OPEN ACCESS Freely available online

Journal of Proteomics & Bioinformatics

Review

Study on the Intercations Between Genes Lincked to Covid-19 and Genes EXpressed in the Human Body

Gustavo Simoes Carnivali*

Federal University of Minas Gerais, Belo Horizonte - Minas Gerais – Brazil

ABSTRACT

Covid 19 is a viral disease that emerged in China in December 2019 and has already spread throughout the world, with a high mortality and a high proliferation rates. It currently consists of a major focus of study, and among recent studies, one presented a set of 94 genes expressed in the human body with influence on Covid 19. The aim of this study is to understand the interaction of these 94 genes with other genes expressed in the human body, characterizing their properties by means of a gene expression network containing thousands of genes expressed in the human body. This characterization may concern a diverse set of properties. This study enabled the discovery of a limited set of genes with a regulatory role and a high influence on the human genes. The analysis also showed that although the 94 genes linked to Covid 19 may integrate several distinct metabolic pathways, they present great participation in only two main groups. Thus, this study has deepened the understanding on the genetic functioning of Covid 19 and may support explanations on the phenomena related to the disease, such as its high proliferation or the various phenotypic side effects caused by it. The author would also like to thank Capes, Fapemig and UFMG for the funding.

INTRODUCTION

At the end of December 2019, the outbreak of a disease caused by a novel coronavirus (Covid-19, formerly known as 2019-nCoV, caused by the SARS-CoV-2 virus) was reported in Wuhan China. Currently, several countries around the globe have been affected by this disease. Covid-19 is a disease with symptomatic treatment, but it can be acute and deadly considering that the severe onset of the illness can result in death due to progressive respiratory failure. On March 12, 2020, there were 125,048 confirmed cases worldwide and the infection presented a mortality rate of approximately 3.7% .

Despite its high mortality and worldwide occurrence, the disease caused by the new coronavirus still does not have an efficient treatment. This paucity creates an urgent need to understand this illness in order to develop effective treatments for Covid-19. This study aimed to develop a genetic analysis of the effects of coronavirus on the set of human genes. Quantitative information between the parasite genes and the host were made explicit since this information not only brings greater understanding regarding the disease, but also provides support for proposing a treatment for Covid-19.

The article involved the study of the set of common genes to the virus and not expressed in the human body, presenting 94 genes expressed in humans that have connections to the Covid-19 virus. This article aims to analyze genetic factors around these 94 genes, through the development of a gene co-expression network of various genes expressed in the human body, including the 94 genes related to Sars-CoV-2. These were analyzed regarding their connections with other human genes, with the goal of finding pathways of genetic connections in order to not only know more about the disease, but also to determine, among these 94 genes, the most likely genes to cause typical phenotypic effects or to be inhibited by different treatments.

Accordingly, three analyses were performed on the interaction between the set of genes linked to Covid-19 and other genes expressed in the human body: analysis of vertices connection to determine which of the 94 genes most influence other genes expressed in humans; analysis to determine the number of vertices reached by the genes linked to Covid 19 in order to examine the possible genetic and phenotypic diversity of the

***Correspondence to:** Gustavo Simoes Carnivali*·Federal University of Minas Gerais, Belo Horizonte Minas Gerais – Brazil, Tel: -5532999340065,Email: gustavocarnivali@gmail.com

Received; August12, 2020; **Accepted:**August30,2021; **Published:**September10,2021

Citation: CARNIVALI, G. S .2021 Study on the interactions between genes linked to Covid-19 and genes expressed in the human body . J Proteomics Bioinform14:p254

disease; and analysis by communities of the genes network expressed in the human body, in order to determine functional sets of genes and their participation in the various symptoms of the disease.

After the analyses, it was possible to identify several important properties of the disease, such as expressed regulatory genes and genes participation in common metabolic pathways.

METHODS

Gene expression can be interpreted as the process in which DNA nucleotide sequences are transcribed into RNA or proteins in a functional genetic product. The speed of gene expression may differ and be influenced by several factors.

Altering the expression of a gene can increase or decrease the expression of other genes. For instance, an increase in the expression of one gene, caused by an anomaly, can generate an increase the expression of another gene which was not initially associated with this anomaly, but that participates in the phenotypic effects of the anomaly. Finding these relationships allows the discovery of unknown metabolic pathways about a disease or anomaly.

