Original Article

HEALTH SCIENCES
MEDICINE

DOI: 10.32322/jhsm.991465

J Health Sci Med 2021; 4(6): 960-966

Evaluation of studies on molecular biology and genetics related to COVID-19 with data mining

Esra Güzel Tanoğlu[1,2], Muhammed Fevzi Esen[3]

[1]University of Health Sciences, Institution of Hamidiye Medical Sciences, Department of Molecular Biology and Genetics, Istanbul, Turkey
[2]University of Health Sciences, Experimental Medicine Research and Application Center, Istanbul, Turkey
[3]University of Health Sciences, Institution of Hamidiye Medical Sciences, Department of Health Information Systems, Istanbul, Turkey

Cite this article as: Güzel Tanoğlu E, Esen MF. Evaluation of studies on molecular biology and genetics related to COVID-19 with data mining. J Health Sci Med 2021; 4(6): 960-966.

ABSTRACT

Aim: The aim of this study was to examine the most common studies about molecular biology and genetics related to COVID-19. In addition, the aim was also to determine the subject focus of studies about COVID-19 during the pandemic with data mining.

Material and Method: Review and research articles, book chapters, conference abstracts, case reports and mini reviews published between March 2020 and July 2021 were included in this study. We retrieved only articles from the genetics discipline. The MeSH heading "genetics [GENET]" was used including the specific fields in the MeSH hierarchy of cytogenetics, genomics, human genetics, immunogenetics, molecular biology, pharmacogenetics, phenomics, radiation genetics, toxicogenetics, gene ontology, microbial genetics, behavioral and population genetics.

Results: A total of 6234 research articles were evaluated in our study. Of the 85966 terms, 5833 met the threshold from title and abstract extraction. We showed that betacoronavirus, viral pneumonia, viral RNA, spike glycoprotein, coronavirus, middle-aged and animals were the most repetitive terms. Clinical laboratory techniques, polymerase chain reaction and reverse transcriptase polymerase techniques were the main focus for the detection of COVID-19. We found that molecular-based COVID-19 studies were most frequently published by the Journal of Medical Virology, Viruses, and PLoS One. We found that the institutes where molecular-based studies investigating COVID-19 were conducted are in the United States (USA), China and England. The USA and China were in the first rank for countries that conducted the most frequent molecular-based COVID-19 studies, and Turkey was in 19th place in terms of published molecular COVID-19 studies.

Conclusion: It is important to identify the issues and mechanisms most frequently investigated in molecular-based studies related to COVID-19. Scientific approaches founded on evidence-based data may be beneficial to find the curative treatment for COVID-19 infection and to effectively prevent this infection.

Keywords: COVID-19, genetics, molecular biology, data mining

INTRODUCTION

With the worldwide COVID-19 pandemic, humanity is facing a global health threat. There are more than one hundred million infected individuals affected by COVID-19 due to the rapid spread of the virus, resulting in the death of more than four million people. (1).

According to the World Health Organization (WHO), approximately 80% of COVID-19 patients are asymptomatic, approximately 20% of them progress with respiratory tract symptoms, and 5% of these patients need respiratory support. Although the respiratory system is primarily affected including severe pneumonia, COVID-19 involvement in the heart, kidney, nervous system, liver and gastrointestinal system was also reported. Although different races, genders and age groups have equal susceptibility to the virus, i the disease has higher prevalence in people over the age of 60 years (2,3). Individuals with comorbidities such as cardiovascular diseases, hypertension, diabetes, asthma, chronic liver and chronic kidney disease have higher mortality rate (4,5).

