Bonn-Aachen International Center for Information Technology

(B-IT)

University of Bonn

Master Program in Life Science Informatics

Master Thesis

Landscaping of COVID-19 Data for the Discovery of Insightful Patterns on Ethnicities – An Attempt

Submitted by…….....Ram Kumar Ruppa Surulinathan

First Reviewer……..……....Prof. Dr. Martin Hofmann-Apitius

Second Reviewer…….Prof. Dr. Juliane Fluck

Internal Supervisors……Dr. Marc Jacobs & Bruce Schultz

A thesis submitted in fulfillment of the requirements for the degree of M.Sc. in Life Science Informatics in the

Bonn-Aachen International Center for Information Technology

(B-IT)

in collaboration with

Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

October 2021

ACKNOWLEDGEMENTS

I am extremely grateful to *Prof. Dr. Martin Hofmann-Apitius*, Head of Department of Bioinformatics at Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin for offering me a place in his group to perform my master thesis on COVID-19 and clinical trials.

I express my deep sense of gratitude to *Dr. Marc Jacobs*, Group leader of Software and Scientific Computing and Deputy Head of Department of Bioinformatics, for initiating and inspiring me into this research work, invaluable guidance, patience and encouragement during the start of the work.

Also, my second supervisor – *Bruce Schultz* for his time and invaluable feedback during the sessions thus mentoring the project.

I thank *Prof. Dr. Juliane Fluck*, who agreed to act as my second reviewer for my thesis.

I thank different researchers from SCAI namely *Dr. Alpha Tom Kodamullil* and *Vinay Bharadwaj* for introducing the SCAIView, *Aliaksandr Masny* for his session in working with the SCAIView APIs.

I also thank *Natascha von Oppenkowski-Schmidt, Meike Kneips and Alina Enns* for helping me with the official works at the SCAI.

Finally, we would like to offer our genuine gratitude to God almighty for blessing us with this opportunity to explore the realms of science and last but not least we would like to dedicate this to our parents who prayed for us and motivated us throughout our work and giving us the strength to prevail.

A SMALL NOTE

At this juncture, I am reminded of the renowned philosopher Ayyan Thiruvalluvar's words, which is stated below:

With rising flood the rising lotus flower its stem unwinds;

The dignity of men is measured by their minds.

- Adapted from *Thirukural* Couplet No. 595

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SARS-CoV-2	**S**evere **A**cute **R**espiratory **S**yndrome **Co**rona**v**irus-**2**
ICTRP	**I**nternational **C**linical **T**rials **R**egistry **P**latform
WHO	**W**orld **H**ealth **O**rganization
ECRIN	**E**uropean **C**linical **R**esearch **I**nfrastructure **N**etwork
MDR	**M**eta**d**ata **R**epository
BEL	**B**iological **E**xpression **L**anguage
SBML	**S**ystems **B**iology **M**arkup **L**anguage
SBGN	**S**ystems Biology Graph Notation
BioPAX	**Bio**logical **Pa**thways E**x**change
OWL	**W**eb **O**ntology **L**anguage
OBDA	**O**ntology **B**ased **D**ata **A**ccess
OBO	**O**pen **B**iological and **B**iomedical **O**ntology
MeSH	**Me**dical **S**ubject **H**eadings
NLP	**N**atural **L**anguage **P**rocessing
NER	**N**amed **E**ntity **R**ecognition
API	**A**pplication **P**rogramming **I**nterface
GWAS	**G**enome **W**ide **A**ssociation Studies
A	**A**denine
C	**C**ytosine
T	**T**hymine
G	**G**uanine
COVID-19 HGI	**COVID-19 H**ost **G**enetics **I**nitiative
SNP	**S**ingle **N**ucleotide **P**olymorphism
FIMM	**I**nstitute of **M**olecular **M**edicine in **F**inland
BiKMi	**Bi**ological **K**nowledge **Mi**ner

ABSTRACT

COVID-19 disease became a global disaster in the 21st century. This disease is influenced by several parameters including the extent of the curfew as it is contagious. Fortunately, the scientific community across the globe collaborated to facilitate COVID-19 research. To unravel the ethnicity specific disease characteristics, certain SNPs found in the population might be a potential cause for the infectivity, drug response etc. The SNPs found in a population can be linked to molecular pathways and the outcome of the trial. We collected the SNP data for distinct populations from COVID-19 Host Genetics Initiative. Then, we integrated the SNP data into the COVID-19 Knowledge graph – Pharmacome. Eventually, this resulted in the creation of four ethnicity specific Pharmacomes namely Europe, Asia, Africa, North America and South America with 1798 SNPs and 239 genes. Chromosome 3 was associated with the largest number of COVID genes. Certain prominent genes like OAS, ACE2, DPP9 and TYK2 were identified in the results. Secondly, we performed the pathway analysis of COVID-19 genes that have a COVID-19 association score of 1.0. This revealed certain key pathways such as spike protein maturation, interferon signaling, interleukin signaling. The Pharmacome was further explored to reveal new drug targets for drug repurposing experiments and revealed many new drugs along with some already available drugs for the treatment. Adding to the sex bias of the disease, in an attempt for a genetics-based drug target mechanism network we identified the top molecular pathways enriched with the gene MECP2.

Keywords: COVID-19, SARS-CoV-2, SNPs, Pharmacome, Host Genetics Initiative, MECP2 gene

1. INTRODUCTION

This chapter includes the scientific questions that were attempted to solve in this study and the organization of the thesis.

1.1 Scientific Questions

The list of the scientific questions that were attempted to solve are listed as following:

- How do different Pharmacomes generated based on ethnicities differ in terms of SNPs?
- How do medical interventions and clinical outcomes differ if they are of different ethnicities?
- What are the significant molecular pathways involved in the COVID-19 risk genes?
- What is the sex bias of COVID-19 in terms of its infection and fatality rate?
- What are the possible COVID-19 drug targets that are worth checking for drug repurposing revealed by the Pharmacome?

1.2 Thesis Organization

The organization of the thesis is as follows:

- Chapter 01 describes the scientific questions of this study.
- Chapter 02 describes background knowledge.
- Chapter 03 describes the materials and methods in the study.
- Chapter 04 describes the results and discussions of the study.
- Chapter 05 describes the conclusions of the study.
- Chapter 06 lists the scientific references of the study.

2. THEORETICAL BACKGROUND

This chapter introduces COVID-19 disease, virus, symptoms, clinical trials, risk genes and knowledge graphs.

2.1 SARS-CoV-2 a novel coronavirus

The Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2, colloquially known as COVID-19) a novel strain and fatal coronavirus was first identified in the Wuhan city of China in December 2019. The common symptoms include increased body temperature, dry cough, nausea and body pains (Esakandari et al., 2020). The rapid spread of the virus posed a threat of life to the global environment. According to the records maintained by Worldometer, the top ten most affected nations include the USA, India, Brazil, France, Russia, UK, Turkey, Italy, Spain and Germany (Worldometer, 2021).

Coronaviruses generally are classified under the family Coronoviridae and subfamily Coronavirinae which is subdivided into four genera namely Alphacoronavirus, Betacoronavirus, Gammacoronavirus and Deltacoronavirus (Mittal et al., 2020). The SARS-CoV-2, a member of Betacoronavirus genera whose sequence is 96% homologous to the bat coronavirus. Its primary reservoir is bats and transmitted to human beings through an intermediate host called Pangolin (Zhao et al., 2020). The SARS-CoV-2 shares 79.5% sequence identity to that of SARS-CoV (Zheng & Song, 2020).

Angiotensin-converting enzyme 2 (ACE2) is a primary functional receptor for the coronaviruses. It is responsible for the process of viral infection and replication inside the host cells (Kai & Kai, 2020). A protease enzyme TMPRSS2 is involved in this process as an activator in the SARS-CoV and SARS-CoV-2 infection (Mousavizadeh & Ghasemi, 2020).

2.2 Structure of COVID-19 virus

The SARS-CoV-2 virus is spherical in structure with a positive single stranded RNA (ssRNA) genome packed inside the nucleocapsid protein (N) and enveloped by the membrane glycoprotein protein (M), an envelope protein (E), and the spike protein (S). The typical virus lengths are between 26.4 and 31.7 kilobases (kb) with the GC content (Guanine-Cytosine) ranging between 32% and 43% (Mousavizadeh & Ghasemi, 2020).

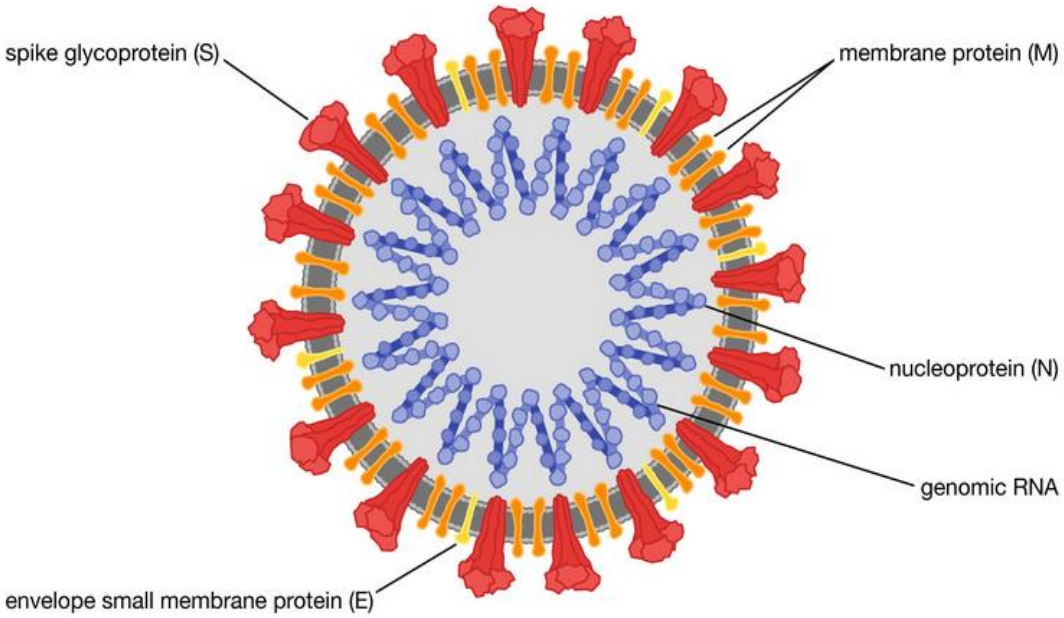

Figure 2.1: Structure of novel coronavirus SARS-CoV-2, Adapted from (Britannica, 2021)

2.2.1 COVID-19 Variants

The virus replicates inside the host cell, thus creating multiple copies for the invasion. It also occurs that some copying errors could occur during this process which is termed mutation (*Coronavirus Variants and Mutations*, 2021). Therefore, a virus becomes more diverse by mutations for its survival in its environment and thus leading to resistance against treatments. Researchers compare the genetic differences between the viruses for the identification of

variants and their relatedness (*About Variants of the Virus That Causes COVID-19*, 2021).

