ORIGINAL RESEARCH
published: 24 May 2021
doi: 10.3389/frma.2021.683212

Profiling COVID-19 Genetic Research: A Data-Driven Study Utilizing Intelligent Bibliometrics

Mengjia Wu[1], *Yi Zhang*[1]*, *Mark Grosser*[2], *Steven Tipper*[2], *Deon Venter*[2], *Hua Lin*[2] *and Jie Lu*[1]

[1]*Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia,* [2]*23Strands, Pyrmont, NSW, Australia*

The COVID-19 pandemic constitutes an ongoing worldwide threat to human society and has caused massive impacts on global public health, the economy and the political landscape. The key to gaining control of the disease lies in understanding the genetics of SARS-CoV-2 and the disease spectrum that follows infection. This study leverages traditional and intelligent bibliometric methods to conduct a multi-dimensional analysis on 5,632 COVID-19 genetic research papers, revealing that 1) the key players include research institutions from the United States, China, Britain and Canada; 2) research topics predominantly focus on virus infection mechanisms, virus testing, gene expression related to the immune reactions and patient clinical manifestation; 3) studies originated from the comparison of SARS-CoV-2 to previous human coronaviruses, following which research directions diverge into the analysis of virus molecular structure and genetics, the human immune response, vaccine development and gene expression related to immune responses; and 4) genes that are frequently highlighted include *ACE2*, *IL6*, *TMPRSS2*, and *TNF*. Emerging genes to the COVID-19 consist of *FURIN*, *CXCL10*, *OAS1*, *OAS2*, *OAS3*, and *ISG15*. This study demonstrates that our suite of novel bibliometric tools could help biomedical researchers follow this rapidly growing field and provide substantial evidence for policymakers' decision-making on science policy and public health administration.

Keywords: COVID-19, bibliometrics, network analytics, genetic research, knowledge discovery

OPEN ACCESS

Edited by:
Koen Jonkers,
European Commission, Luxembourg

Reviewed by:
Peter Kokol,
University of Maribor, Slovenia
Lili Wang,
United Nations University – Maastricht Economic and Social Research Institute on Innovation and Technology, Netherlands

****Correspondence:***
Yi Zhang
Yi.Zhang@uts.edu.au

Specialty section:
This article was submitted to Research Policy and Strategic Management, a section of the journal Frontiers in Research Metrics and Analytics

Received: *20 March 2021*
Accepted: *06 May 2021*
Published: *24 May 2021*

Citation:
Wu M, Zhang Y, Grosser M, Tipper S, Venter D, Lin H and Lu J (2021) Profiling COVID-19 Genetic Research: A Data-Driven Study Utilizing Intelligent Bibliometrics. Front. Res. Metr. Anal. 6:683212. doi: 10.3389/frma.2021.683212

INTRODUCTION

The COVID-19 pandemic has developed into an unprecedented global crisis that impacts daily lives and healthcare services provision of human beings. To stop its spread and efficiently control it, the biomedical research community has responded proactively on multiple fronts, including in the field of genetic research. By deciphering the genetic mechanisms underlying the body's response to SARS-CoV-2 infection, we can arrive at a better understanding of COVID-19 pathogenesis, diagnosis, treatment, and, potentially, prevention, such as the optimization of vaccine development. In practical terms, multiple efforts on COVID-19 genetic research have resulted in a substantial amount of research publications (Chua et al., 2020; Shin et al., 2020; Wrobel et al., 2020; Pairo-Castineira et al., 2021). However, the downside of this productivity is that the quantity of COVID-19 literature proliferates and results in difficulties for researchers in comprehending this field's changing knowledge landscape, particularly regarding the emerging information on various genes involved in COVID-19 response.

Bibliometrics is a subject that deciphers the patterns of scientific activities by quantitively tracking and measuring research activities. Traditional bibliometric approaches exploit statistical models to analyze bibliographic information such as author entities, keywords and citations and measure scientific activities. In the medical domain, such approaches have been successfully examined by multiple studies, such as profiling research landscapes (Thompson and Walker, 2015; Liao et al., 2018; Huang et al., 2019; Zhang L. et al., 2020) and discovering/inferring knowledge associations (Hristovski et al., 2005; Porter et al., 2020; Wu et al., 2021). Nowadays, artificial intelligence (AI) and data science techniques have empowered current bibliometrics with novel capabilities of excavating implicit knowledge and inferring potential knowledge associations from bibliometric data, which we named intelligent bibliometrics (Zhang et al., 2020c).

The intensive growth of COVID-19 publications has triggered considerable attention from the bibliometrics community. As part of those studies, Colavizza et al. (2021) profiled the trending research topics in the early-stage COVID-19 studies to indicate the initial foci of such studies, Chahrour et al. (2020) presented descriptive statistics of publication distribution and raised a call for more observational studies and therapeutic trials, Fry et al. (2020) revealed how COVID-19 has impacted and reshaped the worldwide scientific collaboration landscape, Zhang et al. (2021a) highlighted the disruption and resilience of research topics in coronavirus studies due to the outbreak of COVID-19. Different from other COVID-19 bibliometric studies, the study presented here leverages multiple approaches of intelligent bibliometrics and focuses specifically on the topic of COVID-19 genetic research. It utilizes multiple traditional and novel bibliometric analyses to profile the research landscape of this emerging field. This study particularly addresses the following questions:

1) Who are the key players in COVID-19 genetic research?
2) What research topics are prevalently addressed in COVID-19 genetic studies?
3) How have the foci of genetic studies changed during the COVID-19 crisis?
4) What specific genes are frequently highlighted and which ones are emerging as relatively newly-described entities that may be potentially important in COVID-19 genetic studies?

To answer these questions, we identified 5,632 COVID-19 genetic research articles within PubMed and applied a three-stage analysis, including 1) performing co-occurrence analysis to identify key players and core research topics in the field; 2) employing a topic tracking method named scientific evolutionary pathways (SEP) to trace the changing foci of these research topics over time during the COVID-19 pandemic period; 3) utilizing bio-entity network analytics to identify key genes and emerging genes in COVID-19 genetic research. The results also identify top research institutions in this field and their collaborating patterns, and some potential insights in terms of science, technology and innovation from the rapidly growing body of COVID-19 genetic research.

MATERIALS AND METHODS

The research framework of this study is given in **Figure 1**.

Data Collection

PubMed is a public biomedical literature database developed by the National Library of Medicine (NIH) and comprises over 32 million medical articles and online books.[1,2] Falagas et al. (2008) recommend PubMed as the optimal bibliometric database for medical and life sciences, which exactly coheres with the foci of our study. Its advantage in biomedical information retrieval is providing specialized functions like Medical Subject Heading (MeSH) search and biomedical filters (including the species filter we used in this work) to return precise results. While considering Web of Science (WoS) is also a well-recognized data source in traditional bibliometric studies, we compared the search results of WoS and PubMed using the same search string (the filter and MeSH terms excluded) and noted that 93% of our collected data are indexed by WoS, indicating a wide coverage of PubMed data. Given that circumstances, in our study, we only exploited the PubMed database as our data source.

Using the search strategy below, we collected a dataset of the genetic research performed on COVID-19 and SARS-CoV-2 from PubMed:

> (“"COVID-19/genetics”[MeSH Terms] OR ((“genes”[MeSH Terms] OR “genetics”[MeSH Terms] OR “gene”[All Fields] OR “genes”[All Fields] OR “genome”[All Fields] OR “genetics”[All Fields]) AND (“COVID-19”[All Fields] OR “SARS-Cov-2”[All Fields]))) AND (humans [Filter])
>
> Search date: 08/03/2021

The search yielded 5,632 publications related to COVID-19 genetic research. We restricted the species to human since our primary goal is only to explore the important human genes act in COVID-19 infection.

Key Player Identification and Research Landscape Profiling

Co-occurrence visualization, a type of science map, is a classical and effective way of profiling the research landscape in a certain field (Callon et al., 1983). The co-occurrence of authorship entities like authors, affiliations and countries reflects established collaborations at the individual, institutional or international levels. The co-occurrence of research content information like keywords or scientific terms in the same context indicates similar semantic meanings and their high probability of constituting a research topic. Visualization of

[1] Detailed information on the PubMed database could be found on the website: https://pubmed.ncbi.nlm.nih.gov/

[2] Input data and the results can be assessed through the link: https://github.com/IntelligentBibliometrics/COVID-19-genetic-research. Readers can use VOSViewer to reproduce these interactive maps and explore the detailed connections between a given node and others. The same for **Figure 4**.

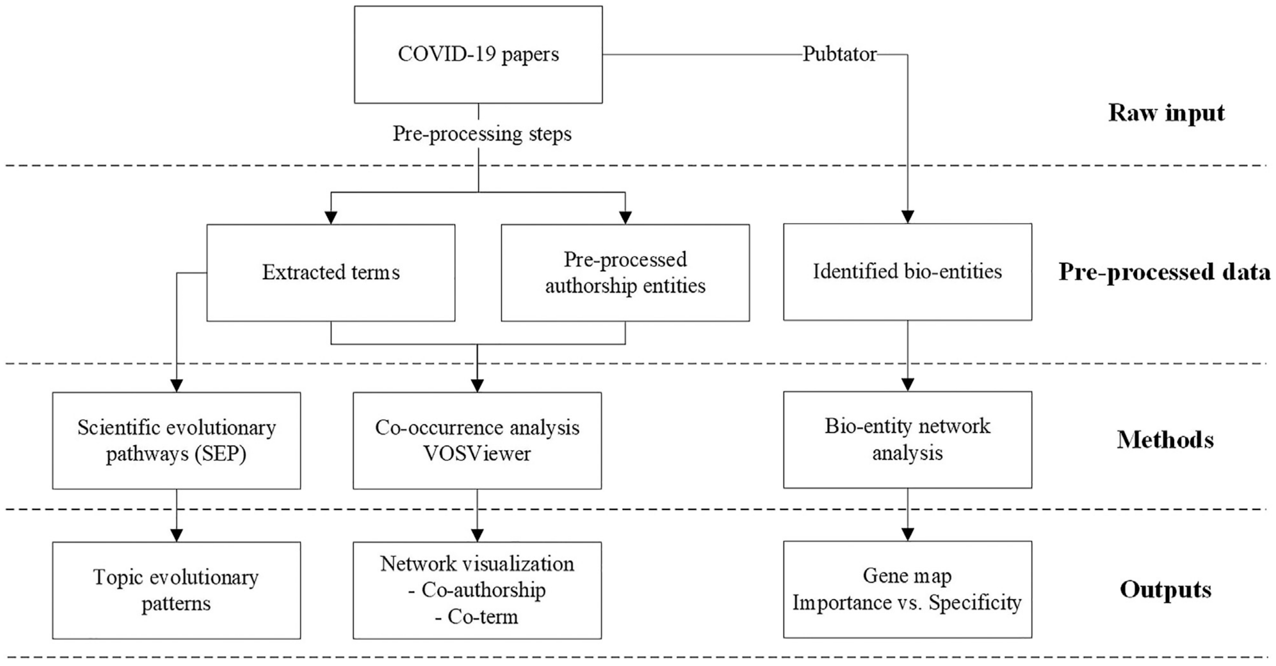
COVID-19 papers
Pubtator
Pre-processing steps
Raw input
Extracted terms
Pre-processed authorship entities
Identified bio-entities
Pre-processed data
Scientific evolutionary pathways (SEP)
Co-occurrence analysis VOSViewer
Bio-entity network analysis
Methods
Topic evolutionary patterns
Network visualization
- Co-authorship
- Co-term
Gene map
Importance vs. Specificity
Outputs

FIGURE 1 | The research framework.

such co-occurrence networks could straightforwardly highlight the key players or research communities in the field and reveal trending research topics in a specific field.