Molecular and genetic mechanisms underlie all these COVID-19-related clinical manifestations and conditions that affect the course of the disease (6). Unfortunately, the molecular, biological and genetic mechanisms of the SARS-CoV-2 virus, which we recently encountered, are still not clearly known. Studies and articles dealing with molecular

Corresponding Author: Esra Güzel Tanoğlu, esra.guzel@sbu.edu.tr

Received: 05.09.2021 **Accepted:** 19.09.2021

and genetic mechanisms, as well as clinical presentations of COVID-19, are being published in increasing numbers. (7, 8). However, the areas of focus in molecular biology and genetics-based COVID-19 studies, which have a very common study subject, that are most researched and which topics involve the most frequently asked questions are unknown. Determining which subjects and areas these studies focus on and which mechanisms they focus on, and taking scientific steps by combining the findings like puzzle pieces is the most powerful and rational course of action to eliminate the COVID-19 infection.

In this current data mining study, the aim was to determine the most common research topics in molecular biology and genetics studies related to COVID-19. Moreover, the aim was also to determine the subject focus of studies about COVID-19 during the pandemic.

MATERIAL AND METHOD

This current study is a computer based data-mining study. There is no need to obtain ethical committee approval. All procedures were carried out in accordance with the ethical rules and the principles of the Declaration of Helsinki.

In this research, the study sample consisted of 6234 articles published between March 2020 and July 2021. We addressed all publications regardless of their number of co-authors. Review and research articles, book chapters, conference abstracts, case reports and mini reviews were included in the sample. We retrieved only articles from molecular studies in genetics discipline. The MeSH heading "Genetics [GENET]" was used, including the specific fields in the MeSH hierarchy of cytogenetics, genomics, human genetics, immunogenetics, molecular biology, pharmacogenetics, phenomics, radiation genetics, toxicogenetics, gene ontology, microbial genetics, behavioral and population genetics.

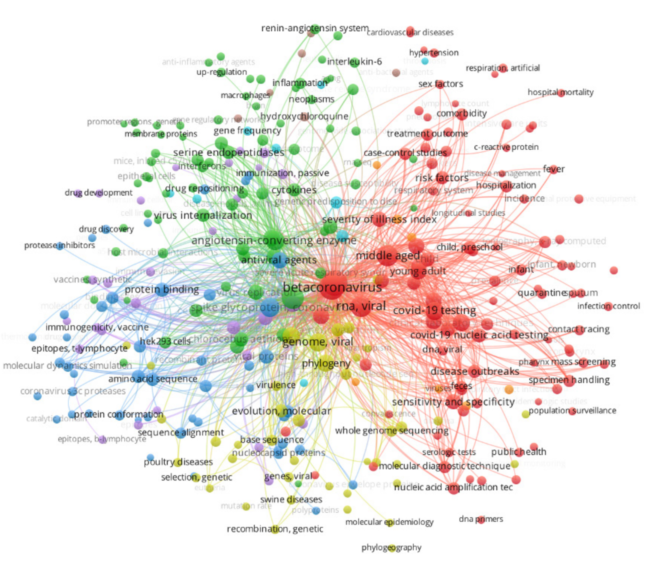

Figure 1. Network diagram of MeSH keywords

We choose the binary counting method to indicate the number of documents in which a term occurs at least once. For bibliographic mapping, the terms were extracted from MeSH headings, title and abstract fields. For title and abstract text analysis, the minimum number of occurrences of a term was set to 5 and relevance scores were calculated. Of the 85966 terms, 5833 met the threshold from title and abstract extraction. Then, the most relevant terms were selected based on the scores. The terms with low relevance scores were filtered out manually in order to focus on more informative terms. The calculation of relevance scores was performed according to Van Eck and Waltman (9).

Co-occurrence analysis was also performed for MeSH keywords. In the analysis, the relatedness of the items was determined based on the number of documents in which the items occur together. Fractional counting was used to determine the weight of a link. The minimum number of occurrences of a keyword was set to 1, so that all of the MeSH keywords (N=5240) met the threshold for co-occurrence analysis.