A “variant” can be termed as a group of viruses, which shares the same set of inherited mutations meaning that they emerged from a common ancestor. These mutations get accumulated in a lineage and are termed as “strains”.

There are several studies for determining the mutation effects in the S-protein, functional domain of SARS-CoV-2. These include various phenotypes namely higher affinity to ACE2 receptors, higher fatality rate, resistance to antibodies and higher infectivity (Bakhshandeh et al., 2021).

The US Government has classified into three classes of SARS-CoV-2 variants namely (*SARS-CoV-2 Variant Classifications and Definitions*, 2021):

a. Variants of Interest

   These variants are associated with increased transmissibility and reduced response for the corresponding treatments. These variants require surveillance and investigations on the spread of the virus. Some examples include B.1.427 (Epsilon), B.1.525 (Eta), B.1.526 (Iota) and B.1.617.1 (Kappa).

b. Variants of Concern

   These variants are associated with a high number of hospitalizations and deaths, reduced response for treatments or even the inability to the diagnosis detection. These variants require notification to the WHO (World Health Organization), local authorities to limit the spread of the variant and subsequent steps towards efficient treatment and diagnostics. Some examples include B.1.1.7 (Alpha), B.1.351 (Beta) and B.1.617.2 (Delta).

c. Variants of High Consequences

   These variants are associated with medical countermeasures (MCMs), which reduce the effectiveness of other variants including the increased infection rates and higher failure of diagnostic detection and treatment. This again

requires notification to the WHO (World Health Organization) to contain the transmission of the variant. Fortunately, there have not been any such variants observed across any country.

2.3 Symptoms of COVID-19 virus

The common symptoms observed during the COVID-19 infection are listed below:

a. Fever
b. Cough
c. Fatigue

The rare symptoms observed during the COVID-19 infection are listed below:

a. Diarrhea
b. Sore throat
c. Rhinorrhea
d. Congestion (Fu et al., 2020)

The primary site of the COVID-19 infection is the lungs which lead to the damage of alveolar epithelial cells. During the infection, the ACE2 receptor is occupied by the SARS-CoV-2 virus thus inhibiting the normal activity. This leads to the increased availability of angiotensin II (Ang II) which damages the tissues, blood vessels, therefore, leading to pathogenesis. Apart from respiratory infection, it also leads to malfunctioning of other organ systems such as the Central Nervous System (CNS) therefore leading to the loss of taste and smell, cardiovascular diseases and renal damage (Esakandari et al., 2020).

During the onset of symptoms, there is an elevated level of infection level of biomarkers namely Tumor Necrosis Factor (TNF), Interleukins (IL), Procalcitonin, Erythrocyte sedimentation rate, C-reactive protein (CRP), Serum ferritin along with the blood cells count (Qin et al., 2020).

2.4 Clinical studies

Clinical studies are experimental studies performed using volunteers to examine different interventions like medical, surgical or behavioral ones. They can be classified into interventional studies and observational studies. These studies are led by a medical doctor assisted by other doctors, nurses carried out in hospitals, universities and research institutes (*Learn About Clinical Studies*, n.d.).

In the year 1747, the very first clinical trials of scurvy were conducted by Dr. James Lind, a Scottish surgeon as the disease resulted in a higher mortality rate in the ship. The treatment of the affected sailors using citric fruits like lemons and oranges resulted in drastic improvements (Bhatt, 2010). This led to the foundation of clinical studies in the future. In recent years, clinical studies have been performed for various diseases including HIV, cancer, epilepsy etc.

2.4.1 Different phases of interventional clinical studies

The clinical trial studies consist of four phases to determine whether the drug could be employed for public usage. They are described as follows:

Phase I: During the initial phase, around 20 to 80 participants are enrolled having no underlying medical complications to evaluate the highest dosage levels that can be administrated without serious side effects.

Phase II: During this phase, around 100 to 300 participants are enrolled to evaluate the effectiveness of the medication along with short time side effects if occurred. This is carried out for up to several years.

Phase III: During this phase, around 3000 participants are enrolled to evaluate the effectiveness of the medication across the diverse population and varied dosage thus studying both drug safety and efficacy. The rare and long time effects are observed during this phase. The medications are approved if the trial results are positive.

Phase IV: During this final phase, long time side effects and efficacy of the approved medications are evaluated across the thousands of participants (*What Are Clinical Trials and Studies?*, n.d.).

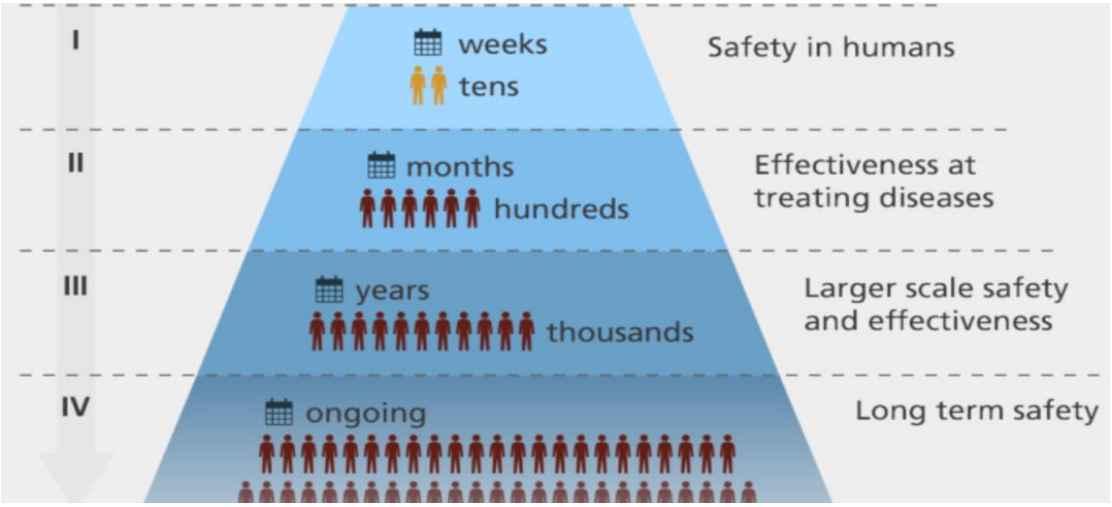

Figure 2.2: Different phases of clinical trial studies. Adapted from (Understanding the Clinical Trial Process, 2016)

2.5 Clinical Trial Registries

The results of the clinical studies are recorded and published in a repository called Clinical Trial Registries. They are open source and their accession is available for the scientific community and public for free. Each nation has its trial registry systems centrally maintained by its government or other approved institution. Some notable registries are described as follows:

2.5.1 ClinicalTrials.gov

ClinicalTrials.gov is a US based trial registry maintained by the National Library of Medicine (NLM) at the National Institute of Health (NIH), made available in February 2000. It holds both interventional and observational studies carried out in 50 US states and 220 countries. The records contain information pertaining to the disease under investigation; the type of interventions employed; meta information like ethnicity of the participants, demographic data, inclusion/exclusion criteria for the study, location, list of comorbidities if any, clinical variables and laboratory variables. The

results are included sometimes if the trials are subject to Section 801 of FDAAA (FDAAA 801) (*ClinicalTrials.Gov Background*, n.d.).

The ClinicalTrials.gov registry could be accessed via https://clinicaltrials.gov/. The registry has currently recorded 6,454 studies for the novel COVID 19 virus as of 27th August 2021.

2.5.2 ICTRP

International Clinical Trials Registry Platform (ICTRP) is a project of the World Health Organization (WHO) to ensure the registration of the *WHO Trial Registration Data Set* and its accessibility to the public. For the clinical study to be considered as fully registered, it must contain a minimum amount of information referred to as Trial Registration Data Set (TRDS). Some of them include the title of the study, disease conditions investigated, participant's location, type of the study, duration of the study and outcome of the study etc. (*About ICTRP*, n.d.).

The ICTRP could be accessed via https://apps.who.int/trialsearch/. It was established in August 2005 in Geneva, Switzerland.

2.5.3 EU Clinical Trials Register

The interventional studies conducted in the European Union (EU) or the European Economic Area (EEA) after May 01, 2004 are recorded in EU Clinical Trials Register. In this registry, the description of phase II to phase IV along with its summary results are available. The summary results include the trial information, endpoints, adverse effects identified in patients if available with the additional information. The registry doesn't contain any information on non-interventional studies, surgical procedures, medical devices and psychotherapeutic procedures (*About the EU Clinical Trials Register*, n.d.). The EU Clinical Trials Register could be accessed via https://www.clinicaltrialsregister.eu/ctr-search/search.

2.5.4 ECRIN-MDR

European Clinical Research Infrastructure Network (ECRIN) is an EU based non-profit organization built to facilitate multinational

clinical research across twelve EU countries. To support COVID-19 research, ECRIN developed the Metadata Repository (MDR). It standardizes the metadata about the clinical studies and thus it could be accessed by a web interface. This portal is an open source enabling researchers to access worldwide with the results directing to the open access journal article or a trial registry entry if the results are publicly available (*Clinical Research Metadata Repository | ECRIN*, n.d.). The ECRIN-MDR could be accessed via https://ecrin.org/tools/clinical-research-metadata-repository.

2.5.5 Global Coronavirus COVID-19 Clinical Trial Tracker

This is a specialized real time dashboard of the clinical trials for COVID-19. The results are gathered from various registries like International Clinical Trials Registry Platform (ICTRP), Chinese Clinical Trial Registry (ChiCTR) in China, ClinicalTrials.gov in the US, EU Clinical Trials Register in EU/EEA, Clinical Research Information Service – Republic of Korea (CRiS) in South Korea, Iranian Registry of Clinical Trials (IRCT) in Iran, Japan Primary Registries Network (JPRN) in Japan and German Clinical Trials Register (DRKS) in Germany. To identify potential clinical studies Artificial Intelligence (AI) based methods are employed. COVID-19 trials are mapped based on geographical and intervention results and visualized in a convincing plot (Thorlund et al., 2020). The real time dashboard could be accessed via https://www.covid-trials.org/.

