BioMed

MDPI

Article

Interoperability of COVID-19 Clinical Phenotype Data with Host and Viral Genetics Data

Anna Bernasconi * and Stefano Ceri

Department of Electronics, Information, and Bioengineering, Politecnico di Milano, 20133 Milan, Italy; stefano.ceri@polimi.it

* Correspondence: anna.bernasconi@polimi.it

Abstract: The outbreak of the COVID-19 epidemic has focused enormous attention on the genetics of viral infection and related disease. Since the beginning of the pandemic, we focused on the collection and integration of SARS-CoV-2 databases, which contain information on the structure of the virus and on its ability to spread, mutate, and evolve; data are made available from several open-source databases. In the past, we gathered experience on human genomics data by building models and integrated databases of genomic datasets (representing, e.g., mutations, gene expression profiles, epigenetic signals). We also coordinated the development of a data dictionary describing the clinical phenotype of the COVID19 disease, in the context of a very large consortium. The main objective of this paper is to describe the content of the data dictionary and the process of data collection and organization. We also argue that—in the context of the COVID-19 disease—interoperability between the three domains of viral genomics, clinical phenotype, and human host genomics is essential for empowering important analysis processes and results. We call for actions that could be performed to link these data.

Keywords: COVID-19; SARS-CoV-2; data interoperability; data dictionary; viral genomics; host genetics; clinical phenotypes

Citation: Bernasconi, A.; Ceri, S. Interoperability of COVID-19 Clinical Phenotype Data with Host and Viral Genetics Data. *BioMed* **2022**, *2*, 69–81. https://doi.org/10.3390/biomed2010007

Academic Editor: Wolfgang Graier

Received: 27 December 2021
Accepted: 25 January 2022
Published: 27 January 2022

1. Introduction

The outbreak of COVID-19 has presented novel challenges to the research community, pushed by the intent of rapidly mitigating the pandemic effects. During these times, we have observed the production of an exorbitant amount of data; the total number of sequences of SARS-CoV-2 available worldwide went from few hundreds in March 2020, up to about one hundred thousand in August 2020, and more than 5 million in December 2021. Inspired by our work on genomic data integration [1,2], we searched for effective ways to help investigate the new phenomenon. We produced concise models to understand and organize data as a basis for building search, visualization, and analysis systems.

In this paper, we overview human and viral genomic data systems in place (Section 2), clarifying our position in the COVID-19 data modeling and tool development community; then, we explore the conceptual model proposed for capturing COVID-19 disease phenotype information within an important global initiative and briefly discuss currently available alternatives (Section 3). Finally, we overview efforts that link COVID-19-related datasets on the viral/host genotype and phenotype levels (Section 4). Limitations of the possible solutions are discussed before we conclude (Section 5).

2. Background

Understanding viruses from a conceptual modeling perspective is very important. In April 2020, we designed the Viral Conceptual Model (VCM, [3]): the sequence of the virus is the central entity, described by three views regarding (i) the information on the

virus and on the infected host; (ii) details of the technology and process used for extracting sequences; (iii) metadata on the project and laboratory managing the sampling, sequencing, and analysis pipelines. Additionally, we modeled sequences' annotated parts (known genes, coding and untranslated regions, etc.) and their nucleotide/amino acids mutations, computed with respect to the reference sequence of the viral species. An enriched model—resulting from the ontological unpacking of the initial VCM—was also proposed [4]. We then described an abstract model that allows representing both the data (thus embedding the VCM) and the external knowledge that is being collected about SARS-CoV-2. This includes notions on variants, their effects (in terms of disease severity, transmissibility, vaccine escape, etc.), their composition (in terms of sets of mutations), the peculiarities of mutations due to their original and alternative nucleotide or amino acid residues, and the definition of particular regions of the genome with given functions. The proposal is preliminarily sketched in CoV2K [5], but we are working towards a much broader definition, allowing performing a targeted search that interconnects data and knowledge.

After our modeling effort, we built solid pipelines for extracting data from the original deposition portals and integrating them within our global model VCM. In doing so, the most difficult aspect we had to consider was the growth of data (which reached more than 5 million genomes in December 2021, in only 1.5 years). Such data continuously need to be mastered by increasingly powerful computing resources with several logical and physical optimizations. ViruSurf [6] is our first system, designed for collecting sequences from the two biggest—completely open—SARS-CoV-2 data sources, e.g., GenBank [7] and COG-UK [8]. We also implemented the dual system ViruSurf-GISAID, storing sequences from GISAID [9], currently hosting the major deposition database (accessing GISAID data is subject to a Data Agreement that is typically granted to users from research institutes and academy). ViruSurf offers a practical interface where each drop-down menu is a metadata attribute describing viral sequences. Possible values are paired with the number of available sequences in the database. Different conditions can be built on the presence of specific mutational patterns (predicating on either nucleotides or amino acid residues). EpiSurf and EpiSurf-GISAID [10] are companion systems for analyzing sequences mutations in the context of specific viral genomic regions, i.e., epitopes. Epitopes, extracted from the Immune Epitope Database (https://www.iedb.org/, last accessed 26 January, 2022), are strings of amino acid residues from a virus protein that can be recognized by antibodies or other host's receptors. In the mentioned interfaces, the results are produced as browsable tables of sequences and epitopes, described by their metadata. They can be downloaded as textual files that are easily embedded in bioinformatic pipelines. A more visual and interactive support is provided by VirusViz [11] can be opened directly from sets of sequences defined within ViruSurf or EpiSurf, as well as from a user-input file of sequences. VirusViz allows partitioning the population of interest into groups and comparatively visualizing their mutation distributions, with several options for highlighting positions, mutational patterns, and regions of interests. VirusViz has a dual close-source solution, i.e., VirusLab [12], commercialized by the Quantia Consulting S.R.L. company and developed in collaboration with our group at Politecnico di Milano, within the EIT project N. 20663. We have just finalized ViruClust [13], a tool for comparing SARS-CoV-2 genomic sequences and lineages in space and time without any computational background. We are currently developing VariantHunter (https://github.com/DEIB-GECO/Docker_VariantHunter, last accessed 26 January, 2022), an application that indicates possible emerging variants, and CoV2K-API (http://gmql.eu/cov2k/api/redoc, last accessed 26 January, 2022), a flexible API for exploring the interplay between SARS-CoV-2 data and knowledge—an ever-growing source of information for variants, their effects, and genetic characteristics.

All mentioned systems are part of the broad architecture illustrated in Figure 1, showing different areas: data sources (orange background), data bases (pink), data models (blue), and systems (gray), such as web applications and tools. Our work on viral resources was very efficient thanks to our background in human genomics. We previously

proposed another conceptual model focused on human genomics [1], based on a central entity representing files of genomic regions, similarly described from various dimensions. We next developed and implemented an integrated database, searchable through the GenoSurf interface [2]. The methods for data loading, integration, and cleaning of viral sequences were adapted from META-BASE [14], a pipeline for data ingestion developed for GenoSurf. Datasets are organized by using the Genomic Data Model [15], which couples the outcome of a biological experiment with clinical and biological information of the studied sample. Open data are retrieved from different sources such as ENCODE [16], TCGA [17], Roadmap Epigenomics [18], and 1000 Genomes [19] and can be used for answering complex biological queries with the GenoMetric Query Language system [20]. The models and systems are general enough to consider many different signals of the human genome, including studies that may be useful to represent COVID-19-related problems.

