Semantic and population analysis of the genetic targets related to COVID-19 and its association with genes and diseases

Louis Papageorgiou[1], Eleni Papakonstantinou[1], Io Diakou[1], Katerina Pierouli[1], Konstantina Dragoumani[1], Flora Bacopoulou[2], George P Chrousos[2], Elias Eliopoulos[1], Dimitrios Vlachakis[1,2,3*]

1. Laboratory of Genetics, Department of Biotechnology, School of Applied Biology and Biotechnology, Agricultural University of Athens, 75 Iera Odos, 11855 Athens, Greece
2. University Research Institute of Maternal and Child Health & Precision Medicine, National and Kapodistrian University of Athens, "Aghia Sophia" Children's Hospital, 11527 Athens, Greece
3. Division of Endocrinology and Metabolism, Center of Clinical, Experimental Surgery and Translational Research, Biomedical Research Foundation of the Academy of Athens, 11527 Athens, Greece

**Correspondence to*: Dr Dimitrios Vlachakis, dimvl@aua.gr, Laboratory of Genetics, Department of Biotechnology, School of Applied Biology and Biotechnology, Agricultural University of Athens, 75 Iera Odos, 11855 Athens, Greece

Keywords: genetic targets, genetic grammar, biomarkers, disease, data mining, semantics, population analysis, COVID-19

Abstract: SARS-CoV-2 is a coronavirus responsible for one of the most serious, modern worldwide pandemics, with lasting and multi-faceted effects. By late 2021, SARS-CoV-2 has infected more than 180 million people and has killed more than 3 million. The virus gains entrance to human cells through binding to ACE2 via its surface spike protein and causes a complex disease of the respiratory system, termed COVID-19. Vaccination efforts are being made to hinder the viral spread and therapeutics are currently under development. Towards this goal, scientific attention is shifting towards variants and SNPs that affect factors of the disease such as susceptibility and severity. This genomic grammar, tightly related to the dark part of our genome, can be explored through the use of modern methods such as natural language processing. We present a semantic analysis of SARS-CoV-2 related publications, which yielded a repertoire of SNPs, genes and disease ontologies. Population data from the 100Genomes Project were subsequently integrated into the pipeline. Data mining approaches of this scale have the potential to elucidate the complex interaction between COVID-19 pathogenesis and host genetic variation; the resulting knowledge can facilitate the management of high-risk groups and aid the efforts towards precision medicine.

Introduction

Coronaviridae is a family of enveloped viruses with a single-stranded RNA genome roughly 25 to 32 kb long. The size of the virion ranges from 118 to 136 nm, while its surface is studded with the characteristic, large spike (S) glycoprotein. The viral family is further divided into subfamilies *Orthocoronavirinae* and *Letovirinae,* with the latter comprising of a single genus, *Alphaletovirus*. The *Orthocoronavirinae* subfamily circumscribes four non-monotypic genera, namely *Alpha*-, *Beta*-, *Gamma*- and *Deltacoronavirus* (1). Alpha- and betacoronaviruses commonly infect mammals, gammacoronaviruses mainly infect avian species, while deltacoronaviruses infect mammals and birds (2). The effects of human coronavirus infection can range from mild, such as in the case of HCoV-229E, to potentially life-threatening, such as in the case of the Middle East respiratory syndrome coronavirus (MERS-CoV) and severe acute respiratory syndrome coronavirus (SARS-CoV-1 and 2) (3).

The standard coronavirus virion comprises of the membrane, envelope and spike proteins, which are all embedded in the viral envelope, as well as the nucleocapsid protein, which interacts with the viral RNA at the virion's core (4). The large coronavirus genome possesses untranslated regions at both ends; two large ORFS at the 5' end, ORF1a and ORF1b, code for non-structural proteins necessary for the formation of the replication and transcription complex (RTC), while ORFs encoding structural and accessory proteins are transcribed from the 3' end (5). During infection, the coronavirus spike (S) protein mediates binding to specific cellular receptors. For example, both the betacoronaviruses SARS-CoV and SARS-CoV-2 recognize the angiotensin-converting enzyme 2 (ACE2) (6), while other betacoronaviruses like MERS-CoV and HKU4 recognize the dipeptidyl peptidase 4 (DPP4) (7). The S protein is a homotrimeric, class I fusion glycoprotein, forming petal-shaped projections on the virion's surface (8). In some coronaviruses, S is cleaved during the maturation process while in others, including SARS-CoV, S is arranged into two domains, S1 and S2, with different functions (9, 10). Within the surface-exposed S1 domain lies the receptor-binding domain (RBD) responsible for interaction with the host cell receptor while the transmembrane S2 domain mediates fusion between viral and host cell membranes (11). In a study by Wang *et al*., the SARS-CoV spike protein, through interaction with murine macrophages, was found to induce IL-6 cytokines and release of TNF-α (12). Interleukin-6 (IL-6) plays a key role in the innate and acquired immune response, inducing the acute-phase response after the occurrence of infection and inflammation (13). In a preprint, Hsu *et al.* proposed that the SARS-CoV-2 spike protein induces significant NF-κB activations as well as production of pro-inflammatory cytokines (14). The described mechanism of action is the stimulation of the MAPK-NF-κB axis through the binding of the S protein to the ACE2 receptor, resulting in the release of cytokines (14). Furthermore, in a recent preprint, modeling and docking studies highlighted a potential interaction between the SARS-CoV-2 spike glycoprotein and nicotinic acetylcholine receptors (nAChRs), proposing an underlying mechanism participating in severe COVID-19 (15).

Since being declared in March 2020, the ongoing COVID-19 pandemic has affected countries on a near global scale, with more than 248 million confirmed cases and more than 5 million deaths by November 2021 (https://www.who.int/), challenging healthcare systems, economies and communities in multiple ways. COVID-19 exhibits strong heterogeneity when it comes to clinical representation, ranging from asymptomatic to severe disease affecting multiple organs (16). Influenza-like symptoms tend to be prevalent, as the main sites of infection are the upper and lower respiratory tract, however other organs such as the heart of kidneys can be affected as sites of ACE2 expression (17). Factors which impact the risk and severity of COVID-19 are continuously being investigated. A 2021 meta-analysis of more than 17 million patient data highlighted common variables linked to adverse outcome, such as older age, severe obesity and active cancer (18, 19). One drug, remdesivir, has been approved by the FDA for treatment of COVID-19 while investigational therapies, such as monoclonal antibodies, are being explored. Chen *et al.* reported the isolation of two lead IgG1 monoclonal antibodies which effectively blocked the binding between ACE2 and the SARS-CoV-2 RBD (20). Out of a set of neutralizing antibodies isolated by Wu *et al.*, two antibodies, B38 and H4, effectively blocked the binding of the RBD to ACE2 (20). CB6, a specific human monoclonal antibody isolated by Shi *et al.*, was shown to hinder SARS-CoV-2 infection in vitro as well as in rhesus monkeys, by targeting an epitope that overlaps with ACE2 binding sites in the SARS-CoV-RBD (21). Non-RBD monoclonal antibodies are also investigated (22).

To curtail the damaging effects of COVID-19 and expedite herd immunity, the scientific community raced to develop vaccines against SARS-CoV-2. Currently available vaccines rely on the spike protein as an immunogen because of its key roles during viral entry; the first category, mRNA and adenoviral vector vaccines, provide genetic information for spike protein synthesis, while the second category, inactivated vaccines, constitute protein-based strategies (23). By November 2021, more than 53% of the world population has received at minimum one dose of a COVID-19 vaccine (24). Nevertheless, vaccine hesitancy is a widespread phenomenon, as evidenced by data stemming from behavior analysis conducted by the Imperial College of London (25). In surveys about citizens' willingness to get vaccinated against COVID-19 in Germany, France, Italy, Australia, Spain and Japan, the share of the surveyed population who were unvaccinated and unwilling to get vaccinated ranged between 12-22% (25). Through global circulation, SARS-CoV-2 variants have and will continue to emerge as a result of selective pressure and continuous viral replication within the population of hosts. One likely selective pressure is for mutations which improve intrinsic fitness, such as in the case of the D164G substitution in the spike protein (26). Increased infection in the upper airway due to D164G has allowed the variant to dominate over the wild-type virus (26, 27). The Delta variant (B.1.617.2), which is becoming the dominant strain globally according to WHO, has been shown to be eightfold less sensitive to vaccine-elicited antibodies in comparison to the wild-type Wuhan-1 bearing D164G in vitro (26). Therefore, the global scientific community is called to keep a close eye on the ever-

changing landscape of the SARS-CoV-2 mutational landscape and its potential effects on the vaccines' effectiveness.

Genome-wide association studies (GWAS) are an important tool in the investigation of disease pathogenesis and enable the characterization of relevant single nucleotide polymorphisms (SNPs) (28). Furthermore, genetic variants which are linked to diseases can shape a polygenic risk score, which characterizes the individual's susceptibility to certain diseases (29, 30). As it has been evidenced, polymorphisms occurring in regions that do not code for proteins are frequent and can have equally potent effects (31, 32). Therefore, when exploring the variation of the individual's genetic makeup, it would be unwise to limit ourselves to the coding regions of the human genome. When analyzing genetic variation under the scope of infection and disease risk, "genomic grammar" can be an appropriate term, since it is not limited to the gene but encompasses factors related to the dark part of the genome which have only recently begun to be investigated (32).

The 1000 Genomes Project provides an invaluable pool of whole genome sequencing data, with a goal of constructing an inventory of genetic variations within the human genome (33). Genomes of more than 2.500 individuals have been mapped for genetic variation (34). During the project's analysis, a specific allele frequency is assigned to each pinpointed variant, calculated by dividing the number of the allele's occurrence in the population by the total sum of copies of all the alleles at the genetic locus of interest. Data provided by the 1000 Genomes Project include – among others - the general allele frequency of the determined variants and the corresponding allele frequencies of five major groups, Europeans (EUR), Africans (AFR), Americans (AMR), East Asians (EAS) and South Asians (SAS) (35). When conducting population analyses, the allele frequency is a key component, since it corresponds to the occurrence of a distinct genetic variant within a population (36). Allele frequencies, which are provided within the range of [0-1], constitute a reflection of genetic diversity; monitoring their changes allows the detection of shifts within the population (37).