2.6 Ontology

Ontologies are developed to capture knowledge about a specific domain of interest like Genes, Clinical Trials, Parkinson's disease, COVID-19 etc. across various disciplines. An ontology best describes the concepts within the specified domain along with the exploration of relationships existing between the concepts. An ontology consists of various components including Individuals, Classes, Attributes, Relations, Restrictions, Rules and Axioms. The previously listed ontologies are based on the *Open Biological and Biomedical Ontology* (OBO) Foundry principles in *Web Ontology Language* (OWL) representations. The biocurators and other

researchers help in building an ontology by curating various entities, relationships, properties in a hierarchical manner (Hoyt, 2020).

Some applications of ontologies include extraction of information from various sources, improvement of communication/interoperability between people and organizations. Guo *et al*., used ontology as a technology for the health statistic data as it is a foundation for analysis of multi-source data and its efficiency in its management (Guo et al., 2017). Ontology Based Data Access (OBDA), a promising approach in clinical research bridges the semantic gaps and makes them available in a convenient RDF format for processing and querying (Kock-Schoppenhauer et al., 2017).

2.6.1 COVID-19 Ontology

COVID-19 ontology is designed to facilitate COVID-19 research by facilitating literature search and thus widely applied in text mining and drug repurposing. In the aspect of text mining, information retrieval and extraction of COVID-19 topics like “symptoms”, “transmission”, “virology” etc., are obtained using the COVID SCAIView interface. New drug targets as candidates for repurposing and their target mechanisms in the context of COVID-19 are advantageous.

For building the ontology, the concepts and entities were derived from several research articles, reviews, COVID-19 related websites which are assembled using Protégé ontology editor on the principles of *Open Biological and Biomedical Ontology* (OBO) (Sargsyan et al., 2021).

Such large ontologies cover a wide range of topics including clinical, transmission, chemical pathways.

This ontology is made available for free to the scientific community on several platforms including BioPortal, a dedicated repository for biomedical ontologies. It is available on the web link https://bioportal.bioontology.org/ontologies/COVID-19.

2.7 COVID-19 Literature mining

The step of analyzing the COVID-19 associated information is the exploration of information from biomedical literature by semantic search and information retrieval system for knowledge discovery. The free search engine PubMed contains citations related to biomedicine, life sciences and other associated fields uses Medical Subject Headings (MeSH) for the annotation of abstracts thus enabling the semantic search by expanding the query terms. Annotation of biological entities in the text corpus (collections of text) is not supported by PubMed. The manual annotation of the large volumes of scientific literature is practically infeasible due to time complexity and laboriousness.

Natural Language Processing (NLP) is an advanced field of artificial intelligence for the automatic extraction of knowledge from unstructured data such as the scientific literature in this scenario. The identification of different biological entities is known as “Named Entity Recognition (NER)” which finds out the names of genes, proteins, chemicals, small molecules etc.

The NLP techniques have been employed in the extraction and analysis of data from various sources namely Electronic Health Records (EHR), social media sites by sentiment analysis for the prediction of a medical condition or outbreak of a disease (Arora et al., 2021). Khanday et al., used machine learning techniques for the classification of clinical reports into four cases namely COVID, ARDS, SARS and Both (Khanday et al., 2020).

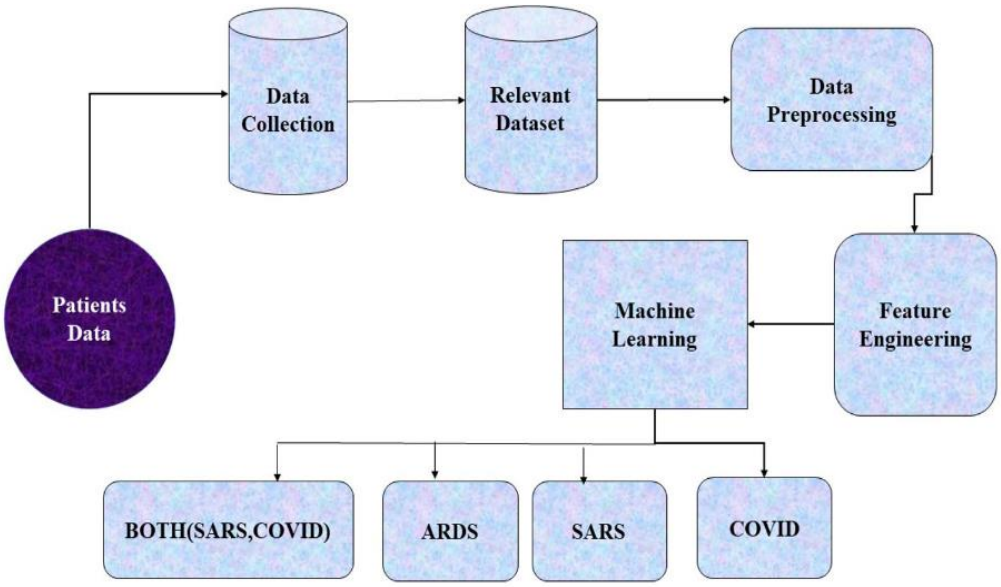

Figure 2.3: The flowchart for the prediction of COVID-19 from clinical reports. Adapted from (Khanday et al., 2020)

2.7.1 SCAIView

SCAIView is a tool that integrates knowledge discovery and semantic search. The software employs machine learning and Named Entity Recognition (NER) for the identification of biological entities from a list of scientific articles. Thus, the tool enables researchers to answer complex scientific queries in a simple and intuitive way as possible by transforming the queries and ontological filtering to narrow the results.

The input of the SCAIView is the search terms combined by boolean operators and the retrieval of documents with their annotations.

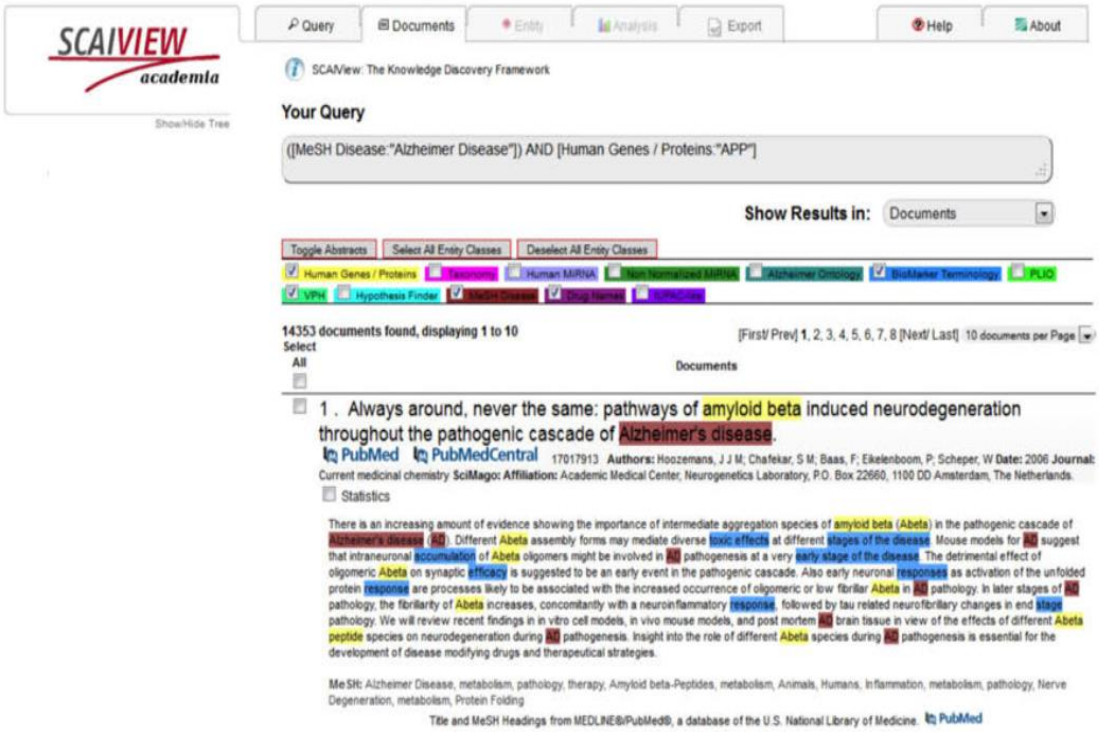

Figure 2.4: Input and output of SCAIView. Adapted from (*Introduction to SCAIView*, n.d.).

The key features of SCAIView include a user friendly query builder, accurate search results sorted by relevance measured by F1 score etc., visualization and ranking of the results, export of the results in

user preferred file formats like .csv, .tsv, .json and programmatic access using an API.

SCAIView also has auto completion features and the results are color coded to highlight different entities within the literature (*Introduction to SCAIView*, n.d.).

For the COVID-19 literature mining, there is a dedicated site called COVID SCAIView at https://covid.scaiview.com/ which is available for free to the scientific community.

2.8 COVID-19 Risk genes and Host genetics initiative

The human genome consisted of nearly 3 billion nucleotide base pairs namely - adenine (A), cytosine (C), thymine (T) and guanine (G). The genetic sequences between any two individuals are very similar, up to 99.9% and the remaining 0.01% difference that makes each individual unique is called genetic variation across populations and species. “Genome Wide Association Studies (GWAS)” are performed to analyze whether the genetic variation in the regions of the genome is associated with a particular disease. In this study, there are two groups of participants, namely the group having the disease under investigation (disease group) and the group not having the disease (normal group). The researchers identify the variants that occur frequently thus indicating to be the risk effect for the disease group while the protective effect on the control group. The results of the GWAS studies are the risk signs that could be prominent for a specified disease in a group of populations (*Explainer: Genome-Wide Association Studies*, n.d.).

Concerning COVID-19, the COVID-19 Host Genetics Initiative (COVID-19 HGI) was developed by Dr. Andrea Ganna from the Institute of Molecular Medicine in Finland (FIMM) to understand the genetic variation and disease susceptibility/severity in COVID-19 patients. The COVID-19 HGI was developed with the objective of generating, sharing and analyzing data related to COVID-19 genetic determinants. The knowledge could help the researchers with drug repurposing and contribute to the global knowledge of COVID-19 (Andrea Ganna, 2020).