RESULTS

As mentioned above, 6234 published articles were included in this data mining study. In **Figure 1** (network diagram of MeSH key terms), betacoronavirus, viral pneumonia, viral RNA, spike glycoprotein, coronavirus, middle-aged and animals are the most repetitive terms and the studies about these terms have the highest link strength. It is noteworthy that almost all studies about COVID-19 are related to each other and the studies concentrate around betacoronavirus and RNA. In addition, studies in the field of COVID-19 genetics are shaped around the keywords shown in **Figure 2**. Clinical laboratory techniques, polymerase chain reaction (PCR) and reverse transcriptase polymerase techniques (RT-PCR) are the main focus for the detection of COVID-19.

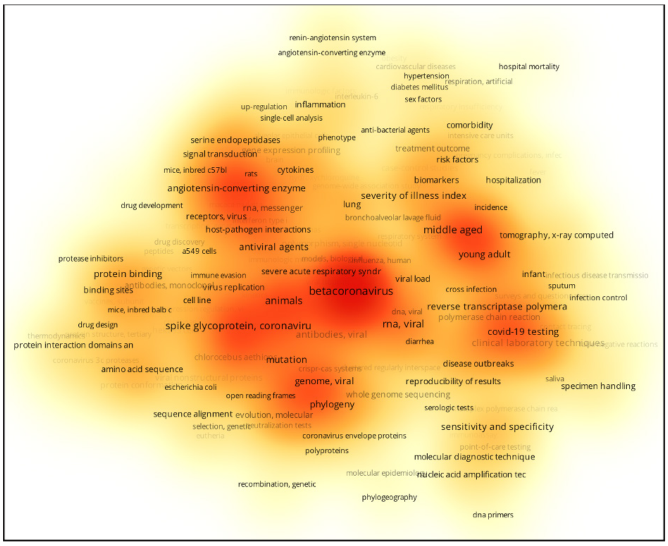

Figure 2. Occurence of MeSH data

Clusters of research fields in the MeSH data are shown **Table 1**. According to **Table 1**, all items are clustered by their links, occurrences and strengths. The subjects in each cluster were studied jointly. Viral pneumonia, betacoronavirus, viral RNA, middle aged and antibodies were in the first cluster. The topics in the first cluster were studied for middle-aged people. The most repeated item in the cluster was betacoronavirus which had the highest link strength, indicating the total strength of the co-authorship with other researchers. Cluster 1 also contained the most studied topics. In cluster 2, animal studies with angiotensin-converting enzyme 2 and peptidyl-dipeptidase-a have the highest links and occurrence. In cluster 3, spike glycoprotein, coronavirus, antiviral agents, protein binding, and viral proteins are included. Phylogeny, mutation, evolution, molecular, genetic variation, and chiroptera terms are included in cluster 4. In cluster 5, COVID-19 vaccines, antibodies, neutralizing, viral vaccines, antigens, viral, antibodies and monoclonal terms are included.

In **Figure 3**, the clustering results for the terms obtained from the titles and abstracts of the studies are given. Accordingly, the studies are gathered around 4 different clusters. The first cluster was carried out in relation to the subjects of receptor binding domain, rbd, phylogenetic analysis, bat, epitobe, viral entry, spike glycoprotein and pedv. In the second cluster, the topics are angiotensin, mouse, tmprss2, IFN, kidney, TNF, ACE2 expression and macrophage.

The studies that make up the 3rd cluster focus on COVID-19 detection, assay, PCR, qPCR and nasopharyngeal swab (**Table 2**).

The most repeated terms in the studies are COVID-19 detection, nasopharyngeal swab and PCR. The distribution of published articles by countries, university departments and number of documents is given in **Table 3**. Accordingly, USA, China and UK are the countries with the most published studies. The list of journals with the highest number of publications is given in **Table 4**. Journal of

Table 1. Clusters of research fields (from MeSH data)