As mentioned previously, COVID-19 exhibits variability across individuals, hinting at a trove of genetic factors which contribute to COVID-19 susceptibility and severity (38). As we wade through the third SARS-CoV-2 wave, the rapidly increasing volume of biomedical and genomic data calls for the implementation of modern techniques, for knowledge to be extracted and incorporated into novel therapeutic strategies. Natural language processing and other machine-learning techniques can make efficient use of the vast COVID-19 related literature, allowing the exploration of the complex architecture behind COVID-19 susceptibility and severity. Herein, we present a pipeline of semantic analysis of COVID-19 literature data for the mining of related SNPs, genes and disease ontologies. In the second phase of our analysis, we integrate population data from the 1000 Genomes Project. Our pipeline can serve as an example of an integrated approach in the research against COVID-19, towards estimating the "key" genomic target and providing beneficial knowledge in the personalization of medicine and the efficient assessment of populations at higher risk of infection and severe disease, on the basis of the genomic grammar and specifically SNPs they harbor.

Methods

Dataset collection and filtering

Using NCBI's Entrez programming utilities, scientific literature in MEDLINE format was collected from Pubmed (39), limiting the search to the term "COVID-19" and publication dates post 2020. The MEDLINE files were collected in text form for subsequent filtering and feeding into the semantic analysis pipeline, which is summarized in Figure 1. A similar approach for the analysis of scientific publications has been described elsewhere (40). The filtering step of the pre-analysis included the removal of articles which were duplicates and unrelated to the subject.

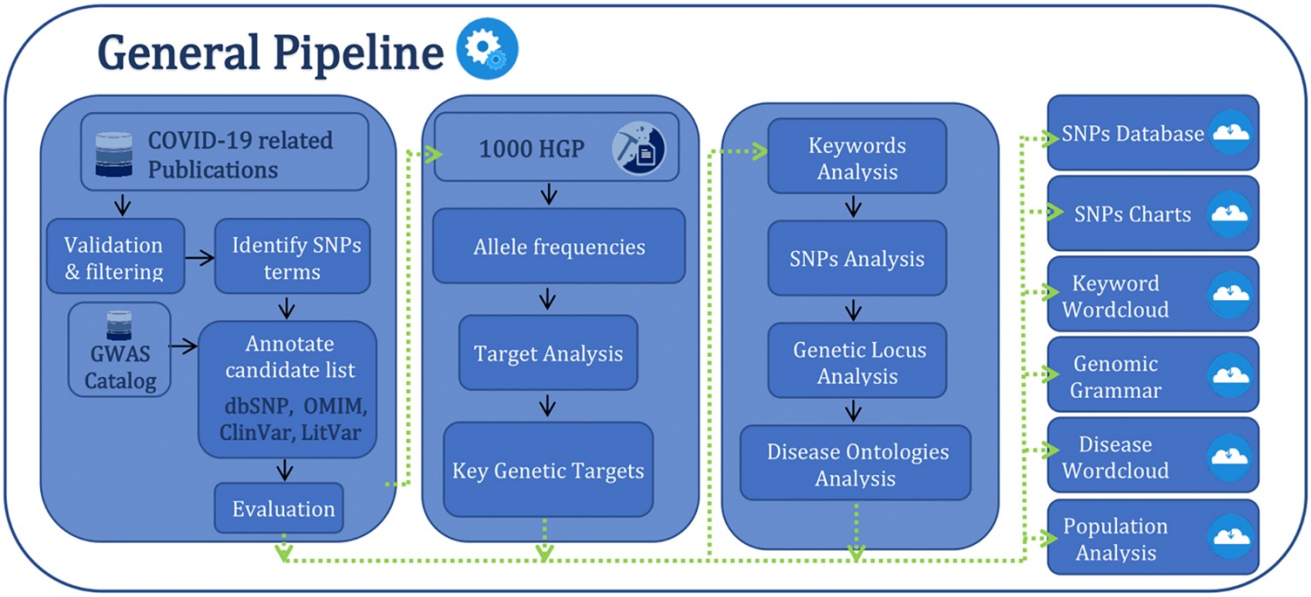

Figure 1. Summarized workflow of the study's analysis.

COVID-19 related SNPs

A search query was built with regular expressions in order to identify the candidate SNPs using the extracted dataset of the related articles with COVID-19. The extracted SNPs within the dataset of scientific articles were stored in a structured database for further analysis. Additionally, each article MEDLINE file was mined for supplementary information, such as MeSH/MEDLINE key terms, ontologies studied for their role in COVID-19 and mutations/polymorphisms. The candidate list was enriched with SNP data and meta-data from the GWAS Catalogue (41) and lastly, duplicates were removed from the candidate list.

Data mining and semantics analysis

The candidate list of COVID-19 related SNPs was annotated through the use of publicly available databases of genetic variation, including dbSNP, ClinVar, LitVar and OMIM (42-45). The construction of the final SNPs

dataset was carefully monitored; the annotated SNPs were evaluated, in order to remove those which were present as entries in the aforementioned databases and in the mined literature, but were actually reported as having no effect regarding COVID-19. Finally, the SNPs were subjected to semantics analysis to extract the desired knowledge, such as key terms, genomic grammar and disease ontologies (40, 46, 47).

Population analysis

Five population groups were studied, focusing on Europeans, Africans, Americans, East Asians and South Asians. Sample sizes and origins of the individual population samples are given in the International Genome Sample Resource (IGSR) [78], which has been developed under the 1000 Human Genomes Project (1000 HGP) [79]. The elaboration of the present study has been performed using human genomes which were contained in the phase three collection of the IGSR on reference assembly GRCh38 (48, 49). Since the 1000 Genomes Project has created call sets of sequence variants for each of the different genomes sequenced, the downloaded data were multi-individual VCFs (50) per chromosome, with genotypes listed for each sample (49). Histograms regarding each population were generated with suitable packages of the Python programming language. Statistical analysis of the results was carried out with the use of the R Biocircos package, which enables the visualization of genomic-related data and is based on the Javascript library developed by Cui et al (51).

Results

COVID-19 related key terms and SNPs

The collection of COVID-19 related biomedical literature enables the identification of keywords as they appear within the MEDLINE files. Through querying of the Pubmed database, 147.396 non-duplicate scientific articles corresponding to the search term “COVID-19” were collected in a final dataset and were subsequently mined for related keywords. A total of 98.497 keywords were assembled, out of which 2.677 were identified as most frequent, providing a first estimation of relation to COVID-19. The most frequently appearing keywords are visualized as a word cloud in Figure 2. The word cloud visualization technique enables the presentation of the results, where the size of the words – in this case the keywords – indicates their frequency within the dataset. SNPs were found to be contain within 147 of the collected articles. Following their extraction, their enrichment through GWAS Catalogue and their annotation, a total of 526 SNPs were collected, to be further subjected to evaluation. Out of them, 339 SNPs related to COVID-19 were identified and collected.

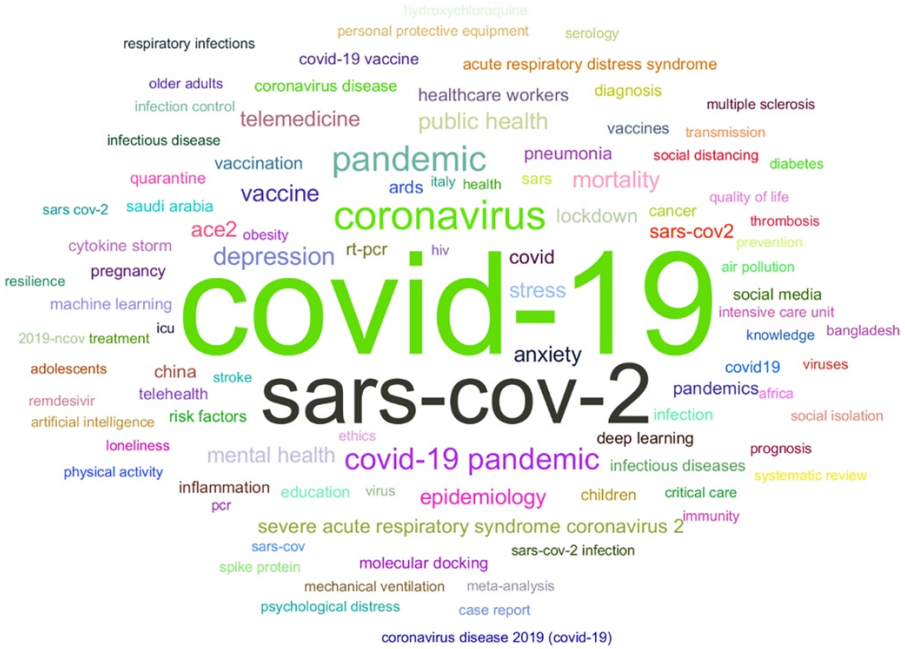

Figure 2. Word cloud presentation of COVID-19 related keywords, according to the input dataset of literature articles. The size of the words in the cloud mirrors their frequency within the dataset.

Genomic grammar

After semantics analysis of the annotated and evaluated COVID-19-related SNPs, the genomic grammar of the disease was constructed. The genomic grammar, as mentioned earlier, constitutes the genomic map of COVID-19 variation, encompassing the genes and non-coding locations which harbor the related SNPs. The results concerning COVID-19's genomic grammar are visualized in word-cloud form in Figure 3. Since the size corresponds to the frequency of the genomic "word", a quick survey of Figure 3 enables the identification of some of the prominent genetic players related to COVID-19. For example, *ACE2* is easily identified as a central gene. In the renin-angiotensin system (RAS), which is important for blood pressure levels and by extension for the proper function of multiple organs, angiotensin I (Ang-II) and angiotensin II (Ang-II) constitute important biomolecules (52). The angiotensin-converting enzyme (ACE) converts Ang-I to Ang-II, which can bind to angiotensin type I receptor (AT1R) and angiotensin type II receptor (AT2R) (53). Interaction between Ang-II and AT1R triggers processes such as vasoconstriction, fibrosis and inflammation, while interaction with AT2R counteracts the AT1R-mediated effects (54). ACE2 is a homologue of ACE and can convert Ang-I and Ang-II to angiotensin-(1-7) (55). Angiotensin-(1-7) acts on the G protein-coupled receptor MAS to carry out processes such as anti-fibrosis, anti-inflammation, generally having an effect opposite to that of Ang-II/AT1R binding (56). As previously mentioned, ACE2 participates in the entry of SARS-CoV-2 into the host cells through interaction with the viral surface protein S (57). Significant ACE2 increase has been documented in patients with severe COVID-19 (58), and since this receptor is highly

expressed in various organs and tissues and is a major component of inflammation, its association with multiple-organ failure syndromes in COVID-19 remains under intense study (59, 60). *ACE2* polymorphisms and their effect on COVID-19 severity, susceptibility and progression are gradually being explored as a tool to monitor the disease outcome on an individual patient level (61-63).