The COVID-19 HGI is a collaboration of 3000 researchers from 46 studies with more than 49,000 patients diagnosed with COVID-19

and 2 million controls. The participants belong to 6 ancestry groups. The major drawback is the participants are majorly belonging to European ancestry (80 %). There were three patient groups with their defined phenotypes:

1. COVID-19 confirmation by a physician or a laboratory or a self-test.
2. Hospitalization with moderate to severe COVID-19.
3. Hospitalization which required respiratory support or dead (Asgari & Pousaz, 2021).

There are many challenges associated with the study of genetic variants:

a. Availability of large cohort size of population
b. Inability of the GWAS statistical methods to deal with the rare variants.
c. Identification of the rare variants and associating it with an exact biological etiology

The summary of the results is provided as a Manhattan plot. It associates the COVID-19 traits with the genetic variants across the entire genome.

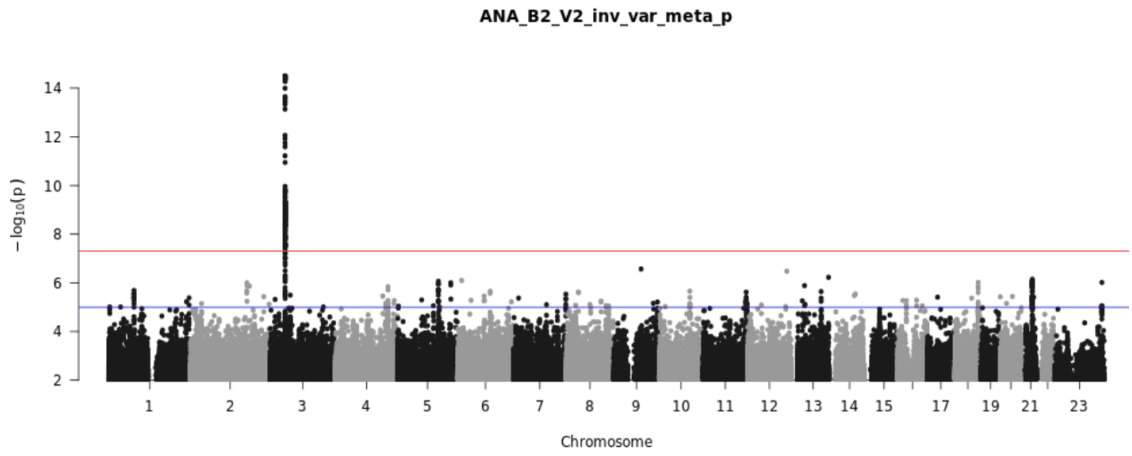

Figure 2.5: Manhattan plot representing a GWAS result. Adapted from (Wolford & Veerapen, 2020)

It contains the chromosome numbers along the X-axis and the log transformed p-value along the Y-axis. The points in the plot denote the p-value between the variant and the disease under study. The most associated genetic variant is the variant that has the highest p-value (Wolford & Veerapen, 2020). The statistical power of the

experiments has resulted from the combined results of all 46 studies.

The COVID-19 HGI results could be directly accessed from the weblink at https://app.covid19hg.org/.

2.9 Biological Expression Language (BEL)

The scientific knowledge available within literature is mostly in the form of long free text and thus capturing the available knowledge is cumbersome. We can overcome by employing the Biological Expression Language (BEL), a computable language that represents the syntactical representation of biological relationships originally designed by Selvanta is a language developed for the representation of knowledge in the life science domains thus paving the way for biocuration (integration of biological data into databases) and enabling in identifying disease mechanisms, pathways, etc. in the literature (Madan et al., 2019).

The other computable languages include Biological Pathways Exchange (BioPAX), Systems Biology Markup Language (SBML), Systems Biology Graph Notation (SBGN). BioPAX is used for capturing the various networks including metabolic, signaling, and genetic interactions. The SBML represents mathematical models of the networks and pathways while the SBGN represents the graphical representation of biochemical and cellular processes (Hoyt et al., 2018).

In BEL language, the statements are transformed into the subject, relation and object. The normalized namespaces for the subject and object are derived from different databases entries such as HGNC for human genes, MGI for mouse genes and ChEBI for chemical entities. The most common biological entities like genes are notated as g(), proteins as p(), mRNA as r().

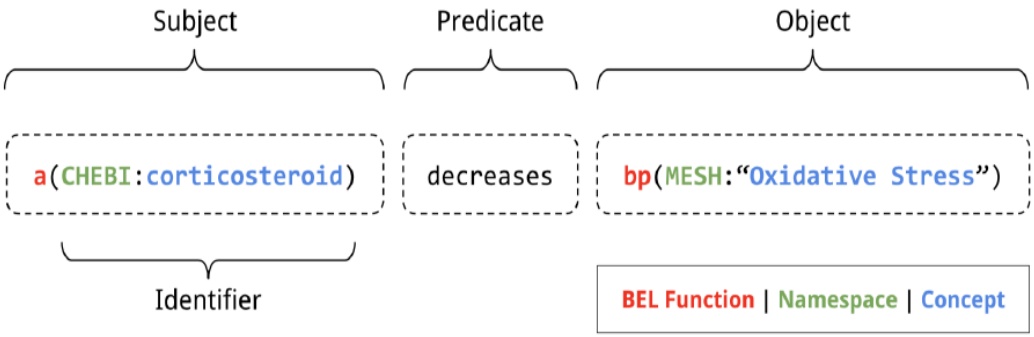

Figure 2.6: Structure of a BEL statement. Adapted from (Madan et al., 2019).

2.10 Knowledge graph

For mining information from large volumes of scientific literature to answer complex queries, a Knowledge Graph (KG) is employed.

A knowledge graph is a directed/non-directed labeled graph consisting of nodes representing entities (e.g., Genes, proteins, chemicals etc.) connected by edges representing the relationship between the nodes (interactions).

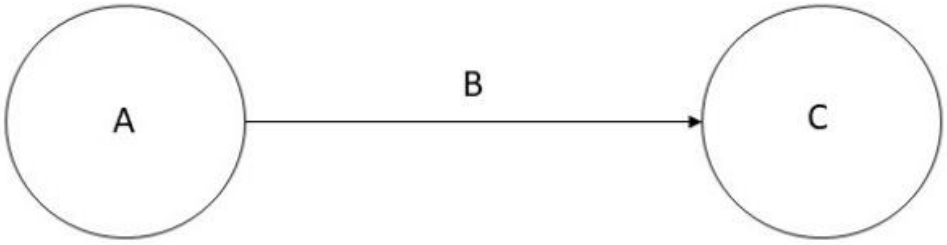

Figure 2.7: A simple Directed labeled graph. Adapted from (What Is a Knowledge Graph?, n.d.)

In the field of life sciences, knowledge graphs are employed in the literature searching for the specific query provided by the user, drug repurposing for identifying drugs effective for the disease etc. (Chatterjee et al., 2021).

2.10.1 COVID-19 Knowledge graph – The Pharmacome

For supporting the COVID-19 research, the scientific community flooded itself with several experiments, hypotheses, disease maps and publications. A single disease map couldn't address the whole understanding of the COVID-19 biology starting from its infection, interaction and pathophysiology. Thus, a unified knowledge graph (COVID-19 Pharmacome) encompassing several public and proprietary knowledge graphs, experimental data was developed (Schultz et al., 2021).

The knowledge graph was specific to COVID-19 drug-target interactions, mechanisms of infection, pathophysiology etc. It consisted of 4,016 nodes (genes, proteins, drugs etc.) and 10,232 relationships as edges (increases, decreases, has_component etc.). The information available within the knowledge graph most pertains to the drug target interactions, proteins and genes associated with COVID-19. Drug target databases such as DrugBank, CHEMBL, PubChem are used for drug target interactions.

The graph was developed from the corpus of 160 scientific literature from various resources including PubMed, LitCovid etc. Then information was manually encoded in BEL which resulted in source, relation, object patterned statements.

A user interface to explore the knowledge graph and query was developed using the Python Django and OrientDB as Biological Knowledge Miner (BiKMi). It is also supported by an API. The web application can be accessed directly for free at https://bikmi.covid19-knowledgespace.de/.

An interesting application of the knowledge graph is the identification of drug candidates for drug repurposing. The drugs, targets and their mechanism within a computable data structure enable a combinatorial treatment approach. The most prominent drugs for the treatment of HIV, Ebola, Malaria like Lopinavir/Ritonavir combination, Remdesivir, Hydroxychloroquine, Oseltamivir respectively are identified within the knowledge graph (Domingo-Fernández et al., 2020).

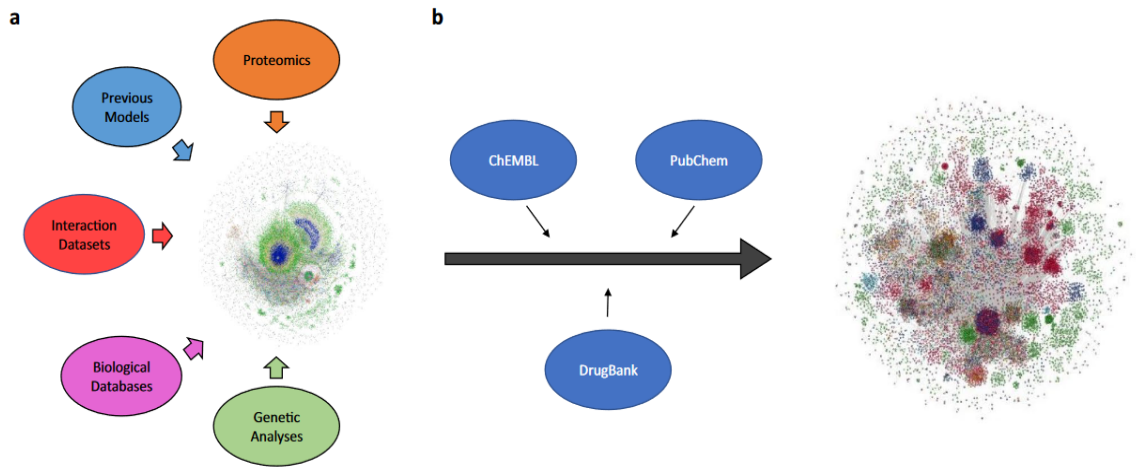

Figure 2.8: COVID-19 Knowledge graph model. Adapted from (Schultz et al., 2021)

3. MATERIALS & METHODS

This chapter describes the various tools and the data used for the analysis.

3.1 Programming Languages and Tools used

The fundamental language used for the analysis is Python. It is an interpreted, object–oriented language, a versatile language known for its application in scientific work, software development, game development, web development, GUI development and system administration (Python Language Reference, 2010). The features of the languages are simple syntax, readability, user-friendly nature, dynamic typing, well maintained documentation and vibrant community forum.