Subject	Cluster	Links	Total Link Strength	Occurence
betacoronavirus	1	312	17695	2290
viral pneumonia	1	313	17186	2253
viral RNA	1	305	8382	1257
middle aged	1	292	8709	1163
reverse transcriptase polymerase chain reaction	1	243	3636	501
viral antibodies	1	272	4371	484
80 and over aged	1	254	3710	431
animals	2	308	10640	1488
angiotensin-converting enzyme 2	2	283	6833	827
peptidyl-dipeptidase-a	2	266	4177	455
virus replication	2	256	2691	342
host-pathogen interactions	2	275	2693	329
lung	2	266	2669	319
chlorocebus aethiops	2	240	2547	264
vero cells	2	237	2471	254
virus receptors	2	234	2277	232
serine endopeptidases	2	204	1792	230
cell line	2	233	1862	223
cytokines	2	240	1593	221
spike glycoprotein, coronavirus	3	298	7829	921
antiviral agents	3	289	3861	503
protein binding	3	224	2786	298
viral proteins	3	243	2056	276
amino acid sequence	3	207	2043	212
phylogeny	4	263	4003	558
mutation	4	277	3712	551
molecular evolution	4	209	1811	243
genetic variation	4	228	1557	216
chiroptera	4	181	1494	171
covid-19 vaccines	5	266	3483	460
antibodies, neutralizing	5	220	2587	273
viral vaccines	5	224	1962	221
viral antigens	5	188	907	99
monoclonal antibodies	5	151	908	90

Table 2. Extracted Items from the titles and abstracts

Label	Cluster	Links	Total Link Strength	Occurence
detection	3	1873	9739	713
assay	3	1642	6417	436
RT PCR	3	1568	5769	382
specimen	3	1175	3178	202
nasopharyngeal swab	3	1035	2969	189
binding	1	1204	2750	237
receptor binding domain	1	1000	2671	231
angiotensin	2	1006	2366	209
rbd (receptor-binding domain)	1	921	2341	191
reaction	3	925	2276	132
phylogenetic analysis	1	938	2162	200
mouse	2	926	2058	172
tmprss2	2	889	2035	182
RT qPCR	3	771	1917	126
s protein	1	881	1880	157
inf (interferon)	2	890	1712	122
bat	1	739	1706	146
woman	4	773	1639	109
adult	4	799	1544	105
epitope	1	642	1430	131
isothermal amplification	3	427	1398	72
reagent	3	537	1330	90
n gene	3	588	1306	81
reverse transcription polymerase chain reaction	3	675	1271	90
kidney	2	660	1247	92

Table 3. Distribution of molecular based COVID-19 studies by country and institution

Institution	Number of Documents	Country
Department of Microbiology, Icahn School of Medicine At Mount Sinai	22	USA
University of Chinese Academy of Sciences	15	China
Departmentof Zoology, University of Oxford	11	UK
Ihu-Méditerranée Infection	11	France
Chan Zuckerberg Biohub	10	USA
Department of Immunology, School of Medicine, Tehran Universityof Medical Sciences	10	Iran
Department of Infectious Diseases and Pathobiology, Vetsuisse Faculty, University of Bern	10	Switzerland
Institute of Evolutionary Biology, University of Edinburgh	10	UK
Laboratório De Flavivírus, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz	9	Brazil
University of Chinese Academy of Sciences	9	China
Africa Health Research Institute, Durban, South Africa	8	South Africa
Broad Institute of Mit and Harvard	8	USA
Department of Biochemistry and Molecular Biology, University of Texas Medical Branch	8	USA