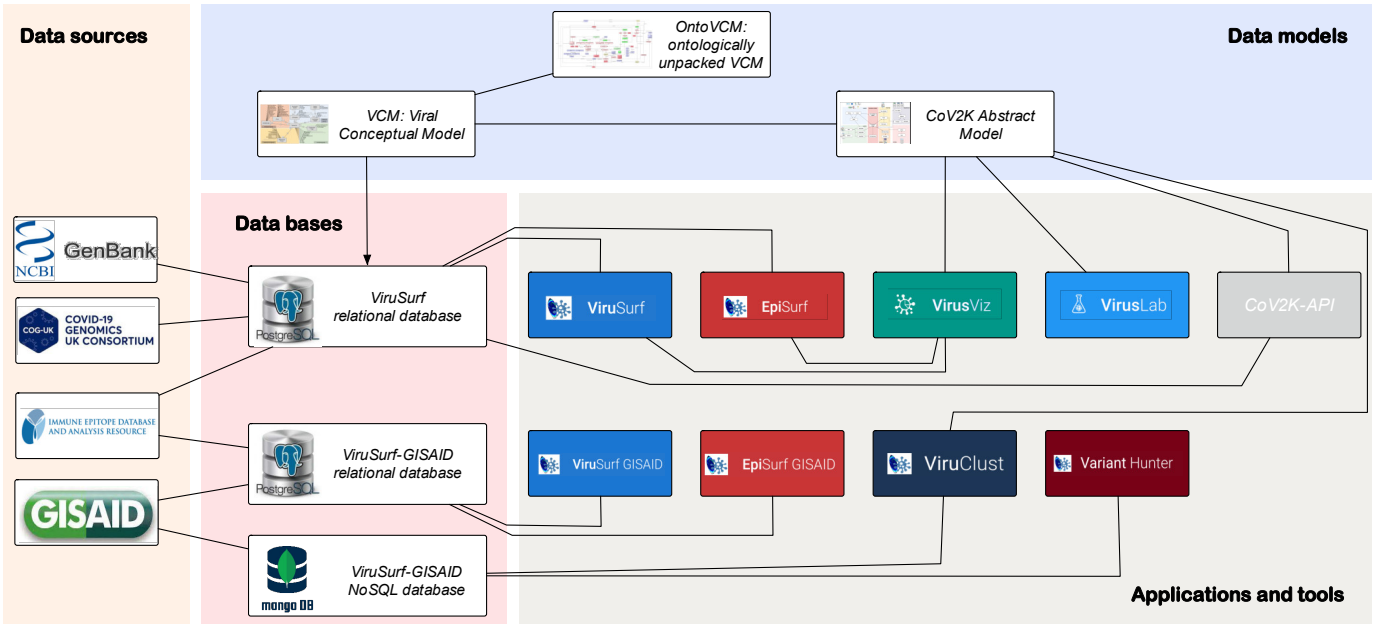

Figure 1. Comprehensive SARS-CoV-2 data sources, models, bases, and tools overview. Undirected links in the schema represent relationships of use between the different actors. For instance, the data extracted from the GenBank source are loaded within the ViruSurf relational database, which is queried by systems such as ViruSurf. In turn, VirusViz, VirusLab, ViruClust, and CoV2K-API employ abstractions defined in the CoV2K abstract model. VariantHunter is the most recent tool, directly using GISAID data retrieved by users that have a specific data access agreement.

3. Clinical Aspects of COVID-19

Several efforts have consolidated practices to gather electronic health record (EHR) and clinical information about COVID-19 patients. Among these, we mention: the COVID19 Research Database, with a schema (https://covid19researchdatabase.org/commondata-schema/, last accessed 26 January, 2022) that includes individuals with claims or electronic records data; the ORCHESTRA network (https://orchestra-cohort.eu/, last accessed 26 January, 2022), focused on building cohorts for clinical, immunological, and epidemiological studies—but also comprising one Work Package on "Biobanking, genomics, and virus–host interaction"; the Consortium for Clinical Characterization of COVID-19 by EHR (4CE [21], https://covidclinical.net/, last accessed 26 January, 2022) promoting EHR data-driven studies of the COVID-19 pandemic. Finally, the COVID-19 Clinical Research Coalition [22] (https://covid19crc.org/, last accessed 26 January, 2022) aims to accelerate COVID-19 clinical research in resource-limited settings—primarily in Africa, Asia, and Latin America.

Several efforts devoted to systematizing clinical data collection and harmonization have been proposed by international organizations (e.g., WHO [23]), national projects (e.g., AllOfUs [24]), or private companies (e.g., 23AndMe, https://www.23andme.com/, last accessed 26 January, 2022).

The main challenges that need to be addressed at the data modeling level correspond to: (i) associating concepts to terminology standard codes provided by organizations (e.g., SNOMED, LOINC, ATC, and ICD); (ii) building appropriate questionnaires to collect data (WHO has provided some guidelines for Case Report Forms at https://www.who.int/teams/health-care-readiness-clinical-unit/covid-19/data-platform, last accessed 26 January, 2022); (iii) building cohorts with statistical significance [25]; (iv) defining phenotypes (shared set of phenotypes to be combined with genomic data for standard GWAS and further meta-analysis).

The collection of homogenized data allows a series of studies (e.g., [26–28]). However, they become more powerful when associated with genetic information of the human host or of the virus, as discussed next.

3.1. The COVID-19 Phenotype Data Dictionary

In addition to the already cited cooperative efforts, we wish to mention the COVID-19 Host Genetics Initiative [29] (https://www.covid19hg.org/, last accessed 26 January, 2022), which aims at gathering an open community of thousands of researchers who produce, share, and analyze data to learn the genetic determinants of COVID-19 susceptibility, severity, and outcomes. Within this international group, of which we share all motivations, we engaged in the design, structuring, and harmonization of a comprehensive data dictionary to help with the submission of individual-level data. For the collection of requirements and cooperative design of the dictionary, we employed a Slack channel dedicated to "covid19-hg-phenotypes" with 1489 members, started on 18 March 2020 by the leaders of the Initiative. The channel hosted 210 posts, plus all the derived public and private replies between active members. In this process, we coordinated about 50 clinicians for cooperatively designing the patients' clinical phenotype definition. The phenotype refers to severe patients who were hospitalized; it has about 200 clinical variables that have been progressively consolidated and annotated, describing demographics, exposure, risk factors, co-morbidities, hospitalization admission and course, and longitudinal encounters with symptoms, treatments, and lab data.

The data dictionary was released on 16 April 2020 (FREEZE 1) and updated on 16 August 2020 (FREEZE 2); both versions are available at http://gmql.eu/phenotype/ (last accessed 26 January, 2022); genetic and related clinical phenotype data are currently being collected and hosted by EGA [30], the European Genome-Phenome Archive of EMBL-EBI. The initiative recommends clinical phenotype data to be submitted following the mentioned data dictionary [31,32]. It has already collected a considerable amount of results, currently reaching ~9.4 K critically ill cases, 25 K hospitalized cases, 125 K reported cases of SARS-CoV-2 infection with almost 3 M controls, as an update [33] of the flagship statement published in Nature at the beginning of 2021 [32], where only 6 K critically ill cases, 13 K hospitalized cases, 50 K reported cases, and 2 M controls were considered.