Figure 3. Word-cloud presentation of COVID-19 related genes and other non-coding regions. The size of the words in the cloud corresponds to their frequency within the respective dataset. Prominent genes and non-coding regions such as ACE2, ABO, IL-6, TMPRSS2, ZTFL1 and LOC107986083 can be identified at a first glance.

Interleukin-6 (IL-6) is a cytokine with multifaceted involvement in the initiation of immune response and inflammatory processes (64). Viral infection triggers its secretion by immune cells such as macrophages, B and T cells, as well other cell types such as endothelial cells and fibroblasts (65). IL-6 is a critical component of the cytokine storm phenomenon in cases of severe COVID-19 and high levels of IL-6, among other cytokines, are one of the hallmarks of severe COVID-19 (66, 67). Polymorphisms in the promoter and regulatory regions of the *IL-6* gene modulate the protein's expression, helping to account for the differences in immune response documented in various ethnic groups against a variety of pathogens, including SARS-CoV-2 (68-70). IL6-AS1 (IL6 Antisense RNA 1) is a long non-coding RNA and has been found to be upregulated in chronic obstructive pulmonary disease, promoting the expression of IL-6 (71). An *IL-6* variant haplotype common in Asian populations was found to be protective against severe COVID-19, associated with lower IL-6 and IL-6-AS1 levels through a disturbance of a binding locus at the IL-6-AS1 enhancer elements (69).

Transmembrane serine protease 2 (TMPRSS2) is an essential host factor for SARS-CoV-2 pathogenicity and an important player in the viral entry into host cells, priming the viral S glycoprotein for viral fusion (72).

TMPRSS2 SNPs have been the object of study for the establishment of disease outcome biomarkers in pathologies such as cancer and severe viral infections such as H1N1 infection (73, 74). A computational analysis aimed at explaining susceptibility differences among populations identified a number of SNPs with predicted effect on protein function (75), while a study in print elected *TMPRSS2* genetic variants as candidate COVID-19 modulators after examining single nucleotide polymorphisms in various ethnic populations (76). Lastly, a common *TMPRSS2* non-synonymous variant, rs12329760, was found to be protective against severe COVID-19 through impacting the enzyme's catalytic ability and thus its role in the viral entry (77).

Variations in the *ABO* gene in chromosome 9 are the basis for the establishment of the conventional ABO blood group (78). The *ABO* locus encodes three alleles, with alleles A and B producing α-1,3-N-acetylgalactosamine transferase, α-1,3-galactosyl transferase B respectively, while allele O exhibits a deletion-caused frameshift and lacks both of the aforementioned enzymatic activities (79). Increased levels of ABO protein in plasma appear to be associated with risk of severe COVID-19 (80). The same study linked COVID-19 risk and severity with the OAS1 gene, which is activated by interferon and participates in the cellular innate antiviral response (81). Although the exact mechanism underlying the effect of the ABO blood group on COVID-19 susceptibility and severity remains under investigation, a recent study reported association of blood groups A and B with increased risk of SARS-CoV-2 infection (82), while another study found B-allele frequencies to be correlated with COVID-19 mortality (83).

The *LZTFL1* gene codes for a leucine zipper protein, which associates with E-cadherin and participates in the circulation of a variety of signaling molecules (84, 85). The gene is expressed in pulmonary epithelial cells, among others, and has been recently identified as a target for a probable causative variant related to COVID-19 risk (86, 87). Rs17714054A, the risk allele of the SNP, was found to target *LZTFL1*'s enhancer region, leading to the gene's upregulation (86). According to NCBI data, LOC107986083 is an uncharacterized non-coding RNA located in chromosome 3 and has been found to be broadly expressed in the testis and thyroid, among other tissues. Positionally, it is associated with the LZTFL1 gene.

Disease ontologies

The semantics analysis of the SNPs, related genes and non-coding regions, enables the subsequent extraction of information regarding disease profiling related to COVID-19. The diseases identified through our analysis are summarized in world-cloud form in Figure 4, where the size of the words, or disease terms in our case, mirrors the frequency of the term. A number of prevalent words, such as neoplasms, chronic hepatitis C or obesity, can thus be easily identified with a first study of the visualized results. Neoplasms are abnormal and excessive tissue growths which, when malignant, are known as cancers (88). Patients with hematologic malignancies were found to be in a significant risk of COVID-19 related death (89). In a preliminary, exploratory analysis, essential thrombocythemia, a type of myeloproliferative neoplasm, was found to be

associated with greater risk of venous thromboembolism in COVID-19 patients (90). Similarly, myeloproliferative neoplasm patients and especially those suffering from myelofibrosis were in high risk of mortality during COVID-19 infection (91). Paraoxonase-1 (PON1) is a lactonase which degrades lipid peroxides in low-density lipoproteins (LDL) and high-density lipoproteins (HDL) (92). The PON1-L55M polymorphism at the *PON1* genetic locus has been identified as a potential risk factor for breast cancer (93). Furthermore, *PON1* genetic variants are under investigation for involvement in predisposition to obesity-associated fatty liver disease (94). A recent study investigating a potential association between *PON1* polymorphisms and COVID-19 pointed at a possible positive correlation between M55 and prevalence and mortality of COVID-19, although these findings remain to be validated in larger-scale studies (95).

Figure 4. Word-cloud presentation of disease ontologies related to COVID-19. Word size within the cloud reflects the frequency of the word, allowing the quick identification of strongly related diseases, such as neoplasms, non-alcoholic fatty liver disease, fibrosis and obesity.

Increased left ventricular strain as a fibrosis marker was found to predict moderate and severe COVID-19 progression with accuracy (96). Myocardial fibrogenesis is induced by transforming growth factor β1 (TGF-β1) (97, 98). An increase in TGF-β1 expression possibly leads to susceptibility to SARS-CoV-2 infection as a result of increased expression of NRP-1 neuropilin-1 (NRP-1), a co-receptor for TGF-β1, in the lungs (99). Lastly, lung inflammation and pulmonary fibrosis can be potentially promoted by SARS-CoV-2 complications, such as Acute Respiratory Distress Syndrome (ARDS) (100).

Non-alcoholic fatty liver disease (NAFLD) is a condition marked by fat accumulation in the parenchymal region of the liver without ties to hepatic fat accumulation factors, such as alcohol abuse or hepatitis B infection (101). NAFLD patients are susceptible to COVID-19 due to an existing layer of risk factors, such as

type 2 diabetes, which promotes susceptibility to infection (102), obesity, which is linked to respiratory complications (103), and cardiovascular disease (104). A meta-analysis study between NAFLD and non-NAFLD patients reported an increased risk of severe COVID-19 infection and ICU admission (105). The added presence of obesity in NAFLD patients appears to increase the severity of COVID-19 (106). In a pooled study of COVID-19 and NAFLD data, the presence of NAFLD was associated with an increased risk of severe COVID-19 (107). Obesity has been associated with various inflammatory mediators such as IL-6 (108, 109), which in our study was found to be part of COVID-19's genomic grammar. Subsets of immune cells in the white adipose tissue lead to a surge in inflammation-promoting cytokines like tumor necrosis factor α (TNFα) and IL-6 (108). TNFα and IL-6 are players in the signaling of the initial phase of cytokine-storm, a phenomenon prevalent in severe COVID-19 (110, 111). The hyperinflammatory state observed in obese individuals may also lead to coagulopathies (112), the hallmark of which are shifts in the levels of D-dimer, a fibrin degradation product (FDP) (113). Correlation has been shown between D-dimer and COVID-19 severity (114).

Chronic hepatitis C is caused by the hepatitis C virus (HCV) following acute infection, with potential complications such as liver damage, cirrhosis and cancer (115). Genetic variation at the Interferon lambda 4 (IFNL4) genetic locus has been studied with the aim to identify viral clearance predictors (116). One polymorphism, rs12979860, has been associated with clearance of hepatitis C virus and other RNA viruses which target the upper respiratory system (117, 118). Additionally, association has been evidenced between this polymorphism and the response to type I IFN treatment efforts in patients with chronic hepatitis C. The T allele of the aforementioned polymorphism was found to be overexpressed in COVID-19 patients, highlight its potential as a risk factor for COVID-19 (119).

Apolipoprotein E (APOE) e4 genotype, which has been linked to high risk for Alzheimer's disease, was described as a potential predictor of severe COVID-19 infection (120). In addition, a study by Taylor *et al.* identified four severe COVID-19 risk-associated genes which had been previously linked to increased risk of developing Alzheimer's, further hinting at possible interplay between the two diseases (121).

Population analysis

Directional change and reversal in allele frequencies has been shown in the 339 COVID-19 related SNPs between the individuals of the major five clusters. The histogram analysis of COVID-19 related SNPs using the 1000 Genome Project dataset shows similar distribution in the five major groups (Figure 5). Although different allele frequencies have been identified in the studied SNPs, some groups appear to have similar distributions with different numbers such as the Africans and East Asians example or the Europeans and Americans example (Figure 5). The five studied groups have accumulated different SNPs totals at the sensitive two extremities of the allele frequencies including the cluster of the "low allele frequencies" (0.1 ≥ SNP allele

frequency) and the cluster of the "high allele frequencies" (0.9 ≤ SNP allele frequency) (Figure 2) (122-124). Our findings are in agreement with the expected observations (125).