The correlation between genes is well represented by a graph, or complex network, which consists of mathematical structures widely used to represent relationships between objects. A graph G= (V, E) is a structure formed by a non-empty set of vertices V and a set of edges E [16]. In this study, the set of vertices V represents the genes that participate in the phenotypic effects of a disease, while the set of edges E represents the relationships between those genes.

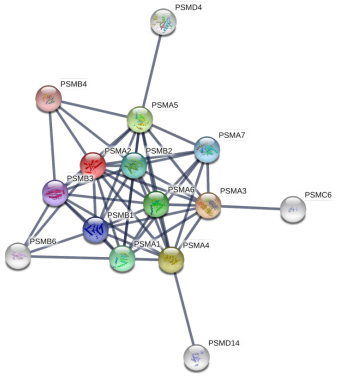

Figure 1: Example of GCN with human genes

A gene co-expression network is a graph that represents the co-expression relationships between genes. A GCN is commonly used both to find a group of genes associated with an anomaly and to create conceptual detailing on the phenotypic effects that each gene engender in a disease. A visual example of GCN with human genes is demonstrated in Figure 1.

An interactome was obtained with the interactions between SARS-CoV-2 and human proteins. The DNA of several Sars-Cov-2 have already been described so several viral proteins that make up the virus are already known (https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/). In short, the interactome contains protein-protein interactions from the of SARS-CoV-2 with known human proteins. The discovered human proteins represent human bodily functionalities influenced by viral proteins. The interactome has a total of 125 proteins. These proteins are already being used for, among other things, greater understanding of Sars-Cov-2 and the reuse of medicines for Covid-19.

The String tool offers a database with thousands of genes expressed in humans (among them the 94 genes linked to Covid-19 found in the study), allowing the creation of a GCN, which is helpful to determine the fundamental properties of the 94 genes in question.

ANALYSIS OF GENES DEGREE LINKED TO COVID-19

This analysis verifies the number of total connections that a gene has in a network (i.e., vertex degree) .The analysis of a vertex degree determines the connectivity of vertices in a network; in a GCN, genes with a high degree have high relevance in the expression of other genes. If their expression is altered, genes with a high degree will cause cascading changes in several other genes, with multiple secondary phenotypic implications caused by these alterations.

The GCN performed is valued at its edges, which means that each edge has a value between 0 and 1, and this value determines the statistical reliability level of the connection's existence. Edges with values close to 1 determine that if a gene's expression is altered, there is a high probability that the gene connected to it will also have its expression altered. In this study, this value was called Score.

Three levels of genes degree linked to Covid-19 were calculated. The first calculation consists of the sum of the Scores of all the edges connected to the gene. Genes connected to other high degree genes also have a high influence probability; considering this, levels 2 and 3 of the measurement considered the sum of the Scores of all edges connected to a gene plus the Score of all edges connected to their neighbors of distance 3 and 5. The distance represents the number of vertices connecting two other vertices in the shortest possible path. For instance, in Figure 1 the PSMC6 gene has distance 1 from the PSMA4 gene and distance 2 from the PSMD14 gene.

The 5 highest scoring genes are listed below, in ascending order for the 3 measurement levels. The parentheses present the score of each gene:

Level 1: PSMA2 (28.193), EEF1A1 (24.821), POLR2B (16.298), ENO1 (8.454), CHEK2 (7.076)

Level 2: EEF1A1 (1448.55), PSMA2 (1359.71), POLR2B (1281.5), ENO1 (258.986), SERPING1 (75.144)

Level 3: EEF1A1 (1509.94), PSMA2 (1506.29), POLR2B (1488.6), RPS17 (871.532), ENO1 (312.495)

A summary of the three highest scoring genes properties regarding each type of calculation is presented below:

PSMA2 - Component of the 20S nucleus proteasome complex involved in the proteolytic degradation of most intracellular

proteins. This complex plays numerous essential roles within the cell, having associations with different regulatory particles. This gene is also connected to cardiovascular problems, a common condition in Covid 19. this gene is also associated with breast cancer. [21]

EEF1A1 - This gene is associated with a protein that promotes ribosome binding during protein biosynthesis. With the participation of other genes, it forms a complex that acts as a specific Th1 cell transcription factor and binds the IFN-gamma promoter to directly regulate its transcription. Therefore, it is importantly involved in the production of Th1 cytokines. This gene is also associated with cardiovascular problems and breast cancer.

