Infection, Genetics and Evolution 90 (2021) 104730

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Different SARS-CoV-2 haplotypes associate with geographic origin and case fatality rates of COVID-19 patients

Manisha Goyal [a], Katrien De Bruyne [b], Alex van Belkum [a], Brian West [c,*]

[a] bioMérieux, Open Innovation and Partnerships, 3 Route de Port Michaud, 38390 La Balme Les Grottes, France
[b] bioMérieux, Applied Maths, Keistraat 120, B-9830 Sint-Martens-Latem, Belgium
[c] bioMérieux, Applied Maths, 13809 Research Blvd., Suite 645, Austin, Texas 78750, USA

ARTICLE INFO

Keywords:
SARS-CoV-2
COVID-19
Genotyping
Haplotypes - geno-to-pheno correlation
Fatality risk

SUMMARY

The current pandemic of COVID-19 is caused by the SARS-CoV-2 virus for which many variants at the Single Nucleotide Polymorphism (SNP) level have now been identified. We show here that different allelic variants among 692 SARS-CoV-2 genome sequences display a statistically significant association with geographic origin ($p < 0.000001$) and COVID-19 case severity ($p = 0.016$). Geographic variation in itself is associated with both case severity and allelic variation especially in strains from Indian origin ($p < 0.000001$). Using an new alternative bioinformatics approach we were able to confirm that the presence of the D614G mutation correlates with increased case severity in a sample of 127 sequences from a shared geographic origin in the US ($p = 0.018$). While leaving open the question on the pathogenesis mechanism involved, this suggests that in specific geographic locales certain genotypes of the virus are more pathogenic than others. We here show that viral genome polymorphisms may have an effect on case severity when other factors are controlled for, but that this effect is swamped out by these other factors when comparing cases across different geographic regions.

1. Introduction

The SARS coronavirus-2 (SARS-CoV-2) causes COVID19 (Kadkhoda, 2020). This disease is now pandemic and it is killing hundreds of thousands of people on a global scale (e.g. Potere et al., 2020). Viruses, especially those with an RNA genome, have a tendency to evolve relatively rapidly during episodes of intense geographic spread. Using modern genomic sequencing technologies the genetic changes associated with global but also more local dissemination can be documented rapidly (Sekizuka et al., 2020). Also, within viral populations variants can be traced due to the quasi-species nature of the SARS-CoV-2 virus (Jary et al., 2020). For SARS-CoV-2 thousands of Single Nucleotide Polymorphisms (SNPs) have already been identified, several of which have become fixed in the more recent, geographically defined viral populations at large (Saha et al., 2020; Sapoval et al., 2020; Kaushal et al., 2020; Yang et al., 2020). Rapid regional spread of SARS-CoV-2 may lead to increased allelic variability during periods of extended transmission (Gudbjartsson et al., 2020). Although not all of these SNPs translate in amino acid variation in coding sequences (CDS), a significant portion does change the structure of important viral proteins. It is currently not clear what the effect of such variations is on viral phenotypes (e.g. its capacity to adhere to target host cells, efficiency of invasion of host cells, rapidity of replication, disease features in infected hosts etc) also because defining such effects is usually performed in artificial in vitro models. Such models are often cumbersome, have an intrinsic infectious risk for those working with it and may not adequately represent the real-life in vivo situation (Lamers et al., 2020; Leibel et al., 2020). Modern bioinformatics tools may add flexibility to such laborious assays and are helpful in defining associations between viral genome variation and differential effects that such viral variants have during infection (e.g. Gallego et al., 2004; Ji et al., 2020).

Many physiological and clinical parameters have been described that significantly contribute to COVID-19 mortality. Among these are advanced age (Papadopoulos et al., 2020), smoking (Grundy et al., 2020), obesity (Hussain et al., 2020), diabetes (Rajpal et al., 2020), hypertension (Zaki et al., 2020), cardio-vascular problems (Mishra et al., 2020) and quite some others (Thompson et al., 2020; Williamson et al., 2020). Relatively little information is available on the contribution to disease severity and mortality by viral variability itself (Pachetti et al., 2020). A physiologically important mutation changing the amino acid sequence of the RNA-dependent RNA polymerase (RdRp) was noted but the effect on disease severity could not be assessed. Furthermore, it was

* Corresponding author.
E-mail address: brian.west@biomerieux.com (B. West).

https://doi.org/10.1016/j.meegid.2021.104730
Received 27 October 2020; Received in revised form 18 January 2021; Accepted 20 January 2021
Available online 26 January 2021

Table 1
SARS-CoV-2 amino acid substitutions giving rise to haplotype variation as defined by genomic locus, position, and inferred date.

