neuro COVID vs. IIH - monoMeta -...	
Name	Rank	PV	Rank	PV	Rank	PV
bicarbonate transp...	1	0.007	1	0.098	1	0.048
gas transport	2	0.007	2	0.098	2	0.048
oxygen transport	3	0.007	3	0.107	3	0.048
hydrogen peroxide ...	4	0.014	4	0.254	4	0.048
drug transport	5	0.017	5	0.254	5	0.048
osteoclast differe...	6	0.017	6	0.266	6	0.048
antibiotic catabol...	7	0.018	7	0.289	7	0.048
antibiotic metabol...	8	0.018	8	0.295	8	0.048
hydrogen peroxide ...	9	0.018	9	0.334	9	0.048
reactive oxygen sp...	10	0.020	10	0.349	10	0.048
cofactor catabolic...	11	0.048	11	0.385	11	0.048
response to stimul...	12	0.104	12	0.493	12	0.048
defense response t...	13	0.153	13	0.517	13	0.048
leukocyte chemotax...	14	0.242	14	0.638	14	0.048
response to extern...	15	0.242	15	0.705	15	0.048
G protein-coupled ...	16	0.263			16	0.048
G protein-coupled ...	17	0.263			17	0.048
adenylate cyclase-...	18	0.263			18	0.050
cell chemotaxis	19	0.263			19	0.053
neutrophil chemota...	20	0.263			20	0.065

(a) EA included in paper 10

GO:0051028: mRNA transport
GO:0006997: nucleus organization
GO:0008277: regulation of G protein-coupled receptor signaling pathway
GO:0042060: wound healing
GO:0043632: modification-dependent macromolecule catabolic process
GO:0043549: regulation of kinase activity
GO:0042327: positive regulation of phosphorylation
GO:0006468: protein phosphorylation
GO:0045116: protein neddylation
GO:0035239: tube morphogenesis
GO:0006366: transcription by RNA polymerase II
GO:0008285: negative regulation of cell population proliferation
GO:0016241: regulation of macroautophagy
GO:0009314: response to radiation
GO:0000972: transcription-dependent tethering of RNA polymerase II gene DNA at nuclear periphery
GO:0044057: regulation of system process
GO:0072006: nephron development
GO:0036293: response to decreased oxygen levels
GO:0046822: regulation of nucleocytoplasmic transport
GO:0031331: positive regulation of cellular catabolic process
0
5
10
15
20
25
30
-log10(P)

