MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220283

A Comparative Analysis of Phenotypes Derived from Genes or Biomedical Literature in COVID-19

Sophie Steenson[a], Christopher Hawthorne[a], Guillermo Lopez-Campos[a]

[a] *Wellcome-Wolfson Institute for Experimental Medicine, Queen's University Belfast, Belfast, Northern Ireland, United Kingdom*

Abstract

Since the emergence of SARS-CoV-2 in November 2019, there has been an exponential production of literature due to worldwide efforts to understand the interactions between the virus and the human body. Using an "in-house" developed script we retrieved gene annotations and identified phenotype enrichments. Human Phenotype Ontology terms were retrieved from the literature using the Onassis R package. This produced both disease-gene and disease-phenotype data as well as data for gene-phenotype interactions. Overall, we retrieved 181 human phenotypes that were identified by both approaches. Further in-depth analysis of these relationships could provide further insights in the molecular mechanisms related with the observed phenotypes, answers and hypotheses for key concepts within COVID-19 research.

Keywords:

COVID-19, Computational Biology, Data Mining.

Introduction

The need to use automated solutions has been further highlighted and stressed in the study of SARS-CoV-2/COVID-19. By May 2021, merely a year and a half after its appearance in November 2019, more than 130,000 publications were available in PubMed and quickly exceeded human capacity for manual annotation and outpaced the capability of individual researchers to follow up the advances and publications in the area. To manage this data deluge, biomedical informatics scrambled along with many other disciplines to develop tools and solutions to support and advance the understanding and development of novel therapies against this disease [1].

It is known that physiological and phenotypical outcomes of SARS-CoV-2 infection involves a combination of both viral and host genomic factors [2]. Expression of host genes results in individual phenotypes, which are understood to be 'visible and measurable characteristics'. This includes both innate phenotypes, and observable and measurable symptoms that appear with illness. In this paper we used some of the methods previously developed to explore the complex relationships between genes, phenotypes and diseases using COVID-19 as a proof of concept. We have combined different biomedical informatics tools and methods to extract genes and phenotypes from the literature corpus around COVID-19. We proceeded to use this information to compare the described phenotypes in the literature with those phenotypes that could be predicted from the gene related information as well as the potential underlying molecular mechanisms.

Methods

Data extraction and analysis was implemented for both gene and phenotype biomarkers via R scripts, within R 4.0.3. In-house scripts aided in the extraction and analyses of gene and phenotype information with the application of both Bioconductor and CRAN repositories.

The methodology consisted in two different and parallel steps. The first of these steps consisted in the retrieval and annotation of gene-disease-phenotype relationships. The second step related to the retrieval of disease-phenotype annotations.

For the retrieval and annotation of disease-gene data we used GAPAL, an in-house developed script for the analysis and retrieval of genes and proteins from the literature and pathways analysis [3] using PubtatorCentral annotations and PubMed queries. For this project, we provided GAPAL with the following input parameters:

- A PubMed query structured as "("COVID-19"[MeSH Terms] OR "SARS-CoV-2"[MeSH Terms])".
- A database with PubtatorCentral annotations (derived from the annotations available on 12/11/2020)
- Pathway analyses based on "g-Profiler" annotations.
- Significance threshold for pathway and phenotype enrichment of adj.p-value <0.05
- Marker frequency threshold of 10. This value determines the minimum number of papers on which a marker appears in order to be considered for the gene set enrichment analysis.

GAPAL analyses ultimately generated a list of genes and "gene-derived" phenotypes based on the Human Phenotype Ontology enrichments.

For the retrieval of the disease-phenotype annotations we used the R package RISMed for the retrieval of PubMed abstracts and used a modified version and limited version of the PubMed query previously mentioned due to computational limitations. This modified query was '("COVID-19[TiAB] OR "SARS-CoV-2"[TiAB] AND ("Symptom" OR "Phenotype")'

Phenotypes for the annotation of the literature were extracted from the Human Phenotype Ontology (HPO) and we used annotations downloaded from HPO website (https://hpo.jax.org/) on 02/02/2021. This file was used to generate a dictionary that was used by the Onassis R [https://github.com/eugeniaeueu/Onassis] package for the phenotype annotation stage generating a "literature-derived" set of HPO annotations. These

HPO terms where then used to retrieve their gene annotations creating a "HPO-derived" gene list.