Substitution	Locus	Codon #	Date
L - > S	ORF8	84	2020-01-12
D - > G	S	614	2020-01-12
P - > L	ORF1b	314	2020-01-13
Q - > H	ORF3a	57	2020-01-23
T - > I	ORF1a	265	2020-02-23
Y - > C	ORF1b	1464	2020-02-23
P - > L	ORF1b	1427	2020-02-23

shown that a 328 basepair deletion in ORF8 clinically associated with a lesser chance for developing hypoxia during COVID-19 (Young et al., 2020). Very recently however, Toyoshima et al. (2020), Nakamichi et al. (2020) and Hodcroft et al. (2020) reported the first viral mutations that associated with fatality rates for COVID-19 and concluded that viral variation, together with host susceptibility and the environment co-define the course of COVID-19.

Using a novel viral typing tool, we here assess SNP-based haplotype variation in a large set of SARS-CoV-2 genome sequences, define the SARS-CoV-2 population structure and dynamics and associate these with clinical findings, including fatality rates, among patients.

2. Materials and methods

2.1. Collection of viral sequence information and database development

SARS-CoV-2 viral genome sequences were collected using the Global Initiative on Sharing Avian Influenza Data (GISAID) database which combined more than 90.000 genome sequences including phenotypic and disease-related metadata. Over 6400 of these sequences included relatively complete dossiers on patient status information. Sequences and metadata were stored, processed, and analyzed in a BIONUMERICS (v8.0) database, with a SQLite backend. Data quality assessment was performed by filtering the GISAID sequences for completeness (>29,000 bp) and by comparing genome sequences to the NC_045512 NCBI reference sequence. Genomic sequences were only analyzed when every CDS was the same length as the matching CDS in the NC_045512 reference sequence, i.e. without insertions or deletions.

2.2. Bioinformatic analysis of viral sequences

The BIONUMERICS SARS-CoV-2 plugin tool (bioMérieux, Applied Maths, Sint-Martens-Latem, Belgium) facilitates the processing and combined analysis of SARS-CoV-2 genomic sequences, whether downloaded from a public data repository or generated locally. The plugin tool is part of the BIONUMERICS platform and can be only used in the context of this software package. Each genomic sequence imported into BIONUMERICS was separated by the plugin tool into subsequences matching the annotated CDSs while ignoring the small fraction of intergenic regions in the NCBI reference sequence for SARS-CoV-2 (NC_045512). Next, each of these sequences was analyzed for SNPs relative to the reference sequence. SNPs were stored in the database as a character type experiment to be used for comparison and strain typing using BIONUMERICS' clustering tools (dendrograms and minimum spanning trees). SNPs were also translated, enabling SNP interpretation based on actual amino acid changes. The "haplotype", as defined in the plugin, was determined by categorization of a set of common missense SNPs translated into amino acids (Sekizuka et al., 2020). This haplotype information was also stored in the database and displayed on the trees and networks for easy group detection.

2.3. Tool modules

After being downloaded from GISAID, FASTA-formatted genomic sequences were imported into the database using a dedicated sequence import routine available in BIONUMERICS. The SARS-CoV-2 plugin applied a BLAST approach to extract 26 subsequences from each genome. The subsequences of sample Wuhan-Hu-1 (NC_045512), installed automatically by the plugin, were used as reference sequences for the BLAST searches. The subsequences extracted from the genomic sequences were stored in the corresponding destination sequence type experiments. These sequence types were identified by ORF and, for ORF1, an additional Nuclear Shuttle Protein (nsp) tag. After the BLAST screening, the following detailed results were reported for each