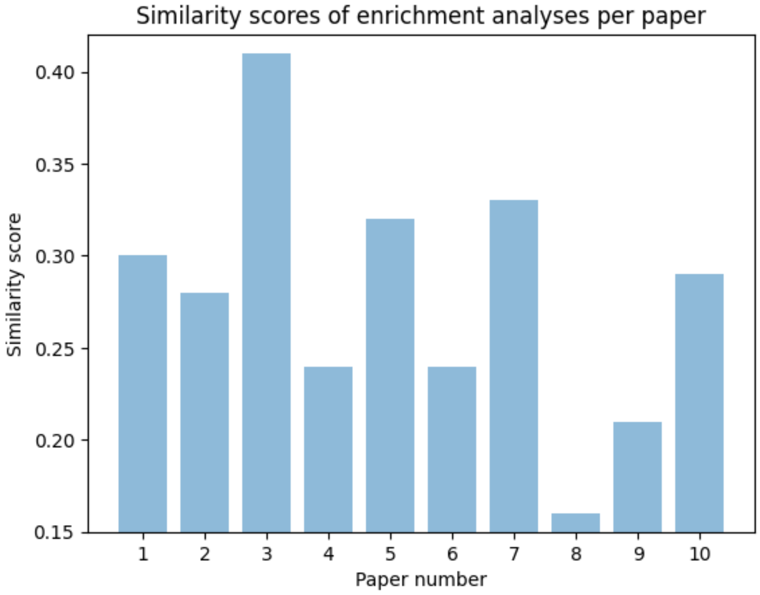
(b) EA re-performed in Metascape

Figure 12: Side-by-side of both EAs performed on the DEG-list from paper 10

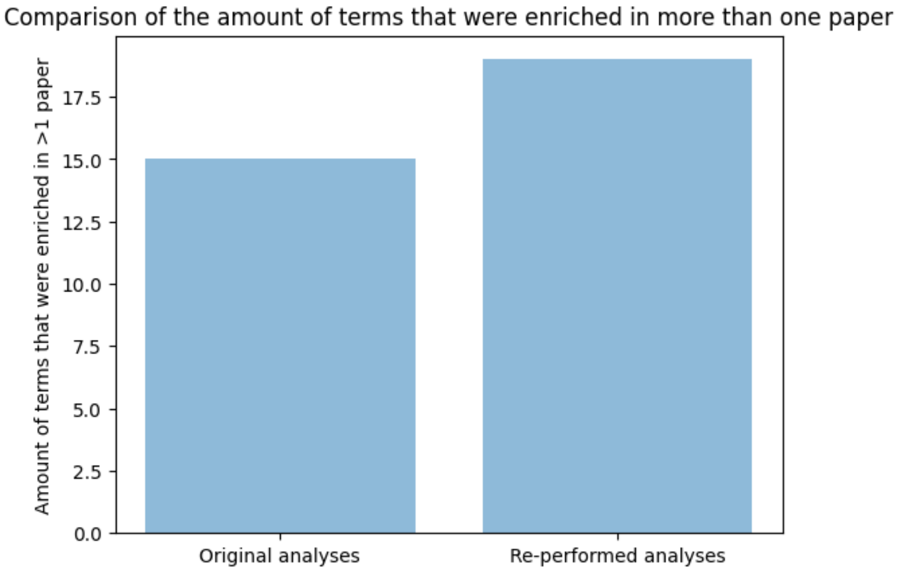
As we can tell, in the original the most enriched terms were related to gas transport. The re-analysed enrichment analysis involved very different terms including terms such as *mRNA transport* and *nucleus organization*.

4.11 Comparing the papers

As the results above indicate, the difference between the enrichment analysis performed within the papers and the enrichment analysis that was done recently show a difference. To determine the effecs of GO evolution, an additional analysis was performed that aimed to calculate the temporal semantic similarity between the results of the different enrichment analyses. The result of this can be seen in the figure below included below. As we can see in this figure, paper 3 was considered the most similar to the original enrichment analysis while paper 8 was considered the least similar to the original enrichment analysis. Paper 3 used a very outdated version of GO, since Enrichr had not been updated since 2018, while paper 8 used a version of GO that was from October, 2020. This could indicate that the enriched terms found in paper 3 had not evolved as much as the enriched terms found in paper 8, the outdated version of GO used in paper 8 only further highlighted the temporal differences that were found in the re-performed analyses.



Another difference between the original analyses and the re-performed analyses were the amount of terms that were considered enriched in more than one paper. Generally, these papers summarised the top terms so it could be possible that the whole dataset of enriched terms would provide more overlap between the enriched terms.



Interestingly, there are no overlapping terms between the terms found in the original enrichment analyses and terms found in the recently performed enrichment analyses. When looking at the terms found in the original enrichment analyses, some evolution can be found. These changes to the parent terms and child terms of these terms have been detailed in the tables below. The only changes to the parent and child terms that have been added to the table are ones that took place in January 2020 or more recent.

An example of a hierarchy chart that shows the changes in hierarchy of a term between the time of publication and early 2023 has been included below. The red arrow indicates that a link has been added between GO:0051607 and GO:0044790.

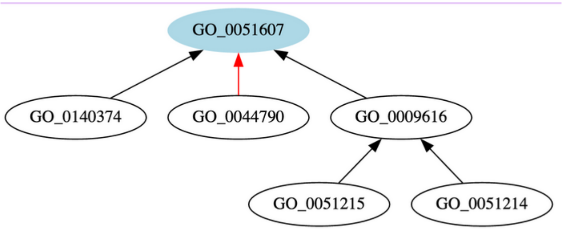


Table 2: Terms that were enriched in > 1 paper: Original analyses

GO-term	Changes to parent terms	Changes to child terms
GO:0051607	Link to GO:0140546 added, links to GO:0098542 & GO:0002252 removed	Link to GO:0044790 added
GO:0060337	Link to GO:0140888 added, link to GO:0019221 removed	No changes
GO:0032755	No changes	Link to GO:2000778 & GO:0045410 removed
GO:0050727	No changes	Link to GO:1900015 & GO:0035490 removed
GO:0072593	No changes	Link to GO:0046209 removed
GO:0032943	No changes	Link to GO:0044565 added
GO:0046651	No changes	Link to GO:0044565 removed
GO:0042119	No changes	No changes
GO:0019080	No changes	No changes
GO:0019083	No changes	No changes
GO:0034340	No changes	No changes
GO:0009615	No changes	No changes
GO:0070661	No changes	No changes
GO:0042110	No changes	No changes
GO:0032635	No changes	No changes