3.1.1. Cooperative Construction of the Dictionary

An initial draft of the survey was inspired by about 15 COVID-19 questionnaires used across different studies, including the UCSF CHIRP clinical intake, the Canada CanPATH questionnaire, the Columbia University COVID-19 questionnaire, the case report form for Confirmed Novel Coronavirus COVID-19 by WHO, handouts of NIH Intramural, 23andMe, all of U.S. and Helix research programs, the University of Michigan and Universities of Chile COVID-19 surveys, and the Finnish institute for health and welfare. A document was circulated among all groups participating in the initiative, requesting active feedback from those who were designing studies that required to record patient in-

formation. A first COVID-19 Phenotype Definition was drafted. Then, variables were organized in sets that correspond to specific clinical questions. In the following, we will outline the principles that were used to evolve the first draft into the FREEZE 1 and FREEZE 2 versions; we acted as moderators of the process.

General Principles. Every variable had a unique Variable Name that could be used to cross-relate variables across multiple studies. Variables were grouped according to the clinical question that they related to; some groups described general variables that should always be present. Categorical variables had a list of possible values. Variables with plural names were used to indicate that they could have multiple values. Variables could be entered by a Contributor (P = Patient, MD = Healthcare Professional, Any) and were divided into the categories One-Time (associated with the patient, never changing) and Visit-Time (associated with a visit/follow-up event, progressively collected after the first patient identification).

Identification. We noted that, without an identifier, patients cannot be related to an organization, nor their clinical progression can be traced. We thus considered the important issue of creating identification mechanisms for both patients and organizations:

- For the hospital taking care of the patient, we chose a simple identification, i.e., the pair country code–city code of the hospital phone contact. Note that this strategy left space for possible extensions, e.g., contributors could add a number or ZIP code for the cities having multiple hospitals or a doctor ID for the contributors who would enable doctors to collect records in the territory.
- For the anonymized patient identifier, we assumed that each contributor would provide her method. When such information was omitted, all the records input for a single patient were treated as uncorrelated.

Document construction: The process of data contribution was very open and collaborative. A "dataspace" was available to contributors, that could add clinical questions or add variables or even add values to existing variables but could not change or drop its content without moderation and should avoid adding variables about overlapping topics or values with overlapping meaning with respect to existing values.

3.1.2. Proposed Model

The final dictionary is illustrated by the entity relationship diagram [34] in the green central rectangle of Figure 2; the Patient is the central concept, whose phenotype information is collected at admission and during the course of hospitalizations, hosted by a given Hospital. For ease of visualization, attributes are clustered within Attribute Groups, and attributes within groups can be further clustered within Subgroups, denoted by white squares (for brevity, these are not further expanded into specific attributes). Groups connected directly to the PATIENT describe:

- Demography&Exposure, including (i) demographic information about ancestry, height, weight, current pregnancy, the highest educational level, physical demands of the job, and vicinity to minors; (ii) COVID-19 exposure information regarding carriers, travels, or medical professional work.
- RiskFactor, such as smoking, alcohol or other substances habits.
- Comorbidity, of many different kinds, outlined by the subgroups *ImmuneSystem* (e.g., HIV, CD4+T cell count, organ transplant), *Respiratory* (e.g., asthma, cystic fibrosis, sleep apnea), *Renal* (e.g., chronic kidney disease), *CardioVascular* (e.g., hypertension, stroke, bypass), *Neurological* (e.g., dementia or neurological/neuropsychiatric disease), *Cancer* (e.g., leukemia, lymphoma, malignant solid tumor), *RareDisorder*, and *DigestiveOrgan*.
- AdmissionSymptom, with admission date, results of COVID-19 test on a given date, and a comprehensive description of *SymptomsDef* at the time of admission (first point of a longitudinal study), including, for instance, cough, fever, temperature, chest pain, nausea, etc.

- HospitalizationCourse, describing the situation of the Patient at the discharge date, including its cause and—possibly—information about the ICU stay. The subgroup *CriticalConditions* includes several parameters. For numerical values, the worst value during the whole course of hospitalization was requested, including, e.g., respiratory rate and frequency, a series of concentration levels (blood oxygen saturation SpO_2%, PaO_2%, FiO_2%), days on ventilation, duration of pneumonia, septic shock, or organ failure.

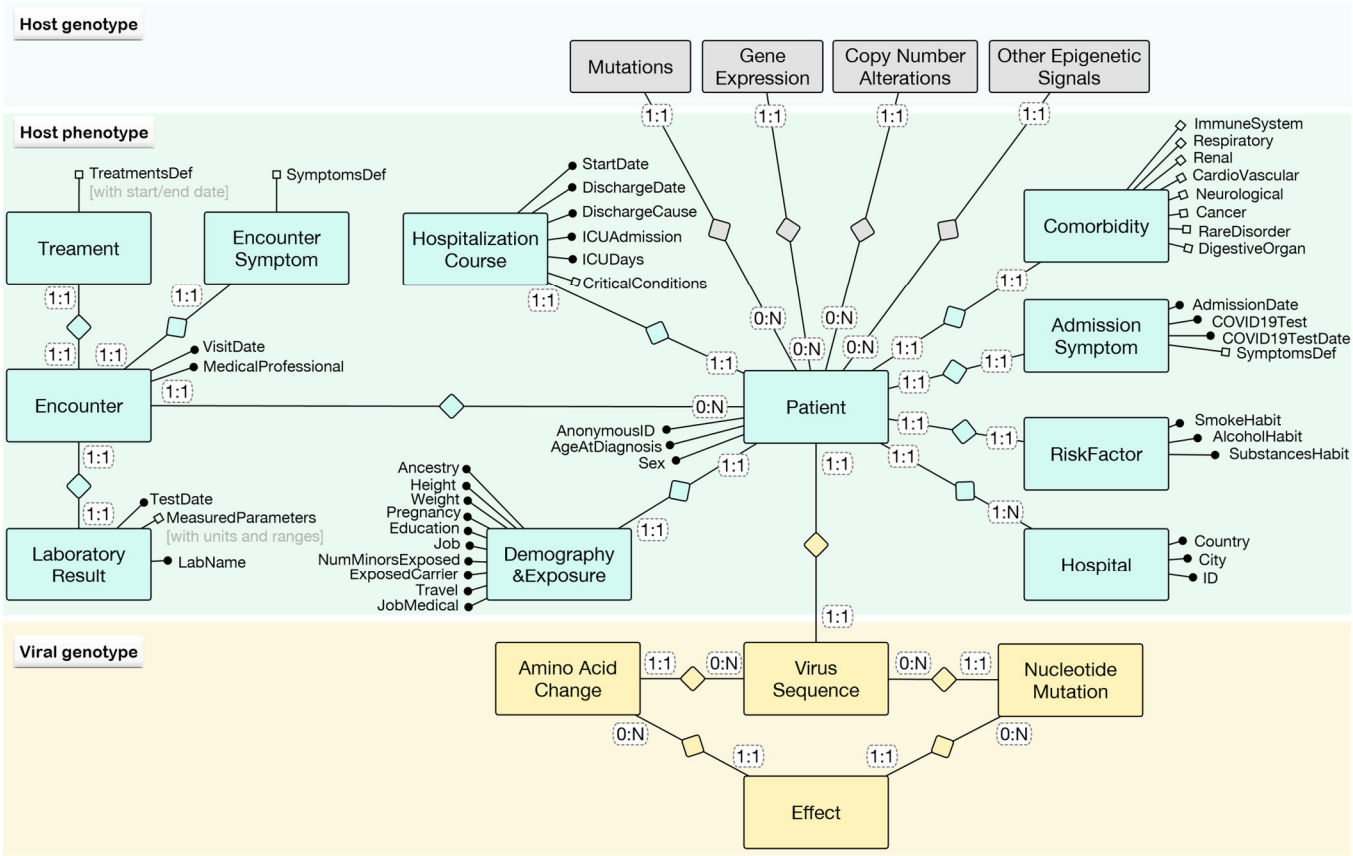

Figure 2. Schema of patient phenotype for a viral disease linked to heterogeneous genomic information and to the sequence of the infecting virus.