A co-occurrence network could be denoted as $G = (V, E)$ where V represents the set of entities (authorship entities, terms, etc.) and E means the set of co-occurrences between entities. The co-occurrence network could also be represented as an adjacency matrix A with:

$$A_{i,j\,(i\neq j)} = \begin{cases} CF(i,j) & if\ \ entity\ \ i\ \ and\ \ j\ \ co-occur\ \ in\ \ at\ \ least\ \ one\ \ document \\ 0 & otherwise \end{cases}$$

where $CF(i, j)$ means the frequency (i.e., number of publications) that entity i and j co-occur.

In the co-authorship network, considering the scale variation of different author entities, we will normalize $A_{i,j}$ by calculating the Jaccard Coefficient of i and j (Salton and McGill, 1986):

$$A_{i,j\,(i\neq j)} = \begin{cases} \dfrac{CF(i,j)}{F(i) + F(j) - CF(i,j)} & if\ \ entity\ \ i\ \ and\ \ j\ \ co-occur\ \ in\ \ at\ \ least\ \ one\ \ document \\ 0 & otherwise \end{cases}$$

where $F(i)$ means the occurring frequency (i.e., number of publications) of entity i.

VOSViewer (Van Eck and Waltman, 2010) is utilized to visualize the co-occurrence networks of co-term and co-authorship. Note that all the parameters in VOSViewer are following its default settings (e.g., attraction and repulsion for layout, resolution and minimum cluster size for clustering). The local moving algorithm integrated in VOSViewer is simultaneously applied to identify communities in a network, which is reflected as the grouping and coloring strategy for its involved nodes.

Scientific Evolutionary Pathways

Scientific Evolutionary Pathway (SEP), developed by Zhang et al. (2017), is a method of tracking the change of research topics retrieved from a collection of scholarly literature. The design of SEP adopts the assumption that scientific innovation results from the accumulative changes and recombination of existing knowledge (Fleming, 2001; Fleming and Sorenson, 2004). Algorithmically, SEP deems a document a collection of scientific terms and represents the document set as a document-term matrix. By dividing the literature dataset into sequential slices according to their publication gaps (year or month), SEP simulates scientific documents as streaming vectors of terms with a given vocabulary collected from the dataset and 1) extracts research topics via a K-means clustering algorithm, 2) captures semantic drifts of topics to identify topic evolution, and 3) represents topic evolutionary patterns by predecessor-descendant relationships between topics. The stepwise algorithm is given as follows:

Concept definition: A topic is denoted as T and defined as a collection of articles represented by term vectors. Every topic has two attributes: 1) the centroid c, which is represented by the mean vector of all the articles that belong to the topic; and 2) the radius r, which is represented by the largest Euclidean distance of the centroid to every article in T.

Step 1: Initialize the document-term matrix and divide the documents into time slices according to their publishing dates.

Step 2: Group articles in the first time slice (Slice 0) to an initial topic T_i and label it with the top-frequency term, then calculate its radius r and centroid c.

Step 3: For every article a in the next time slice (Slice 1), calculate the article's Euclidean distance $E(a,c)$ between a and the topic's centroid c. If $(E(a,c)-r)/r>\sigma$ (σ is a given threshold which is set as 0.1 by default), it will be recognized as a drifted article to T_i, otherwise the article will be classified as belonging to T_i.

Step 4: Once all articles in Slice 0 have been processed, update the centroid and radius of T_i with these newly added articles. For the drifted articles, use a K-means clustering algorithm to form a new topic set $\{T_n\}$. All topics in $\{T_n\}$ will be deemed as the descendent topics of T_i.

Step 5: For the forthcoming time slices, iterate Steps 3–4 until the last slice of the entire dataset. However, starting from Slice 2, we will measure an article's similarity with not only T_i but all the existing topics (e.g., $\{T_n\}$), and assign the article to the most similar topic based on Salton's cosine similarity.

The outcome of SEP analysis is a set of topics with a time stamp indicating when it was born and with directed links representing their predecessor-descendent relationships.

Bio-Entity Network Analysis

Bio-entity network analysis is a method for literature-based discovery proposed by Wu et al. (2021), which is used to discover and infer genetic knowledge for a specific disease. The method extracts bio-entity concepts from literature and constructs a heterogeneous bio-entity co-occurrence network. Further, it presents the extracted bio-entities comprehensively based on their topological importance and specificities to the target disease. In our case, we select disease, chemical, gene and genetic variant as four representative bio-entities, and the co-occurrence network is denoted as:

$$G=\left(V^{d,c,g,v},E^{pairwise\left(d,c,g,v\right)}\right)$$

where d,c,g,v respectively represent the nodes of diseases, chemicals, genes and genetic variants.

$$w_{V_iV_j\,(i\neq j)}=\begin{cases}CF\left(V_i,V_j\right) & \textit{if } V_i \textit{ and } V_j \textit{ co}-\textit{occur in at least one document}\\ 0 & \textit{otherwise}\end{cases}$$

where $w_{V_iV_j\,(i\neq j)}$ represent the weight of edge linking V_i and V_j.

Three node centralities are employed to measure the importance of individual nodes in the whole network. Referring to the discussion given by Zhang et al. (2021), among the three centralities, *degree centrality (DC)* measures a node's ability to aggregate information, representing its local influence within a network; *closeness centrality (CC)* measures a node's ability to disseminate information, as well as its global influence on all other nodes within a network; *betweenness centrality (BC)* represents a node's ability to act as a bridge between diverse information content, which calculates the proportion of shortest paths going through the node. The three centralities can be calculated as follows:

$$DC(V_i)=\frac{\sum_{k\in\{d,c,g,v\}}^{k}\sum_{j=1}^{|V^k|}\left(1\ \textit{if}\ \ w_{V_iV_j^k\,(i\neq j)}\ \ \textit{else}\ \ 0\right)}{|V^{d,c,g,v}|-1}$$

where V_i is the target node, V_j^k is the jth node in category $k\in\{d,c,g,v\}$, $\left|V^k\right|$ is the number of nodes in the node set V^k, $\left|V^{d,c,g,v}\right|$ is the number of nodes in the total node set.

$$CC(V_i)=\frac{\left|V^{d,c,g,v}\right|-1}{\sum_{k\in\{d,c,g,v\}}^{k}\sum_{j=1}^{|V^k|}d_{V_iV_j^k}}$$

where $d_{V_iV_j^k}$ denotes the shortest topological distance between V_i and V_j^k.

$$BC(V_i)=\frac{2\sum_{x,y\in\{d,c,g,v\}}^{x,y}\sum_{a=1}^{|V^x|}\sum_{b=1}^{|V^y|}\frac{\sigma\left(V_a^xV_b^y\right)_{V_i}}{\sigma\left(V_a^xV_b^y\right)}}{(|V^{d,c,g,v}|-1)(|V^{d,c,g,v}|-2)}\left(V_i\neq V_a^x\neq V_b^y\right)$$

where $\sigma(V_a^xV_b^y)$ denotes the number of shortest paths between V_a^x and V_b^y, $\sigma(V_a^xV_b^y)_{V_i}$ is the number of shortest paths between V_a^x and V_b^y that pass through V_i.

Apart from the centralities that reflect a node's importance, the method also utilizes another indicator named intersection ratio to measure whether a gene is specific to the target disease. The intersection ratio is calculated as follows:

$$IR(V_i)=\frac{w_{V_iV_d^t}}{\sum_{j=1}^{V^d}w_{V_iV_j^d}}$$

where V_d^t is the target disease, V^d is the set of disease nodes.

Three centralities together constitute the topological importance of a node in a network. To comprehensively rank this importance, we exploit an entropy-based algorithm (Grupp, 1990) to combine those centralities in a data-driven way. Briefly, we initially normalize the values of each centrality using the min-max normalization approach and then calculate the entropy of each normalized centrality and get its weight, respectively. The combined centrality of every individual node can eventually be finalized with those data-driven weights. Using DC, the value set of the degree centrality, as an example, we calculate the weight of DC as follows:

Step 1: Normalize each centrality value set:

$$nDC(V_i)=\frac{DC(V_i)-\min(DC)}{\max(DC)-\min(DC)}$$

where DC is the value set of the degree centrality, $\max(DC)$ and $\min(DC)$ is respectively the maximum and minimum value in this set.

Step 2: Calculate the entropy of each centrality:

$$H_{DC}=\frac{1}{\ln|\mathrm{V}|}\sum_{i}^{|\mathrm{V}|}nDC(V_i)\ln(nDC(V_i))$$

where V is the collection of all nodes in the given network.