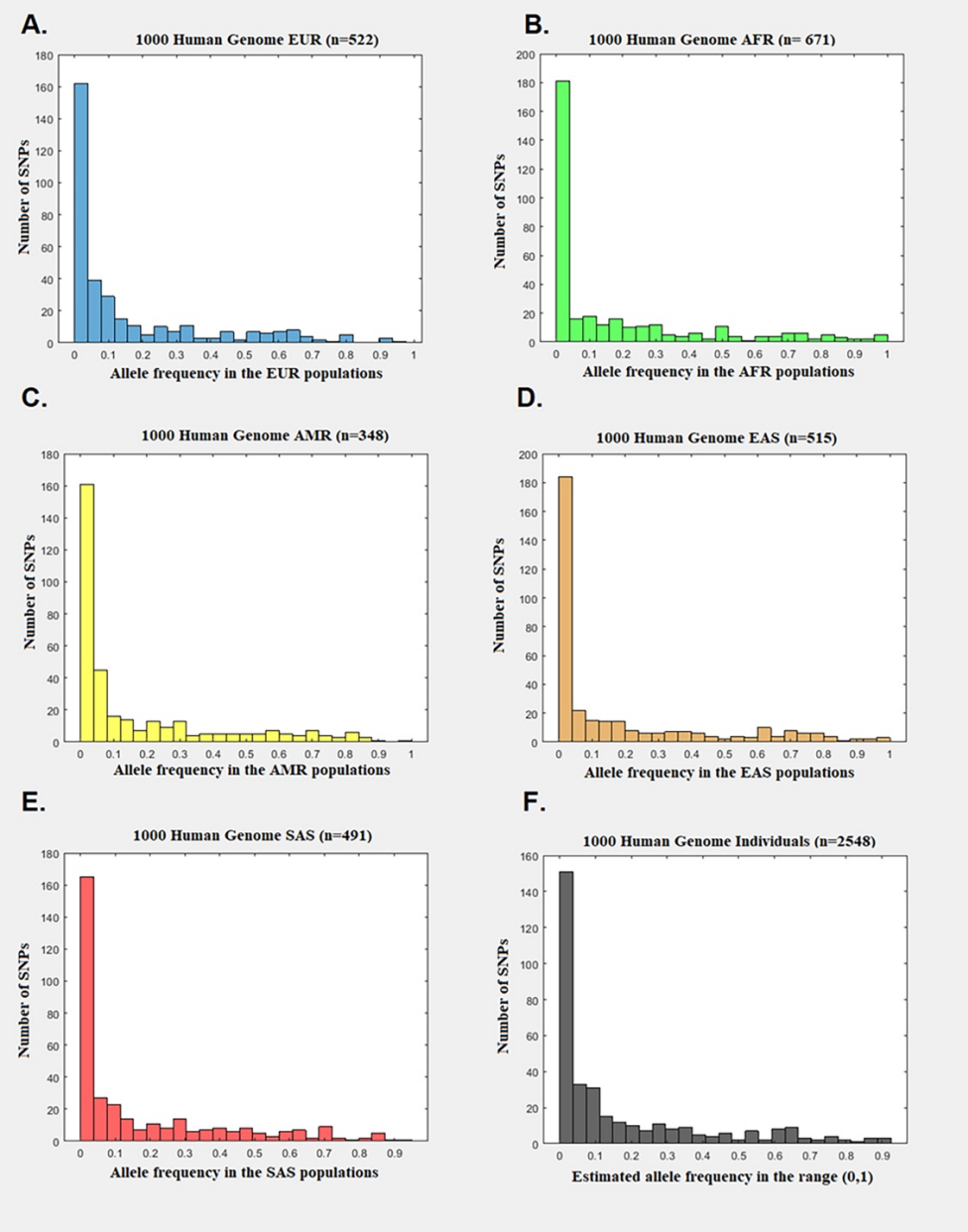

Figure 5. Histogram analysis of the Histogram of the allele frequencies of the COVID-19 related SNPs as extracted from the 1000 Human Genomes project. (A) Histogram of the SNPs allele frequencies for the separated group of the Europeans. (B) Histogram of the SNPs allele frequencies for the separated group of the Africans. (C) Histogram of the SNPs allele frequencies for the separated group of the Americans. (D) Histogram of the SNPs allele frequencies for the separated group of the East Asians. (E) Histogram of the SNPs allele frequencies for the separated group of the South Asians. (F) Histogram of the SNPs allele frequencies for total number of the studied individuals.

Different totals and ids of SNPs have been accumulated in the low and high clusters between the population groups (Figure 5). The American group has the largest sample of SNPs with low allele frequencies followed by Europeans, South Asians, East Asians, and Africans (Figure 5 C, A, E, F, B). On the other hand, the African group has the largest sample of SNPs with high allele frequencies followed by the East Asian group (Figure 5 B, D). The South Asian, American and European groups show significant fewer totals in SNPs with high allele frequencies (Figure 5 E, C, A). Although some population groups are shown some similarities in the ids of the identified SNPs in the low and high clusters, the overall distribution of the COVID-19 related SNPs and their genetic locus per chromosome in the five studied population groups shows a significant differentiation. A general conclusion to be drawn from the results is that the genomic grammar of Africans and East Asians contains more COVID-19 related SNPs in the two sensitive clusters (low and high) than the other groups (Figures 5,6).

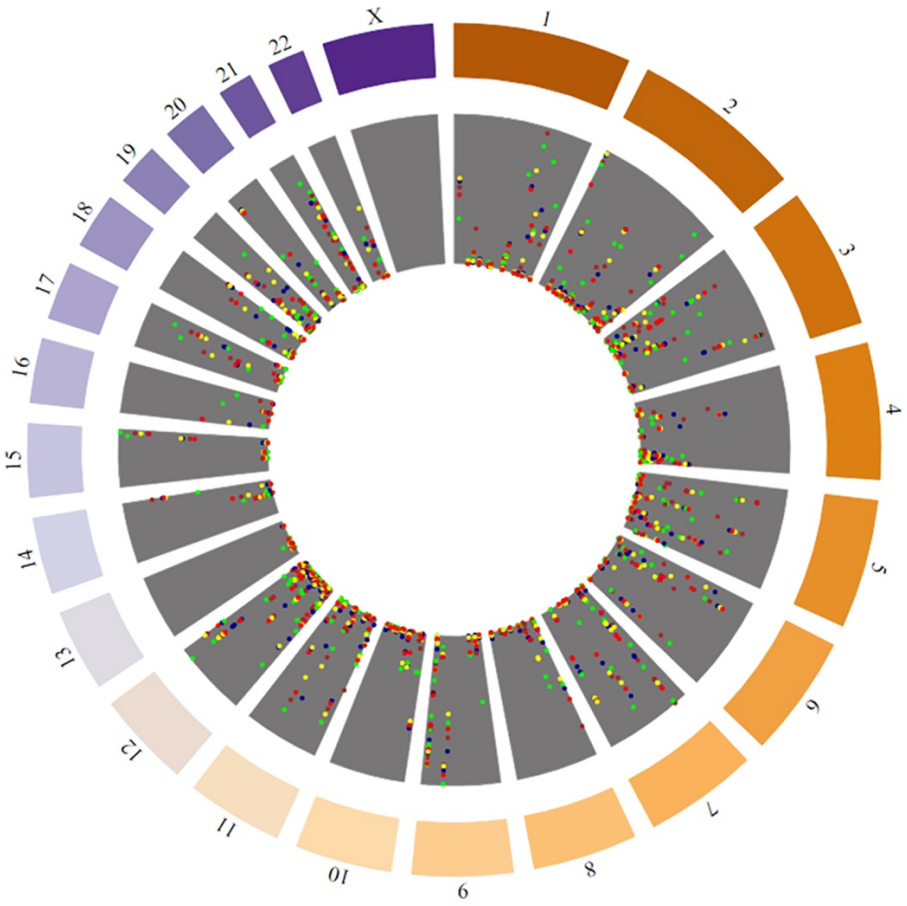

Figure 6. Circos-like visualizations of the genomic data of the five major population groups in the identified COVID-19 related SNPs with low allele frequencies (Blue dots = Europeans, green dots = Africans, yellow dots = Americans, brown dots = East Asians and red dots = South Asians).

Table 1. Distribution of the COVID-19 related SNPs of the high allele frequencies across the five studied population groups.

SNP name	EUR [A]	AFR [B]	AMR [C]	EAS [D]	SAS [E]
rs6775748	✓	✓		✓	✓
rs6054661	✓				
rs11127334	✓	✓	✓	✓	
rs2795384	✓	✓			✓
rs4766676		✓			
rs1800795		✓		✓	
rs1800797		✓		✓	
rs4702		✓			
rs1126579		✓			
rs180040		✓			
rs6127				✓	

COVID-19 related SNPs at the high allele frequency cluster were identified for each population and trends in their distribution were studied. Table 1 summarizes results regarding the high-frequency cluster. The largest number of COVID-19 related SNPs with high allele frequency can be observed within the African population, while the smallest number can be observed within the American population. Europeans and East Asians appear to share a similar sum of COVID-19 related SNPs with high allele frequency, although there is distinction between the specific high-frequency allele SNPs which they harbor.

Population-specific percentages of SNPs of the high and low allele frequency class against the total number of COVID-19 related SNPs are presented in Table 2. It can be observed that low-frequency SNP alleles exhibit a high percentage - more than 60% across all populations - while high-frequency range between 0,5 and 2,7%. These results could be the basis for designing a common treatment for endometriosis with significant discrepancies depending on the population group (126, 127).

Table 2. Percentage of COVID-19 related SNPs with high and low allele frequencies within the five studied populations.