Table 3: Terms that were enriched in > 1 paper: Re-performed analyses

GO-term	Changes to parent terms	Changes to child terms
GO:0010564	No changes	Links to GO:1903379, GO:0007063, GO:0046602, GO:1905213, GO:1903117, GO:0010946, GO:0010520, GO:0032954, GO:0071342, GO:0060629, GO:1903722, GO:0030997, GO:1903221, GO:0032213, GO:0060622, GO:1903023, GO:0034223, GO:0120313, GO:0120264 & GO:0140433 added, links to GO:0071156, GO:1903504 GO:1904289 and GO:0090235 removed
GO:0002683	No changes	Links to GO:0039532, GO:0034152, GO:0034132 & GO:0061060 added, links to GO:0070425, GO:0002906 & GO:0033030 removed
GO:0006468	Link to GO:0036211 added, link to GO:0006464 removed	Links to GO:0016572, GO:0023014, GO:0100002 & GO:0006975 removed
GO:0031347	No changes	Link to GO:0002759 added, links to GO:1904415 & GO:0110132 removed
GO:0070848	Link to GO:0009719 added, link to GO:0010033 removed	No changes
GO:0050900	No changes	Links to GO:0072676 & GO:0036336 removed
GO:0009617	No changes	Links to GO:0140460 & GO:0140459 added
GO:0032680	No changes	Links to GO:0042534 & GO:1904467 removed
GO:0008283	Link to GO:0009987 added, link to GO:0008150 removed	No changes
GO:0032103	No changes	Link to GO:1901672 added
GO:0001775	Link to GO:0032501 added	No changes
GO:0006913	No changes	Link to GO:0043280 removed
GO:0042327	No changes	Link to GO:1903862 removed
GO:0035239	No changes	No changes
GO:0006954	No changes	No changes
GO:0071345	No changes	No changes
GO:0030099	No changes	No changes
GO:0043065	No changes	No changes
GO:0000278	No changes	No changes

From the enriched terms found in the original analyses, one of the terms that evolved the most after January, 2020 was GO:0051607. This term is the term *defense response to virus*, the changes to this term could have been caused by research done to COVID-19 since it seems closely related to COVID-19. Another notable change to the ancestry of the terms commonly found in the original enrichment analyses is the addition and removal of GO:0044565. This term (*dendritic cell proliferation*) was added as a child term to GO:0032943 (*mononuclear cell proliferation*) and removed as a child term from GO:0046651 (*lymphocyte proliferation*).

From the enriched terms found in the re-performed analyses, a term that evolved a lot was GO:0002683 (*negative regulation of immune system process*); 4 child terms were added while 3 child terms were removed. Another term with quite a lot of changes to its child terms was GO:0031347 (*regulation of defense response*). However, the term that evolved most was GO:0010564 (*regulation of cell cycle process*). 20 child terms were added while 4 child terms were removed.

5 Discussion

5.1 Summary of results

From the results, we can tell that the evolution of GO has a clear effect on the interpretation of Omics results. We find that the most similar enrichment analyses had a similarity score of 0.42 (paper 3), while the least similar enrichment analyses had a similarity score of 0.16 (paper 8).

Even though both the original enrichment analyses and the re-analyzed results for the purpose of this thesis included terms that were considered enriched in more than one paper, there was no overlap between these two sets of terms. Additionally, in the enriched terms found in multiple papers of the original enrichment analyses some clear recent changes in ancestry were found.

A further look at the enriched terms found in more than one paper showed that some of the most evolved terms were terms such as *negative regulation of immune system process, regulation of defense response* and *regulation of cell cycle process.* These terms were more common in the re-performed analyses, this could be due to the large amount of changes to the child and/or the parent terms of these terms.

5.2 Interpretation of results

This thesis aimed to answer the question of what the impact of the GO evolution was on COVID-19 results. The results showcase that the enrichment analyses performed with the most recent version of the GO are vastly different to the original enrichment analyses. These changes sometimes include enriched biological processes that were not found originally at all, which could lead to a change in interpretation of the Omics results. Paper 8 had the lowest similarity score and used a GO-version from October, 2020. Since this paper was published in May, 2021 it is possible that the GO had already evolved enough in that time to cause missing results and interpretations.