Step 3: Calculate the weight of each centrality according to the entropy, when we get the entropy of all three centralities:

$$w_{DC}\ \ =\frac{1-H_{DC}}{3-\ (H_{DC}+H_{BC}+H_{CC})}\ \ (0\leq w_{DC}\leq 1)$$

Step 4: Finalize the centrality combination value for every individual node:

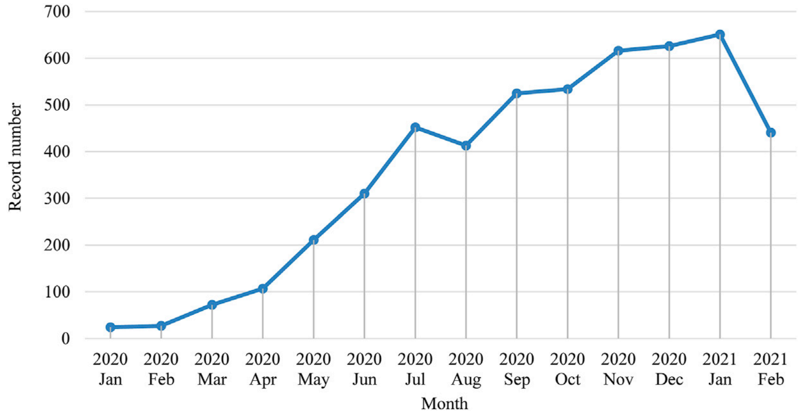

FIGURE 2 | Monthly trend of the number of publications.

TABLE 1 | Top 20 prolific countries, research institutions and journals in this emerging field.

Ranking	Country	Research Institution	Journal
1	United States (1811)	University of California (183)—United States	Journal of Medical Virology (150)
2	China (1,030)	University of Texas (104)—United States	PLoS One (130)
3	United Kingdom (560)	University of Hong Kong (93)—China	Scientific Reports (100)
4	Italy (534)	University of Oxford (92)—United Kingdom	Nature (88)
5	Germany (359)	Wuhan University (76)—China	Viruses (82)
6	India (320)	University of Washington (76)—United States	Nature Communication (70)
7	France (299)	University of Pennsylvania (71)—United States	Frontiers in Immunology (70)
8	Canada (274)	Stanford University (63)—United States	International Journal of Infectious Diseases (68)
9	Spain (214)	University College London (59)—United Kingdom	Science (67)
10	Australia (196)	University of Cambridge (58)—United Kingdom	Emerging Microbes & Infections (64)
11	Brazil (173)	Massachusetts General Hospital (58)—United States	Journal of Clinical Virology (62)
12	Japan (150)	University of Chinese Academy of Sciences (55)—China	Medical Hypotheses (62)
13	Switzerland (141)	Tongji Hospital (54)—China	International Journal of Molecular Sciences (57)
14	Iran (140)	Columbia University (52)—United States	Cell (50)
15	South Korea (115)	University of Toronto (52)—Canada	Journal of Clinical Microbiology (46)
16	Belgium (105)	University of Edinburgh (51)—United Kingdom	Infection, Genetics and evolution (45)
17	Turkey (98)	University of Milan (51)—Italy	Signal Transduction and Targeted Therapy (44)
18	Sweden (88)	Brigham and Women's Hospital (51)—United States	Proceedings of the National Academy of Sciences of the United States of America (42)
19	Saudi Arabia (67)	Peking Union Medical College (50)—China	BMC Infectious Diseases (41)
20	Austria (67)	Fudan University (48)—China	Eurosurveillance (39)

$$Centralities\ \ combination\left(V_i\right) = w_{DC}nDC\left(V_i\right) + w_{CC}nCC\left(V_i\right) + w_{BC}nBC\left(V_i\right)$$

RESULTS

To conduct the analyses above in a systematic manner, we utilized a one-stop platform that integrates multiple bibliometric methods developed in our pilot studies (Zhang et al., 2020d; Wu et al., 2020). This platform contains six modules, including PubMed/PMC data import filter, bibliographic information statistics, co-occurrence analysis, research collaboration prediction, semantic similarity-based document search and SEP analysis. By feeding the raw data from PubMed into our bibliometric platform, we could efficiently generate results and then present these graphically with the aid of multiple visualizing tools, such as Gephi and VOSViewer.

Key Players and Topics Identification

Apart from 623 publications that either miss publication month information or published after February 2021, we presented the monthly change of the numbers of publications in **Figure 2**. We could see a clear overall blooming trend of publications since January 2020. The downward trend starting in August 2020 may indicate that the stage shifts of COVID pandemic status has an influence on related academic research.

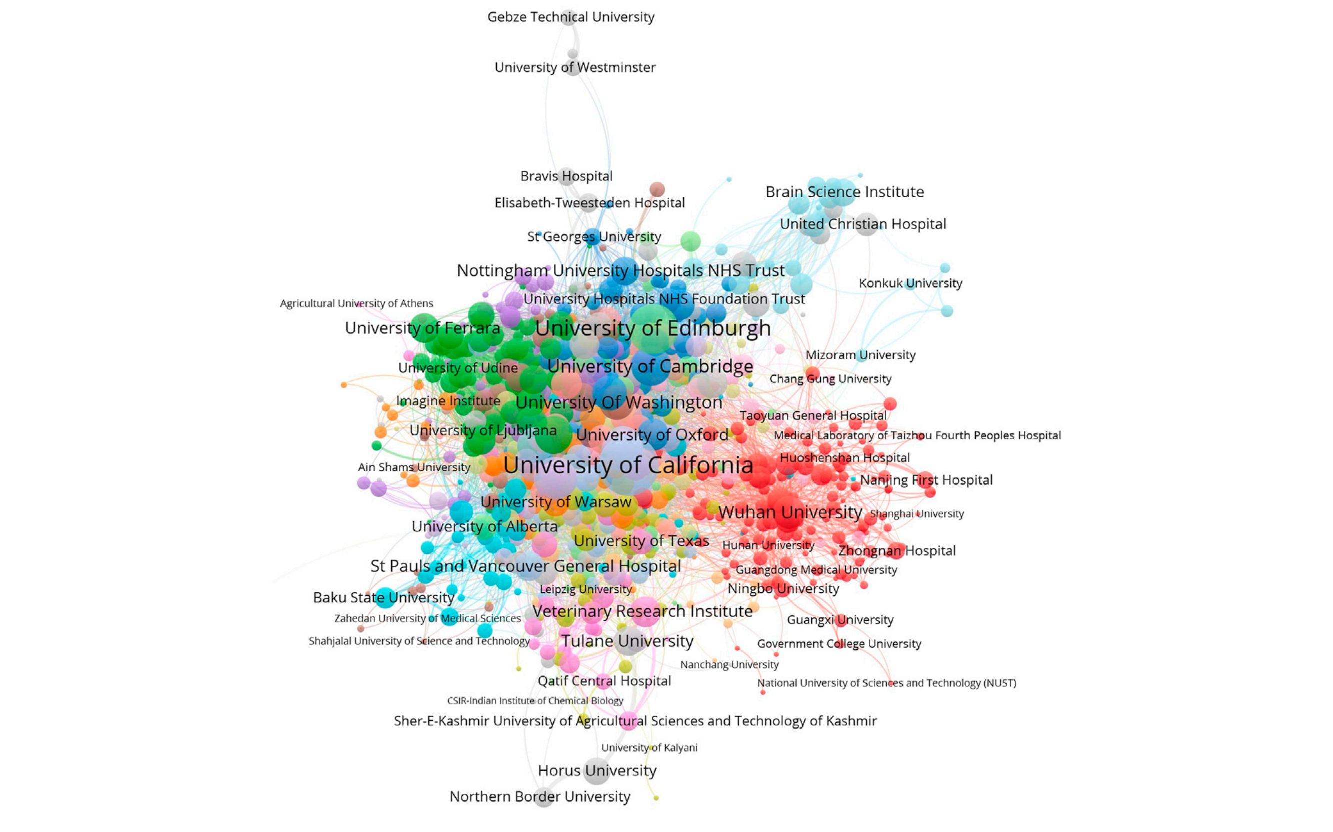

FIGURE 3 | The co-authorship network of research institutions (normalized by Jaccard Coefficient).[2]

We implemented a set of pre-processing and disambiguation approaches to clean affiliation names and terms retrieved from combined titles and abstracts. Ranking by the total number of publications, we present the top 20 countries, research institutions and journal names in **Table 1**. At the country level, the United States and China are still leading this domain, followed by the United Kingdom, Italy, Germany, etc. From the institutional perspective, the top 20 most prolific institutions consist of eight institutions from the United States, six from China, four from the United Kingdom, and one each from Canada and Italy (the other major pandemic center in early 2020, besides China). Interestingly, compared to China's leadership in pioneering general COVID-19 research at the early stage of the COVID-19 pandemic (Fry et al., 2020), prestigious universities from the United States and United Kingdom have become leaders in this domain, particularly in COVID-19 genetic research. The journal distribution of COVID-19 genetic research reflects that virology journals published the largest number of papers, and prestigious journals such as *Nature*, *Science*, *Cell* and *PNAS* also show an inclined favor for such studies.

Then we conducted co-occurrence analysis to cleaned affiliations and terms and generated a co-authorship network and a co-term network. With the aid of VOSviewer, the visualizations of the co-authorship network and word co-occurrence network are shown in **Figures 3**, **4**.

As shown in **Figure 3**, we can observe global collaborating patterns on COVID-19 genetic research. When the size of each node is measured by the sum of Jaccard Coefficients that the represented institution has with all its collaborators, it then reflects its representing institution's collaborative strength in this field. Located at the center of this network, United States universities, such as the University of California systems and the University of Washington, have established broad and close international collaborations with worldwide institutions during the crisis. Research institutions from Europe and the United Kingdom share intensive partnerships with United States institutions. Comparably, Chinese institutions show strong domestic collaborations but have relatively sparse links with their international counterparts. Political issues might be a hidden reason behind this phenomenon, but the national border restrictions may play a crucial role in limiting collaborations within relatively small but well-established groups (Fry et al., 2020; Cai et al., 2021).

In **Figure 4**, the co-term network is partitioned into four communities with different colors: Virus infection mechanisms (red), virus testing (yellow), gene expression related mainly to the immune response to COVID-19 (blue) and COVID-19 clinical manifestation (green).