Population	***High %***	***Low %***
EUR [A]	1,17	67,25
AFR [B]	2,65	63,12
AMR [C]	0,29	64,30
EAS [D]	1,47	64,60
SAS [E]	0,58	62,24

GWAS Catalogue reports our "high"-SNP rs6775748 as an intergenic variant, which was identified in a study investigating genetic and nongenetic COVID-19 associations (128). Genomic regions closest to the intergenic variant include the *SI* gene, which codes for the sucrase-isomaltase protein, and LINC01324, a long intergenic non-protein coding RNA (ncRNA) 1324 (129-131). The same study which reported rs6776748 also reported rs6054661, another intergenic variant mapped between the *BMP2* gene, which codes for bone morphogenetic protein 2, and LINC01428, a long intergenic ncRNA (128). Rs11127334 is an intron variant which is mapped to *MYT1L*, the gene which codes for myelin transcription factor 1 like protein, and the intergenic lncRNA LINC01250 (128, 132). Rs2795384 constitutes an intergenic variant, and its mapped genomic regions include *ARL2BPP7*, the ADP ribosylation factor like GTPase 2 binding protein pseudogene, and *MTND3P4*, MT-ND3 pseudogene 4 (128, 133). Rs4766676 is mapped in the OAS1 gene, more specifically as an intron variant. Intronic variants can influence the process of alternative splicing through interference with the recognition of the splice site, potentially leading to production of malfunctioning protein products (134). As mentioned, 2′-5′-oligoadenylate synthetase 1 (OAS1) is a key player against viral infections, as this interferon-activated enzyme degrades viral RNA in partnership with RNase L (135). According to GWAS Catalogue, rs1800795 is an intron variant of the *IL6* gene, its antisense RNA 1 (*IL6-AS1*), and *STEAP1B*. The latter may encode two different transcripts, STEAP1B2, which is overexpressed in prostate cancer cells, and STEAP1B1 (136). *IL6*, *IL6-AS1* and *STEAP1B* are also mapped to the rs1800797 polymorphism, a non-coding transcript exon variant. This variant has been linked to two immune-related pathologies, asthma and systemic lupus erythematosus (137, 138). Rs4702 is a variant located in the 3 prime untranslated region (UTR) of the *FURIN* gene, which codes for FURIN, a pro-protease convertase bound to host membranes (139). This SNP has been shown to influence alveolar and neuron infection by SARS-CoV-2 *in vitro* (140). The SARS-CoV-2 spike harbors a FURIN cleavage site, which promotes entry into lung cells and is absent from SARS-CoV (141, 142). Rs1126579 is an SNP identified at the *CXCR2* gene, leading to a 3 prime UTR variant. The gene codes for C-X-C motif chemokine receptor 2 (CXCR2), a key stimulator of immune cell migration, which binds to interleukin 8 (IL8) and chemokine ligand 1 (143). Another "high" SNP, rs180040, is mapped to the *CYP1B1*

gene, which encodes an enzyme of the cytochrome P450 family of monooxygenases that catalyze reactions of lipid synthesis and drug metabolism (144). Drug clearance is an important element in COVID-19 patient treatment; the state of hyperinflammation which is often observed in COVID-19 can potentially alter the function of cytochrome P450 enzymes in critical organs, thus affecting drug clearance and the course of therapeutic regiments (145). Lastly, rs6127 is mapped to the *SELP* gene, which codes for selectin P, a cell adhesion molecule (146). In a recent whole exome sequencing study, the polymorphism was found to be associated with thrombosis and COVID-19 severity in male patients (147). Overall, the COVID-19 related SNPs which fall into the cluster of high allele frequency are located in varying genomic regions, from introns and exons to intergenic regions. These results provide insight into key genetic targets within the studied population groups, with the potential to inform and guide policies of management and treatment according to the population-specific COVID-19 related SNPs and their corresponding clusters of allele frequencies.

Discussion

Human SNPs and their influence on enhanced resistance or susceptibility to viral disease have been the subject of intensive research, applied to pathogens of global concern, such as influenza or HIV (148-150). Similarly, human genomic variants related to COVID-19 severity and sensitivity to infection are being evaluated as tools to guide and adjust therapeutic strategies, such as the choice of administered drugs (151). The wide range of pathologies that already exist within the global population inevitably lead to the formation of a complex web, with layers of potential comorbidities between them and COVID-19. As we move into the age of "omics" and the COVID-19 related data becomes vast and heterogeneous, the modern framework of computational systems lends itself to researchers as a powerful tool. The natural-language processing pipeline proposed herein enables the effective search for potential connections between COVID-19, genes and other diseases, using the trove of characterized SNPs as our guiding light. With a combination of semantic analysis and machine learning we have drawn COVID-19's "genomic grammar", i.e the associated genomic regions which house the SNPs. Furthermore, we have examined disease profiling ontologies in connection with COVID-19, such as neoplasms, chronic hepatitis C and Alzheimer's. This firstly allows the identification of risk groups and secondly, may inform efforts towards personalized medicine, where the patient's genomic makeup determines the therapeutic approach. Lastly, we sought to expand our results to a populational scale, encompassing data from the 1000 Genomes Project, to gain insight into key genetic targets for potential exploration in studied population groups. A major portion of the scientific community's effort is dedicated to studying SARS-CoV-2's genome and its coding products in the search of effective ways to control and face the ongoing pandemic. Simultaneously, the genetic background of the human host – and its variations - constitute a source of invaluable knowledge, which can complete the picture and inform the search for effective, safe and targeted COVID-19 therapeutics.

Acknowledgments

Not applicable.

Funding

The authors would like to acknowledge funding from the following organizations: i) AdjustEBOVGP-Dx (RIA2018EF-2081): Biochemical Adjustments of native EBOV Glycoprotein in Patient Sample to Unmask target Epitopes for Rapid Diagnostic Testing. A European and Developing Countries Clinical Trials Partnership (EDCTP2) under the Horizon 2020 'Research and Innovation Actions' DESCA; and ii) 'MilkSafe: A novel pipeline to enrich formula milk using omics technologies', a research co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T2EDK-02222).

References

1. Mavrodiev EV, Tursky ML, Mavrodiev NE, Ebach MC and Williams DM: On Classification and Taxonomy of Coronaviruses (Riboviria, Nidovirales, Coronaviridae) with the special focus on severe acute respiratory syndrome-related coronavirus 2 (SARS-Cov-2). bioRxiv 2020.2010.2017.343749, 2020.

2. Woo P, Lau S, Lam C, Lau C, Tsang A, Lau J, Bai R, Teng J, Tsang C-C*, et al.*: Discovery of Seven Novel Mammalian and Avian Coronaviruses in the Genus Deltacoronavirus Supports Bat Coronaviruses as the Gene Source of Alphacoronavirus and Betacoronavirus and Avian Coronaviruses as the Gene Source of Gammacoronavirus and Deltacoronavirus. Journal of Virology 86: 3995-4008, 2012.

3. Korsman SNJ, van Zyl GU, Nutt L, Andersson MI and Preiser W: Human coronaviruses. Virology 94-95, 2012.

4. Wang Y, Grunewald M and Perlman S: Coronaviruses: An Updated Overview of Their Replication and Pathogenesis. In: Coronaviruses: Methods and Protocols. Maier HJ and Bickerton E (eds). Springer US, New York, NY, pp1-29, 2020.

5. Lai MMC and Cavanagh D: The Molecular Biology of Coronaviruses. In: Advances in Virus Research. Vol 48. Maramorosch K, Murphy FA and Shatkin AJ (eds). Academic Press, pp1-100, 1997.

6. Li F: Receptor recognition mechanisms of coronaviruses: a decade of structural studies. Journal of virology 89: 1954-1964, 2015.

7. Millet JK, Jaimes JA and Whittaker GR: Molecular diversity of coronavirus host cell entry receptors. FEMS Microbiology Reviews 2020.

8. Duan L, Zheng Q, Zhang H, Niu Y, Lou Y and Wang H: The SARS-CoV-2 Spike Glycoprotein Biosynthesis, Structure, Function, and Antigenicity: Implications for the Design of Spike-Based Vaccine Immunogens. Frontiers in Immunology 11: 2593, 2020.

9. Izaguirre G: The Proteolytic Regulation of Virus Cell Entry by Furin and Other Proprotein Convertases. Viruses 11: 837, 2019.

10. Xiao X, Chakraborti S, Dimitrov A, Gramatikoff K and Dimitrov D: The SARS-COV S glycoprotein: Expression and functional characterization. Biochemical and Biophysical Research Communications 312: 1159-1164, 2003.

11. Simmons G, Reeves J, Rennekamp A, Amberg S, Piefer A and Bates P: Characterization of severe acute respiratory syndrome-associated coronavirus (SARS-CoV) spike glycoprotein-mediated viral entry. Proceedings of the National Academy of Sciences of the United States of America 101: 4240-4245, 2004.

12. Wang W, Ye L, Ye L, Li B, Gao B, Zeng Y, Kong L, Fang X, Zheng H*, et al.*: Up-regulation of IL-6 and TNF-α induced by SARS-coronavirus spike protein in murine macrophages via NF-κB pathway. Virus Research 128: 1-8, 2007.

13. Dienz O and Rincon M: The effects of IL-6 on CD4 T cell responses. Clinical Immunology 130: 27-33, 2009.

14. Hsu ACY, Wang G, Reid AT, Veerati PC, Pathinayake PS, Daly K, Mayall JR, Hansbro PM, Horvat JC*, et al.*: SARS-CoV-2 Spike protein promotes hyper-inflammatory response that can be ameliorated by Spike-antagonistic peptide and FDA-approved ER stress and MAP kinase inhibitors <em>in vitro</em&gt. bioRxiv 2020.2009.2030.317818, 2020.

15. Farsalinos K, Eliopoulos E, Leonidas D, Papadopoulos G, Tzartos S and Poulas K: Title. Molecular modelling and docking experiments examining the interaction between SARS-CoV-2 spike glycoprotein and neuronal nicotinic acetylcholine receptors. 2020.

16. Hu B, Guo H, Zhou P and Shi Z-L: Characteristics of SARS-CoV-2 and COVID-19. Nature Reviews Microbiology 19: 141-154, 2021.

17. Harrison AG, Lin T and Wang P: Mechanisms of SARS-CoV-2 Transmission and Pathogenesis. Trends in Immunology 41: 1100-1115, 2020.

18. Booth ATC, Reed AB, Ponzo S, Yassaee A, Aral M, Plans D, Labrique AB and Mohan D: Population risk factors for severe disease and mortality in COVID-19: A global systematic review and meta-analysis. PLoS ONE 16: 2021.

19. Papakonstantinou E, Dragoumani K, Efthimiadou A, Palaiogeorgou AM, Pierouli K, Mitsis T, Chrousos GP, Bacopoulou F and Vlachakis D: Haematological malignancies implications during the times of the COVID-19 pandemic. Oncol Lett 22: 856, 2021.

20. Chen X, Li R, Pan Z, Qian C, Yang Y, You R, Zhao J, Liu P, Gao L*, et al.*: Human monoclonal antibodies block the binding of SARS-CoV-2 spike protein to angiotensin converting enzyme 2 receptor. Cellular & Molecular Immunology 17: 2020.