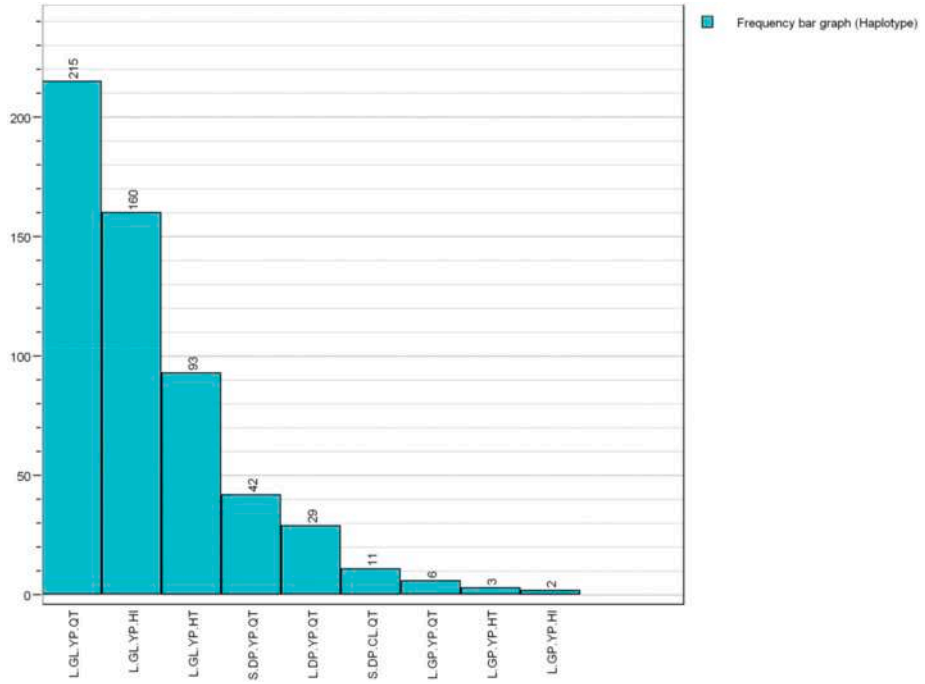

Fig. 1. SARS-CoV-2 haplotype counts among samples included in this study providing adequate patient status assessment.

Table 2
Patient status transformation into a numerical score of case severity.

Patient Status	Case Severity
Asymptomatic	1
Mild case/Outpatient/Retirement home/Symptomatic	2
Alive/Released/Recovered	3
Hospitalized	4
Severe/ICU	5
Deceased	6

destination sequence type (Locus column): whether or not a BLAST hit was found, its position on the genome sequence (Start and Stop), sequence identity (Identity (%)) and sequence overlap (Length (%)), the length of the retrieved subsequence, the number of mismatches with the reference sequence (Mismatches) and the number of gaps (Open gaps) and length correction (if applied).

2.4. Haplotype determination

In the second step of the process, the haplotypes were determined for each sample. The haplotype, as defined in the SARS-CoV-2 plugin, consists of a set of high-frequency amino acid substitutions which are summarized in Table 1. Three pairs of these substitutions were observed to be in linkage disequilibrium (DP/GL, YP/CL, and QT/HI). The substitutions are ordered on the basis of the date on which they first appeared, as inferred by Nextstrain (Hadfield et al., 2018) and with the most frequent ones being: S.DP.YP.QT, S.DP.CL.QT, L.DP.YP.QT, L.GL.YP.QT, L.GL.YP.HT, L.GL.YP.HI, L.GP.YP.QT and L.GP.YP.HT (see Fig. 1 for a review on their relative abundance among isolates of SARS-CoV-2).

2.5. SNP calculation

After extraction, the plugin screened each subsequence for SNPs by automating the built-in BIONUMERICS SNP analysis tool. The resulting SNP set was filtered based on the relaxed (non-ACGT bases allowed) SNP filtering template and the retained SNPs were stored in the SNP character experiment.

2.6. Clustering SNP data into dendrograms

Entries to be clustered were selected based on suitability. In the first step, all selected entries were screened for the presence of the subsequences extracted in the prior processing step. Entries for which one or more subsequences are missing have an incomplete SNP character set and were excluded from the comparison. A similarity matrix was calculated based on the SNP experiment, using the categorical (differences) similarity coefficient, and displayed in the similarities panel. A dendrogram was then calculated based on the complete linkage (furthest neighbor) clustering algorithm (Sneath and Sokal, 1973). A minimum spanning tree (MST) was then calculated in the advanced cluster analysis window of BIONUMERICS, using default priority rule settings. The SNPs stored in the SNP experiment of the selected entries were translated and the amino acids stored in the SNP_TRANSL experiment file.