Each patient is characterized by multiple instances of Encounter; attribute groups of encounters describe an EncounterSymptom (with their subgroup *SymptomsDef*), a Treatment (with a *TreatmentsDef* subgroup that has a start and an end date), and a LaboratoryResult (with a test date, name of the laboratory, and a subgroup of *MeasuredParameters* with specific measure units and their typical ranges).

Researchers can extract the patient phenotype and differentiate cases and controls in a number of ways. For example, one of the COVID-19 Host Genetics Initiative analyses discriminates between mild, severe, or critical COVID-19 disease severity based on a set of EncounterSymptoms and HospitalizationCourse conditions, whilst another analysis distinguishes cases and controls based on Comorbidities and AdmissionSymptoms.

4. Human/Viral Genomics Interoperability

4.1. Host Genotype and Host Phenotype

To find important genotype–phenotype correlations, well-defined phenotypes need to be ascertained in a quantitative and reproducible way [35]. Since the very first months of the COVID-19 pandemic, large efforts are being conducted for linking the genotype of the human host to the COVID-19 phenotype.

Genome-wide association studies (GWAS) [36] and multi-omic approaches [37] (e.g., gene expression and proteomics) can uncover common variants and networks underlying the host response. In particular, GWASs have reached interesting results [38], associating human genetic variants with severe COVID-19 [39], COVID-19 requiring intensive care-unit admission [40], or respiratory failure (the 3p21.31 gene cluster and the ABO locus, a marker identifying type A blood [41]). Zeberg and Paabo focused on genomic regions that protect against severe COVID-19 [42], finding that the gene cluster identified in [41] is located in an area of chromosome 3 inherited from Neanderthals, possibly explaining statistical differences pointing to population groups (European, Asian) that are hit harder by COVID-19 than others. Exploiting the UK Biobank data (https://www.ukbiobank.ac.uk/, last accessed 26 January, 2022) several studies have been conducted, showing, e.g., variants that increase the risk of COVID-19-related mortality [36], lifestyle factors such as obesity—associated with impaired pulmonary functions—that are related to illness severity [43], specific genotypes (ApoE e4e4) associated with COVID-19 test positivity [44], and the impact of sex-related genetic differences on the disease [45,46].

Several international organizations were set up to allow the agreement between researchers and studies around the world. The COVID-19 Host Genetics Initiative has contributed several important findings among the mentioned ones [40,41]; in their flagship paper [32], they described three GWAS meta-analyses comprising 49,562 COVID-19 patients from 46 studies across 19 countries worldwide. The results included 15 genome-wide significant loci associated with SARS-CoV-2 infection or severe illness, several in genes previously associated with autoimmune and inflammatory diseases. The implications of these findings for treatment or prevention will require further evaluation.

Other initiatives contributing to studying the host genetics related to COVID-19 include GenOMICC (https://genomicc.org/, last accessed 26 January, 2022), ISARIC4C (https://isaric4c.net/, last accessed 26 January, 2022), BRACOVID, Gen-COVID (https://sites.google.com/dbm.unisi.it/gen-covid/, last accessed 26 January, 2022), and the COVID Human Genetic Effort [47] (https://www.covidhge.com/, last accessed 26 January, 2022). The latter is focused on the role of single-gene, inborn errors of immunity in severe COVID-19; it has been found that altered activity of the type 1 interferon system—due either to variants of component genes or to antibodies to their products—is involved in perhaps 10% of severe COVID-19 pneumonia [48].

The most recent results concern the definition of phenotypes related to COVID-19 outcomes (focusing on susceptibility, given exposure, mild clinical manifestations, and an aggregate score of symptom severity) [49] by the AncestryDNA Science Team [50]. Evidence has been gathered that the expression of the ACE2 gene has an influence on COVID-19 risk and is also predictive of severe disease [51]. The GenOMICC (Genetics Of Mortality In Critical Care) performed GWAS in 2244 critically ill patients with COVID-19 from 208 UK intensive care units, able to identify robust genetic signals that relate to key host antiviral defense mechanisms and act as mediators of inflammatory organ damage in patients with the disease [40].

COVID-19 hospitalized patients have been characterized by means of clinical and molecular parameters [52]. An organized and systematic approach to biobanking [53] has allowed a broad coverage of clinical and genetic data for the GEN-COVID Multicenter Study, massively advancing COVID-19 research. A post-Mendelian genetic model has been proposed in [54], designing an Integrated PolyGenic Score to measure the combined effect of common and rare variants. Outcome severity has been associated with blood type and a gene-rich locus on chr3p21.31 in [55]; life-threatening COVID-19 has been connected to inborn errors of type I IFN immunity [48]; other conditions have been linked to male and female characteristics [56–58], from an age-dependent [59] or an age-independent [60] perspective. COVID-19 susceptibility genetic factors have been studied using symptom-based case predictions [31].

Although research on COVID-19 host genomics is still in progress, all these collaborative efforts show that the human genetics scientific community has been able to gather

from around the world to jointly address the pandemic. As a consequence, the identification of new host genetic factors associated with COVID-19 is driven and encouraged by new collaborative platforms, data sharing, joint analyses, and publications of findings.

From the point of view of open data, Gene Expression Omnibus [61] provides nearly 900 samples on SARS-CoV-2 (search query https://www.ncbi.nlm.nih.gov/geo/browse/?view=samples&search=sars-cov-2, last accessed 26 January, 2022). In Figure 2, see the grey area, where each PATIENT can be connected to her genome information, possibly genomic samples describing single-nucleotide MUTATIONS, GENEEXPRESSION profiles, COPYNUMBERALTERNATIONS, or OTHEREPIGENETICSIGNALS.

4.2. Viral Genotype and Host Conditions

GISAID [9], the most adopted database for viral sequence depositions, provides an annotation type regarding the "patient status" (e.g., "ICU; Serious", "Hospitalized; Stable", "Released", "Discharged") for a restricted number of sequences (only 4% of the entire collection in December 2021). In addition, 2019nCoVR [62] has only stored 208 clinical records (https://bigd.big.ac.cn/ncov/clinic, last accessed 26 January, 2022) related to specific sequences since the beginning of the pandemic; these include key-value pairs about the onset date, travel/contact history, clinical symptoms, and tests. Some early findings connected virus sequences with human phenotype, including very small datasets (e.g., [63–66]. We are currently not aware of open data sources or datasets sharing viral sequences linked to the phenotype data of the host organism. Such combination would enable interesting and comprehensive queries concerning the impact of sequence variants on the clinical phenotype of patients affected by COVID-19.

From a data-driven perspective, the only information that is shared concerns the general effects of mutations on epidemiological or immunological aspects of the SARS-CoV-2 infection, without a link to specific patients. This general information is collected, for instance, by COG-UK Mutation Explorer [8] (http://sars2.cvr.gla.ac.uk/cog-uk/, last accessed 26 January, 2022) in tabular format for each specific mutation and by CoVariants [67], CDC [68], ECDC [69] in relation to given Variants of Concern or Interest or to variants that have raised the research community attention. These are provided in textual form. Our own work [5] has attempted to systematize such effects into a taxonomy described in https://github.com/DEIB-GECO/cov2k_data_collector/blob/master/CoV2K_Effects_Taxonomy.pdf (last accessed 26 January, 2022).