The network shows the relatedness and relationships within and between these four main areas of scientific efforts. The red community mostly covers molecular concepts related to the

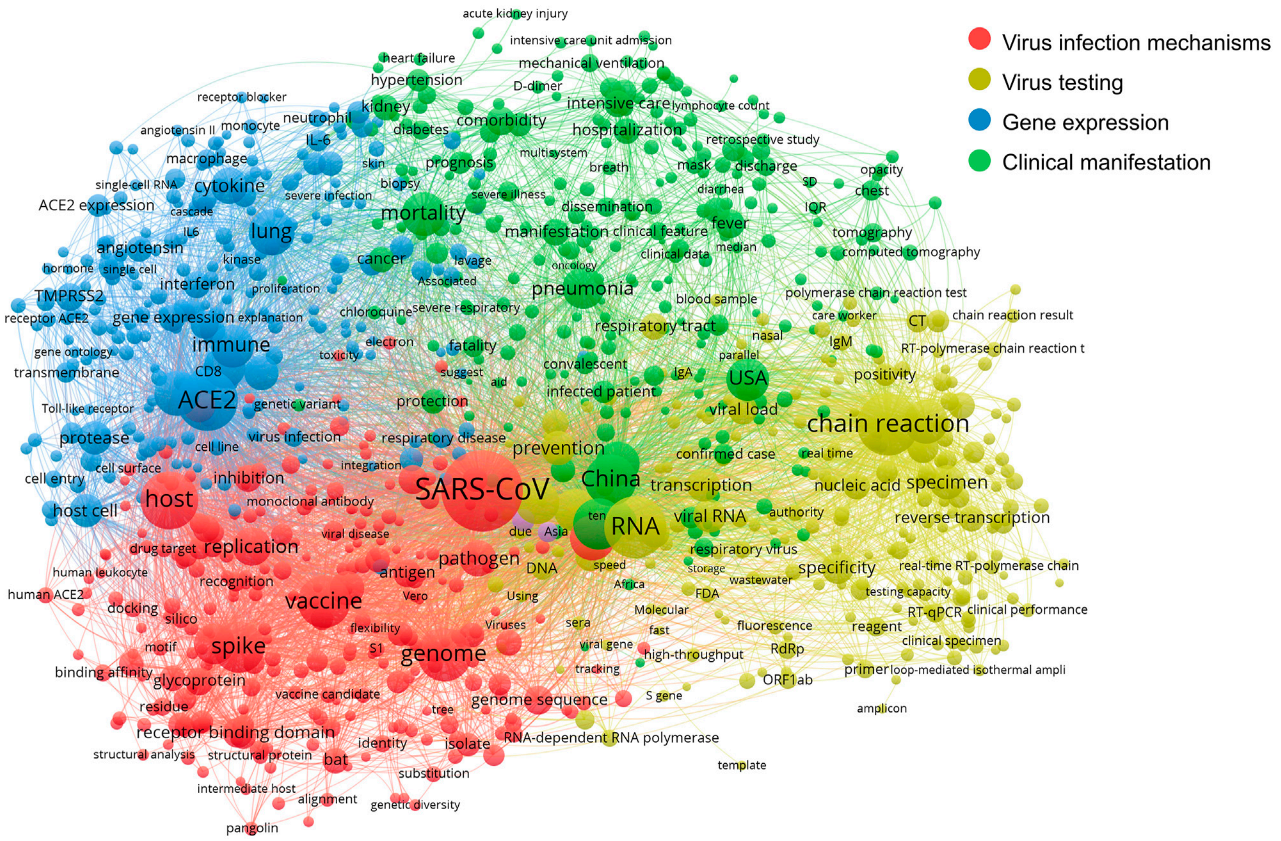

FIGURE 4 | The co-occurrence network of scientific terms.

mechanism of virus infection, like the virus *spike protein S1*, *antigen*, *pathogen*, *host* and *vaccine*. The yellow community groups terms related to testing techniques such as *RNA*, *nucleic acid*, *RT-qPCR* and *chain reaction*. The blue community indicates gene expression, explicitly demonstrating genes and proteins involved in this broad area of research, including immune system cell types (macrophage; neutrophil), cell components (cell surface), cell surface markers (CD8), and immune response genes (IL6; interferon). Lastly, terms in the green community mostly describe the clinical phenotypes and manifestation of COVID-19 and epidemiological information, including *China*, the *United States* and *mortality*.

Topic Evolution

The SEP approach was applied to the 5,166 publications before April 2021. With the aid of Gephi (Bastian et al., 2009), we visualized the SEP on COVID-19 genetic research between January 2020 and April 2021 and present it in **Figure 5**. In this figure, 85 nodes are generated and linked by directed edges, representing their predecessor-descendant relationships, i.e., evolutionary patterns. Every topic name is followed by a bracketed time label, indicating when the topic was born. The descriptive statistics of the SEP is given in **Table 2**. We further applied an approach of community detection integrated in Gephi to group these topics into communities (with colors).

As given in **Figure 5**, during the 15 months, topics on COVID-19 genetic research have evolved into five communities, that is, #1 virus molecular structure, #2 virus testing, #3 human immune response, #4 vaccine development, and #5 gene expression.

Tracing back to the early stage of COVID-19 genetic research, most of the studies focus on the genetic similarity of COVID-19 with previous prevalent coronaviruses like severe acute respiratory syndrome coronavirus (*SARS-CoV*) and Middle East respiratory syndrome coronavirus (*MERS-CoV*). With this as a starting point, community #1 (the green branch) maps research on the molecular structure of viruses, including its probable source of origin (*bat*), *genome sequence*, how the virus invades the *host cell* and how the *spike protein* acts during the infection process, etc. Community #2 (the purple branch) details another research direction of COVID-19 testing and screening and covers a wide variety of topics like clinical manifestations (*fever*, *cough*, *cell count*), testing method (*viral RNA* testing) and techniques (*reverse transcription-polymerase chain reaction test*, which is regarded as the "gold standard" of virus RNA testing), test *sensitivity*, etc. This community reflects that genetic research on virus testing gathers substantial attention in the mid-term research of COVID-19. Community #3 (the red branch) on the right side concentrates on the body reaction of humans to COVID-19 infection. This community consists of concepts detailing the clinical manifestations and immune

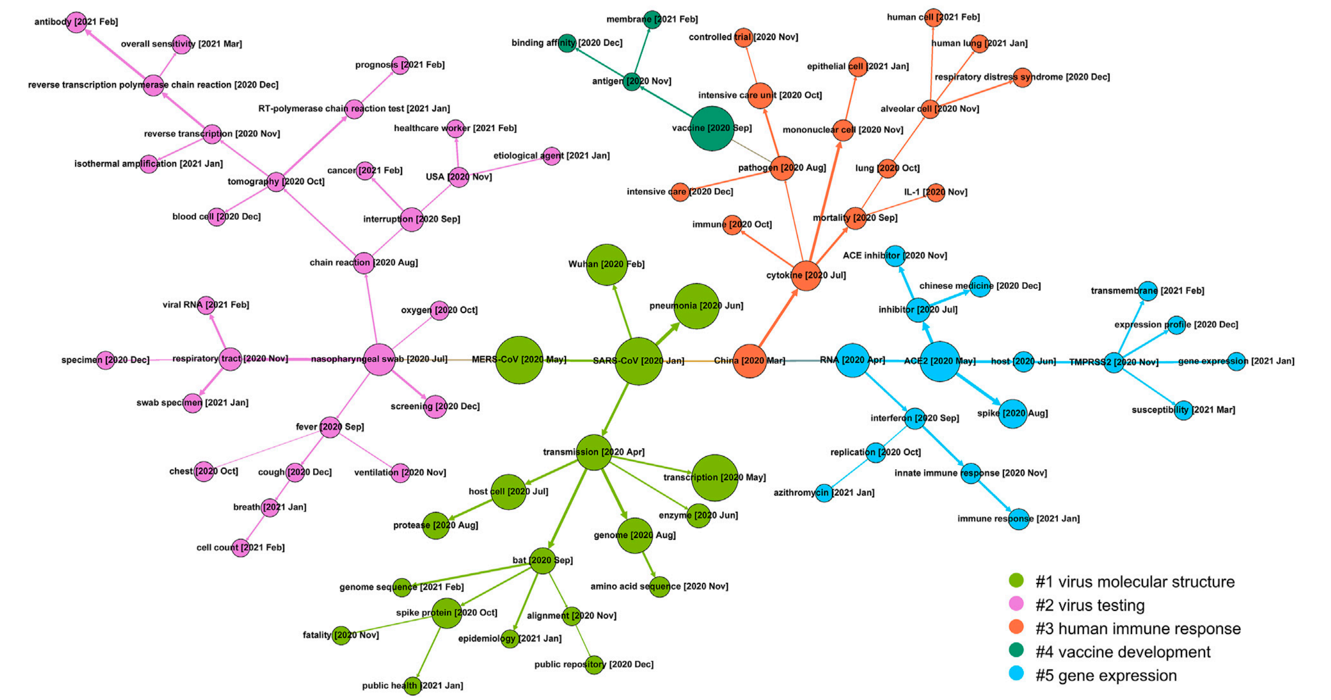

FIGURE 5 | The SEP on COVID-19 genetic research between January 2020 and April 2021.

TABLE 2 | The basic statistics of 85 topics.

Node number	Maximum publication number[a]	Minimum publication number[a]	Standard deviation
85	321 MERS-CoV [2020 January]—321 SARS-CoV [2020 January]—320 Transcription [2020 May]—309 Pneumonia [2020 June]—208	3 Etiological agent [2021 January]—3 Human lung [2021 January]—3 Cell count [2021 February]—3 Healthcare worker [2021 February]—3	84.374

[a]*We list the top four topics that contain the largest/smallest numbers of publications in the table.*

reactions in human cells. For example, the *cytokine* storm, which is postulated to be one of the major reasons for severe disease progression, is the result of an overreaction of the human immune system. Community #4 *vaccine* development is derived from the branch of immunological research and becomes a relatively outlying research focus. Along with this timeline, the specific process within human cells is also frequently studied as presented in community #5 gene expression, where *ACE2* and *TMPRSS2*, two of the key genes involved in the viral infection process, are highlighted.

Bio-Entity Network Analysis

Entity Extraction and Pre-processing

We utilized Pubtator (Wei et al., 2019), a deep learning-based entity extraction tool developed by the National Library of Science (NLM), to extract biomedical concepts from titles and abstracts of the collected research papers. The raw output from Pubtator is lists of biomedical concepts with exclusive identifiers. We then employed multiple biomedical dictionaries, including Medical Subject Headings,[3] NCBI gene dictionary of *Homo sapiens*,[4] as well as a single nucleotide polymorphism (SNP) dictionary from the dbSNP database,[5] to map those concepts to unified bio-entities.

The extraction process resulted in 48,201 raw biomedical concepts, including diseases, chemicals, genes and genetic variants. We then mapped every concept to its related dictionary and applied two cleaning steps to remove noise (Step 1) and consolidate synonyms (Step 2). 2,573 unique bio-entities were retrieved after the cleaning steps, with the stepwise results presented in **Table 3**. The 2,573 bio-entities were then used to construct a heterogeneous co-occurrence network.

[3] https://www.ncbi.nlm.nih.gov/mesh/

[4] https://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz

[5] https://www.ncbi.nlm.nih.gov/snp/

TABLE 3 | Stepwise results of the pre-processing procedure.