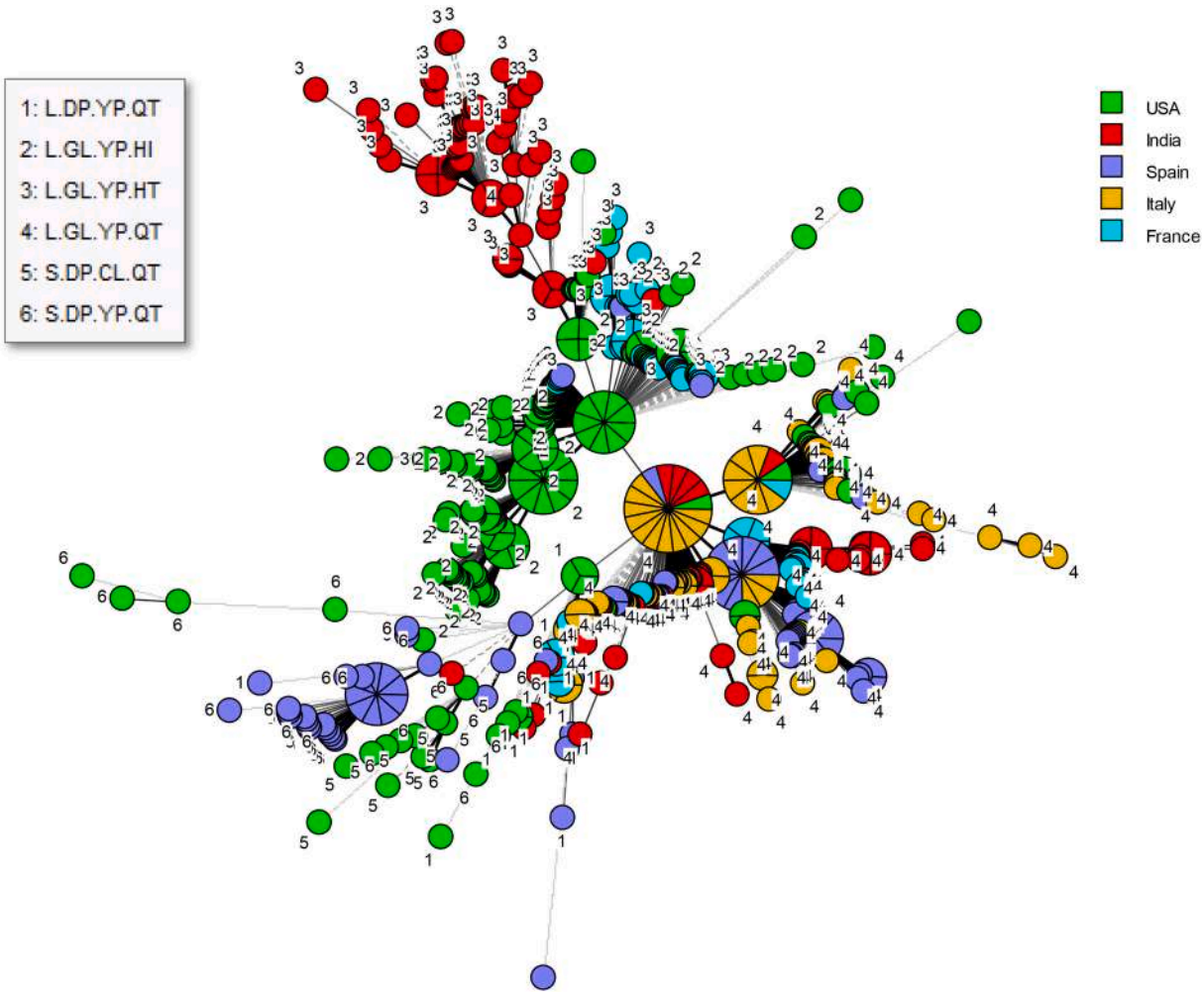

Fig. 2. Minimum spanning tree for all SARS-CoV-2 genomes included in the present study. Genomes are labeled by haplotype and color-coded by country of origin.