In Figure 2, in the yellow rectangle, each Patient can be connected to the information on the ViralSequence by which he/she is infected; such sequence may hold mutations on two different levels, i.e., Amino Acid Changes or Nucleotide Mutations, which yield Effects on multiple levels.

4.3. Host Genetics, Host Clinical Phenotype, and Viral Genome

A viral disease is a complex system including the virus's genotype and the host's genotype and phenotype, as captured in Figure 2. The databases for each of these systems are so far curated by different communities of scientists; links connecting patients to their phenotype and viral sequences are normally missing. If such links existed in public databases, patients could be connected to their genetic profiles (including several signals such as mutations and gene expressions) and to the sequence of the virus that infected them, yielding at most one-to-one relationships between the corresponding databases (see Figure 2). However, genome evolution can be traced, as it happens for tumors, and longitudinal studies of viral sequences are starting to emerge for COVID-19, showing how the virus mutates when repeatedly sampled from the same human [70]. Therefore, the clinical course of a patient should be linked to multiple sequencing events of both the human genome and the virus.

Very few works relate the three systems of viral genomics and host genotype and phenotype in the literature. An interesting approach to study the interactome of viruses with their host was proposed by Messina et al. [71] and further studied in [72].

Infrastructures that provide support for handling all these different kinds of data are already emerging. Among these, we mention the National COVID Cohort Collaborative (N3C [73], https://github.com/National-COVID-Cohort-Collaborative, last accessed 26 January, 2022), leveraging on an organized and inclusive workstream of data extraction, harmonization, and analysis, and the secure SCOR infrastructure [74], ready to be deployed to support trackable data sharing and facilitate automated legally compliant federated analyses on an international scale. A broad collection of datasets related to COVID-19 data has been accomplished within on FAIRsharing [75] and can be reached at https://fairsharing.org/collection/TDRCOVID19Participantleveldatasharingplatformsregistries (last accessed Jan 26th, 2022).

5. Discussion

In general, there is a strong need to connect human genotype, human phenotype, and viral genotype so as to build a complete and fully encompassing scenario for data analysis. When links are absent in the data, they could be learned through computational methods. With such connections available, viral mutations could be linked to the effects in specific organs or to global disease severity (e.g., requiring intensive care) and be seen as an aspect of clinical practice. While some ongoing studies are connecting human genetics to COVID-19 (e.g., [76]), few studies so far connected COVID-19 to the viral sequence, each of them based on very few patients (e.g., [63,65]); studies encompassing combinations of variables from the three systems have not been performed yet.

We conclude with the belief that a better linking among databases (and, correspondingly, improved communication between specialists in the various disciplines) will help us to better understand infectious diseases and to empower a richer precision medicine. For example, it has been studied that the co-occurrence of certain lab-generated mutations modifies the antigenicity of the SARS-CoV-2 virus and, therefore, its sensitivity to specific neutralizing monoclonal antibodies [77]. This kind of knowledge, when properly structured and applied in a hospital context dealing with COVID-19 patients, can practically inform treatment decisions of clinicians on given sets of patients, whose clinical profile or infecting viral strain is known. Based on our contributions described in Section 2, we built a strong background for the three described systems: we developed both GenoSurf and ViruSurf and, within the COVID-19 Host Genetics Initiative, we coordinated the cooperative design of COVID-19's patient phenotype. We envision an environment where a tool such as GenoSurf could be fueled with the results of research on the genetic and genomic determinants of the COVID-19 disease and, in particular, with datasets highlighting results from GWAS (linking mutations to specific traits of COVID-19 disease) from exome sequencing studies, revealing the role and impact of mutations localized within specific genes, considered either individually or collectively (e.g., sets of co-occurring mutations). Then, ViruSurf would include up-to-date sequences of SARS-CoV-2 as deposited by laboratories over the world. Finally, a clinical phenotype database, based on the described data dictionary, would collect information on the patients and the course of their disease.

In the context of the COVID-19 disease, interoperability between the three domains of viral genomics, clinical phenotype, and human host genomics is essential. We call for actions that can be performed to link these data and make them widely accessible.

Author Contributions: Data dictionary conceptualization, A.B. and S.C.; original draft preparation, A.B.; funding acquisition, S.C. All authors have read and agreed to the published version of the manuscript.

Funding: Human and viral genomic research reported in this paper was funded by the ERC AdG GeCo (data-driven Genomic Computing) n. 693174.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The systems of the viral genomics suite are described at http://www.bioinformatics.deib.polimi.it/geco/?virus (last accessed 26 January, 2022) and made available at http://www.bioinformatics.deib.polimi.it/geco/?try_virus (last accessed 26 January, 2022). The data dictionary is available at http://gmql.eu/phenotype/ (last accessed 26 January, 2022).