The list of python modules and libraries used includes

- Pandas (McKinney, 2010),
- NumPy (Harris et al., 2020),
- Matplotlib (Hunter, 2007),
- Requests (Chandra & Varanasi, 2015),
- Pycountry (Theune, 2008),
- UpSetPlot (Lex et al., 2014).

3.1.1 Pandas

Pandas is an open source library for working with datasets and implementing statistical models. This library provides “DataFrame” and “Series” as primary data structures with the similar functionality of R to work with tabular data (McKinney, 2010). It is suitable to work with tabular data (Excel spreadsheet or others), time series data and matrix data. Several methods include data cleaning and munging, manipulating the dataframe, handling missing data, date/time handling, creating visualizations and data analysis. It is built on top of Numpy and well suited to integrate into other third party libraries.

3.1.2 NumPy

The primary array programming language in python is NumPy. It is an open source library that provides multidimensional python array objects known as NumPy array. It contains a pointer to memory along with its data type. The elements within the array must be with the same datatype and occupy the same memory. Vectorization aids in manipulating the multidimensional arrays quickly without the absence of *for-loops* and indexing. A plethora of python packages are built using NumPy as a foundation (Harris et al., 2020).

3.1.3 Matplotlib

Matplotlib is a python library built for making 2D plots using NumPy and other libraries. The code for Matplotlib consisted of three parts namely pylab interface, Matplotlib frontend and backend. This library is useful for creating several kinds of visualizations with defined parameters in a few lines of code (Hunter, 2007).

3.1.4 Requests

Requests is a simple Hypertext Transfer Protocol (HTTP) library for python to communicate with a webserver. The simple syntax of the library allows the users to make efficient API calls with minimal effort (Chandra & Varanasi, 2015).

3.1.5 pycountry

A python package for the conversion between country codes, continent names and ISO country names (Theune, 2008).

3.1.6 UpSetPlot

UpSet is a novel visualization technique for the quantitative analysis of sets including their intersections and aggregates obtained by intersections as Venn diagrams become inconvenient when the size of the sets increases. The plot contains a matrix with rows indicating a segment in a Venn diagram and columns

indicating each set. The matrix cells containing filled elements means that they are participating in the intersection. It supports multiple intersections to make aggregates for visualizations. This tool is useful in making interactive visualizations across scientific domains namely genetics, genomics, pharmacology etc (Lex et al., 2014).

3.2 COVID-19 clinical trials

The clinical trials of COVID-19 were retrieved from the ClinicalTrials.gov trials repository. It had around 6029 clinical trials covering different parts of the world. 60% of the obtained trials were interventional and the rest 40% were observational trials. The different entities were clinical trial identifier, name of the region(s), the official title of the study, list of medical interventions.

3.2.1 Clinical trial information using SCAIView API

The clinical trial registry doesn't contain all the information such as medical interventions. It is retrieved using SCAIView API. This was done by selecting the right keywords and the right ontology.

3.3 COVID-19 haplotype specific risk alleles

The COVID-19 haplotype specific risk alleles were obtained from COVID-19 Host genetics initiative databases (HGI). Round 5 was released on 18th January 2021 with three different phenotypes A2 (very severe respiratory confirmed COVID-19 vs population), B1 (hospitalized COVID-19 vs non hospitalized) and C2 (laboratory confirmed COVID-19) covering the countries within the continents including Asia, Africa, Europe, North America and South America.

Each file with the following column identifiers was obtained from the database:

Column headers	**Description**
# CHR	chromosome number
POS	chromosome position
REF	reference allele

ALT	alternative allele
SNP	#CHR:POS:REF:ALT
all_meta_AF	allele frequency
rsid	SNP identifier

Table 3.1: Column headers and description of an HGI result file.

The problem associated with this approach is the non-availability of participants from diverse backgrounds and genetic information linked to the COVID-19 trials.

3.3.1 SNP identifier to Gene identifier conversion

The obtained SNPs from the COVID-19 Host Genetics Initiative Database (HGI) were mapped to their respective HGNC names using the dbSNP variation services API.

3.3.2 Continent Information using pycountry package

The continent information for a clinical study and the HGI result was obtained from the country listed by using the pycountry package.

3.3.3 Integration of the haplotype specific information to the knowledge graph

The data obtained from the COVID-19 HGI were integrated into the COVID-19 Pharmacome by Bruce Schultz who is one of the developers and maintainers.

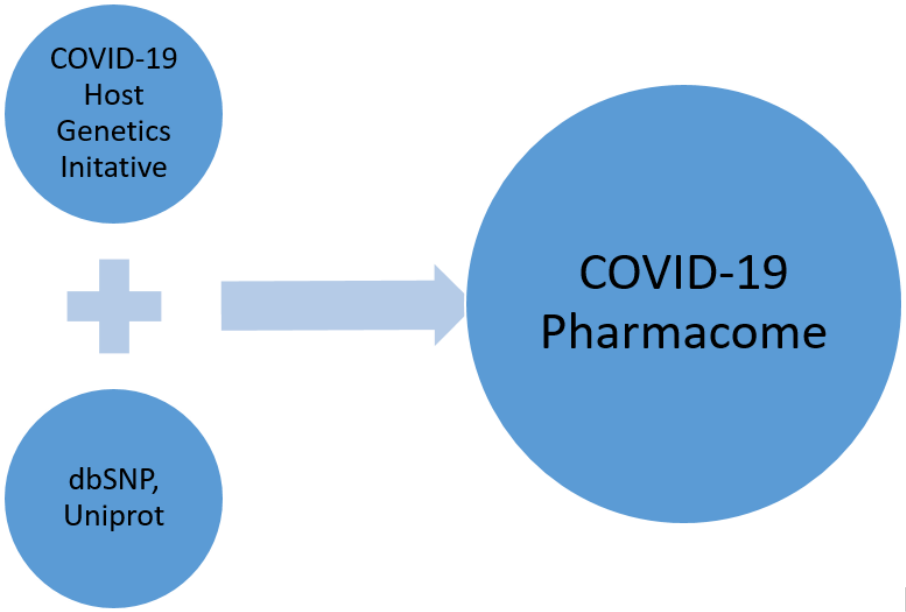

Figure 3.1 Integration of data into Pharmacome

The SNP data from the Pharmacome mapped with its gene name from dbSNP and protein id from Uniprot databases were integrated into the Pharmacome.

3.4 COVID-19 Data Portal

The COVID-19 Data Portal was released on 20th April 2020 to facilitate COVID-19 research by achieving open data sharing. This portal was developed by European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI). It consists of seven main categories including viral sequences, host sequences, expression, proteins, biochemistry, literature and imaging. The data can be obtained by the user interface and API services (Harrison et al., 2021).

3.4.1 COVID-19 Association Score

The COVID-19 Data Portal contains the COVID-19 Association score for each gene target related to the disease. This score is a weighted harmonic sum of association scores of all data source scores. It ranges between 0 and 1. The maximum the score the stronger its association with the disease (Ochoa et al., 2021).

There are 2228 gene targets for which an association score is defined.

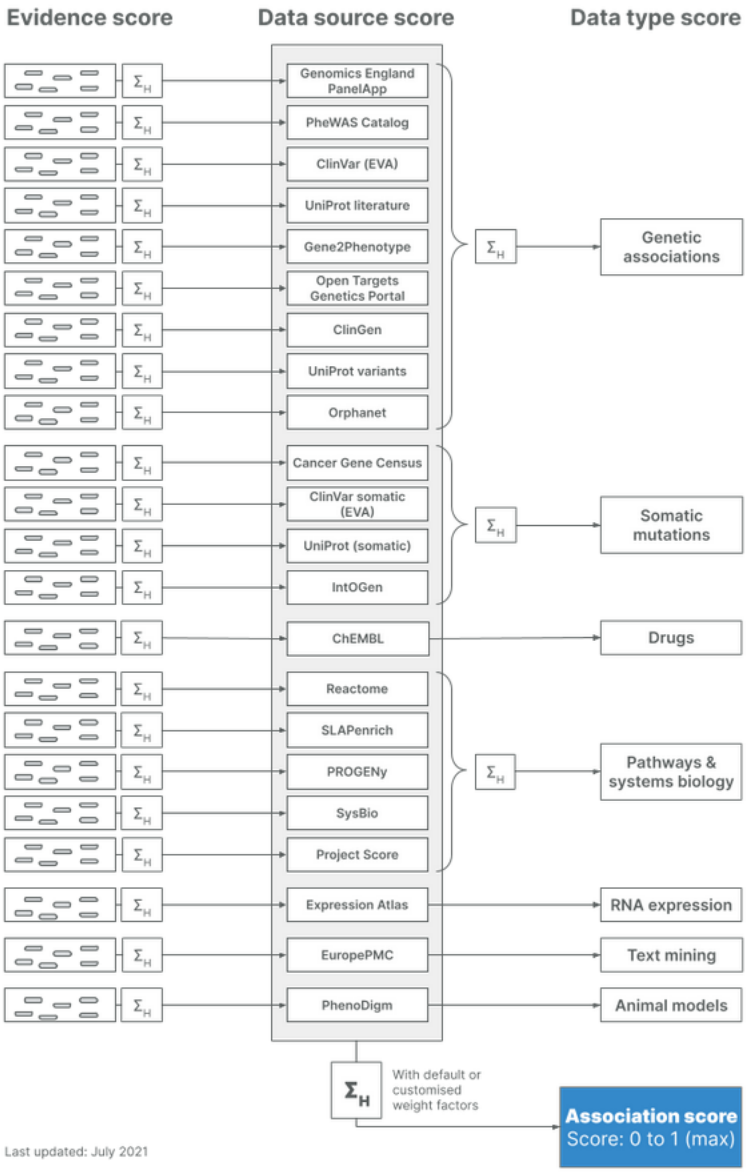

Figure 3.2 Definition of COVID-19 Association Score. Adapted from (Ochoa et al., 2021)

The COVID-19 Association Score for the genes in the Pharmacome is obtained using the API. 918 Pharmacome genes were found in the portal with a maximum score of 1.0 and a minimum score of 0.04.

4. RESULTS AND DISCUSSION

This chapter describes the results obtained in the study with the key discussions.