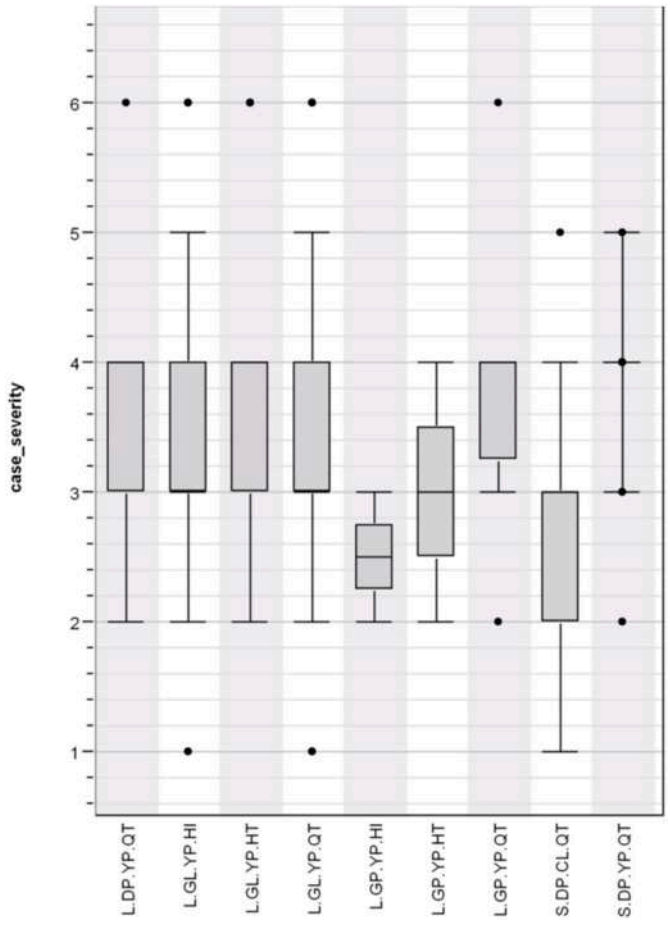

Fig. 3. COVID-19 case severity by haplotype distribution (H = 2.360; p = 0.016743).

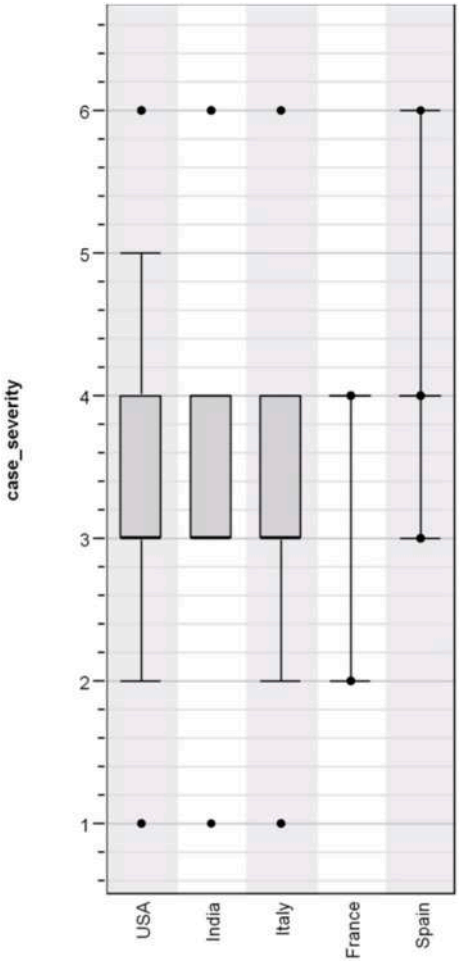

Fig. 4. Overview of COVID-19 case severity by country of origin (H = 58.285; p = 0.000000).

2.7. Case severity

The patient status information for each genome sequence was imported as a category (e.g. "asymptomatic", "hospitalized", "deceased"). Each patient's status was evaluated sometime between when the sample taken and when it was submitted, and does not necessarily reflect the case's outcome. We created a decision network in BIONUMERICS to convert each category to an integer value representing increasing case severity, on a scale from 1 to 6 (Table 2).

2.8. Statistical analysis

Tables of contingencies between two different categories (e.g. haplotypes and countries) were evaluated for unexpected frequencies with the chi-squared test. Distributions of case severity rankings across three or more categories (e.g. haplotypes or countries) were evaluated with the Kruskal-Wallace H test by ranks. Distributions of case severity rankings across two categories were evaluated with the Mann-Whitney test by sum of ranks.

3. Results

We extracted 692 SARS-CoV-2 genomic sequences originating from the USA, India, Italy, France and Spain from the GISAID database. These regions were chosen for being well represented among sequences with complete patient status metadata. The MST for these sequences shows a high degree of genotypic heterogeneity within each country although clusters representing local dissemination of closely related genotypes were obviously observed as well (Fig. 2). Fig. 2 also illustrates that strains deriving from the USA and India show global representation as well. Of note, certain types are genuinely pandemic whereas others are more geographically restricted.

Table 3
Contingency table for haplotype by country, with SARS-CoV-2 sequence counts shown; Chi square = 597.170, *P* = 0.000000.