Acknowledgments: We are grateful to Arif Canakoglu and Pietro Pinoli for their long-lasting collaboration on human and viral genomics, to Andrea Ganna (Univ. Helsinki) for giving us the coordinating role in the data dictionary construction, to Francesca Mari and Alessandra Renieri (Univ. Siena) for sharing the coordination decisions, and to Sulggi Lee (UCSF), Kathrin Aprile von Hohenstaufen Puoti (independent), Sara Pigazzini (UniMiB), and Catherine Moermans (CHU Liege) for their contributions to the data dictionary production.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bernasconi, A.; Ceri, S.; Campi, A.; Masseroli, M. Conceptual Modeling for Genomics: Building an Integrated Repository of Open Data. In *International Conference on Conceptual Modeling. ER 2017.* Springer International Publishing: Cham, Switzerland, 2017; pp. 325–339. https://doi.org/10.1007/978-3-319-69904-2_26.
2. Canakoglu, A.; Bernasconi, A.; Colombo, A.; Masseroli, M.; Ceri, S. GenoSurf: Metadata driven semantic search system for integrated genomic datasets. *Database* **2019**, *2019*. https://doi.org/10.1093/database/baz132.
3. Bernasconi, A.; Canakoglu, A.; Pinoli, P.; Ceri, S. Empowering Virus Sequence Research Through Conceptual Modeling. In *International Conference on Conceptual Modeling. ER 2020.* Springer International Publishing: Cham, Switzerland, 2020; pp. 388–402. https://doi.org/10.1007/978-3-030-62522-1_29
4. Guizzardi, G.; Bernasconi, A.; Pastor, O.; Storey, V.C. Ontological Unpacking as Explanation: The Case of the Viral Conceptual Model. In *International Conference on Conceptual Modeling. ER 2021.* Springer International Publishing: Cham, Switzerland, 2021; pp. 356–366. https://doi.org/10.1007/978-3-030-89022-3_28.
5. Al Khalaf, R.; Alfonsi, T.; Ceri, S.; Bernasconi, A. CoV2K: A Knowledge Base of SARS-CoV-2 Variant Impacts. In *International Conference on Research Challenges in Information Science. RCIS 2021.* Springer International Publishing: Cham, Switzerland, 2021; pp. 274–282. https://doi.org/10.1007/978-3-030-75018-3_18
6. Canakoglu, A.; Pinoli, P.; Bernasconi, A.; Alfonsi, T.; Melidis, D.P.; Ceri, S. ViruSurf: An integrated database to investigate viral sequences. *Nucleic Acids Res.* **2021**, *49*, D817–D824.
7. Sayers, E.W.; Cavanaugh, M.; Clark, K.; Ostell, J.; Pruitt, K.D.; Karsch-Mizrachi, I. GenBank. *Nucleic Acids Res.* **2019**, *47*, D94–D99.
8. The COVID-19 Genomics UK (COG-UK) Consortium. An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe* **2020**, *1*, e99–e100.
9. Shu, Y.; McCauley, J. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **2017**, *22*.
10. Bernasconi, A.; Cilibrasi, L.; Al Khalaf, R.; Alfonsi, T.; Ceri, S.; Pinoli, P.; Canakoglu, A. EpiSurf: Metadata-driven search server for analyzing amino acid changes within epitopes of SARS-CoV-2 and other viral species. *Database* **2021**, *2021*. https://doi.org/10.1093/database/baab059
11. Bernasconi, A.; Gulino, A.; Alfonsi, T.; Canakoglu, A.; Pinoli, P.; Sandionigi, A.; Ceri, S. VirusViz: Comparative analysis and effective visualization of viral nucleotide and amino acid variants. *Nucleic Acids Res.* **2021**, *49*, e90–e90 https://doi.org/10.1093/nar/gkab478.
12. Pinoli, P.; Bernasconi, A.; Sandionigi, A.; Ceri, S. VirusLab: A Tool for Customized SARS-CoV-2 Data Analysis. *BioTech* **2021**, *10*, 27.
13. Cilibrasi, L.; Pinoli, P.; Bernasconi, A.; Canakoglu, A.; Chiara, M.; Ceri, S. ViruClust: Direct comparison of SARS-CoV-2 genomes and genetic variants in space and time. *Bioinformatics*, btac030, **2022**. https://doi.org/10.1093/bioinformatics/btac030
14. Bernasconi, A.; Canakoglu, A.; Masseroli, M.; Ceri, S. META-BASE: A Novel Architecture for Large-Scale Genomic Metadata Integration. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *PP*, 1–1. https://doi.org/10.1109/tcbb.2020.2998954.
15. Masseroli, M.; Canakoglu, A.; Ceri, S.; Integration and Querying of Genomic and Proteomic Semantic Annotations for Biomedical Knowledge Extraction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *13*, 209–219. https://doi.org/10.1109/tcbb.2015.2453944.
16. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74.
17. Grossman, R.L.; Heath, A.P.; Ferretti, V.; Varmus, H.E.; Lowy, D.R.; Kibbe, W.A.; Staudt, L.M. Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **2016**, *375*, 1109–1112. https://doi.org/10.1056/nejmp1607591.
18. Kundaje, A.; Roadmap Epigenomics Consortium; Meuleman, W.; Ernst, J.; Bilenky, M.; Yen, A.; Heravi-Moussavi, A.; Kheradpour, P.; Zhang, Z.; Wang, J.; et al. Integrative analysis of 111 reference human epigenomes. *Nature* **2015**, *518*, 317–330. https://doi.org/10.1038/nature14248.

19. Genomes Project Consortium. A global reference for human genetic variation. *Nature* **2015**, *526*, 68.
20. Masseroli, M.; Canakoglu, A.; Pinoli, P.; Kaitoua, A.; Gulino, A.; Horlova, O.; Nanni, L.; Bernasconi, A.; Perna, S.; Stamoulakatou, E.; et al. Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. *Bioinformatics* **2018**, *35*, 729–736. https://doi.org/10.1093/bioinformatics/bty688.
21. Brat, G.A.; Weber, G.M.; Gehlenborg, N.; Avillach, P.; Palmer, N.P.; Chiovato, L.; Cimino, J.; Waitman, L.R.; Omenn, G.S.; Malovini, A.; et al. International electronic health record-derived COVID-19 clinical course profiles: The 4CE consortium. *npj Digit. Med.* **2020**, *3*, 1–9. https://doi.org/10.1038/s41746-020-00308-0.
22. Xu, S.; Li, Y. Global coalition to accelerate COVID-19 clinical research in resource-limited settings. *Lancet* **2020**, *395*, 1322–1325.
23. World Health Organization. (2020). Revised Case Report Form for Confirmed Novel Coronavirus COVID-19 (Report to WHO within 48 h of Case Identification): data dictionary, 27 February 2020. Available online: https://apps.who.int/iris/handle/10665/336099 (accessed on 31 December 2021).
24. Collins, F.S.; Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **2015**, *372*, 793–795. https://doi.org/10.1056/nejmp1500523.
25. Kohane, I.S.; Aronow, B.J.; Avillach, P.; Beaulieu-Jones, B.K.; Bellazzi, R.; Bradford, R.L.; Brat, G.; Cannataro, M.; Cimino, J.J.; García-Barrio, N.; et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. *J. Med. Internet Res.* **2021**, *23*, e22219. https://doi.org/10.2196/22219.
26. Bonfante, F.; Costenaro, P.; Cantarutti, A.; Di Chiara, C.; Bortolami, A.; Petrara, M.R.; Carmona, F.; Pagliari, M.; Cosma, C.; Cozzani, S.; others. Mild SARS-CoV-2 infections and neutralizing antibody titers. *Pediatrics* **2021**, *148*, e2021052173.
27. Brand, I.M.; Gilberg, L.; Bruger, J.M.; Garí, M.; Wieser, A.; Eser, T.M.; Frese, J.; Ahmed, M.I.; Rubio-Acero, R.; Guggenbuehl Noller, J.M.; et al. Broad T cell targeting of structural proteins after SARS-CoV-2 infection: High throughput assessment of T cell reactivity using an automated interferon gamma release assay. *Front. Immunol.* **2021**, *12*, 1825.
28. Antonelli, M.; Penfold, R.S.; Merino, J.; Sudre, C.H.; Molteni, E.; Berry, S.; Canas, L.S.; Graham, M.S.; Klaser, K.; Modat, M.; et al. Risk factors and disease profile of post-vaccination SARSCoV-2 infection in UK users of the COVID Symptom Study app: A prospective, communitybased, nested, case-control study. *Lancet Infect. Dis.* **2022**, *22*, 43–55.
29. Initiative, C.H.G.; 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. *Eur. J. Hum. Genet.* **2020**, *28*, 715.
30. Flicek, P.; Birney, E. The European Genotype Archive: Background and Implementation [White paper], 2007 March 30. Available online: https://ega-archive.org/files/ega_whitepaper.pdf (accessed on 31 December 2021)
31. van Blokland, I.V.; Lanting, P.; Ori, A.P.S.; Vonk, J.M.; Warmerdam, R.C.A.; Herkert, J.C.; Boulogne, F.; Claringbould, A.; Lopera-Maya, E.A.; Bartels, M.; et al. Using symptom-based case predictions to identify host genetic factors that contribute to COVID-19 susceptibility. *PLoS ONE* **2021**, *16*, e0255402. https://doi.org/10.1371/journal.pone.0255402.
32. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature* **2021**, 600, 472–477. https://doi.org/10.1038/s41586-021-03767-x.
33. COVID-19 Host Genetics Initiative.; Ganna, A. Mapping the human genetic architecture of COVID-19: An update. *medRxiv* **2021**. https://doi.org/10.1101/2021.11.08.21265944.
34. Chen, P.P.-S. The entity-relationship model—Toward a unified view of data. *ACM Trans. Database Syst.* **1976**, *1*, 9–36. https://doi.org/10.1145/320434.320440.
35. Murray, M.F.; Kenny, E.E.; Ritchie, M.D.; Rader, D.J.; Bale, A.E.; Giovanni, M.A.; Abul-Husn, N.S. COVID-19 outcomes and the human genome. *Genet. Med.* **2020**, *22*, 1175–1177. https://doi.org/10.1038/s41436-020-0832-3.
36. Hu, J.; Li, C.; Wang, S.; Li, T.; Zhang, H. Genetic variants are identified to increase risk of COVID-19 related mortality from UK Biobank data. *Hum. Genom.* **2021**, *15*, 1–10. https://doi.org/10.1186/s40246-021-00306-7.
37. Overmyer, K.A.; Shishkova, E.; Miller, I.J.; Balnis, J.; Bernstein, M.N.; Peters-Clarke, T.M.; Meyer, J.G.; Quan, Q.; Muehlbauer, L.K.; Trujillo, E.A.; et al. Large-Scale Multi-omic Analysis of COVID-19 Severity. *Cell Syst.* **2020**, *12*, 23–40.e7. https://doi.org/10.1016/j.cels.2020.10.003.
38. Ellinghaus, D.; Degenhardt, F.; Bujanda, L.; Buti, M.; Albillos, A.; Invernizzi, P.; Fernández, J.; Prati, D.; Baselli, G.; Asselta, R.; et al. Genomewide association study of severe Covid-19 with respiratory failure. *N. Engl. J. Med.* **2020**, *383*, 1522–1534.
39. Carter-Timofte, M.E.; Jørgensen, S.E.; Freytag, M.R.; Thomsen, M.M.; Andersen, N.-S.B.; Al-Mousawi, A.; Hait, A.S.; Mogensen, T.H. Deciphering the Role of Host Genetics in Susceptibility to Severe COVID-19. *Front. Immunol.* **2020**, *11*, 1606. https://doi.org/10.3389/fimmu.2020.01606.
40. Pairo-Castineira, E.; Clohisey, S.; Klaric, L.; Bretherick, A.D.; Rawlik, K.; Pasko, D.; Walker, S.; Parkinson, N.; Fourman, M.H.; Russell, C.D.; et al. Genetic mechanisms of critical illness in COVID-19. *Nature* **2021**, *591*, 92–98. https://doi.org/10.1038/s41586-020-03065-y.
41. The Severe Covid-19 GWAS Group. Faculty Opinions recommendation of Genomewide Association Study of Severe COVID-19 with Respiratory Failure. *N. Engl. J. Med.* **2020**. https://doi.org/10.3410/f.738155517.793575780.
42. Zeberg, H.; Pääbo, S. A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. *Proc. Natl. Acad. Sci. USA* **2021**, *118*. https://doi.org/10.1073/pnas.2026309118.
43. Yates, T.; Razieh, C.; Zaccardi, F.; Davies, M.J.; Khunti, K. Obesity and risk of COVID-19: Analysis of UK biobank. *Prim. Care Diabetes* **2020**, *14*, 566–567. https://doi.org/10.1016/j.pcd.2020.05.011.
44. Kuo, C.-L.; Pilling, L.C.; Atkins, J.L.; Masoli, J.A.H.; Delgado, J.; Kuchel, G.A.; Melzer, D. ApoE e4e4 Genotype and Mortality With COVID-19 in UK Biobank. *J. Gerontol. Ser. A Boil. Sci. Med Sci.* **2020**, *75*, 1801–1803. https://doi.org/10.1093/gerona/glaa169.