21. Wu Y, Wang F, Shen C, Peng W, Li D, Zhao C, Li Z, Li S, Bi Y*, et al.*: A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. Science 368: 1274-1278, 2020.

22. Brouwer PJM, Caniels TG, van der Straten K, Snitselaar JL, Aldon Y, Bangaru S, Torres JL, Okba NMA, Claireaux M*, et al.*: Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. Science 369: 643, 2020.

23. Heinz FX and Stiasny K: Distinguishing features of current COVID-19 vaccines: knowns and unknowns of antigen presentation and modes of action. npj Vaccines 6: 104, 2021.

24. Mathieu E, Ritchie H, Ortiz-Ospina E, Roser M, Hasell J, Appel C, Giattino C and Rodés-Guirao L: A global database of COVID-19 vaccinations. Nature Human Behaviour 5: 947-953, 2021.

25. Jones SP: Imperial College London Big Data Analytical Unit and YouGov Plc. 2020, Imperial College London YouGov Covid Data Hub, v1.0. YouGov Plc 2020.

26. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, Zhang X, Muruato AE, Zou J*, et al.*: Spike mutation D614G alters SARS-CoV-2 fitness. Nature 592: 116-121, 2021.

27. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T*, et al.*: Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell 182: 812-827.e819, 2020.

28. Cano-Gamez E and Trynka G: From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Frontiers in Genetics 11: 424, 2020.

29. Lewis CM and Vassos E: Polygenic risk scores: from research tools to clinical instruments. Genome Medicine 12: 44, 2020.

30. Vlachakis D, Papakonstantinou E, Sagar R, Bacopoulou F, Exarchos T, Kourouthanassis P, Karyotis V, Vlamos P, Lyketsos C*, et al.*: Improving the Utility of Polygenic Risk Scores as a Biomarker for Alzheimer's Disease. Cells 10: 2021.

31. Jiang K, Zhu L, Buck MJ, Chen Y, Carrier B, Liu T and Jarvis JN: Disease-Associated Single-Nucleotide Polymorphisms From Noncoding Regions in Juvenile Idiopathic Arthritis Are Located Within or

Adjacent to Functional Genomic Elements of Human Neutrophils and CD4+ T Cells. Arthritis Rheumatol 67: 1966-1977, 2015.

32. Giral H, Landmesser U and Kratzer A: Into the Wild: GWAS Exploration of Non-coding RNAs. Frontiers in Cardiovascular Medicine 5: 181, 2018.

33. Altshuler D, Durbin R, Abecasis G, Bentley D, Chakravarti A, Clark A, Collins F, De La Vega F, Donnelly P*, et al.*: A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073, 2010.

34. Gazal S, Sahbatou M, Babron M-C, Génin E and Leutenegger A-L: High level of inbreeding in final phase of 1000 Genomes Project. Sci Rep 5: 17453-17453, 2015.

35. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P*, et al.*: A global reference for human genetic variation. Nature 526: 68-74, 2015.

36. Taioli E, Pedotti P and Garte S: Importance of allele frequency estimates in epidemiological studies. Mutation Research/Reviews in Mutation Research 567: 63-70, 2004.

37. Chen N, Juric I, Cosgrove EJ, Bowman R, Fitzpatrick JW, Schoech SJ, Clark AG and Coop G: Allele frequency dynamics in a pedigreed natural population. Proceedings of the National Academy of Sciences 116: 2158, 2019.

38. Vlachakis D, Papakonstantinou E, Mitsis T, Pierouli K, Diakou I, Chrousos G and Bacopoulou F: Molecular mechanisms of the novel coronavirus SARS-CoV-2 and potential anti-COVID19 pharmacological targets since the outbreak of the pandemic. Food Chem Toxicol 146: 111805, 2020.

39. Roberts RJ: PubMed Central: The GenBank of the published literature. Proceedings of the National Academy of Sciences 98: 381, 2001.

40. Papageorgiou L, Zervou MI, Vlachakis D, Matalliotakis M, Matalliotakis I, Spandidos DA, Goulielmos GN and Eliopoulos E: Demetra Application: An integrated genotype analysis web server for clinical genomics in endometriosis. Int J Mol Med 47: 115, 2021.

41. Buniello A, MacArthur J, Cerezo M, Harris L, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E*, et al.*: The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research 47: 2018.

42. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM and Sirotkin K: dbSNP: the NCBI database of genetic variation. Nucleic acids research 29: 308-311, 2001.

43. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D*, et al.*: ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research 46: D1062-D1067, 2018.

44. Allot A, Peng Y, Wei C-H, Lee K, Phan L and Lu Z: LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. Nucleic acids research 46: W530-W536, 2018.

45. Hamosh A, Scott AF, Amberger JS, Bocchini CA and McKusick VA: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research 33: D514-D517, 2005.

46. Merelli I, Calabria A, Cozzi P, Viti F, Mosca E and Milanesi L: SNPranker 2.0: a gene-centric data mining tool for diseases associated SNP prioritization in GWAS. BMC Bioinformatics 14: S9, 2013.

47. Papageorgiou L, Alkenaris H, Zervou MI, Vlachakis D, Matalliotakis I, Spandidos DA, Bertsias G, Goulielmos GN and Eliopoulos E: Epione application: An integrated web-toolkit of clinical genomics and personalized medicine in systemic lupus erythematosus. Int J Mol Med 49: 8, 2022.

48. Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P and Genomes Project C: Alignment of 1000 Genomes Project reads to reference assembly GRCh38. GigaScience 6: 1-8, 2017.

49. Lowy-Gallego E, Fairley S, Zheng-Bradley X, Ruffier M, Clarke L, Flicek P and Genomes Project C: Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. Wellcome open research 4: 50, 2019.

50. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT*, et al.*: The variant call format and VCFtools. Bioinformatics 27: 2156-2158, 2011.

51. Cui Y, Chen X, Luo H, Fan Z, Luo J, he S, Yue H, Peng Z and Chen R: BioCircos.js: an Interactive Circos JavaScript Library for Biological Data Visualization on Web Applications. Bioinformatics (Oxford, England) 32: 2016.

52. Santos RAS, Sampaio WO, Alzamora AC, Motta-Santos D, Alenina N, Bader M and Campagnole-Santos MJ: The ACE2/Angiotensin-(1–7)/MAS Axis of the Renin-Angiotensin System: Focus on Angiotensin-(1–7). Physiological Reviews 98: 505-553, 2017.

53. Iwai M and Horiuchi M: Devil and angel in the renin–angiotensin system: ACE–angiotensin II–AT1 receptor axis vs. ACE2–angiotensin-(1–7)–Mas receptor axis. Hypertension Research 32: 533-536, 2009.

54. Carey RM: AT2 Receptors: Potential Therapeutic Targets for Hypertension. American Journal of Hypertension 30: 339-347, 2017.

55. Keidar S, Kaplan M and Gamliel-Lazarovich A: ACE2 of the heart: From angiotensin I to angiotensin (1–7). Cardiovascular Research 73: 463-469, 2007.

56. Santos R, Ferreira A, Verano-Braga T and Bader M: Angiotensin-converting enzyme 2, angiotensin-(1-7) and Mas: New players of the renin-angiotensin system. The Journal of endocrinology 216: 2012.

57. Jackson CB, Farzan M, Chen B and Choe H: Mechanisms of SARS-CoV-2 entry into cells. Nature Reviews Molecular Cell Biology 2021.

58. Reindl-Schwaighofer R, Hödlmoser S, Eskandary F, Poglitsch M, Bonderman D, Strassl R, Aberle J, Oberbauer R, Zoufaly A*, et al.*: Angiotensin-Converting Enzyme 2 (ACE2) Elevation in Severe COVID-19. American Journal of Respiratory and Critical Care Medicine 203: 2021.

59. Iwasaki M, Saito J, Zhao H, Sakamoto A, Hirota K and Ma D: Inflammation Triggered by SARS-CoV-2 and ACE2 Augment Drives Multiple Organ Failure of Severe COVID-19: Molecular Mechanisms and Implications. Inflammation 44: 13-34, 2021.

60. loganathan S, Kuppusamy M, Wankhar W, Gurugubelli KR, Mahadevappa VH, Lepcha L and Choudhary Ak: Angiotensin-converting enzyme 2 (ACE2): COVID 19 gate way to multiple organ failure syndromes. Respiratory Physiology & Neurobiology 283: 103548, 2021.

61. Suryamohan K, Diwanji D, Stawiski EW, Gupta R, Miersch S, Liu J, Chen C, Jiang Y-P, Fellouse FA*, et al.*: Human ACE2 receptor polymorphisms and altered susceptibility to SARS-CoV-2. Communications Biology 4: 475, 2021.

62. Möhlendick B, Schönfelder K, Breuckmann K, Elsner C, Babel N, Balfanz P, Dahl E, Dreher M, Fistera D*, et al.*: ACE2 polymorphism and susceptibility for SARS-CoV-2 infection and severity of COVID-19. Pharmacogenetics and Genomics 31: 2021.

63. Khayat A, Assumpcao P, Khayat B, Araújo T, Batista-Gomes J, Imbiriba L, Ishak G, Assumpção P, Moreira F*, et al.*: ACE2 polymorphisms as potential players in COVID-19 outcome. PLOS ONE 15: e0243887, 2020.

64. Tanaka T, Narazaki M and Kishimoto T: IL-6 in inflammation, immunity, and disease. Cold Spring Harb Perspect Biol 6: a016295-a016295, 2014.

65. Velazquez-Salinas L, Verdugo-Rodriguez A, Rodriguez LL and Borca MV: The Role of Interleukin 6 During Viral Infections. Frontiers in Microbiology 10: 2019.

66. Brábek J, Jakubek M, Vellieux F, Novotný J, Kolář M, Lacina L, Szabo P, Strnadová K, Rösel D*, et al.*: Interleukin-6: Molecule in the Intersection of Cancer, Ageing and COVID-19. International Journal of Molecular Sciences 21: 2020.

67. Zhu Z, Cai T, Fan L, Lou K, Hua X, Huang Z and Gao G: Clinical value of immune-inflammatory parameters to assess the severity of coronavirus disease 2019. International Journal of Infectious Diseases 95: 332-339, 2020.