	Raw	Step 1	Cleaned	Step 2	Nodes
Disease	31,974	Removed noisy concepts like "cardioembolic", "JAGS", "nonvitamin", etc. that could not be mapped to MeSH	31,963	MeSH	801
Chemical	4,494		3,724		678
Gene	11,211	Excluded genes that do not belong to *Homo-sapiens*	8,781	NCBI Gene	968
Gene variant					
DNA mutation	69	Removed variants with unclear loci (i.e., could not be mapped to an SNP ID)	17	dbSNP	126
Protein mutation	349		91		
SNP	104	—	104		
Total	48,201	—	44,680	—	2,573

TABLE 4 | Counts of the different types of edges.

	Disease	Chemical	Gene	Genetic variant
Disease	8,231	4,872	6,966	499
Chemical	4,872	2,121	2,268	37
Gene	6,966	2,268	5,692	385
Genetic Variant	499	37	385	777

The co-occurrence network of the 2,573 entities contains 31,848 edges. The counts of different types of edges are given in **Table 4**.

Profiling Bio-Entities

Based on the frequencies of those identified bio-entities, we list the top 10 highly frequent bio-entities in **Table 5**. The monthly changes of those entities regarding their frequency are provided in **Figure 6**. Genetic variation is not given due to the relatively small amount of data in this category. We only traced the data to January 2021 since the latter collection is incomplete due to the publishing lags.

In general, the frequencies of the top 10 entities in each category keep increasing during the pandemic. Predictably, ACE2 is the top mentioned gene with a noticeable frequency gap with the following genes in **Figure 6A**. This is mainly because ACE2 is the primary functional receptor for the SARS-CoV-2 virus in human cells (Zheng Y.-Y. et al., 2020). Next, **Figure 6B** profiles co-morbidities discussed in COVID-19 genetic studies, besides those symptoms and COVID-19 manifestations, such as cough, respiratory distress syndrome and inflammation, key co-morbidities in COVID-19 genetic studies include neoplasms, diabetes and hypertension. Noteworthy, **Figure 6C** presents multiple prevalent drug treatments that were utilized and trialed for COVID-19, including hydroxychloroquine, azithromycin, remdesivir and the lopinavir-ritonavir drug combination, indicating that the pharmacogenomics of those drugs is also a major interest of this domain.

Emerging Gene Discovery

We applied the approach of bio-entity network analytics and attributed each node of genes with two indicators: centrality combination and intersection ratio. We then normalized the two indicators as X-axis Intersection Ratio and Y-axis Centrality Combination and located all genes in a coordinate system in **Figure 7**. According to our design, a high value of centrality combination may indicate the importance/impact of a given gene to relatively broad domains of the target disease, while a high value of intersection ratio may represent the specialty of a given gene to the target disease.

When COVID-19 is the target disease of our study, the centrality combination indicates the strength of a gene's contribution to broad COVID-19 related concepts, while the intersection ratio reveals the strength of a gene's specific relationship to COVID-19. From the perspective of centrality combination, we could identify *ACE2*, *IL6*, *TMPRSS2*, and *TNF* as the set of frequently highlighted genes. ACE2 is the major functional receptor for the SARS-CoV-2 virus (Zheng Y.-Y. et al., 2020). TMPRSS2 is an enzyme that primes the spike S

TABLE 5 | The top 10 entities ranked by the raw frequency.

Ranking	Disease	Chemical	Gene	Genetic variant
1	Death	Oxygen	*ACE2*	rs2285666
2	Pneumonia	Hydroxychloroquine	*TMPRSS2*	rs12329760
3	Inflammation	Remdesivir	*IL6*	rs4646116
4	Fever	Serine	*CRP*	rs11385942
5	Neoplasms	Chloroquine	*TNF*	rs12252
6	Respiratory distress syndrome, adult	Lipids	*CD4*	rs1244687367
7	Cough	Azithromycin	*ACE*	rs143936283
8	Diabetes mellitus	Lopinavir-ritonavir drug combination	*CD8A*	rs73635825
9	Hypertension	Nitrogen	*IFNG*	rs8176746
10	Zoonoses	Aldosterone	*FURIN*	rs8176719

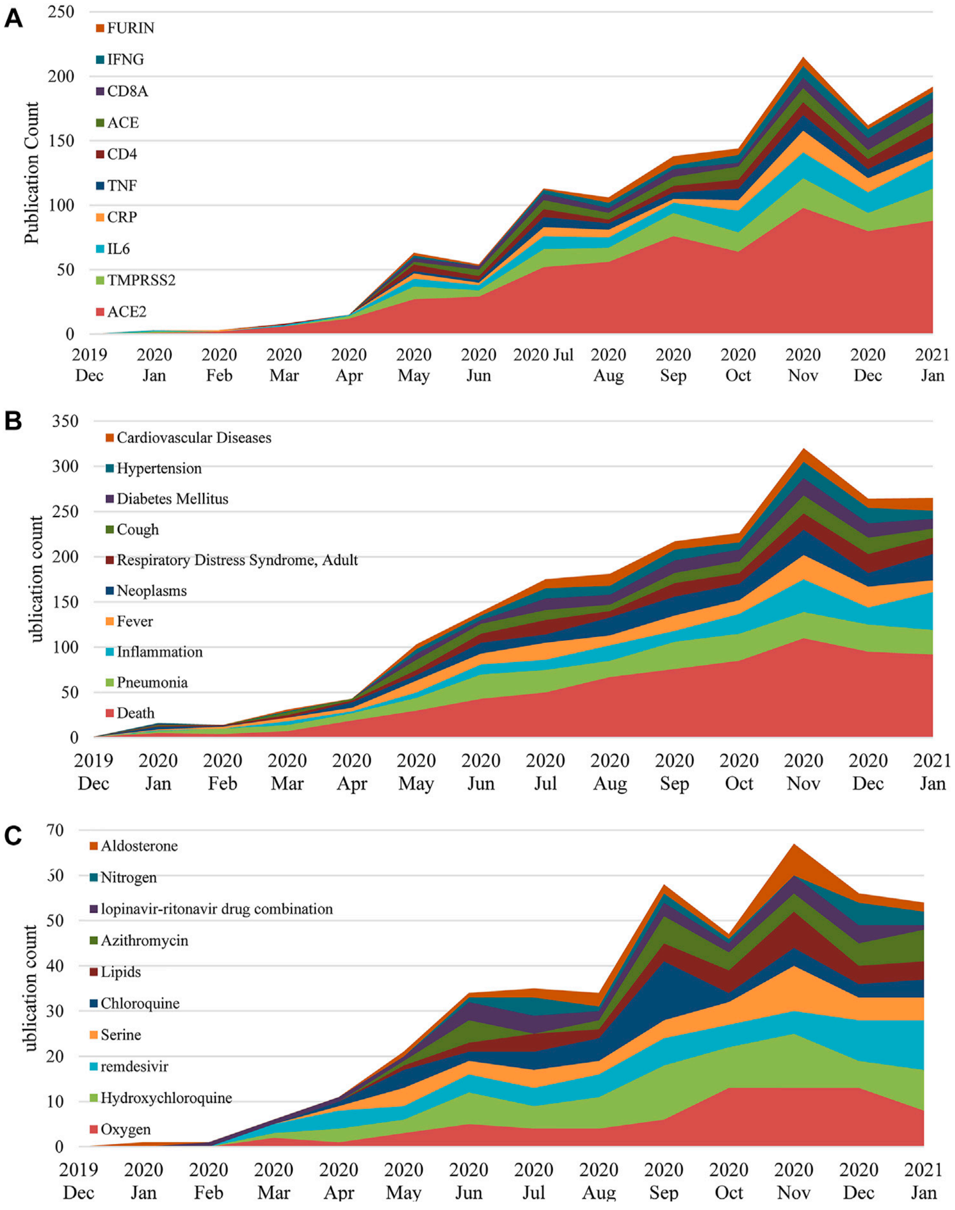

FIGURE 6 | Top 10 genes **(A)**, co-morbidities **(B)** and chemicals **(C)** in COVID-19 genetic research between December 2019 and January 2021.

protein of the SARS-CoV-2 virus to promote virus entry (Hoffmann et al., 2020b), IL6 and TNF are pro-inflammatory cytokines that are found generally elevated in severe COVID-19 patients (Cao, 2020). From the perspective of intersection ratio, we present **Figure 8** to zoom into those emerging genes.

Figure 8 zooms in to some recently mentioned genes in the literature, which own both high intersection ratio and centrality combination in the first quadrant of **Figure 7**. The potential role of these "emerging" genes and their products in COVID-19 is discussed below:

FURIN: FURIN is an essential cleavage enzyme for the spike protein of SARS-CoV-2 in the virus infection process. From the biochemical perspective, Klimstra et al. (2020) identified the association between a putative furin cleavage signal generated by a novel insertion of the SARS-CoV-2 spike S glycoprotein and the expanded host range. Wrobel et al. (2020) discovered that the cleavage at the furin-cleavage site decreases the overall stability of SARS-CoV-2 S and facilitates the adoption of the open conformation required for the viral S (spike) protein to bind to the ACE2 receptor. From the treatment perspective, Hoffmann et al. (2020a) highlighted that obtaining a S1/S2 multibasic cleavage site was essential for COVID-19 infection and indicated furin as a potential target for therapeutic intervention. A similar finding was also presented by Sallenave and Guillot (2020), whose study identified a furin-like cleavage site in SARS-CoV-2 to facilitate the S protein priming, they also

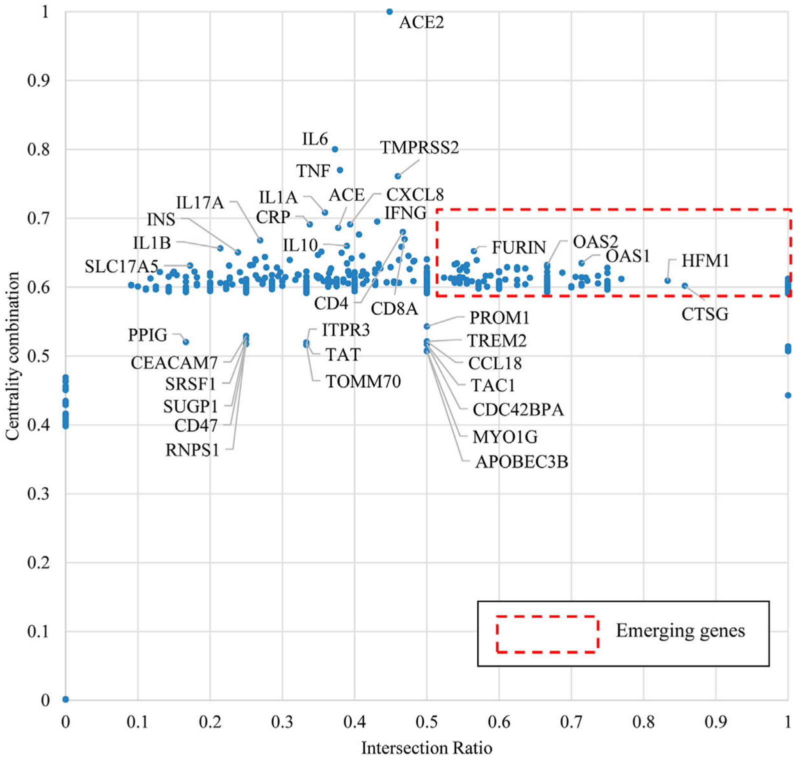

FIGURE 7 | Emerging gene discovery for COVID-19 genetic research.

claimed that furin inhibitors can be targeted as potential drug therapies for SARS-CoV-2.