45. Penna, C.; Mercurio, V.; Tocchetti, C.G.; Pagliaro, P. Sex-related differences in COVID-19 lethality. *Br. J. Pharmacol.* **2020**, *177*, 4375–4385.
46. Van Der Made, C.I.; Simons, A.; Schuurs-Hoeijmakers, J.; Heuvel, G.V.D.; Mantere, T.; Kersten, S.; Van Deuren, R.C.; Steehouwer, M.; Van Reijmersdal, S.V.; Jaeger, M.; et al. Presence of Genetic Variants Among Young Men with Severe COVID-19. *JAMA* **2020**, *324*, 663. https://doi.org/10.1001/jama.2020.13719.
47. Casanova, J.L.; Su, H.C.; Abel, L.; Aiuti, A.; Almuhsen, S.; Arias, A.A.; Bastard, P.; Biggs, C.; Bogunovic, D.; Boisson, B.; et al. A global effort to define the human genetics of protective immunity to SARS-CoV-2 infection. *Cell* **2020**, *181*, 1194–1199.
48. Zhang, Q.; Bastard, P.; Liu, Z.; Le Pen, J.; Moncada-Velez, M.; Chen, J.; Ogishi, M.; Sabli, I.K.D.; Hodeib, S.; Korol, C.; et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science* **2020**, *370*, eabd4570. https://doi.org/10.1126/science.abd4570.
49. Roberts, G.H.; Partha, R.; Rhead, B.; Knight, S.C.; Park, D.S.; Coignet, M.V.; Zhang, M.; Berkowitz, N.; Turrisini, D.A.; Gaddis, M.; others. Novel COVID-19 phenotype definitions reveal phenotypically distinct patterns of genetic association and protective effects. *medRxiv* **2021**. https://doi.org/10.1101/2021.01.24.21250324.
50. Roberts, G.H.L.; Park, D.S.; Coignet, M.V.; McCurdy, S.R.; Knight, S.C.; Partha, R.; Rhead, B.; Zhang, M.; Berkowitz, N.; Haug Baltzell, A.K.; et al. AncestryDNA COVID-19 Host Genetic Study Identifies Three Novel Loci. *medRxiv* **2020**. https://doi.org/10.1101/2020.10.06.20205864.
51. Horowitz, J.E.; Kosmicki, J.A.; Damask, A.; Sharma, D.; Roberts, G.H.; Justice, A.; Banerjee, N.; Coignet, M.V.; Yadav, A.; Leader, J.B.; others. Genome-wide analysis in 756,646 individuals provides first genetic evidence that ACE2 expression influences COVID-19 risk and yields genetic risk scores predictive of severe disease. *medRxiv* **2021**. https://doi.org/10.1101/2020.12.14.20248176.
52. Benetti, E.; Giliberti, A.; Emiliozzi, A.; Valentino, F.; Bergantini, L.; Fallerini, C.; Anedda, F.; Amitrano, S.; Conticini, E.; Tita, R.; et al. Clinical and molecular characterization of COVID-19 hospitalized patients. *PLoS ONE* **2020**, *15*, e0242534. https://doi.org/10.1371/journal.pone.0242534.
53. Daga, S.; GEN-COVID Multicenter Study; Fallerini, C.; Baldassarri, M.; Fava, F.; Valentino, F.; Doddato, G.; Benetti, E.; Furini, S.; Giliberti, A.; et al. Employing a systematic approach to biobanking and analyzing clinical and genetic data for advancing COVID-19 research. *Eur. J. Hum. Genet.* **2021**, *29*, 745–759. https://doi.org/10.1038/s41431-020-00793-7.
54. Post-Mendelian Genetic Model in COVID-19. *Cardiol. Cardiovasc. Med.* **2021**, *5*, 673–694.
55. Shelton, J.F.; Shastri, A.J.; Ye, C.; Weldon, C.H.; Filshtein-Somnez, T.; Coker, D.; Symons, A.; Esparza-Gordillo, J.; Aslibekyan, S.; Auton, A.; et al. Trans-ethnic analysis reveals genetic and non-genetic associations with COVID-19 susceptibility and severity. *MedRxiv* **2020**. https://doi.org/10.1101/2020.09.04.20188318.
56. Monticelli, M.; Mele, B.H.; Benetti, E.; Fallerini, C.; Baldassarri, M.; Furini, S.; Frullanti, E.; Mari, F.; GEN-COVID Multicenter Study; Andreotti, G.; et al. Protective Role of a *TMPRSS2* Variant on Severe COVID-19 Outcome in Young Males and Elderly Women. *Genes* **2021**, *12*, 596. https://doi.org/10.3390/genes12040596.
57. Fallerini, C.; Daga, S.; Mantovani, S.; Benetti, E.; Picchiotti, N.; Francisci, D.; Paciosi, F.; Schiaroli, E.; Baldassarri, M.; Fava, F.; et al. Association of Toll-like receptor 7 variants with life-threatening COVID-19 disease in males: Findings from a nested case-control study. *eLife* **2021**, *10*. https://doi.org/10.7554/elife.67569.
58. Baldassarri, M.; Picchiotti, N.; Fava, F.; Fallerini, C.; Benetti, E.; Daga, S.; Valentino, F.; Doddato, G.; Furini, S.; Giliberti, A.; et al. Shorter androgen receptor polyQ alleles protect against life-threatening COVID-19 disease in European males. *eBioMedicine* **2021**, *65*, 103246–103246. https://doi.org/10.1016/j.ebiom.2021.103246.
59. Nakanishi, T.; Pigazzini, S.; Degenhardt, F.; Cordioli, M.; Butler-Laporte, G.; Maya-Miles, D.; Bujanda, L.; Bouysran, Y.; Niemi, M.E.; Palom, A.; et al. Age-dependent impact of the major common genetic risk factor for COVID-19 on severity and mortality. *J. Clin. Investig.* **2021**, *131*. https://doi.org/10.1172/jci152386.
60. Zanella, I.; Zacchi, E.; Piva, S.; Filosto, M.; Beligni, G.; Alaverdian, D.; Amitrano, S.; Fava, F.; Baldassarri, M.; Frullanti, E.; et al. *C9orf72* Intermediate Repeats Confer Genetic Risk for Severe COVID-19 Pneumonia Independently of Age. *Int. J. Mol. Sci.* **2021**, *22*, 6991. https://doi.org/10.3390/ijms22136991.
61. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for functional genomics data sets–update. *Nucleic Acids Res.* **2012**, *41*, D991–D995.
62. Zhao, W.-M.; Song, S.-H.; Chen, M.-L.; Zou, D.; Ma, L.-N.; Ma, Y.-K.; Li, R.-J.; Hao, L.-L.; Li, C.-P.; Tian, D.-M.; et al. The 2019 novel coronavirus resource. *Yi Chuan = Hereditas* **2020**, *42*, 212–221.
63. Lescure, F.-X.; Bouadma, L.; Nguyen, D.; Parisey, M.; Wicky, P.-H.; Behillil, S.; Gaymard, A.; Bouscambert-Duchamp, M.; Donati, F.; Le Hingrat, Q.; et al. Clinical and virological data of the first cases of COVID-19 in Europe: A case series. *Lancet Infect. Dis.* **2020**, *20*, 697–706. https://doi.org/10.1016/s1473-3099(20)30200-0; Correction in *Lancet Infect. Dis.* **2020**, *20*, e148.
64. Lu, R.; Zhao, X.; Li, J.; Niu, P.; Yang, B.; Wu, H.; Wang, W.; Song, H.; Huang, B.; Zhu, N.; et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* **2020**, *395*, 565–574. https://doi.org/10.1016/s0140-6736(20)30251-8.
65. Böhmer, M.M.; Buchholz, U.; Corman, V.M.; Hoch, M.; Katz, K.; Marosevic, D.V.; Böhm, S.; Woudenberg, T.; Ackermann, N.; Konrad, R.; et al. Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: A case series. *Lancet Infect. Dis.* **2020**, *20*, 920–928. https://doi.org/10.1016/s1473-3099(20)30314-5.
66. Tang, X.; Wu, C.; Li, X.; Song, Y.; Yao, X.; Wu, X.; Duan, Y.; Zhang, H.; Wang, Y.; Qian, Z.; others. On the origin and continuing evolution of SARS-CoV-2. *Natl. Sci. Rev.* **2020**, *7*, 1012–1023.