	L. DP. YP. QT	L. GL. YP. HI	L. GL. YP. HT	L. GL. YP. QT	L. GP. YP. HI	L. GP. YP. HT	L. GP. YP. QT	S. DP. CL. QT	S. DP. YP. QT
USA	8	131	14	22	2	2	1	11	8
India	4	1	62	46	0	1	0	0	1
Spain	5	2	1	44	0	0	5	0	33
Italy	6	0	0	81	0	0	0	0	0
France	6	26	16	22	0	0	0	0	0

Table 4
ANOVA tests on numerical case severity versus SARS-CoV-2 haplotype.

Country	Statistic	*P*-value
USA	H = 11.222	0.129228
India	H = 0.557	0.756956
France	H = 2.383	0.496744
Italy	Sum of ranks: L.GL.YP.QT: 3429 L.DP.YP.QT: 399	0.023739
Spain	H = 14.210	0.006653

Table 5
SARS-CoV-2 haplotype counts for geographic divisions.

	L.DP. YP.QT	L.GL. YP.HI	L.GL. YP.HT	L.GL. YP.QT	L.GP.YP.HI	L.GP.YP.HT	L.GP.YP.QT	S.DP.CL.QT	S.DP.YP.QT
California	8	76	14	18	2	2	1	6	6
Gujarat	2	1	62	44	0	1	0	0	0
Ile de France	6	26	16	22	0	0	0	0	0
Louisiana	0	40	0	0	0	0	0	0	0
Abruzzo	0	0	0	23	0	0	0	0	0
Basque Country	0	0	0	13	0	0	3	0	5
Lombardy	0	0	0	19	0	0	0	0	0
Texas	0	9	0	3	0	0	0	1	2
Friuli Venezia Giulia	0	0	0	12	0	0	0	0	0
Apulia	0	0	0	12	0	0	0	0	0
Andalusia	2	1	0	0	0	0	1	0	7
Aragon	1	0	1	7	0	0	0	0	1
Galicia	0	0	0	5	0	0	0	0	4
La Rioja	0	0	0	0	0	0	0	0	9
Castilla	0	0	0	2	0	0	0	0	5
Campania	0	0	0	7	0	0	0	0	0
Puerto Rico	0	3	0	1	0	0	0	3	0
Lazio	6	0	0	1	0	0	0	0	0
Veneto	0	0	0	6	0	0	0	0	0
Melilla	0	1	0	4	0	0	0	0	0
Catalunya	0	0	0	4	0	0	1	0	0
Madrid	1	0	0	4	0	0	0	0	0
Telangana	2	0	0	2	0	0	0	0	0
Navarra	0	0	0	3	0	0	0	0	0
Comunitat Valenciana	0	0	0	1	0	0	0	0	2
Canarias	1	0	0	1	0	0	0	0	0
South Carolina	0	2	0	0	0	0	0	0	0
Florida	0	1	0	0	0	0	0	0	0
Marche	0	0	0	1	0	0	0	0	0
None	0	0	0	0	0	0	0	0	1
Montana	0	0	0	0	0	0	0	1	0

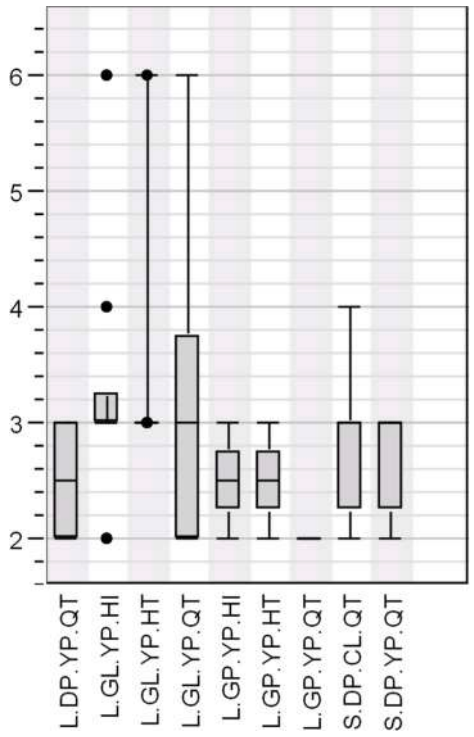

Fig. 5. COVID case severity versus haplotype in California, USA (H = 12.514; p = 0.129694).