68. Karcioglu Batur L and Hekim N: Correlation between interleukin gene polymorphisms and current prevalence and mortality rates due to novel coronavirus disease 2019 (COVID-2019) in 23 countries. Journal of Medical Virology 93: 5853-5863, 2021.

69. Chen T, Lin Y-X, Zha Y, Sun Y, Tian J, Yang Z, Lin S-W, Yu F, Chen Z-S*, et al.*: A Low-Producing Haplotype of Interleukin-6 Disrupting CTCF Binding Is Protective against Severe COVID-19. mBio 12: e0137221-e0137221, 2021.

70. Merkhofer R, O'Neill M, Xiong D, Hernandez-Santos N, Dobson H, Fites J, Shockey A, Wuethrich M, Pepperell C*, et al.*: Investigation of Genetic Susceptibility to Blastomycosis Reveals Interleukin-6 as a Potential Susceptibility Locus. mBio 10: 2019.

71. Yi E, Zhang J, Zheng M, Zhang Y, Liang C, Hao B, Hong W, Lin B, Pu J*, et al.*: Long noncoding RNA IL6-AS1 is highly expressed in chronic obstructive pulmonary disease and is associated with interleukin 6 by targeting miR-149-5p and early B-cell factor 1. Clinical and Translational Medicine 11: e479, 2021.

72. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, Wu N-H*, et al.*: SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell 181: 271-280.e278, 2020.

73. Cheng Z, Zhou J, To K, Chu H, Li C, Wang D, Yang D, Zheng S, Hao K*, et al.*: Identification of TMPRSS2 as a susceptibility gene for severe 2009 pandemic A(H1N1) influenza and A(H7N9) influenza. Journal of Infectious Diseases 212: 2015.

74. Luostari K, Hartikainen JM, Tengström M, Palvimo JJ, Kataja V, Mannermaa A and Kosma V-M: Type II transmembrane serine protease gene variants associate with breast cancer. PloS one 9: e102519-e102519, 2014.

75. Paniri A, Hosseini MM and Akhavan-Niaki H: First comprehensive computational analysis of functional consequences of TMPRSS2 SNPs in susceptibility to SARS-CoV-2 among different populations. J Biomol Struct Dyn 39: 3576-3593, 2021.

76. Asselta R, Paraboschi EM, Mantovani A and Duga S: ACE2 and TMPRSS2 variants and expression as candidates to sex and country differences in COVID-19 severity in Italy. Aging (Albany NY) 12: 10087-10098, 2020.

77. David A, Parkinson N, Peacock TP, Pairo-Castineira E, Khanna T, Cobat A, Tenesa A, Sancho-Shimizu V, Gen OI*, et al.*: A common TMPRSS2 variant protects against severe COVID-19. medRxiv 2021.2003.2004.21252931, 2021.

78. Clausen H, Bennett EP and Grunnet N: Molecular genetics of ABO histo-blood groups. Transfusion Clinique et Biologique 1: 79-89, 1994.

79. Yamamoto F-i, McNeill PD and Hakomori S-i: Genomic organization of human histo-blood group ABO genes. Glycobiology 5: 51-58, 1995.

80. Hernández Cordero AI, Li X, Milne S, Yang CX, Bossé Y, Joubert P, Timens W, van den Berge M, Nickle D*, et al.*: Multi-omics highlights ABO plasma protein as a causal risk factor for COVID-19. Hum Genet 140: 969-979, 2021.

81. Ibsen MS, Gad HH, Thavachelvam K, Boesen T, Desprès P and Hartmann R: The 2'-5'-oligoadenylate synthetase 3 enzyme potently synthesizes the 2'-5'-oligoadenylates required for RNase L activation. Journal of virology 88: 14222-14231, 2014.

82. Rana R, Ranjan V and Kumar N: Association of ABO and Rh Blood Group in Susceptibility, Severity, and Mortality of Coronavirus Disease 2019: A Hospital-Based Study From Delhi, India. Frontiers in Cellular and Infection Microbiology 11: 1071, 2021.

83. Delanghe JR, De Buyzere ML and Speeckaert MM: ABO Blood Groups and Coronavirus Disease 2019 (COVID-19). Clinical Infectious Diseases 72: e917-e917, 2021.

84. Promchan K and Natarajan V: Leucine zipper transcription factor-like 1 binds adaptor protein complex-1 and 2 and participates in trafficking of transferrin receptor 1. PLOS ONE 15: e0226298, 2020.

85. Seo S, Zhang Q, Bugge K, Breslow DK, Searby CC, Nachury MV and Sheffield VC: A Novel Protein LZTFL1 Regulates Ciliary Trafficking of the BBSome and Smoothened. PLOS Genetics 7: e1002358, 2011.

86. Downes DJ, Cross AR, Hua P, Roberts N, Schwessinger R, Cutler AJ, Munis AM, Brown J, Mielczarek O*, et al.*: Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus. Nature Genetics 53: 1606-1615, 2021.

87. Ravindra NG, Alfajaro MM, Gasque V, Habet V, Wei J, Filler RB, Huston NC, Wan H, Szigeti-Buck K*, et al.*: Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium. bioRxiv : the preprint server for biology 2020.2005.2006.081695, 2020.

88. Markert CL: Neoplasia: A Disease of Cell Differentiation. Cancer Research 28: 1908, 1968.

89. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, Curtis HJ, Mehrkar A, Evans D*, et al.*: Factors associated with COVID-19-related death using OpenSAFELY. Nature 584: 430-436, 2020.

90. Barbui T, De Stefano V, Alvarez-Larran A, Iurlo A, Masciulli A, Carobbio A, Ghirardi A, Ferrari A, Cancelli V*, et al.*: Among classic myeloproliferative neoplasms, essential thrombocythemia is associated with the greatest risk of venous thromboembolism during COVID-19. Blood Cancer Journal 11: 21, 2021.

91. Barbui T, Vannucchi AM, Alvarez-Larran A, Iurlo A, Masciulli A, Carobbio A, Ghirardi A, Ferrari A, Rossi G*, et al.*: High mortality rate in COVID-19 patients with myeloproliferative neoplasms after abrupt withdrawal of ruxolitinib. Leukemia 35: 485-493, 2021.

92. Camps J, Marsillach Lopez J and Joven J: The paraoxonases: Role in human diseases and methodological difficulties in measurement. Critical reviews in clinical laboratory sciences 46: 83-106, 2009.

93. Pan X, Huang L, Li M, Mo D, Liang Y, Liu Z, Huang Z, Huang L, Liu J*, et al.*: The Association between PON1 (Q192R and L55M) Gene Polymorphisms and Risk of Cancer: A Meta-Analysis Based on 43 Studies. BioMed Research International 2019: 5897505, 2019.

94. Diels S, Cuypers B, Tvarijonaviciute A, Derudas B, Van Dijck E, Verrijken A, Van Gaal LF, Laukens K, Lefebvre P*, et al.*: A targeted multi-omics approach reveals paraoxonase-1 as a determinant of obesity-associated fatty liver disease. Clinical Epigenetics 13: 158, 2021.

95. Saadat M: Prevalence and mortality of COVID-19 are associated with the L55M functional polymorphism of Paraoxonase 1. Proceedings of Singapore Healthcare 20101058211040582, 2021.

96. Mustroph J, Hupf J, Baier MJ, Evert K, Brochhausen C, Broeker K, Meindl C, Seither B, Jungbauer C*, et al.*: Cardiac Fibrosis Is a Risk Factor for Severe COVID-19. Frontiers in Immunology 12: 4456, 2021.

97. Meng X-m, Nikolic-Paterson DJ and Lan HY: TGF-β: the master regulator of fibrosis. Nature Reviews Nephrology 12: 325-338, 2016.

98. Ramos-Mondragón R, Galindo CA and Avila G: Role of TGF-beta on cardiac structural and electrical remodeling. Vasc Health Risk Manag 4: 1289-1300, 2008.

99. Mayi BS, Leibowitz JA, Woods AT, Ammon KA, Liu AE and Raja A: The role of Neuropilin-1 in COVID-19. PLoS Pathog 17: e1009153-e1009153, 2021.

100. Michalski JE, Kurche JS and Schwartz DA: From ARDS to pulmonary fibrosis: the next phase of the COVID-19 pandemic? Translational Research 2021.

101. Benedict M and Zhang X: Non-alcoholic fatty liver disease: An expanded review. World J Hepatol 9: 715-732, 2017.

102. Carey IM, Critchley JA, DeWilde S, Harris T, Hosking FJ and Cook DG: Risk of Infection in Type 1 and Type 2 Diabetes Compared With the General Population: A Matched Cohort Study. Diabetes Care 41: 513, 2018.

103. De Jong A, Molinari N, Pouzeratte Y, Verzilli D, Chanques G, Jung B, Futier E, Perrigault PF, Colson P*, et al.*: Difficult intubation in obese patients: incidence, risk factors, and complications in the operating theatre and in intensive care units. British Journal of Anaesthesia 114: 297-306, 2015.

104. Fargion S, Porzio M and Fracanzani AL: Nonalcoholic fatty liver disease and vascular disease: state-of-the-art. World J Gastroenterol 20: 13306-13324, 2014.

105. Singh A, Hussain S and Antony B: Non-alcoholic fatty liver disease and clinical outcomes in patients with COVID-19: A comprehensive systematic review and meta-analysis. Diabetes Metab Syndr 15: 813-822, 2021.

106. Zheng KI, Gao F, Wang X-B, Sun Q-F, Pan K-H, Wang T-Y, Ma H-L, Chen Y-P, Liu W-Y*, et al.*: Letter to the Editor: Obesity as a risk factor for greater severity of COVID-19 in patients with metabolic associated fatty liver disease. Metabolism 108: 154244-154244, 2020.

107. Sachdeva S, Khandait H, Kopel J, Aloysius MM, Desai R and Goyal H: NAFLD and COVID-19: a Pooled Analysis. SN Compr Clin Med 1-4, 2020.

108. Kern L, Mittenbühler MJ, Vesting AJ, Ostermann AL, Wunderlich CM and Wunderlich FT: Obesity-Induced TNFα and IL-6 Signaling: The Missing Link between Obesity and Inflammation-Driven Liver and Colorectal Cancers. Cancers (Basel) 11: 24, 2018.