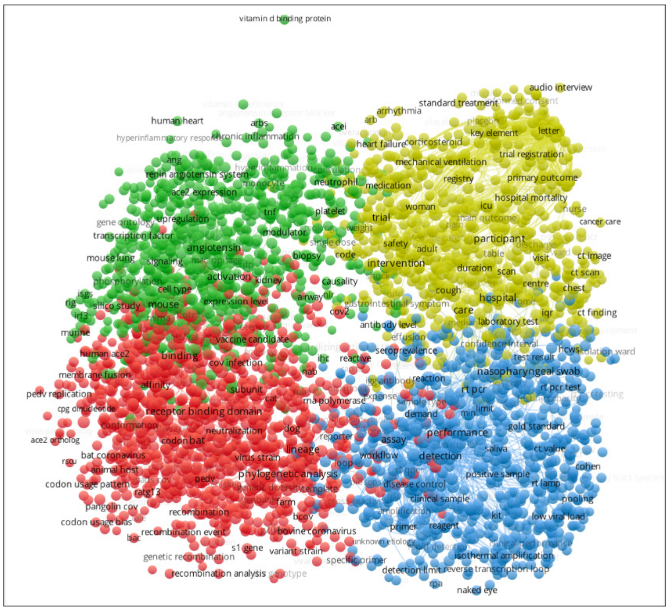

Figure 3. Clusters from title and abstract **extraction**

Medical Virology, Viruses, PLoS One, Scientific Reports, and Nature were found to be the journals with the most publications in 2020, 2021 and in total. Lastly, the list of journals which published molecular-based COVID-19 studies by year is found in **Table 5.**

Table 4. List of journals which published molecular based COVID-19 studies by years

Journal Name	2020	2021	Total
Journal of Medical Virology	95	104	199
Viruses	83	94	177
PLoS One	98	71	169
Scientific Report	70	80	150
Nature	61	56	117
Nature Communications	50	50	100
Frontiers in Immunology	52	32	84
International Journal of Molecular Sciences	52	23	75
Emerging Microbes & Infection	58	17	75
International Journal of Infectious Diseases	41	34	75
Infection, Genetics and Evolution	46	23	69
Science	38	31	69
Cell	36	26	62
Proceedings of the National Academy of Sciences	29	29	58
Journal of Clinical Virology	48	8	56
Journal of Virology	33	14	47
Signal Transduction and Targeted Therapy	32	19	51
Journal of Clinical Microbiology	41	6	47
BMC Infectious Diseases	30	17	47
Medical Hypotheses	34	12	46
Virus Research	29	14	43
Eurosurveillance	32	10	42
Emerging Infectious Diseases	30	12	42
Clinical Microbiology and Infection	25	15	40
Journal of Virological Methods	13	25	38
Archives of Virology	24	12	36
PLoS Pathogens	10	25	35
The New England Journal of Medicine	19	15	34
Nature Medicine	25	9	34
JAMA	17	17	34
The journal of Infectious Diseases	23	11	34
Biochem Biophys Res Commun	8	24	32
Genes (Basel)	19	11	30
Cell Host Microbe	15	15	30
Frontiers in Cellular and Infection Microbiology	20	9	29
British Medical Journal	4	23	27

Table 5. Distribution of molecular based COVID-19 studies by countries

#	Country	Study number
1	USA	2816
2	China	2267
3	Italy	637
4	Germany	353
5	France	342
6	United Kingdom	263
7	India	210
8	Canada	116
9	Japan	113
10	Spain	107
11	Korea	98
12	Switzerland	92
13	Netherlands	86
14	Australia	88
15	Saudi Arabia	72
16	Israel	57
17	Brazil	53
18	Denmark	43
19	Turkey	37
20	South Africa	37
21	Vietnam	37
22	Poland	31
23	Sweden	30
24	Pakistan	30
25	Russia	22
26	Finland	16

Note: In studies with multiple authors, if the countries in which the institutions are located are not the same, the study is equally distributed to the countries of the authors.

DISCUSSION

COVID-19 is a dangerous infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), for which we have no curative treatment (10). COVID-19 first emerged in late December 2019 in the Chinese province of Wuhan and soon spread globally around the world (11). It led to the declaration of the COVID-19 pandemic by the WHO. As of August 2021, it has caused approximately 200 million confirmed cases and approximately 4 million deaths (1). Due to the current number of cases, death rates and worldwide prevalence, it has become the pandemic of the century and is a major problem affecting humanity worldwide. In the last 1.5 years, scientific studies in almost every discipline around the world have focused on COVID-19. In this process, many studies were carried out about the molecular structure, biology and molecular mechanisms of COVID-19 (12). It is critical to elucidate the molecular structure and mechanism of action of COVID-19 to find a curative treatment and prevent the disease. To the best of our knowledge this current study is one of the most comprehensive studies in the literature.