4.1 COVID-19 HGI results overview

4.1.1 Chromosomes vs SNPs

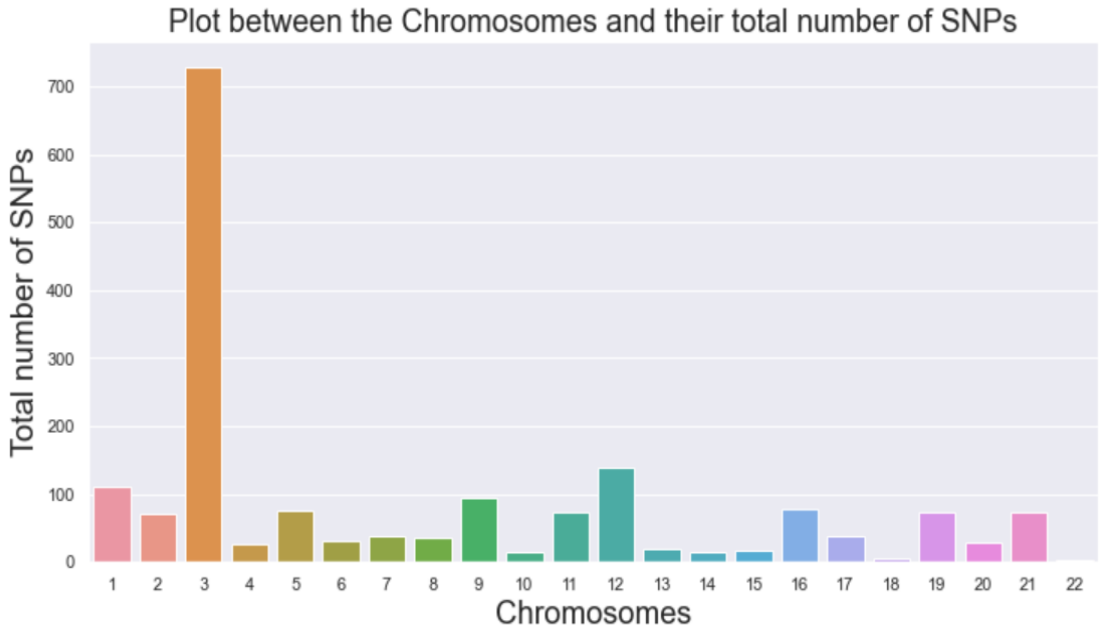

Figure 4.1: Barplot of chromosomes and the total number of SNPs

The comparison of different chromosomes with their respective number of SNPs identified that chromosome 3 has a direct link with the COVID-19 infection. This goes in parallel with the several published studies including the work performed on 72 European patients by an Italian group (Valenti et al., 2021).

4.1.2 Chromosomes vs genes

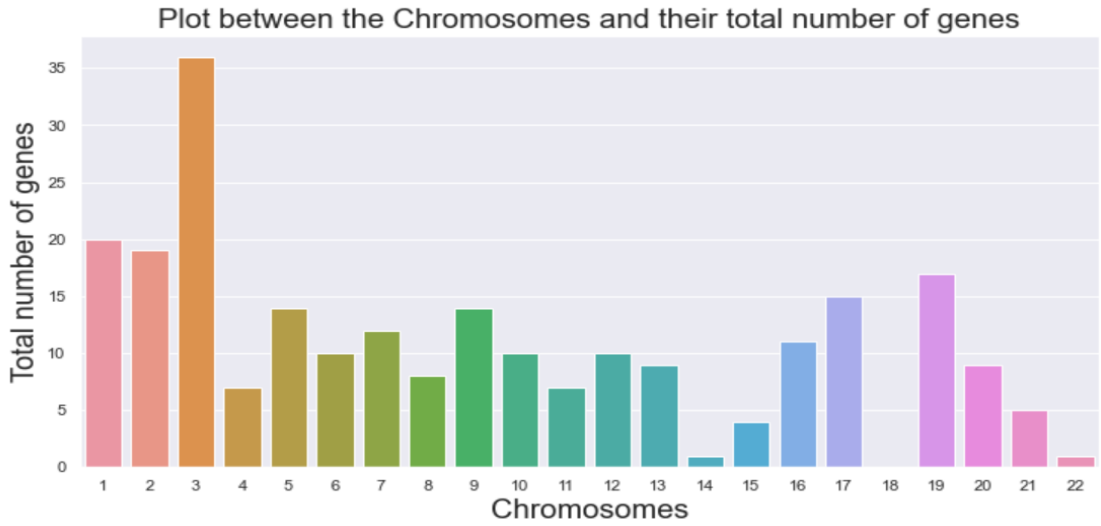

Figure 4.2: Barplot of chromosomes and the total number of genes

The analysis of chromosomes and the SNP mapped HGNC gene name illustrates that chromosome 3 contains the most genes pertaining to COVID-19. Few notable COVID-19 genes that are associated with the COVID-19 risk (Callaway, 2021) such as ABO, SLC6A20, IFNAR2, DPP9, TYK2 are identified in the different chromosomes.

Eventually from the COVID-19 HGI database, the round 5 results yielded 1798 unique SNPs and 239 HGNC genes.

4.2 Generation of ethnicity specific Pharmacomes

The ethnicity specific Pharmacomes were generated based on the assumption that each country participating in the study contributed to the results in an equal fashion. This assumption is because the individual country specific results were not available in the database which is the main disadvantage along with the strong bias on the participants from European ancestry. Eventually, this led to the generation of the following ethnicity specific Pharmacomes namely:

a. North America Pharmacome
b. South America Pharmacome
c. Europe Pharmacome
d. Asia Pharmacome
e. Africa Pharmacome

4.2.1 Ethnicity specific Pharmacomes – Big Picture

The comparison of available different Pharmacomes was done by creating an “Upset” plot which is a tool for visualizing sets of overlapping elements.

Eventually, two such plots were created:

a. Comparison of COVID-19 SNPs and Ethnicity

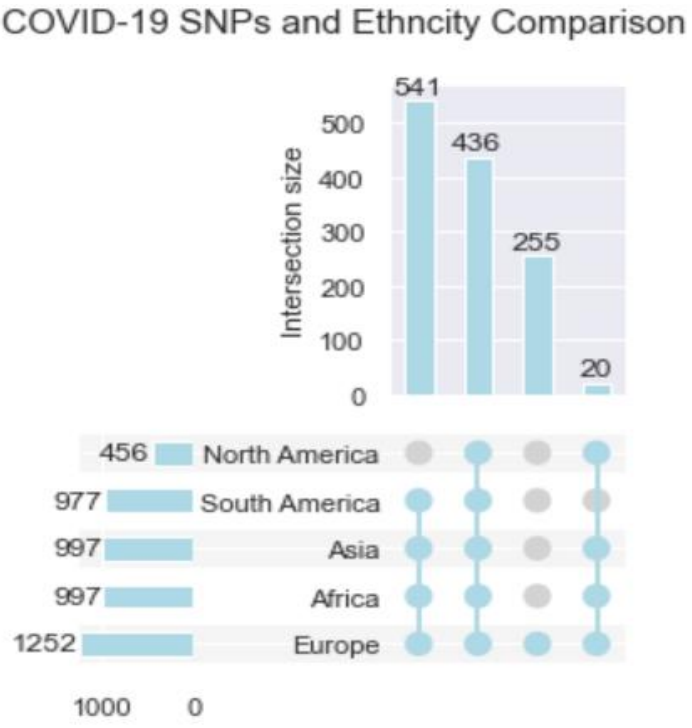

Figure 4.3: Comparison of COVID-19 SNPs and Ethnicity

The barplot represents the different ethnicities in the X axis and the common SNPs present in the Y axis. The number of SNPs that are unique to the European ancestry is 255 while 436 SNPs are present common in all ethnicities. 20 SNPs were common in North American, Asian, African and European ancestry. 541 SNPs were common in South American, Asian, African and European ancestry.

b. Comparison of COVID-19 Genes and Ethnicity

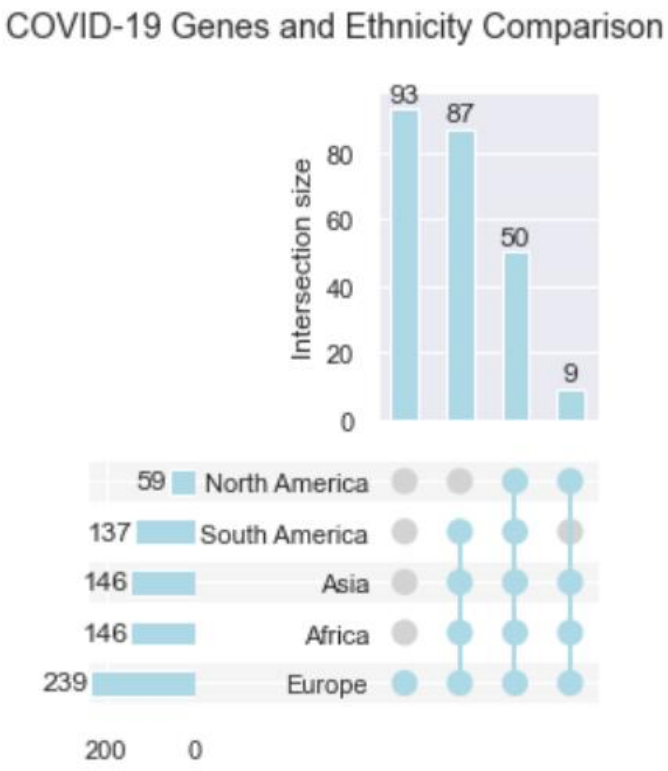

Figure 4.4: Comparison of COVID-19 Genes and Ethnicity

The barplot represents the different ethnicities in the X axis and the common genes present in the Y axis. The number of genes that are unique to European ancestry is 93 while 50 genes are present common in all ethnicities. 9 genes were common in North American, Asian, African and European ancestry. 87 genes were common in South American, Asian, African and European ancestry.

The figures show the overlapping SNPs and genes across the different ethnicities including Asian, European, North American, South American and African.

4.3 COVID-19 Pathway Analysis

The pathway analysis of a subset of Pharmacome genes with the COVID-19 Association score of 1.0 was performed using the Reactome tool (Jassal et al., 2020) and revealed the following top results:

S.No	Pathway name	Pathway identifier	No. of genes involved	p-value
1.	Maturation of spike protein	R-HSA-9694548	19	1.11e-16
2.	Translation of structural proteins	R-HSA-9683701	19	1.11e-16
3.	Translation of structural proteins	R-HSA-9694635	30	1.11e-16
4.	SARS-CoV-1 Infection	R-HSA-9678108	35	1.11e-16
5.	SARS-CoV-2 Infection	R-HSA-9694516	44	1.11e-16

Table 4.1: The top molecular pathways of COVID-19

This indicates that the 47 genes with the COVID-19 Association score of 1.0 were involved in some key molecular pathways of COVID-19 including the maturation of spike protein and translation of structural proteins. Other immunological pathways such as interferon signaling, interleukin signaling and chemokines related pathways were observed in the results.