CXCL10: *CXCL10* is a frequently studied gene in multiple COVID-19 genetic studies (Bermejo-Martin et al., 2020; Chua et al., 2020; Hou et al., 2020; Parkinson et al., 2020; Xiong et al., 2020; Han et al., 2021; Tan et al., 2021). Among those studies, a paper published on *Nature Biotechnology* identified that critical COVID-19 cases had shown stronger interactions between epithelial and immune cells which includes inflammatory macrophages expressing CXCL10 (Chua et al., 2020). Bermejo-Martin et al. (2020) reported that viral RNA load in plasma correlates with higher chemokines levels, including CXCL10 and CCL2. Xiong et al. (2020) also indicated the association between COVID-19 pathogenesis and excessive cytokine release, including CXCL10/IP-10.

OAS1, OAS2, OAS3, IFIT1, IFIT3, IFI44, IFI44L and IFITM1: Current COVID-19 genetic studies incline to analyze those genes together. In a paper published in *Nature*, Pairo-Castineira et al. (2021) identified a significant genetic variant rs10735079 associated with critical illness of COVID-19 in the gene cluster encodes *OAS1*, *OAS2*, and *OAS3*. Interestingly, recent work on archaic human (Neandertal) DNA has identified an additional haplotype in the region of Chromosome 12 containing *OAS1*, *OAS2*, and *OAS3* that protects against severe COVID-19 (Zeberg and Paabo, 2020). Klaassen et al. (2020) identified six genetic variants in innate immunity-related genes, including *OAS1* (p.Arg130His), which might have predictive value for COVID-19 infection. Besides, *IRF9, IFIT1, IFITM1, MX1, OAS2, OAS3, IFI44* and *IFI44L* were found to be upregulated in the COVID-19 infected normal human bronchial epithelial cells (Vishnubalaji et al., 2020). Similarly, Prasad et al. (2020) also found that some interferon-stimulated genes can be considered as potential candidates for drug targets in COVID-19 treatment. Those genes include *IFIT1, IFITM1, IRF7, ISG, MX1*, and *OAS2*. Shi et al. (2021) showed that COVID-19 infections are generally restricted by IFITM1, IFITM2 and IFITM3 using gain- and loss-of-function approaches.

ISG15: The findings of ISG15 are mostly related to the papain-like proteases (PLpro) encoded by the SARS-CoV-2 coronavirus. A paper published in *Nature* revealed a unique preference of SARS-CoV-2 coronavirus of cleaving ubiquitin-like interferon-stimulated gene 15 protein (ISG15), which is different from SARS-CoV (Shin et al., 2020). This study also indicated that SARS-CoV-2 papain-like protease contributes to the cleavage of ISG15 from interferon responsive factor 3 (IRF3) and attenuates type I interferon responses. Klemm et al. (2020) specified that the structure of the SARS-CoV-2 PLpro reveals that S1 ubiquitin-binding site is responsible for high ISG15 activity, while the S2 binding site provides Lys48 chain specificity and cleavage efficiency. Freitas et al. (2020) evaluated the biochemical activity of SARS-CoV-2 PLpro and ISG15 with its counterparts in MERS-CoV and SARS-CoV. They indicated that naphthalene based PLpro inhibitors are shown to be effective at halting SARS-CoV-2 PLpro activity as well as SARS-CoV-2 replication.

IRF3: Zheng et al. (2020b) found that the interaction of COVID-19 M-protein with RIG-1, MAVS, and TBK1 inhibits the formation of multiprotein complexes of those proteins encoded

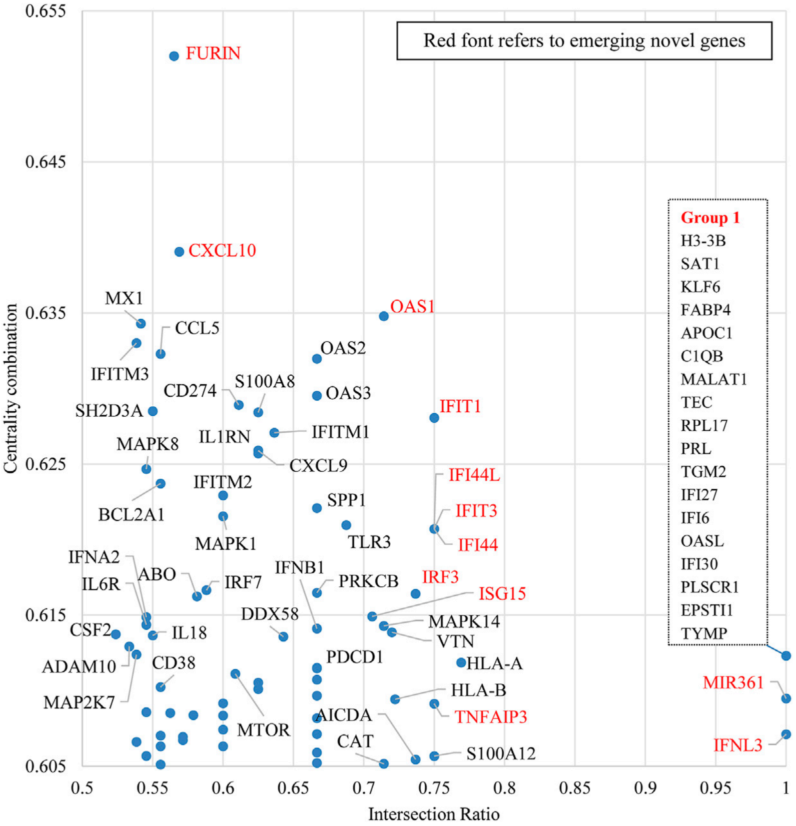

FIGURE 8 | Emerging gene discovery for COVID-19 genetic research–detailed partial view. Note: This figure is a zoom-in map for the red broken-line box in **Figure 7**.

by these genes and subsequently prevents the activation of IRF3. Moustaqil et al. (2020) Identified COVID-19 NSP3 and NSP5 protease's new functions of specifically and selectively cleaving IRF-3, NLRP12 and TAB1. Lévy et al. (2021) reported a case of IFN-α2a therapy in two patients with inborn errors of TLR3 and IRF3 infected with COVID-19. Yin et al. (2020) found that IRF3, IRF5 and NF-κB/p65 are the key transcription factors regulating the IFN response during SARS-CoV-2 infection.

TNFAIP3: Protein and protein interaction analysis from Islam et al. (2020) indicated that *TNFAIP3* is one of the kye hub genes that have good binding affinities with repurposed COVID-19 drug candidates which includes dabrafenib, radicicol and AT-7519. Li et al. (2021) observed the bimodal gene expression of *TNFAIP3* in various immune cells from severely infected COVID-19 patients.

The overlapping genes in Group 1 are investigated in a single paper (Shaath et al., 2020). They identified neutrophils (*IFITM2, IFITM1, H3-H3B, SAT1* and *S100A8*) and macrophage cluster-1 (*CCL8, CCL3, CCL2, KLF6* and *SPP1*) as the main immune cell subsets associated with severe COVID-19 cases. They also found that some upstream regulators (IFNG, PRL, TLR7, PRL, TGM2, TLR9, IL1B, TNF, NFKB, IL1A, STAT3, CCL5) were enriched in bronchoalveolar lavage cells in severe COVID-19 cases compared to the mild cases. Besides, a number of genes found in both mild and severe COVID-19 cases (*IFI27, IFITM3, IFI6, IFIT3, MX1, IFIT1, OASL, IFI30, OAS1*) and genes only in severe cases (*S100A8, IFI44, IFI44L, CXCL8, CCR1, PLSCR1, EPSTI1, FPR1, OAS2, OAS3, IL1RN, TYMP, BCL2A1*) are reported as well.

MIR361: miRNAs are important regulators of viral pathogenesis, particularly among RNA viruses. Pierce et al. (2020) verified the biological plausibility of the predicted miRNA-target RNA interactions, in which miRNA361 binds to the SARS-CoV-2 IFN-α 3′-UTR. Li et al. (2020) showed that hsa-miR-361-3P is one of the top upregulated or downregulated genes in COVID-19 patients compared to the healthy controls.

IFNL3: Stevenson et al. (2021) claimed IFNL3 as one of the predictive markers for severe symptoms of COVID-19 based on analysis of serum chemokines and cytokines from COVID-19 patents, while another pharmacogenomic study did not find the potential of IFNL3 in modifying treatments (Sugiyama et al., 2020). Instead, it identifies CYP2D6 and CYP2C19 as the two likely best targets for treatment modification, especially by ondansetron, oxycodone, and clopidogrel.

DISCUSSION AND CONCLUSION

The COVID-19 pandemic remains a worldwide threat to human health, the global economy and the political landscape.

Controlling the pandemic and strengthening disease prevention and treatment could be the top priority of the entire globe. Despite hopes from vaccine rollouts, there is still substantial knowledge about this virus waiting for discovery and understandings. Whereas a large number of studies on COVID-19 genetic research has been done since the beginning of the COVID-19 pandemic, keeping up with this rapid change is challenging scientific researchers and policymakers. This paper comprehensively analyzed research papers on COVID-19 genetic research published during this pandemic period. We incorporated co-occurrence analysis, scientific evolutionary pathways, and bio-entity network analytics to conduct multi-dimensional tasks of knowledge discovery, particularly with the following four research questions:

Q1: Who are the key players in COVID-19 genetic research?