In line with the data obtained from our study, we determined that "betacoronaviruses" are mostly investigated in molecular studies. In their study on betacoronavirus, Letko et al. (13) confirmed that lineage B betacoronaviruses can enter human cells through an unknown receptor and that human ACE2 is the receptor for SARS-CoV-2. Li et al. (14) suggested that betacoronaviruses may have a much more complex recombination mechanism than our current knowledge.

In molecular studies about COVID-19, "viral pneumonia" is the second most common topic. Tianyu et al. (15) showed that Xuebijing agent inhibits COVID-19 and reduces lung involvement by acting on the AKT1 pathway, a serine-threonine protein kinase protein that is effective in the inflammatory response. In another study, it was shown that when the damage-associated molecular pattern (DAMPs) from the coronavirus is combined with other risk factors such as air pollution, smoking or advanced age, the disease progresses more seriously and causes fatal coronavirus pneumonia (16).

Regarding viral RNAs, which is the third most common molecular study subject, Zhang et al. (17) showed that CoV nonstructural protein 14 (nsp14) functions as (guanine-N7)-methyltransferase (N7-MTase) involved in RNA cap formation. They suggested that it would be an ideal method for designing live attenuated vaccines for coronavirus by eliminating the viral RNA N7-MTase activity. In the study by Jesus et al. (18), they suggested that antisense RNA-mediated gene editing would increase the success of treatment and provide a cost-effective approach to treat COVID-19.

Another leading research topic is the reverse transcriptase polymerase chain reaction method used in the diagnosis of COVID-19. A common mutation in the spike protein of SARS-CoV-2, called D614G (A23403G), is known to occur (19). Al-Jaf et al. (20) reported that the qRT-PCR method is a suitable diagnostic method for the detection of this mutation because it is fast, effective and cost-effective. In a meta-analysis study by Sopp et al. (21) with COVID-19 data, they reported that SARS-CoV-2 RNA tested by qRT-PCR was rarely found in conjunctival samples. In a review article describing the production and distribution of mRNA vaccines in the COVID-19 process, production scales of SARS-CoV-2 RNA vaccines and mRNA vaccine production against new agents were mentioned. In this review, the topicality of the mRNA vaccine was emphasized (22).

While the United States and China are in the top ranking for countries that conducted molecular-based COVID-19 studies, Italy, Germany and France among European countries follow this ranking. Among the reasons for this ranking are the first detection of the virus in China, the population and the budget allocated to research. Our country of Turkey, on the other hand, ranks 19th in terms of published molecular COVID-19 studies, and an increasing number of comprehensive studies were accepted for publication in reputable journals. When molecular-based studies examining COVID-19 are evaluated on an institute basis, the USA, China and England share the top three places. The reason for this may be that there are sufficient devices and equipment on an institutional basis, experienced researchers and sufficient research budgets.

CONCLUSION

It is an obvious fact that clinical and molecular studies conducted during the COVID-19 pandemic will continue after the pandemic. However, it is important to determine in the subjects and areas where the molecular-based studies about COVID-19 are clustered and which mechanisms were investigated. Taking scientific steps according to the evidence-based data obtained will be the most beneficial and rational approach to find curative treatment for COVID-19 infection and to effectively prevent this infection.

ETHICAL DECLARATIONS

Ethics Committee Approval: This current study is a computer based data-mining study. There is no need to obtain ethical committee approval.

Informed Consent: Because of the study design no written informed consent form was obtained from patients.

Referee Evaluation Process: Externally peer-reviewed.