109. Han MS, White A, Perry RJ, Camporez J-P, Hidalgo J, Shulman GI and Davis RJ: Regulation of adipose tissue inflammation by interleukin 6. Proceedings of the National Academy of Sciences 117: 2751, 2020.

110. Ye Q, Wang B and Mao J: The pathogenesis and treatment of the `Cytokine Storm' in COVID-19. J Infect 80: 607-613, 2020.

111. Jia F, Wang G, Xu J, Long J, Deng F and Jiang W: Role of tumor necrosis factor-α in the mortality of hospitalized patients with severe and critical COVID-19 pneumonia. Aging (Albany NY) 13: 23895-23912, 2021.

112. De Pergola G and Pannacciulli N: Coagulation and fibrinolysis abnormalities in obesity. Journal of Endocrinological Investigation 25: 899-904, 2002.

113. Singhania N, Bansal S, Nimmatoori DP, Ejaz AA, McCullough PA and Singhania G: Current Overview on Hypercoagulability in COVID-19. American Journal of Cardiovascular Drugs 20: 393-403, 2020.

114. Li Y, Zhao K, Wei H, Chen W, Wang W, Jia L, Liu Q, Zhang J, Shan T*, et al.*: Dynamic relationship between D-dimer and COVID-19 severity. Br J Haematol 190: e24-e27, 2020.

115. Millman AJ, Nelson NP and Vellozzi C: Hepatitis C: Review of the Epidemiology, Clinical Care, and Continued Challenges in the Direct Acting Antiviral Era. Curr Epidemiol Rep 4: 174-185, 2017.

116. Rugwizangoga B, Andersson ME, Kabayiza J-C, Nilsson MS, Ármannsdóttir B, Aurelius J, Nilsson S, Hellstrand K, Lindh M*, et al.*: IFNL4 Genotypes Predict Clearance of RNA Viruses in Rwandan Children With Upper Respiratory Tract Infections. Frontiers in Cellular and Infection Microbiology 9: 340, 2019.

117. Thomas DL, Thio CL, Martin MP, Qi Y, Ge D, O'Huigin C, Kidd J, Kidd K, Khakoo SI, *et al.*: Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. Nature 461: 798-801, 2009.

118. Prokunina-Olsson L, Muchmore B, Tang W, Pfeiffer RM, Park H, Dickensheets H, Hergott D, Porter-Gill P, Mumy A, *et al.*: A variant upstream of IFNL3 (IL28B) creating a new interferon gene IFNL4 is associated with impaired clearance of hepatitis C virus. Nature Genetics 45: 164-171, 2013.

119. Saponi-Cortes JMR, Rivas MD, Calle-Alonso F, Sanchez JF, Costo A, Martin C and Zamorano J: IFNL4 genetic variant can predispose to COVID-19. Sci Rep 11: 21185, 2021.

120. Kuo C-L, Pilling LC, Atkins JL, Masoli JAH, Delgado J, Kuchel GA and Melzer D: APOE e4 Genotype Predicts Severe COVID-19 in the UK Biobank Community Cohort. J Gerontol A Biol Sci Med Sci 75: 2231-2232, 2020.

121. Taylor K, Das S, Pearson M, Kozubek J, Pawlowski M, Jensen CE, Skowron Z, Møller GL, Strivens M, *et al.*: Analysis of Genetic Host Response Risk Factors in Severe COVID-19 Patients. medRxiv 2020.2006.2017.20134015, 2020.

122. Eberle MA, Rieder MJ, Kruglyak L and Nickerson DA: Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. PLoS Genet 2: e142, 2006.

123. Norton N, Williams NM, Williams HJ, Spurlock G, Kirov G, Morris DW, Hoogendoorn B, Owen MJ and O'Donovan MC: Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. Hum Genet 110: 471-478, 2002.

124. Fredman D, Sawyer SL, Stromqvist L, Mottagui-Tabar S, Kidd KK, Wahlestedt C, Chanock SJ and Brookes AJ: Nonsynonymous SNPs: validation characteristics, derived allele frequency patterns, and suggestive evidence for natural selection. Human mutation 27: 173-186, 2006.

125. Kobayashi Y, Yang S, Nykamp K, Garcia J, Lincoln SE and Topper SE: Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. Genome Medicine 9: 13, 2017.

126. Gerlinger C, Faustmann T, Hassall JJ and Seitz C: Treatment of endometriosis in different ethnic populations: a meta-analysis of two clinical trials. BMC women's health 12: 9, 2012.

127. Bougie O, Yap MI, Sikora L, Flaxman T and Singh S: Influence of race/ethnicity on prevalence and presentation of endometriosis: a systematic review and meta-analysis. BJOG : an international journal of obstetrics and gynaecology 126: 1104-1115, 2019.

128. Shelton JF, Shastri AJ, Ye C, Weldon CH, Filshtein-Sonmez T, Coker D, Symons A, Esparza-Gordillo J, Chubb A, *et al.*: Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. Nature Genetics 53: 801-808, 2021.

129. Rodriguez IR, Taravel FR and Whelan WJ: Characterization and function of pig intestinal sucrase-isomaltase and its separate subunits. European Journal of Biochemistry 143: 575-582, 1984.

130. Hauri HP, Quaroni A and Isselbacher KJ: Biogenesis of intestinal plasma membrane: posttranslational route and cleavage of sucrase-isomaltase. Proceedings of the National Academy of Sciences of the United States of America 76: 5183-5186, 1979.

131. Fan C-N, Ma L and Liu N: Systematic analysis of lncRNA–miRNA–mRNA competing endogenous RNA network identifies four-lncRNA signature as a prognostic biomarker for breast cancer. Journal of Translational Medicine 16: 2018.

132. Mall M, Kareta MS, Chanda S, Ahlenius H, Perotti N, Zhou B, Grieder SD, Ge X, Drake S*, et al.*: Myt1l safeguards neuronal identity by actively repressing many non-neuronal fates. Nature 544: 245-249, 2017.

133. Del-Aguila JL, Beitelshees AL, Cooper-Dehoff RM, Chapman AB, Gums JG, Bailey K, Gong Y, Turner ST, Johnson JA*, et al.*: Genome-wide association analyses suggest NELL1 influences adverse metabolic response to HCTZ in African Americans. Pharmacogenomics J 14: 35-40, 2014.

134. Cooper DN: Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes. Hum Genomics 4: 284-288, 2010.

135. Yang E and Li MMH: All About the RNA: Interferon-Stimulated Genes That Interfere With Viral RNA Processes. Frontiers in Immunology 11: 3195, 2020.

136. Gomes I, Santos C and Maia C: Expression of STEAP1 and STEAP1B in prostate cell lines, and the putative regulation of STEAP1 by post-transcriptional and post-translational mechanisms. Genes & cancer 5: 142-151, 2014.

137. Han Y, Jia Q, Jahani PS, Hurrell BP, Pan C, Huang P, Gukasyan J, Woodward NC, Eskin E*, et al.*: Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. Nature Communications 11: 1776, 2020.

138. Langefeld CD, Ainsworth HC, Graham DSC, Kelly JA, Comeau ME, Marion MC, Howard TD, Ramos PS, Croker JA*, et al.*: Transancestral mapping and genetic load in systemic lupus erythematosus. Nature Communications 8: 16021, 2017.

139. Declercq J and Creemers JWM: Chapter 725 - Furin. In: Handbook of Proteolytic Enzymes (Third Edition). Rawlings ND and Salvesen G (eds). Academic Press, pp3281-3285, 2013.

140. Dobrindt K, Hoagland DA, Seah C, Kassim B, O'Shea CP, Murphy A, Iskhakova M, Fernando MB, Powell SK*, et al.*: Common Genetic Variation in Humans Impacts In Vitro Susceptibility to SARS-CoV-2 Infection. Stem Cell Reports 16: 505-518, 2021.

141. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG and Decroly E: The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res 176: 104742-104742, 2020.

142. Hoffmann M, Kleine-Weber H and Pöhlmann S: A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. Molecular Cell 78: 779-784.e775, 2020.

143. Zhang X, Guo R, Kambara H, Ma F and Luo HR: The role of CXCR2 in acute inflammatory responses and its antagonists as anti-inflammatory therapeutics. Curr Opin Hematol 26: 28-33, 2019.

144. Werck-Reichhart D and Feyereisen R: Cytochromes P450: a success story. Genome Biology 1: reviews3003.3001, 2000.

145. El-Ghiaty MA, Shoieb SM and El-Kadi AOS: Cytochrome P450-mediated drug interactions in COVID-19 patients: Current findings and possible mechanisms. Med Hypotheses 144: 110033-110033, 2020.

146. Chen M and Geng J-G: P-selectin mediates adhesion of leukocytes, platelets, and cancer cells in inflammation, thrombosis, and cancer growth and metastasis. Archivum Immunologiae et Therapiae Experimentalis 54: 75-84, 2006.

147. Fallerini C, Daga S, Benetti E, Picchiotti N, Zguro K, Catapano F, Baroni V, Lanini S, Bucalossi A*, et al.*: SELP Asp603Asn and severe thrombosis in COVID-19 males: implication for anti P-selectin monoclonal antibodies treatment. medRxiv 2021.2005.2025.21257803, 2021.

148. Kenney AD, Dowdle JA, Bozzacco L, McMichael TM, St Gelais C, Panfil AR, Sun Y, Schlesinger LS, Anderson MZ*, et al.*: Human Genetic Determinants of Viral Diseases. Annu Rev Genet 51: 241-263, 2017.

149. Ciancanelli MJ, Huang SXL, Luthra P, Garner H, Itan Y, Volpi S, Lafaille FG, Trouillet C, Schmolke M*, et al.*: Infectious disease. Life-threatening influenza and impaired interferon amplification in human IRF7 deficiency. Science 348: 448-453, 2015.

150. Samson M, Libert F, Doranz B, Rucker J, Liesnard C, Farber C, Saragosti S, Lapouméroulie C, Cognaux J*, et al.*: Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 382: 722-725, 1996.

151. Hou Y, Zhao J, Martin W, Kallianpur A, Chung MK, Jehi L, Sharifi N, Erzurum S, Eng C*, et al.*: New insights into genetic susceptibility of COVID-19: an ACE2 and TMPRSS2 polymorphism analysis. BMC Medicine 18: 216, 2020.