Universities from the United States are leading this domain, and research institutions from China, the United Kingdom, and Canada closely follow. Investigations on their collaboration patterns reveal among those leading countries, China tends to have strong domestic collaborations while other countries incline to establish international collaborations with each other.

Q2: What are the key research topics prevalently addressed in COVID-19 genetic research?

Figure 4 reveals four major research topics in this field, that is, virus infection mechanisms, virus testing, gene expression related to the immune response to COVID-19 and COVID-19 clinical manifestation.

Q3: How do the foci of genetic studies change during the COVID-19 crisis?

As indicated in **Figure 5**, the initial focus of COVID-19 genetic research is comparative studies between this novel coronavirus and previously discovered coronaviruses, including SARS-CoV and MERS-CoV. Starting from the mid-term of the COVID-19 crisis, this research focus diverges into the analysis of virus molecular structure and the human immune response. Further, derived from the direction of human immune response, vaccine development and gene expression related immunology become two emerging directions in this field.

Q4: Which specific genes are frequently highlighted and which ones are potentially emerging to COVID-19 genetic research?

In **Figures 7**, **8**, we identified genes *ACE2, IL6, TMPRSS2*, and *TNF* as frequently highlighted ones in COVID-19 genetic research, and also identified genes, such as *FURIN, CXCL10, OAS1, OAS2, OAS3, ISG15*, etc., as emerging genes that may require further attention from the research community.

In terms of technical implications, our study provides a suite of intelligent bibliometric tools for biomedical researchers to conduct medical knowledge discovery. For example, it could profile the research landscape of a given medical case, with identified key players, research topics and highlighted associations between genes and diseases. Compared with other bibliometrics conducted on COVID-19, this work provides a systematic and adaptable research framework to profile the research landscape and exploit disease genetics-relate knowledge from literature. Additionally, this study specifically focused on COVID-19 genetic research and targeted a set of frequently highlighted genes and emerging genes on COVID-19, which could then turn to be the clue for COVID-19 prevention and treatment. The results of this study could benefit 1) clinical researchers with longitudinal analyses on COVID-19 genetic research, and 2) policymakers with insights in recognizing potential threats from the COVID-19 and providing pre-emptive actions on national strategies, science policy, and public health and administration for gene-level prevention and treatments.

There are also some limitations for future studies. From the methodological perspective, we designed a data-driven method to identify primary genes and emerging genes from the literature. However, integrating multiple data sources, such as clinical trials and curated medical knowledge databases, may gain value-added benefits. From the perspective of validation measurements, we employed evidence from the literature to interpret and support our findings with assistance from our medical experts. Nonetheless, the qualitative validation may be integrated with multiple quantitative measures, e.g., historical data-based validation, and expert knowledge-based scoring.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found at https://github.com/IntelligentBibliometrics/COVID-19-genetic-research.

AUTHOR CONTRIBUTIONS

Designed research: MW, YZ, HL, MG, JL. Data collection: MG, MW, HL. Data analysis: MW, YZ, DV. Results interpretation and validation: DV, ST. Toolkit development: MW, YZ, HL, MG. Wrote the paper: MW, DV, YZ.

FUNDING

This work is financially supported by a series of collaborative projects between the University of Technology Sydney and the 23 Strands Pty Ltd; and the Australian Research Council under Discovery Early Career Researcher Award (DECRA) DE190100994.

ACKNOWLEDGMENTS

We gratefully acknowledge the exceptional support and suggestions from Professor Cristobal dos Remedios from the University of Sydney.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frma.2021.683212/full#supplementary-material

REFERENCES

Bastian, M., Heymann, S., and Jacomy, M. (2009 March). Gephi: An open-source software for exploring and manipulating networks. In Proceedings of the International AAAI Conference on Web and Social Media, 3 (1).

Bermejo-Martin, J. F., González-Rivera, M., Almansa, R., Micheloud, D., Tedim, A. P., Domínguez-Gil, M., et al. (2020). Viral RNA Load in Plasma Is Associated with Critical Illness and a Dysregulated Host Response in COVID-19. *Crit. Care* 24 (1), 1–13. doi:10.1186/s13054-020-03398-0

Cai, X., Fry, C. V., and Wagner, C. (2021). International Collaboration during the COVID-19 Crisis: Autumn 2020 Developments. *Scientometrics* 126, 3683. doi:10.1007/s11192-021-03873-7

Callon, M., Courtial, J.-P., Turner, W. A., and Bauin, S. (1983). From Translations to Problematic Networks: An Introduction to Co-word Analysis. *Soc. Sci. Inf.* 22 (2), 191–235. doi:10.1177/053901883022002003

Cao, X. (2020). COVID-19: Immunopathology and its Implications for Therapy. *Nat. Rev. Immunol.* 20 (5), 269–270. doi:10.1038/s41577-020-0308-3

Chahrour, M., Assi, S., Bejjani, M., Nasrallah, A. A., Salhab, H., Fares, M., et al. (2020). A Bibliometric Analysis of COVID-19 Research Activity: a Call for Increased Output. *Cureus* 12 (3), e7357. doi:10.7759/cureus.7357

Chua, R. L., Lukassen, S., Trump, S., Hennig, B. P., Wendisch, D., Pott, F., et al. (2020). COVID-19 Severity Correlates with Airway Epithelium-Immune Cell Interactions Identified by Single-Cell Analysis. *Nat. Biotechnol.* 38 (8), 970–979. doi:10.1038/s41587-020-0602-4

Colavizza, G., Costas, R., Traag, V. A., Van Eck, N. J., Van Leeuwen, T., and Waltman, L. (2021). A Scientometric Overview of CORD-19. *PLoS One* 16 (1), e0244839. doi:10.1371/journal.pone.0244839

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., and Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and Weaknesses. *FASEB j.* 22 (2), 338–342. doi:10.1096/fj.07-9492lsf

Fleming, L. (2001). Recombinant Uncertainty in Technological Search. *Manag. Sci.* 47 (1), 117–132. doi:10.1287/mnsc.47.1.117.10671

Fleming, L., and Sorenson, O. (2004). Science as a Map in Technological Search. *Strateg. Manag. J.* 25 (8-9), 909–928. doi:10.1002/smj.384

Freitas, B. T., Durie, I. A., Murray, J., Longo, J. E., Miller, H. C., Crich, D., et al. (2020). Characterization and Noncovalent Inhibition of the Deubiquitinase and deISGylase Activity of SARS-CoV-2 Papain-like Protease. *ACS Infect. Dis.* 6 (8), 2099–2109. doi:10.1021/acsinfecdis.0c00168

Fry, C. V., Cai, X., Zhang, Y., and Wagner, C. S. (2020). Consolidation in a Crisis: Patterns of International Collaboration in Early COVID-19 Research. *PLOS One* 15 (7), e0236307. doi:10.1371/journal.pone.0236307

Grupp, H. (1990). The Concept of Entropy in Scientometrics and Innovation Research: An Indicator for Institutional Involvement in Scientific and Technological Developments. *Scientometrics* 18 (3-4), 219–239. doi:10.1007/bf02017763

Han, K., Blair, R. V., Iwanaga, N., Liu, F., Russell-Lodrigue, K. E., Qin, Z., et al. (2021). Lung Expression of Human Angiotensin-Converting Enzyme 2 Sensitizes the Mouse to SARS-CoV-2 Infection. *Am. J. Respir. Cel Mol Biol.* 64 (1), 79–88. doi:10.1165/rcmb.2020-0354oc

Hoffmann, M., Kleine-Weber, H., and Pöhlmann, S. (2020a). A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Mol. Cel* 78 (4), 779–784.e775. doi:10.1016/j.molcel.2020.04.022

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., et al. (2020b). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 181 (2), 271–280.e278. doi:10.1016/j.cell.2020.02.052

Hou, X., Zhang, X., Wu, X., Lu, M., Wang, D., Xu, M., et al. (2020). Serum Protein Profiling Reveals a Landscape of Inflammation and Immune Signaling in Early-Stage COVID-19 Infection. *Mol. Cell Proteomics* 19 (11), 1749–1759. doi:10.1074/mcp.rp120.002128

Hristovski, D., Peterlin, B., Mitchell, J. A., and Humphrey, S. M. (2005). Using Literature-Based Discovery to Identify Disease Candidate Genes. *Int. J. Med. Inform.* 74 (2-4), 289–298. doi:10.1016/j.ijmedinf.2004.04.024

Huang, Y., Porter, A., Zhang, Y., and Barrangou, R. (2019). Collaborative Networks in Gene Editing. *Nat. Biotechnol.* 37, 1107. doi:10.1038/s41587-019-0275-z

Islam, T., Rahman, M. R., Aydin, B., Beklen, H., Arga, K. Y., and Shahjaman, M. (2020). Integrative Transcriptomics Analysis of Lung Epithelial Cells and Identification of Repurposable Drug Candidates for COVID-19. *Eur. J. Pharmacol.* 887, 173594. doi:10.1016/j.ejphar.2020.173594

Klaassen, K., Stankovic, B., Zukic, B., Kotur, N., Gasic, V., Pavlovic, S., et al. (2020). Functional Prediction and Comparative Population Analysis of Variants in Genes for Proteases and Innate Immunity Related to SARS-CoV-2 Infection. *Infect. Genet. Evol.* 84, 104498. doi:10.1016/j.meegid.2020.104498

Klemm, T., Ebert, G., Calleja, D. J., Allison, C. C., Richardson, L. W., Bernardini, J. P., et al. (2020). Mechanism and Inhibition of the Papain-like Protease, PLpro, of SARS-CoV-2. *EMBO J.* 39 (18), e106275. doi:10.15252/embj.2020106275

Klimstra, W. B., Tilston-Lunel, N. L., Nambulli, S., Boslett, J., McMillen, C. M., Gilliland, T., et al. (2020). SARS-CoV-2 Growth, Furin-Cleavage-Site Adaptation and Neutralization Using Serum from Acutely Infected Hospitalized COVID-19 Patients. *J. Gen. Virol.* 101 (11), 1156–1169. doi:10.1099/jgv.0.001481

Lévy, R., Bastard, P., Lanternier, F., Lecuit, M., Zhang, S.-Y., and Casanova, J.-L. (2021). IFN-α2a Therapy in Two Patients with Inborn Errors of TLR3 and IRF3 Infected with SARS-CoV-2. *J. Clin. Immunol.* 41 (1), 26–27. doi:10.1007/s10875-020-00933-0