From the current results, we can interpret that the evolution of the GO has an impact on COVID-19 results. Especially when examining the terms that were considered enriched in more than one paper, since many of these terms showed clear recent changes in ancestry. The terms that changed most were found in the re-performed enrichment analyses. We can conclude that the changes made to these terms had a large effect on the results of Omics research, and in particular COVID-19 results.

5.3 Implications of results

The research question that this thesis aims to explain is important since it could help guide researchers to choose a tool that uses the most recent version of the GO. As we can tell, using an outdated version of the GO will lead to different interpretations of Omics results which could mean that the researchers miss out on important findings that could be used to treat other diseases than COVID-19.

The results show some of these changes in interpretations. An example of this is paper 3, this paper looked at a COVID-19 induced protein network to identify drugs that could be repurposed to treat

COVID-19, using a GO-version that was not 3 years out of date might have led the researchers to more drugs that could have potentially been repurposed. Another example of this is paper 7, which also examined whether drugs could have been repurposed to treat COVID-19, however the version of the GO used here was roughly 6 months out of date. Moreover, the changes in the ancestry terms such as GO:0010564 and GO:0002683 could also have changed the interpretation of the results found in the papers.

5.4 Limitations

Some limitations of this thesis have to do with the re-usability of the subset of papers themselves. Sometimes it was difficult to generate the exact same set of DEGs as were found within the original papers, due to several reasons. The primary reasons was that many of the datasets used within the paper had been updated months or even years after the publishing of the paper, there was no clear changelog of these updates and it was difficult to know whether these updates on the dataset would have a large effect on an enrichment analysis that was performed on a different version of the dataset than the original. Another reason this was difficult was due to the accessibility of the code used originally, sometimes the code used to process the raw sequencing data was unavailable which made it difficult to know exactly which methods of processing were used to replicate this.

For the purpose of this thesis, a selection of 10 papers was used to create the initial workflow. This subset was chosen based on the fact that these papers included enrichment analyses that were generally easier to re-perform than in other papers; the data was more readily available and allowed for simpler comparison. A wider study looking at a larger subset of papers could yield more promising results. However, the topics in the subset of papers are all quite different. While all papers are related to COVID-19, some papers examine the effects of COVID-19 itself while various other papers look at possible ways to treat COVID-19 or examine the similarities of COVID-19 to other diseases such as Dengue. Using papers that would be more focused on just one of these areas would likely generate more similarities in the recently originally performed analyses, allowing them to be compared more easily.

Another possible limitation was the method used to calculate the semantic similarity. Applying different Sem sim methods could also yield different results. It could be possible that the similarity score would be slightly higher if this were the case.

5.5 Future studies

As noted above, this thesis only looked at a subset of 10 papers from the larger analysis that examined all enrichment analyses found in COVID-19 related papers. An important addition to this thesis would be a larger analysis that took into account all found papers, this would likely yield a clearer list of terms that were considered enriched in more than one paper as well as identify clear changes in interpretation of the original results reported by the papers. Another option would be to use the tool used to calculate the semantic similarity to identify papers where large changes have occurred to specifically identify evolved areas of the GO.

Additionally, COVID-19 is just one area where the evolution of GO can be examined and it could be interesting to apply these methodologies on a set of papers relating to another illness to find other interpretations. Perhaps this method could be used to help uncover new interpretations using the newest version of GO as opposed to the ones used within the paper.

Furthermore, research done about the evolution of GO could help create a set of guidelines that can help developers when developing new tools used for enrichment analysis. This would allow developers to develop tools that could withstand the changing nature of the GO and yield the most meaningful results.

References

[ABB$^+$00] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, and et al. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[AMK$^+$21] Anna C. Aschenbrenner, Maria Mouktaroudi, Benjamin Krämer, Marie Oestreich, Nikolaos Antonakos, Melanie Nuesch-Germano, Konstantina Gkizeli, Lorenzo Bonaguro, Nico Reusch, Kevin Baßler, and etal. Disease severity-specific neutrophil signatures in blood transcriptomes stratify covid-19 patients. *Genome Medicine*, 13(1), 2021.

[BJHA19] Aurélien Brionne, Amélie Juanchich, and Christelle Hennequet-Antier. Viseago: A bioconductor package for clustering biological functions using gene ontology and semantic similarity. *BioData Mining*, 12(1), 2019.