POLR2B - Second largest component of RNA polymerase II, which synthesizes mRNA precursors and many non-functional RNAs. It is considered as contributor to the polymerase catalytic activity. This gene is also associated with cardiovascular problems and breast cancer.

This result demonstrates that irrespective of the type of calculation of the vertices degree, little change was observed between the genes with the highest scores in the three types of calculation, with the first three genes presenting higher scores in comparison to other genes.

Analysis of the genes range linked to Covid-19

This second analysis of the genes range linked to Covid-19 was performed in order to analyze the number of affected genes in case a gene has its expression altered by SARS-CoV-2. This analysis enabled to determine the influence caused by the virus on different patterns of the human phenotype. If a high range gene is altered by Covid-19, it will lead to an alteration in several other genes which were not initially associated with Covid-19.

Equation 1 was performed to calculate the gene range, through the calculation of the range of vertex a (P(a)) as the sum of the ratio of the distance value from vertex a to the reached vertex b, considering N as the number of vertices reached by a. Equation 1 increases the score of vertices that reach several other vertices but, at the same time, raises the score slightly in case the reached vertex is too far from the starting vertex.

$$P(a) = \sum_{i=0}^{N} \left(\frac{1}{Distance(a,b)} \right) \quad \text{Eq (1)}$$

The list of the highest to the lowest scored vertices of the genes linked to Covid 19 is presented below:

EEF1A1, PSMA2, POLR2B, RPS17, ENO1, ACE2, SERPING1, H2AFY2

As expected, the genes with the highest range are also the genes with the highest degree, demonstrating the relevance of these genes in the human genetic network. This result reiterates the previous result presented aforementioned, demonstrating that these genes may be associated with various phenotypic effects present in Covid 19.

ANALYSIS BY COMMUNITIES OF GENES LINKED TO COVID-19

This last analysis performed on the genes linked to Covid-19 was the community analysis. The problem of community detection is to find, in a graph, groups of vertices that have one or more characteristics in common. One of these characteristics may be the common neighbourhood between vertices, which can be given by the number of edges that the vertices of the same group (i.e., a community) share with each other [23]. In short, in this study, the GCN was divided into communities in order to establish functional groups of genes. The groups that the genes linked to Covid-19 participate are highlighted and the results provide information that will be discussed.

For dividing the GCN into communities, a variation of the Louvain Method was performed. These studies present a comparative survey of several community detection methods, among which the Louvain Method stands out for quickly and effectively identifying communities in very large graphs. The variation achieves an acceleration of LM processing time, maintaining the quality of the communities found by the LM. The modification is used because this study's graph has more than 19,000 vertices.

LM was applied to GCN with 19,354 genes grouped in 11,560 communities. The high number of communities in relation to the number of genes is possibly due to the low number of edges of the graph. Only edges with scores greater than 0.3 were considered in order to maintain the reliability of the connections. At the end of this process, the GCN had a total of 291,400 edges connecting the genes.

11,560 communities were found by the LM, from which 47 contained some gene linked to Covid-19. It is a high value considering the total of 94 genes investigated which were linked to Covid-19. Of the 47 communities with a gene linked to Covid-19, a total of 38 communities were composed of only one of these genes. This result may explain the great variety in symptoms that this disease has [22], since its genes are linked to several distinct functional groups.

Analyzing the 47 communities that contain some gene linked to the disease, few communities have a high number of these genes, which can be due to two factors: a characteristic of the disease, presenting few well expressed symptoms and many poorly expressed symptoms; or a characteristic of the study that discovered the 94 genes linked to Covid-19, which allowed the inclusion of genes that have similar functional roles within the cell.

The most populous communities regarding the genes linked to the Covid-19 are listed below:

Community 1: EEF1A1, EIF3F, BTF3, RPS20, PFDN5, CAMLG, YWHAE, SGTA, PSMA2, BZW2, RPS17, BAP1, PPIG, FKBP1A

Community 2: PPIH, UBE2I, MNAT1, ATP6V1G1, CHMP2B, NAE1, TBCB, N4BP2L2, HGS, NCOA5, MARK2

Community 3: EIF4B, LAS1L, POLR2B, DDX5

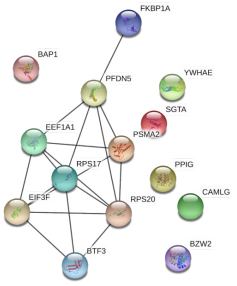

Figure 2: Network of the largest community found with genes linked to Covid-19

Only two communities had a high number of genes linked to Covid-19. Furthermore, the first community has the two highest degree genes demonstrated aforementioned in this study. A visual representation of the graph generated with the genes linked to Covid-19 and community 1 is shown in Figure 2.