Finally, both sets of phenotype annotations (gene-derived and literature-derived) were compared.

Results

A summary of the methodology developed in this study and the results obtained is presented in Figure 1.

The first step using GAPAL resulted in the identification of 64,447 PubMed documents and 11,100 COVID-19 related genes were identified within these documents; this included 8012 human genes and 3088 related to COVID-19 infection in animals and the SARS-CoV-2 viral genes. Following gene annotation, the next step was the geneset enrichment analysis based on the retrieved genes. As previously mentioned we used g:Profiler Human Phenotype Ontology annotations. These analyses allowed us to identify 801 different statistically significant gene-derived HPO terms (adj. p-value <0.05). The use of Onassis, generated a list containing phenotypes and modifier terms together with their frequencies. This list contained 640 literature-derived phenotypes together with the frequency.

The comparison between both approaches resulted in 181 phenotypes identified by both approaches (table 1)

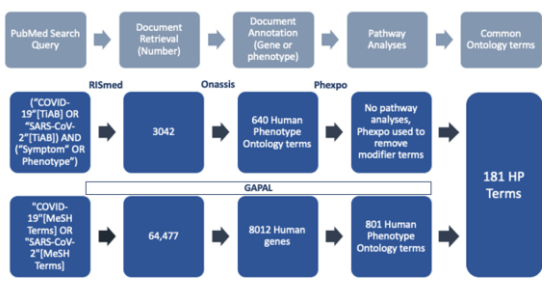

Figure 1– A schematic of data mining and analysis for the bioinformatic tools used for both gene and phenotype searches.

Table 1– List of the top 10 overlapping phenotypes the ranking was based on a combination of adj.pvalues and document frequency

Term	Term name	Term	Term name
HP:0001945	Fever	HP:0002315	Headache
HP:0002090	Pneumonia	HP:0000822	Hypertension
HP:0002094	Dyspnea	HP:0002014	Diarrhea
HP:0012378	Fatigue	HP:0001888	Lymphopenia
HP:0012531	Pain	HP:0002013	Vomiting

Conclusions

With an ever-increasing production of scientific literature, development and use of biomedical informatic tools is an essential aspect of scientific discovery. This has been exemplified with the unprecedented amount of literature produced following the emergence of SARS-CoV-2. We have designed a biomedical informatics based approach that allowed us to successfully extract and bring together human gene and phenotype information contained in the vast amount of COVID-19 literature identifying relevant disease-gene, disease-phenotype and gene-phenotype interactions. This information can be used to provide further insight about the underlying molecular mechanisms that might be driving the individual phenotypes associated with COVID-19 and could potentially support the development of new therapies.

Acknowledgements

Mr. C Hawthorne is supported by a Northern Ireland Department of the Economy (DfE) postgraduate studentship award. Dr. G. Lopez Campos contributions to this work was partially funded by the Health and Social Care (Northern Ireland) Research and Development.

References

[1] E.M. Hechenbleikner, D.V. Samarov, and E. Lin, Data explosion during COVID-19: A call for collaboration with the tech industry & data scrutiny, EClinicalMedicine 23 (2020), 100377.

[2] M. LoPresti, D.B. Beck, P. Duggal, D.A.T. Cummings, and B.D. Solomon, The Role of Host Genetic Factors in Coronavirus Susceptibility: Review of Animal and Systematic Review of Human Literature, Am J Hum Genet 107 (2020), 381-402.

[3] G. Lopez-Campos, E. Bonner, and L. McClements, An Integrative Biomedical Informatics Approach to Elucidate the Similarities Between Pre-Eclampsia and Hypertension, Stud Health Technol Inform 264 (2019), 988-992.

Address for correspondence

Guillermo Lopez Campos, Address: Wellcome-Wolfson Institute for Experimental Medicine, 97 Lisburn Road, Belfast BT9 7BL, United Kingdom. Email: g.lopezcampos@qub.ac.uk..