[bou]

[CVW] Yi Chen, Fons Verbeek, and Katherine Wolstencroft. *FAIR Functional Enrichment: Assessing and Modelling Provenance in Omics Results.*

[dPSD11] L. du Plessis, N. Skunca, and C. Dessimoz. The what, where, how and why of gene ontology–a primer for bioinformaticians. *Briefings in Bioinformatics*, 12(6):723–735, Feb 2011.

[FCL$^+$20] Paolo Fagone, Rosella Ciurleo, Salvo Danilo Lombardo, Carmelo Iacobello, Concetta Ilenia Palermo, Yehuda Shoenfeld, Klaus Bendtzen, Placido Bramanti, and Ferdinando Nicoletti. Transcriptional landscape of sars-cov-2 infection dismantles pathogenic pathways activated by the virus, proposes unique sex-specific differences and predicts tailored therapeutic strategies. *Autoimmunity Reviews*, 19(7):102571, May 2020.

[FLR$^+$21] Mariana G. Ferrarini, Avantika Lal, Rita Rebollo, Andreas J. Gruber, Andrea Guarracino, Itziar Martinez Gonzalez, Taylor Floyd, Daniel Siqueira de Oliveira, Justin Shanklin, Ethan Beausoleil, and et al. Genome-wide bioinformatic analyses predict key host and viral factors in sars-cov-2 pathogenesis. *Communications Biology*, 4(1), 2021.

[Gen]

[GOF22] Jan 2022.

[HHT$^+$21] Namshik Han, Woochang Hwang, Konstantinos Tzelepis, Patrick Schmerer, Eliza Yankova, Méabh MacMahon, Winnie Lei, Nicholas M. Katritsis, Anika Liu, Ulrike Felgenhauer, and et al. Identification of sars-cov-2–induced pathways reveals drug repurposing strategies. *Science Advances*, 7(27), 2021.

[KSGH20] Maxat Kulmanov, Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. Semantic similarity and machine learning with ontologies. *Briefings in Bioinformatics*, 22(4), 2020.

[LLL+21] Dan-Yang Liu, Jia-Chen Liu, Shuang Liang, Xiang-He Meng, Jonathan Greenbaum, Hong-Mei Xiao, Li-Jun Tan, and Hong-Wen Deng. Drug repurposing for covid-19 treatment by integrating network pharmacology and transcriptomics. *Pharmaceutics*, 13(4):545, 2021.

[LLP+21] Lingjie Luo, Wenhua Liang, Jianfeng Pang, Gang Xu, Yingying Chen, Xinrong Guo, Xin Wang, Yi Zhao, Yangdian Lai, Yang Liu, and et al. Dynamics of tcr repertoire and t cell function in covid-19 convalescent individuals. *Cell Discovery*, 7(1), 2021.

[PAN21] Peter Natesan Pushparaj, Angham Abdulrahman Abdulkareem, and Muhammad Imran Naseer. Identification of novel gene signatures using next-generation sequencing data from covid-19 infection models: Focus on neuro-covid and potential therapeutics. *Frontiers in Pharmacology*, 12, 2021.

[PH21] Amber Park and Laura K. Harris. Gene expression meta-analysis reveals interferon-induced genes associated with sars infection in lungs. *Frontiers in Immunology*, 12, 2021.

[TH10] Hannah Tipney and Lawrence Hunter. An introduction to effective use of enrichment analysis software. *Human Genomics*, 4(3):202, Jan 2010.

[TMW+18] Aurelie Tomczak, Jonathan M. Mortensen, Rainer Winnenburg, Charles Liu, Dominique T. Alessi, Varsha Swamy, Francesco Vallania, Shane Lofgren, Winston Haynes, Nigam H. Shah, and et al. Interpretation of biological experiments changes with evolution of the gene ontology and its annotations. *Scientific Reports*, 8(1), 2018.

[Wil]

[Yu20] Guangchuang Yu. Gene ontology semantic similarity analysis using gosemsim. *Methods in Molecular Biology*, page 207–215, 2020.

[ZWL+21] Wenjiang Zheng, Hui Wu, Chengxin Liu, Qian Yan, Ting Wang, Peng Wu, Xiaohong Liu, Yong Jiang, and Shaofeng Zhan. Identification of covid-19 and dengue host factor interaction networks based on integrative bioinformatics analyses. *Frontiers in Immunology*, 12, 2021.