CONCLUSION AND DISCUSSION

This study presented an analysis of a set of human genes linked to the disease caused by the Sars-Cov-2 virus, called Covid-19. The disease began in December 2019 in China and in a few months spread to several countries. Because it is a recent illness, little is known about it; however, both its high mortality and proliferation drive the continuous advance in studies on this disease.

Based on gene expression networks, with the aim of analyzing the genes linked to the disease, this study analyzed the connectivity of a set of 94 human genes linked to Covid-19 with another set of thousands of human genes not related to Covid-19, aiming to analyze these 94 genes regarding their relationships.

The degree of relationship between the genes linked to Covid-19 and other genes expressed in humans was analysed, as well as the number of genes that those linked to Covid 19 can influence. Both analyses were performed in order to find the genes that most influence other genes in the human body. The genes with the greatest range and degree were found, and the characteristics of a limited group of these genes were presented. The analysis of these characteristics enabled the observation that they all participated in the cell's regulatory factors, which may explain a wide range of symptoms caused by Covid-19.

Regarding the analysis of the highest Scored genes' properties, it is worth noting that the aforementioned 3 genes demonstrated an important role in gene regulation processes. This role was expected, since genes linked to regulatory processes are necessarily present in several cellular processes and are connected to several other genes, explaining their high degree of connection. Thus, these genes connect several others to the genes that constitute the Covid-19, necessarily having a significant importance within the disease and its still unknown phenotypic factors.

This result may also explain the high mortality and the high number of mild symptoms caused by Covid-19, already presented in other studies [22, 28, 32]. As the virus alters the expression of both regulatory and high degree genes, this implies that its action in the human body can be diverse and present on several distinct metabolic pathways.

In addition, it was observed that these 3 genes in common influence cardiovascular factors in the body, an evil common to Covid 19. This finding may allow a greater knowledge of the disease in addition to a greater ability to fight the disease, by knowing three genes as similar functionality in the human body influenced by Sars-Cov-2.

GCN was divided into communities as a variation of a classical community detection method. This analysis showed that the genes linked to Covid-19 may play a role in several distinct metabolic pathways. In addition, the results also indicate the existence of two highly expressed metabolic pathways of this disease. This previous result allows a greater understanding of unknown factors of this disease, although new therapeutic targets are needed for the treatment of this disease.

Future studies are necessary in order to analyze other properties of this network, as well as to deepen the analysis of the highest scoring components found. Future studies may also consider the use of network synthesizers other than String and the analysis of other genes besides the 94 studied in this article, as well as use other community detectors, including methods that perform divisions which also take into consideration biological factors of the genes.

The codes and data used in the experiments, as well as other information in the text, are available at: https://www.gscarnivali.com/publications/study-on-the-interactions-between-genes-linked-to-covid-19-and-genes-expressed-in-the-human-body/.

REFERENCES

1. [1] Wu, Fan, et al. "A new coronavirus associated with human respiratory disease in China."; Nature 579.7798 (2020): 265-269.
2. [2] Huang, Chaolin, et al. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China." The Lancet 395.10223 (2020): 497-506.
3. [3] Sohrabi, Catrin, et al. "World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19)." International Journal of Surgery (2020).
4. [4] Chan, Jasper Fuk-Woo, et al. "A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster." The Lancet .
5. [5] World Health Organization. "Coronavirus disease 2019 (COVID-19): situation report, 72." (2020).
6. [6] Zhou, Yadi, et al. "Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2." Cell discovery 6.1 (2020): 1-18.
7. [7] Srinivasan, Suhas, et al. "Structural Genomics of SARS-CoV-2 Indicates Evolutionary Conserved Functional Regions of Viral Proteins." Viruses 12.4 (2020): 360.
8. [8] Shimkets, Richard A., ed. Gene expression profiling. Vol. 258. Springer Science & Business Media, 2004.
9. [9] Parmigiani, Giovanni, et al. "The analysis of gene expression data: an overview of methods and software." The analysis of gene expression data. Springer, New York, NY,2003. 1-45.