4.4 Sex Bias of COVID-19

The 8180 COVID-19 genes in the Pharmacome were mapped to the X and Y chromosome. Eventually, 267 genes were mapped to the X chromosome and 3 genes to the Y chromosome. Several crucial genes like ACE2, TLR7, TLR8, IRAK1, interleukin, interferon genes are observed in the Pharmacome that are pertinent to the infection.

Comparison of the male and female patients infected with the COVID-19, it is observed that males are highly affected by the disease in terms of the infection and fatality rate (Peckham et al., 2020).

The overrepresentation analysis of the proteins encoded by the X chromosome from the Pharmacome performed using the Reactome tool (Jassal et al., 2020) revealed the following top results:

S.No	Pathway name	Pathway identifier	p-value
1.	Loss of MECP2 binding ability to 5hmC-DNA	R-HSA-9022534	3.77e-04
2.	Vpr-mediated induction of apoptosis by mitochondrial outer membrane permeabilization	R-HSA-180897	0.001
3.	Loss of MECP2 binding ability to 5mC-DNA	R-HSA-9022538	0.002
4.	Loss of phosphorylation of MECP2 at T308	R-HSA-9022535	0.004
5.	Loss of MECP2 binding ability to the NCoR/SMRT complex	R-HSA-9022537	0.006

Table 4.2: The top molecular pathways of proteins associated with the X- chromosome in COVID-19

The results indicated the X chromosome proteins of COVID-19 are enriched by the pathways involving the gene “MECP2”. It is a methyl –CpG protein, involved in X-inactivation and in several neurological disorders such as Rett Syndrome (Ballestar et al., 2000). Transcriptional induction of MECP2 may be important for the COVID-19 infection (Mamoor, 2020).

This is due to molecular mechanistic differences between the male and the female. The ACE2 gene which is prominent for the recognition of the COVID-19 virus present in the X chromosome avoids the X-inactivation thereby leading to the phenotype differences between the sexes. The genes encoded in the X chromosome and the sex hormone are the base for the varied phenotypes between males and females (Li et al., 2020).

4.5 Recommendations for COVID-19 drug targets

On exploring the drugs in the COVID-19 Pharmacome, it has recorded 5530 drugs. Around 468 drugs could be mapped to the genes that are of X chromosome and no drugs could be mapped to the Y chromosome.

It also featured various prominent drugs that are used today for the treatment of COVID-19 which are tabulated below, although they are not originally designed for the treatment of COVID-19.

S.No	Drugbank ID	Drug Name
1.	DB00207	Azithromycin
2.	DB01611	Hydroxycholoroquine
3.	DB08877	Ruxolitinib
4.	DB01234	Dexamethasone
5.	DB03496	Flavopiridol

Table 4.3: Some Pharmacome drugs used for the COVID-19 treatment

Also, the Pharmacome suggests a number of possible drug targets that could be employed for drug repurposing.

The top ten recommendations for the trial investigations are listed below:

S.No	Drugbank ID	Drug Name	Number of X chromosome targets
1.	DB00398	Sorafenib	17
2.	DB06595	Midostaurin	16
3.	DB08877	Ruxolitinib	16
4.	DB03496	Alvocidib	15
5.	DB12141	Gilteritinib	13
6.	DB15408	Silmitasertib	13
7.	DB09330	Osimertinib	13
8.	DB09063	Ceritinib	12
9.	DB12978	Pexidartinib	12
10.	DB08896	Regorafenib	12

Table 4.4: Top drug recommendations for the drug repurposing

These drug recommendations are based on the number of X chromosome targets those drugs possessed in the Pharmacome. The larger the number of targets, the more worth it could be escalated for the new experimentations.

However, the Pharmacome drug space is completely biased on the drugs targeting the X chromosome genes, the possible recommendation also could be the drugs targeting the Y chromosome genes like SRY which is a gene involved in the male sex determination.

4.6 Major Mechanisms of COVID-19 infection causing loci

The GWAS results of the Host Genetics Initiative team revealed 13 loci that are associated with the infection of COVID-19 (Andrea Ganna, 2021). The risk genes and the neighboring genes in the locus density region were obtained and they were mapped into the COVID-19 Pharmacome.

The overrepresentation analysis of the risk genes performed using the Reactome tool (Jassal et al., 2020) revealed the following top results:

S.No	**Pathway name**	**Pathway identifier**	**p-value**
1.	Interferon alpha/beta signaling	R-HSA-909733	8.46e-09
2.	Interferon gamma signaling	R-HSA-877300	1.25e-05
3.	Interferon signaling	R-HSA-913531	1.34e-05
4.	OAS antiviral response	R-HSA-8983711	1.01e-04
5.	Regulation of IFNA signaling	R-HSA-912694	0.008

Table 4.5: The top molecular pathways of proteins associated with risk genes identified by the HGI

The molecular pathways indicate the immunological pathways that are triggered as a result of host response of COVID-19 infection including the interferon signaling, oligoadenylate synthase (OAS).

5. CONCLUSION AND OUTLOOK

Thanks to the global community for their collaborative effort in the research of COVID-19 as a result which resulted in various data repositories relevant to the COVID-19 such as LitCovid (COVID-19 literature), COVID-19 Data Portal by EMBL- EBI consisting of sequences, structures, expression data and many other.

Thanks to the innovation of the COVID-19 Knowledge graph called Pharmacome. It represents a constantly updated unified knowledge hub encompassing the various aspects for the understanding of the COVID-19 disease including its various pathways mechanisms. This graph not just holds the pre-existing knowledge discovered by the global scientists but also serves as a platform for coming up with new hypotheses such as the new potential drug targets for the drug repurposing.

There were many challenges faced during the project. They are listed as follows:

1. HGI Data availability

   The HGI data of round 5 was highly biased i.e. most of the participants were of European ancestry (~80%) and absolutely no Australian data. This led to the results being completely biased on the available data.

2. Clinical outcomes availability

   The clinical outcomes from the clinical registries were not clearly defined. If a drug “XYZ” is employed in a trial, the registries don’t give direct information whether it is successful or failed (binary/boolean format) so that it could be mapped into the Pharmacome. A large number of the trials have no results recorded as it is still ongoing.

Perhaps, we may end up depending on a sophisticated "trial parser" for coming up with the dedicated results.

This study was designed to understand the different ethnicity specific Pharmacomes and identify unique SNPs in a certain population that may help to explain why the certain population is highly susceptible and why certain drugs work in population "A" but not in population "B". Therefore, we relied on the COVID-19 Host Genetics Initiative to collect the SNP data for the distinct populations. The hypothesis is that there exists a strong association between the SNPs involved in a certain population and the molecular pathways triggered and the outcome of the trial. We built ethnicity specific Pharmacomes namely Asian, African, North American, South American and European. Unfortunately, due to the non-availability of the genetics data, we were not able to go further. As a result, we were unable to associate SNPs and specific ethnicities.

The indications of molecular pathways involved in the COVID-19 infection such as interferon signaling, interleukin signaling, spike protein maturation and so on will help to understand the biology of COVID-19 and build new hypothesis with the aid of the Pharmacome.

The sex specific dominance of COVID-19 targeted by the drugs affecting the X chromosome in the Pharmacome will pave a way for postulating new potential drugs that could be tested in labs and also sex specific clinical trials to meet the demands.

The availability of data from the diversified participants of all ethnic backgrounds will be a foundation to further understand the COVID-19 better in terms of ethnicity specific infection, fatality, medical interventional response etc. This will also lay a foundation for the designing of better clinical trials i.e. medical interventions specific to the ethnicity and specific to the biological sexes.

6. SCIENTIFIC REFERENCES

About ICTRP. (n.d.). Retrieved April 22, 2021, from https://www.who.int/clinical-trials-registry-platform/about

About the EU Clinical Trials Register. (n.d.). Retrieved April 22, 2021, from https://www.clinicaltrialsregister.eu/about.html

About Variants of the Virus that Causes COVID-19. (2021). https://www.cdc.gov/coronavirus/2019-ncov/variants/variant.html

Andrea Ganna. (2020). The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *European Journal of Human Genetics*, *28*(6), 715–718. https://doi.org/10.1038/s41431-020-0636-6

Andrea Ganna. (2021). Mapping the human genetic architecture of COVID-19. *Nature*. https://doi.org/10.1038/s41586-021-03767-x

Arora, G., Joshi, J., Mandal, R. S., Shrivastava, N., & Virmani, R. (2021). Artificial Intelligence in Surveillance , Diagnosis , Drug Discovery and Vaccine Development against COVID-19. *Pathogens*, *10*(8), 1048. https://doi.org/https://doi.org/10.3390/pathogens10081048

Asgari, S., & Pousaz, L. A. (2021). Genetic clues to COVID susceptibility and severity. *Springer Nature*. https://doi.org/10.1038/d41586-021-01773-7

Bakhshandeh, B., Jahanafrooz, Z., Abbasi, A., Goli, M. B., Sadeghi, M., Mottaqi, M. S., & Zamani, M. (2021). Mutations in SARS-CoV-2; Consequences in structure, function, and pathogenicity of the virus. *Microbial Pathogenesis*, *154*(January). https://doi.org/10.1016/j.micpath.2021.104831

Ballestar, E., Yusufzai, T. M., & Wolffe, A. P. (2000). Effects of rett syndrome mutations of the Methyl-CpG binding domain of the transcriptional repressor MeCP2 on selectivity for association with methylated DNA. *Biochemistry*, *39*(24), 7100–7106. https://doi.org/10.1021/bi0001271

Bhatt, A. (2010). Evolution of Clinical Research: A History Before and Beyond James Lind. *Perspect Clin Res*, *1*(1), 6–10.

Britannica. (2021). *Coronavirus*. https://www.britannica.com/science/coronavirus-virus-group

Callaway, E. (2021). *Gene Variants Linked to COVID RISK*. *595*, 346–348. https://media.nature.com/original/magazine-assets/d41586-021-01827-w/d41586-021-01827-w.pdf

Chandra, R. V., & Varanasi, B. S. (2015). *Python Requests Essentials*.