Li, C., Hu, X., Li, L., and Li, J. h. (2020). Differential microRNA Expression in the Peripheral Blood from Human Patients with COVID-19. *J. Clin. Lab. Anal.* 34 (10), e23590. doi:10.1002/jcla.23590

Li, Y., Duche, A., Sayer, M. R., Roosan, D., Khalafalla, F. G., Ostrom, R. S., et al. (2021). SARS-CoV-2 Early Infection Signature Identified Potential Key Infection Mechanisms and Drug Targets. *BMC Genomics* 22 (1), 1–13. doi:10.1186/s12864-021-07433-4

Liao, H., Tang, M., Luo, L., Li, C., Chiclana, F., and Zeng, X.-J. (2018). A Bibliometric Analysis and Visualization of Medical Big Data Research. *Sustainability* 10 (1), 166. doi:10.3390/su10010166

Moustaqil, M., Ollivier, E., Chiu, H.-P., Van Tol, S., Rudolffi-Soto, P., Stevens, C., et al. (2020). SARS-CoV-2 Proteases PLpro and 3CLpro Cleave IRF3 and Critical Modulators of Inflammatory Pathways (NLRP12 and TAB1): Implications for Disease Presentation across Species. *Emerging Microbes & Infections* 10, 178. doi:10.1080/22221751.2020.1870414

Pairo-Castineira, E., Clohisey, S., Clohisey, S., Klaric, L., Bretherick, A. D., Rawlik, K., et al. (2021). Genetic Mechanisms of Critical Illness in Covid-19. *Nature* 591 (7848), 92–98. doi:10.1038/s41586-020-03065-y

Parkinson, N., Rodgers, N., Fourman, M. H., Wang, B., Zechner, M., Swets, M. C., et al. (2020). Dynamic Data-Driven Meta-Analysis for Prioritisation of Host Genes Implicated in COVID-19. *Scientific Rep.* 10 (1), 1–12. doi:10.1038/s41598-020-79033-3

Pierce, J. B., Simion, V., Icli, B., Pérez-Cremades, D., Cheng, H. S., and Feinberg, M. W. (2020). Computational Analysis of Targeting SARS-CoV-2, Viral Entry Proteins ACE2 and TMPRSS2, and Interferon Genes by Host MicroRNAs. *Genes* 11 (11), 1354. doi:10.3390/genes11111354

Porter, A. L., Zhang, Y., Huang, Y., and Wu, M. (2020). Tracking and Mining the COVID-19 Research Literature. *Front. Res. Metrics Analytics* 5, 12. doi:10.3389/frma.2020.594060

Prasad, K., Khatoon, F., Rashid, S., Ali, N., AlAsmari, A. F., Ahmed, M. Z., et al. (2020). Targeting Hub Genes and Pathways of Innate Immune Response in COVID-19: A Network Biology Perspective. *Int. J. Biol. Macromolecules* 163, 1–8. doi:10.1016/j.ijbiomac.2020.06.228

Sallenave, J.-M., and Guillot, L. (2020). Innate Immune Signaling and Proteolytic Pathways in the Resolution or Exacerbation of SARS-CoV-2 in Covid-19: Key Therapeutic Targets?. *Front. Immunol.* 11, 1229. doi:10.3389/fimmu.2020.01229

Salton, G., and McGill, M. J. (1986). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

Shaath, H., Vishnubalaji, R., Elkord, E., and Alajez, N. M. (2020). Single-Cell Transcriptome Analysis Highlights a Role for Neutrophils and Inflammatory Macrophages in the Pathogenesis of Severe COVID-19. *Cells* 9 (11), 2374. doi:10.3390/cells9112374

Shi, G., Kenney, A. D., Kudryashova, E., Zani, A., Zhang, L., Lai, K. K., et al. (2021). Opposing Activities of IFITM Proteins in SARS-CoV-2 Infection. *EMBO J.* 40 (3), e106501. doi:10.15252/embj.2020106501

Shin, D., Mukherjee, R., Grewe, D., Bojkova, D., Baek, K., Bhattacharya, A., et al. (2020). Papain-like Protease Regulates SARS-CoV-2 Viral Spread and Innate Immunity. *Nature* 587 (7835), 657–662. doi:10.1038/s41586-020-2601-5

Stevenson, J. M., Alexander, G. C., Palamuttam, N., and Mehta, H. B. (2021). Projected Utility of Pharmacogenomic Testing Among Individuals Hospitalized with COVID-19: A Retrospective Multicenter Study in the United States. *Clin. Transl Sci.* 14 (1), 153–162. doi:10.1111/cts.12919

Sugiyama, M., Kinoshita, N., Ide, S., Nomoto, H., Nakamoto, T., Saito, S., et al. (2020). Serum CCL17 Level Becomes a Predictive Marker to Distinguish between Mild/moderate and Severe/critical Disease in Patients with COVID-19. *Gene* 766, 145145. doi:10.1016/j.gene.2020.145145

Tan, S., Chen, W., Xiang, H., Kong, G., Zou, L., and Wei, L. (2021). Screening Druggable Targets and Predicting Therapeutic Drugs for COVID-19 via Integrated Bioinformatics Analysis. *Genes Genom* 43 (1), 55–67. doi:10.1007/s13258-020-01021-8

Thompson, D. F., and Walker, C. K. (2015). A Descriptive and Historical Review of Bibliometrics with Applications to Medical Sciences. *Pharmacotherapy* 35 (6), 551–559. doi:10.1002/phar.1586

Van Eck, N. J., and Waltman, L. (2010). Software Survey: VOSviewer, a Computer Program for Bibliometric Mapping. *Scientometrics* 84 (2), 523–538. doi:10.1007/s11192-009-0146-3

Vishnubalaji, R., Shaath, H., and Alajez, N. M. (2020). Protein Coding and Long Noncoding RNA (lncRNA) Transcriptional Landscape in SARS-CoV-2 Infected Bronchial Epithelial Cells Highlight a Role for Interferon and Inflammatory Response. *Genes* 11 (7), 760. doi:10.3390/genes11070760

Wei, C.-H., Allot, A., Leaman, R., and Lu, Z. (2019). PubTator central: Automated Concept Annotation for Biomedical Full Text Articles. *Nucleic Acids Res.* 47 (W1), W587–W593. doi:10.1093/nar/gkz389

Wrobel, A. G., Benton, D. J., Xu, P., Roustan, C., Martin, S. R., Rosenthal, P. B., et al. (2020). SARS-CoV-2 and Bat RaTG13 Spike Glycoprotein Structures Inform on Virus Evolution and Furin-Cleavage Effects. *Nat. Struct. Mol. Biol.* 27 (8), 763–767. doi:10.1038/s41594-020-0468-7

Wu, M., Zhang, Y., Lu, J., Lin, H., and Grosser, M. (2020). Recommending scientific collaborators: Bibliometric networks for medical research entities. In Developments of Artificial Intelligence Technologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference (FLINS 2020), 480–487. doi:10.1142/9789811223334_0058

Wu, M., Zhang, Y., Zhang, G., and Lu, J. (2021). Exploring the Genetic Basis of Diseases through a Heterogeneous Bibliometric Network: A Methodology and Case Study. *Technol. Forecast. Soc. Change* 164, 120513. doi:10.1016/j.techfore.2020.120513

Xiong, Y., Liu, Y., Cao, L., Wang, D., Guo, M., Jiang, A., et al. (2020). Transcriptomic Characteristics of Bronchoalveolar Lavage Fluid and Peripheral Blood Mononuclear Cells in COVID-19 Patients. *Emerging Microbes & Infections* 9 (1), 761–770. doi:10.1080/22221751.2020.1747363

Yin, X., Riva, L., Pu, Y., Martin-Sancho, L., Kanamune, J., Yamamoto, Y., et al. (2020). MDA5 Governs the Innate Immune Response to SARS-CoV-2 in Lung Epithelial Cells. *Cel Rep.* 34 (2), 108628.

Zeberg, H., and Pääbo, S. (2020). A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. Proceedings of the National Academy of Sciences, 118 (9). doi:10.1073/pnas.2026309118

Zhang, L., Zhao, W., Sun, B., Huang, Y., and Glänzel, W. (2020a). How Scientific Research Reacts to International Public Health Emergencies: a Global Analysis of Response Patterns. *Scientometrics* 124, 747–773. doi:10.1007/s11192-020-03531-4

Zhang, Y., Cai, X., Fry, C. V., Wu, M., and Wagner, C. (2021a). opic evolution, disruption and resilience in early COVID-19 research. *Scientometrics* 1–29.

Zhang, Y., Porter, A. L., Cunningham, S., Chiavetta, D., and Newman, N. (2020c). Parallel or Intersecting Lines? Intelligent Bibliometrics for Investigating the Involvement of Data Science in Policy Analysis. *IEEE Trans. Eng. Manage.* 1–13. doi:10.1109/TEM.2020.2974761

Zhang, Y., Wu, M., Lin, H., Tipper, S., Grosser, M., Zhang, G., et al. (2020d). Framework of Computational Intelligence-Enhanced Knowledge Base Construction: Methodology and A Case of Gene-Related Cardiovascular Disease. *Int. J. Comput. Intelligence Syst.* 13 (1), 1109–1119. doi:10.2991/ijcis.d.200728.001

Zhang, Y., Wu, M., Miao, W., Huang, L., and Lu, J. (2021). Bi-layer Network Analytics: A Methodology for Characterizing Emerging General-Purpose Technologies. Available at SSRN 3830937. https://ssrn.com/abstract=3830937, (Accessed April 24, 2021). doi:10.2139/ssrn.3830937

Zhang, Y., Zhang, G., Zhu, D., and Lu, J. (2017). Scientific Evolutionary Pathways: Identifying and Visualizing Relationships for Scientific Topics. *J. Assoc. Inf. Sci. Technol.* 68 (8), 1925–1939. doi:10.1002/asi.23814

Zheng, Y.-Y., Ma, Y.-T., Zhang, J.-Y., and Xie, X. (2020a). COVID-19 and the Cardiovascular System. *Nat. Rev. Cardiol.* 17 (5), 259–260. doi:10.1038/s41569-020-0360-5

Zheng, Y., Zhuang, M.-W., Han, L., Zhang, J., Nan, M.-L., Zhan, P., et al. (2020b). Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Membrane (M) Protein Inhibits Type I and III Interferon Production by Targeting RIG-I/MDA-5 Signaling. *Signal. Transduction Targeted Ther.* 5 (1), 1–13. doi:10.1038/s41392-020-00438-7

Conflict of Interest: HL, MG, ST, DV are employees of Strands Pty Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.