10. [10] Lee, Cheol-Koo, et al. "Gene expression profile of aging and its retardation by caloric restriction."Science 285.5432 (1999): 1390-1393.
11. [11] van Noort, Vera, Berend Snel, and Martijn A. Huynen. "Predicting gene function by conserved co-expression." TRENDS in Genetics 19.5 (2003): 238-242.
12. [12] Zhu, Qun, et al. "Enhanced protection against fungal attack by constitutive co–expression of chitinase and glucanase genes in transgenic tobacco." Bio/technology 12.8 (1994): 807-812.
13. [13] Emilsson, Valur, et al. "Genetics of gene expression and its effect on disease." Nature 452.7186 (2008): 423-428.
14. [14] Smeets, Cleo JLM, and D. S. Verbeek. "Cerebellar ataxia and functional genomics: identifying the routes to cerebellar neurodegeneration." Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 1842.10 (2014): 2030-2038.
15. [15] Stuart, Joshua M., et al. "A gene-coexpression network for global discovery of conserved genetic modules." science 302.5643 (2003): 249-255.
16. [16] Paulo Feofiloff, Yoshiharu Kohayakawa, and Yoshiko Wakabayashi. Uma introdução sucinta à teoria dos grafos. 2011.
17. [17] Smeets, Cleo JLM, and D. S. Verbeek. "Cerebellar ataxia and functional genomics: identifying the routes to cerebellar neurodegeneration." Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 1842.10 (2014): 2030-2038.
18. [18] Srinivasan, Suhas, et al. "Structural Genomics of SARS-CoV-2 Indicates Evolutionary Conserved Functional Regions of Viral Proteins." Viruses 12.4 (2020): 360.
19. [19] Wheeler, David L., et al. "Database resources of the national center for biotechnology information." Nucleic acids research 36.suppl_1 (2007): D13-D21.
20. [20] Snel, Berend, et al. "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene." Nucleic acids research 28.18 (2000): 3442-3444.
21. [21] UniProt Consortium. "The universal protein resource (UniProt)." Nucleic acids research 36.suppl_1 (2007): D190-D195.
22. [22] Surveillances, Vital. "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)—China, 2020." China CDC Weekly 2.8 (2020): 113-122.
23. [23] Carnivali, Gustavo S., et al. "Método Rápido de Agrupamento de Vértices para Detecção de Comunidades em Redes Complexas de Larga-escala." Anais do XVII Workshop em Desempenho de Sistemas Computacionais e de Comunicação. SBC, 2018.
24. [24] Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." Journal of statistical mechanics: theory and experiment 2008.10 (2008): P10008.
25. [25] Carnivali, Gustavo S., et al. "CoVeC: Coarse-Grained Vertex Clustering for Efficient Community Detection in Sparse Complex Networks." Information Sciences (2020).
26. [26] Lancichinetti, Andrea, Santo Fortunato, and Filippo Radicchi. "Benchmark graphs for testing community detection algorithms." Physical review E 78.4 (2008): 046110.
27. [27] Leskovec, Jure, Kevin J. Lang, and Michael Mahoney. "Empirical comparison of algorithms for network community detection." Proceedings of the 19th international conference on World wide web. 2010.
28. [28] Tian, Sijia, et al. "Characteristics of COVID-19 infection in Beijing." Journal of Infection (2020).
29. [29] Carnivali, Gustavo S. "Analisando características da rede genética gerada por genes vinculados ao Covid-19." InterAmerican Journal of Medicine and Health (IAJMH) (2020).
30. [30] Maranon, David G., et al. "The interface between coronaviruses and host cell RNA biology: Novel potential insights for future therapeutic intervention." Wiley Interdisciplinary Reviews: RNA (2020): e1614.
31. [31] Zhou, Qiongqiong, et al. "Potential therapeutic agents and associated bioassay data for COVID-19 and related human coronavirus infections." ACS Pharmacology & Translational Science (2020).
32. [32] Khodadadi, Ehsaneh, et al. "Study of combining virtual screening and antiviral treatments of the Sars-CoV-2 (Covid-19)." Microbial Pathogenesis (2020): 104241.
33. [33] Clerkin, Kevin J., et al. "COVID-19 and cardiovascular disease." Circulation 141.20 (2020): 1648-1655.