Overall, there was a significant association between haplotype and case severity with haplotype (H = 2.360; p = 0.016743) (Fig. 3). There was also a strong association (H = 58.285; p = 0.000000) between case severity and country (Fig. 4). Furthermore, a contingency table shows a highly significant association (Chi square = 597.170, P = 0.000000) between haplotype and country (Table 3). It shows that L.GL.YP.QT is widespread but predominates in Italy; that L.GL.YP.HT is found primarily in India; that S.DP.YP.QT is prominent mostly in Spain; and that L.GL.YP.HI predominates in the United States. An examination of case severity versus haplotype within each country showed mixed results; only data from Italy and Spain showed a significant association (Table 4).

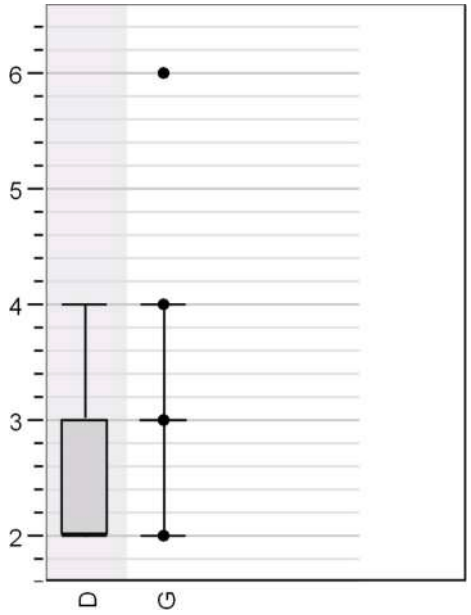

Fig. 6. COVID case severity versus the D614G mutation (Sum of ranks: G 7913.5, D 997.5; p = 0.031085).

To minimize geographic factors while maximizing genetic diversity,

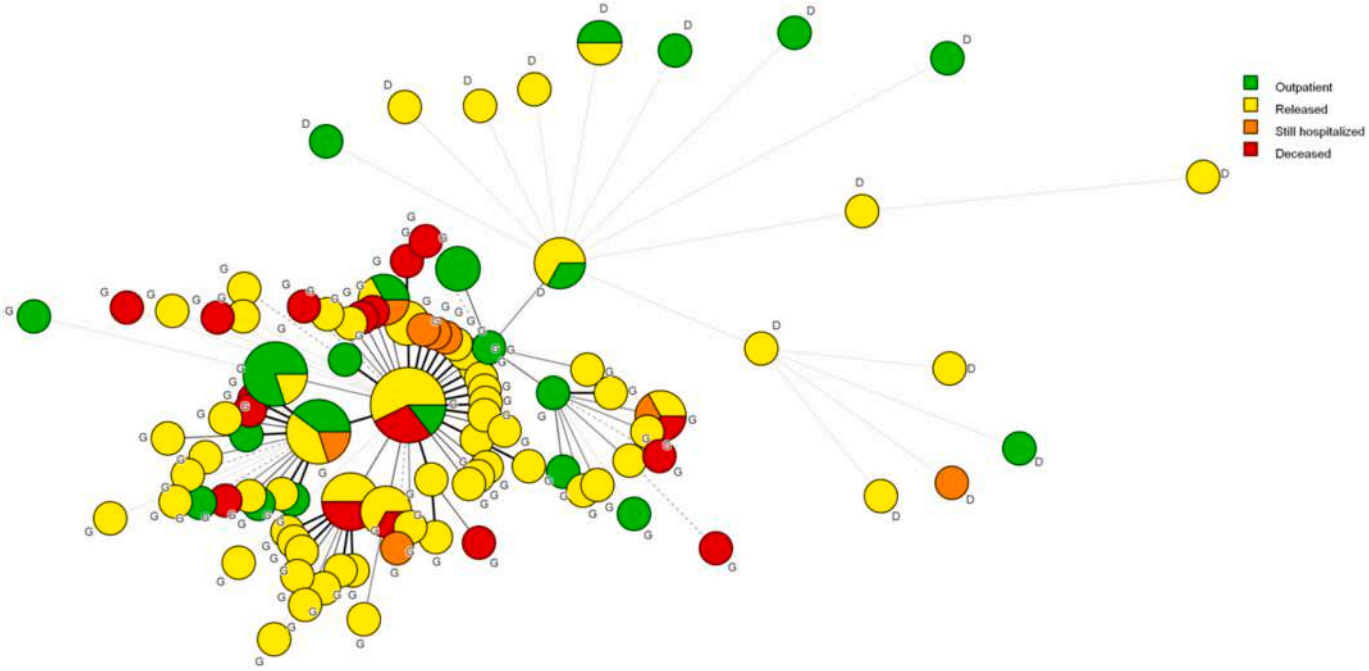

Fig. 7. Minimum spanning tree covering haplotype diversity at the D614G level in association with disease severity. Note that deceased patients are entirely in the G cluster, as are all but one of the still hospitalized patients.