67. Hodcroft, E.B. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. 2021. Available online: https://covariants.org/ (accessed on 13 December 2021).
68. Centers for Disease Control and Prevention. SARS-CoV-2 Variant Classifications and Definitions. 2021. Available online: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html. (accessed on 31 December 2021).
69. European Centre for Disease Prevention and Control. SARS-CoV-2 variants of concern. 2021. https://www.ecdc.europa.eu/en/covid-19/variants-concern (accessed on 13 December 2021).
70. Rose, R.; Nolan, D.J.; Moot, S.; Feehan, A.; Cross, S.; Garcia-Diaz, J.; Lamers, S.L. Intrahost site-specific polymorphisms of SARS-CoV-2 is consistent across multiple samples and methodologies. *medRxiv* **2020**. https://doi.org/10.1101/2020.04.24.20078691.
71. Messina, F.; Giombini, E.; Agrati, C.; Vairo, F.; Bartoli, T.A.; Al Moghazi, S.; Piacentini, M.; Locatelli, F.; Kobinger, G.; Maeurer, M.; others. COVID-19: Viral–host interactome analyzed by network based-approach model to study pathogenesis of SARS-CoV-2 infection. *J. Transl. Med.* **2020**, *18*, 1–10.
72. Gordon, D.E.; Hiatt, J.; Bouhaddou, M.; Rezelj, V.V.; Ulferts, S.; Braberg, H.; Jureka, A.S.; Obernier, K.; Guo, J.Z.; Batra, J.; et al. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science* **2020**, *370*. https://doi.org/10.1126/science.abe9403.
73. Haendel, M.A.; Chute, C.G.; Bennett, T.D.; Eichmann, D.A.; Guinney, J.; Kibbe, W.A.; Payne, P.R.; Pfaff, E.R.; Robinson, P.N.; Saltz, J.H.; et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 427–443.
74. Raisaro, J.L.; Marino, F.; Troncoso-Pastoriza, J.; Beau-Lejdstrom, R.; Bellazzi, R.; Murphy, R.; Bernstam, E.V.; Wang, H.; Bucalo, M.; Chen, Y.; et al. SCOR: A secure international informatics infrastructure to investigate COVID-19. *J. Am. Med Inform. Assoc.* **2020**, *27*, 1721–1726. https://doi.org/10.1093/jamia/ocaa172.
75. Maxwell, L.; Shreedhar, P.; Dauga, D.; McQuilton, P.; Terry, R.; Denisiuk, A.; Molnar-Gabor, F.; Saxena, A.; Sansone, S.A. FAIR, ethical, and coordinated data sharing for COVID-19 response: A review of COVID-19 data sharing platforms and registries. PREPRINT (Version 1) available at Research Square 2021. https://doi.org/10.21203/rs.3.rs-1045632/v1.
76. Benetti, E.; Tita, R.; Spiga, O.; Ciolfi, A.; Birolo, G.; Bruselles, A.; Doddato, G.; Giliberti, A.; Marconi, C.; Musacchia, F.; et al. ACE2 gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population. *Eur. J. Hum. Genet.* **2020**, *28*, 1602–1614. https://doi.org/10.1038/s41431-020-0691-z.
77. Li, Q.; Wu, J.; Nie, J.; Zhang, L.; Hao, H.; Liu, S.; Zhao, C.; Zhang, Q.; Liu, H.; Nie, L.; et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* **2020**, *182*, 1284–1294.