Conflict of Interest Statement: The authors have no conflicts of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

Author Contributions: All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. Jamsheela O. A Study of the correlation between the dates of the first COVID case and the first COVID death of 25 selected countries to know the virulence of the COVID-19 in different tropical conditions. Ethics Med Public Health 2021; 19: 100707.
2. Liu K, Chen Y, Lin R, Han K. Clinical features of COVID-19 in elderly patients: A comparison with young and middle-aged patients. J Infect 2020; 80: e14-e8.
3. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020; 395(10223): 507-13.
4. Wang L, He W, Yu X, et al. Coronavirus disease 2019 in elderly patients: Characteristics and prognostic factors based on 4-week follow-up. J Infect 2020; 80: 639-45.
5. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA 2020.
6. Zhu C, He G, Yin Q, et al. Molecular biology of the SARs-CoV-2 spike protein: A review of current knowledge. J Med Virol 2021; 93: 5729-41.
7. Ozcelik F, Tanoglu A, Guven BB, Keskin U, Kaplan M. Assessment of severity and mortality of COVID-19 with anti-A1 and anti-B IgM isohaemagglutinins, a reflection of the innate immune status. Int J Clin Pract 2021: e14624.

8. Beyazit F, Beyazit Y, Tanoglu A, Haznedaroglu IC. Ankaferd hemostat (ABS) as a potential mucosal topical agent for the management of COVID-19 syndrome based on its PAR-1 inhibitory effect and oestrogen content. Med Hypotheses 2020; 143: 110150.
9. Van Eck NJ, & Waltman, L. Text mining and visualization using VOSviewer. ISSI Newsletter 2011; 7: 50-4.
10. Viruses CSGotICoTo. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 2020; 5: 536-44.
11. Alshammary AF, Al-Sulaiman AM. The journey of SARS-CoV-2 in human hosts: a review of immune responses, immunosuppression, and their consequences. Virulence 2021; 12: 1771-94.
12. James N, Menzies M. Trends in COVID-19 prevalence and mortality: A year in review. Physica D 2021; 425: 132968.
13. Letko M, Marzi A, Munster V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nat Microbiol 2020; 5: 562-9.
14. Li LL, Wang JL, Ma XH, et al. A novel SARS-CoV-2 related coronavirus with complex recombination isolated from bats in Yunnan province, China. Emerg Microbes Infect 2021; 10: 1683-90.
15. Tianyu Z, Liying G. Identifying the molecular targets and mechanisms of xuebijing injection for the treatment of COVID-19 via network parmacology and molecular docking. Bioengineered 2021; 12: 2274-87.
16. Land WG. Role of DAMPs in respiratory virus-induced acute respiratory distress syndrome-with a preliminary reference to SARS-CoV-2 pneumonia. Genes Immun 2021; 22: 141-60.
17. Zhang Z, Liu Q, Sun Y, et al. Live attenuated coronavirus vaccines deficient in N7-Methyltransferase activity induce both humoral and cellular immune responses in mice. Emerg Microbes Infect 2021; 10: 1626-37.
18. de Jesus SF, Santos LI, Rodrigues Neto JF, Vieira TM, Mendes JB, D'angelo MFSV, et al. Therapeutic perceptions in antisense RNA-mediated gene regulation for COVID-19. Gene 2021; 800: 145839.
19. Badua CLDC, Baldo KAT, Medina PMB. Genomic and proteomic mutation landscapes of SARS-CoV-2. J Med Virol 2021; 93: 1702-21.
20. Al-Jaf SMA, Niranji SS, Mahmood ZH. Rapid, inexpensive methods for exploring SARS CoV-2 D614G mutation. Meta Gene 2021; 30: 100950.
21. Sopp NM, Sharda V. An Eye on COVID-19: A Meta-analysis of Positive Conjunctival Reverse Transcriptase-Polymerase Chain Reaction and SARS-CoV-2 Conjunctivitis Prevalence. Optom Vis Sci 2021; 98: 429-36.
22. Tanoğlu EG. Production and distribution of mRNA vaccines: SARS-CoV-2 experience. J Mol Virol Immunol 2020; 1: 27-34.