Chatterjee, A., Nardi, C., Oberije, C., & Lambin, P. (2021). Knowledge graphs for covid-19: An exploratory review of the current landscape. In *Journal of Personalized Medicine* (Vol. 11, Issue 4). https://doi.org/10.3390/jpm11040300

Clinical Research Metadata Repository | ECRIN. (n.d.). Retrieved April 22, 2021, from https://ecrin.org/clinical-research-metadata-repository

ClinicalTrials.gov Background. (n.d.). Retrieved April 21, 2021, from https://clinicaltrials.gov/ct2/about-site/background

Coronavirus Variants and Mutations. (2021). The New York Times. https://www.nytimes.com/interactive/2021/health/coronavirus-variant-tracker.html

Domingo-Fernández, D., Baksi, S., Schultz, B., Gadiya, Y., Karki, R., Raschka, T., Ebeling, C., Hofmann-Apitius, M., & Kodamullil, A. T. (2020). COVID-19 knowledge graph: A computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *BioRxiv*, 19–21. https://doi.org/10.1101/2020.04.14.040667

Esakandari, H., Nabi-Afjadi, M., Fakkari-Afjadi, J., Farahmandian, N., Miresmaeili, S. M., & Bahreini, E. (2020). A comprehensive review of COVID-19 characteristics. *Biological Procedures Online*, *22*(1), 1–10. https://doi.org/10.1186/s12575-020-00128-2

Explainer: Genome-Wide Association Studies. (n.d.). Retrieved May 8, 2021, from https://www.broadinstitute.org/visuals/explainer-genome-wide-association-studies

Fu, L., Wang, B., Yuan, T., Chen, X., & Ao, Y. (2020). Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: A systematic review and meta-analysis. *Journal of Infection*, *80*(January), 656–665. https://doi.org/10.1016/j.jinf.2020.03.041

Guo, M., Hu, H., & Lei, X. (2017). Application of ontology technology in health statistic data analysis. *Studies in Health Technology and*

Informatics, *245*, 915–919. https://doi.org/10.3233/978-1-61499-830-3-915

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Harrison, P. W., Lopez, R., Rahman, N., Allen, S. G., Aslam, R., Buso, N., Cummins, C., Fathy, Y., Felix, E., Glont, M., Jayathilaka, S., Kadam, S., Kumar, M., Lauer, K. B., Malhotra, G., Mosaku, A., Edbali, O., Park, Y. M., Parton, A., … Apweiler, R. (2021). The COVID-19 Data Portal: Accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. *Nucleic Acids Research*, *49*(W1), W619–W623. https://doi.org/10.1093/nar/gkab417

Hoyt, C. T. (2020). *How to Build an Ontology: Examples and Guidelines*. https://cthoyt.com/2020/05/12/building-an-ontology.html

Hoyt, C. T., Domingo-fernández, D., & Hofmann-apitius, M. (2018). *Original article BEL Commons : an environment for exploration and analysis of networks encoded in Biological Expression Language*. *3*, 1–11. https://doi.org/10.1093/database/bay126

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Introduction to SCAIView. (n.d.). Retrieved May 7, 2021, from https://www.scaiview.com/en/introduction.html

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., … D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, *48*(D1), D498–D503. https://doi.org/10.1093/nar/gkz1031

Kai, H., & Kai, M. (2020). Interactions of coronaviruses with ACE2, angiotensin II, and RAS inhibitors—lessons from available evidence and insights into COVID-19. *Hypertension Research*, *43*(7), 648–654. https://doi.org/10.1038/s41440-020-0455-8

Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Mohi Ud Din, M. (2020). Machine learning based approaches for detecting COVID-19 using clinical text data. *International Journal of Information Technology (Singapore)*, *12*(3), 731–739. https://doi.org/10.1007/s41870-020-00495-9

Kock-Schoppenhauer, A. K., Kamann, C., Ulrich, H., Duhm-Harbeck, P., & Ingenerf, J. (2017). Linked Data Applications Through Ontology Based Data Access in Clinical Research. *Studies in Health Technology and Informatics*, *235*, 131–135. https://doi.org/10.3233/978-1-61499-753-5-131

Learn About Clinical Studies. (n.d.). Retrieved April 21, 2021, from https://www.clinicaltrials.gov/ct2/about-studies/learn

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248

Li, Y., Jerkic, M., Slutsky, A. S., & Zhang, H. (2020). Molecular mechanisms of sex bias differences in COVID-19 mortality. *Critical Care*, *24*(1), 4–9. https://doi.org/10.1186/s13054-020-03118-8

Madan, S., Szostak, J., Komandur Elayavilli, R., Tsai, R. T. H., Ali, M., Qian, L., Rastegar-Mojarad, M., Hoeng, J., & Fluck, J. (2019). The extraction of complex relationships and their conversion to biological expression language (BEL) overview of the BioCreative VI (2017) BEL track. *Database*, *2019*(1). https://doi.org/10.1093/database/baz084

Mamoor, S. (2020). Induction of Methyl-cpg-binding Domain Proteins in Coronavirus Infection. *OSF Preprints*. https://doi.org/10.31219/osf.io/3xs25

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, *1*(Scipy), 56–61. https://doi.org/10.25080/majora-92bf1922-00a

Mittal, A., Manjunath, K., Kumar, R., Id, R., Kaushikid, S., Kumar, S., & Vermaid, V. (2020). *COVID-19 pandemic: Insights into structure, function, and hACE2 receptor recognition by SARS-CoV-2*. https://doi.org/10.1371/journal.ppat.1008762

Mousavizadeh, L., & Ghasemi, S. (2020). Genotype and phenotype of COVID-19: Their roles in pathogenesis. In *Journal of Microbiology,*

Immunology and Infection. Elsevier Ltd. https://doi.org/10.1016/j.jmii.2020.03.022

Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., Miranda, A., Fumis, L., Carvalho-Silva, D., Spitzer, M., Baker, J., Ferrer, J., Raies, A., Razuvayevskaya, O., Faulconbridge, A., Petsalaki, E., Mutowo, P., MacHlitt-Northen, S., Peat, G., … McDonagh, E. M. (2021). Open Targets Platform: Supporting systematic drug-target identification and prioritisation. *Nucleic Acids Research*, *49*(D1), D1302–D1310. https://doi.org/10.1093/nar/gkaa1027

Peckham, H., de Gruijter, N. M., Raine, C., Radziszewska, A., Ciurtin, C., Wedderburn, L. R., Rosser, E. C., Webb, K., & Deakin, C. T. (2020). Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. *Nature Communications*, *11*(1), 1–10. https://doi.org/10.1038/s41467-020-19741-6

Python Language Reference. (2010). *Python Software Foundation*. Python Language Reference, Version 3.6.1. https://www.python.org/

Qin, C., Zhou, L., Hu, Z., Zhang, S., Yang, S., Tao, Y., Xie, C., Ma, K., Shang, K., Wang, W., & Tian, D. S. (2020). Dysregulation of immune response in patients with coronavirus 2019 (COVID-19) in Wuhan, China. *Clinical Infectious Diseases*, *71*(15), 762–768. https://doi.org/10.1093/cid/ciaa248

Sargsyan, A., Kodamullil, A. T., Baksi, S., Darms, J., Madan, S., Gebel, S., Keminer, O., Jose, G. M., Balabin, H., DeLong, L. N., Kohler, M., Jacobs, M., & Hofmann-Apitius, M. (2021). The COVID-19 Ontology. *Bioinformatics*, *36*(24), 5703–5705. https://doi.org/10.1093/bioinformatics/btaa1057

SARS-CoV-2 Variant Classifications and Definitions. (2021). https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html

Schultz, B., Zaliani, A., Ebeling, C., Reinshagen, J., Bojkova, D., Lage-Rupprecht, V., Karki, R., Lukassen, S., Gadiya, Y., Ravindra, N. G., Das, S., Baksi, S., Domingo-Fernández, D., Lentzen, M., Strivens, M., Raschka, T., Cinatl, J., DeLong, L. N., Gribbon, P., … Hofmann-Apitius, M. (2021). A method for the rational selection of drug repurposing candidates from multimodal knowledge harmonization. *Scientific*

Reports, *11*(1), 1–10. https://doi.org/10.1038/s41598-021-90296-2

Theune, C. (2008). *pycountry*. https://pypi.org/project/pycountry/

Thorlund, K., Dron, L., Park, J., Hsu, G., Forrest, J. I., & Mills, E. J. (2020). A real-time dashboard of clinical trials for COVID-19. *The Lancet Digital Health*, *2*(6), e286–e287. https://doi.org/10.1016/S2589-7500(20)30086-8

Understanding the Clinical Trial Process. (2016). https://mstranslate.com.au/understanding-clinical-trial-process/

Valenti, L., Griffini, S., Lamorte, G., Grovetti, E., Uceda Renteria, S. C., Malvestiti, F., Scudeller, L., Bandera, A., Peyvandi, F., Prati, D., Meroni, P., & Cugno, M. (2021). Chromosome 3 cluster rs11385942 variant links complement activation with severe COVID-19. *Journal of Autoimmunity*, *117*(January), 337–339. https://doi.org/10.1016/j.jaut.2021.102595

What Are Clinical Trials and Studies? (n.d.). Retrieved April 21, 2021, from https://www.nia.nih.gov/health/what-are-clinical-trials-and-studies

What is a Knowledge Graph? (n.d.). Retrieved May 8, 2021, from https://web.stanford.edu/class/cs520/2020/notes/What_is_a_Knowledge_Graph.html

Wolford, B., & Veerapen, K. (2020). *COVID-19 HGI Results for Data Freeze 4*. https://www.covid19hg.org/blog/2020-11-24-covid-19-hgi-results-for-data-freeze-4-october-2020/

Worldometer. (2021). *COVID Live Update - Worldometer*. https://www.worldometers.info/coronavirus/

Zhao, J., Cui, W., & Tian, B. (2020). The Potential Intermediate Hosts for SARS-CoV-2. *Frontiers in Microbiology*, *11*, 2400. https://doi.org/10.3389/fmicb.2020.580137

Zheng, M., & Song, L. (2020). Novel antibody epitopes dominate the antigenicity of spike glycoprotein in SARS-CoV-2 compared to SARS-CoV. *Cellular and Molecular Immunology*, *17*(5), 536–538. https://doi.org/10.1038/s41423-020-0385-z

7. DECLARATION OF AUTHORSHIP

I herewith certify that this material is my own work, that I used only those sources and resources referred to in the thesis, and that I have identified citations as such.

Bonn, October 13, 2021

Ram Kumar Ruppa Surulinathan