we selected the sequences from California for further analysis. As shown in Table 5, these 133 sequences included all nine haplotypes, 20 of which were "D" types. A single CA sequence was submitted by Naval Health Research Center. A Kruskal-Wallis test by ranks did not show a statistically significant association between haplotype and case severity (Fig. 5). However, there was an apparent trend with regard to the D614G mutation (Fig. 6). By grouping the haplotypes into "D" and "G" types, a Mann-Whitney test revealed a significant association between the D614G genotypes and case severity ($p = 0.031085$). This is once more reflected in the MST (Fig. 7) where all of the deceased patients are shown to fall within the G allele group.

4. Discussion

Several studies have addressed the relevance of human genetic polymorphism in severity and mortality of COVID-19 (Bosso et al., 2020; Li et al., 2020; Lu et al., 2020; Asselta et al., 2020; McCoy et al., 2020). Host variation is usually associated with pathogen adaptation and evolution. The relevance of viral variation in this respect has been studied by Parlikar et al. (2020) who analyzed 167 SARS-CoV-2, 312 SARS-CoV, and 5 Pangolin CoV genomes to help understand their origin and evolution. The phylogeny of the subgenus Sarbecovirus confirmed the fact that SARS-CoV-2 strains evolved from their common ancestors putatively residing in bat or pangolin hosts. These authors predicted a few country-specific patterns of relatedness but failed to document any relatedness between genotypes and disease phenotypes in human patients. Two other recent publications again touch upon a lack of viral variation in the development of more or less serious disease. In the review by Callaway et al. (2020) it is concluded that viral mutations do not contribute to mortality and that more likely than not environmental conditions have a more significant clinical impact than viral variation. Zhang et al. (2020) conclude similarly, based on the bioinformatic analyses of experimentally defined genome sequences. In this study, the number of clinical isolates sequenced may have been a limiting factor.

We have here set out to correlate viral genotypes with host phenotypes in more detail using a large number of SARS-CoV-2 genome sequences from a broader geographic origin. We show that genotypic variants across multiple geographic regions are associated with variation in case severity. Given the likelihood that both genotype and case severity are influenced by other geographic factors, we controlled for geographic variation by focusing on one region with a relatively high degree of genotypic variation. Within this region, we showed a significant association between the D614G mutation and case severity. We also demonstrated that controlling for confounding parameters had a big effect on retrieving significant correlations between viral types and pathogenicity within patients.

The D614G mutation has received a great deal of attention with respect to its rapid global dissemination (Dearlove et al., 2020) and its significant influence on the spike protein's affinity for the ACE2 receptor. Recent studies demonstrated that in situ images of S trimer conformational changes were affected by the D614G substitution (Ke et al., 2020). This mutation abolishes a salt bridge to K854 and may reduce folding of the 833–854 loop. It has been suggested (Korber et al., 2020) that this mutation increases the virus' transmissibility, without necessarily increasing its virulence, thereby explaining its rapid spread in multiple locations. A counterargument (Grubaugh et al., 2020) has proposed that genetic drift and founder effects could also explain this pattern. More recently, the D614G mutation was identified as a marker associated with fatality rate at a countrywide level (Toyoshima et al., 2020). Our current results support these findings independently, using a completely different set of sequences and an alternative bioinformatic approach, and here show that this mutation could in fact result in increased case severity. However, we cannot rule out the possibility that transmissibility and virulence are not independent. Even if 614G is not more virulent than its D614 ancestor, ease of transmission could lead to higher viral loads in actual patients, thereby increasing the likelihood of severe cases. The polymorphisms we have identified in this project may have an effect on case severity when other factors are controlled, but that this effect is swamped out by these other factors when comparing cases across different geographic regions. Future studies should investigate the relationships among genotype, viral load, and patient outcome to sort out the underlying mechanisms.

Although this study focused on genotypes that were of particular interest at the time the data were gathered, our approach could be adapted easily to novel variants such as B.1.1.7, first observed in the UK (Public Health England, 2020). A recent update to the BIONUMERICS

SARS-CoV-2 plugin includes a tool to identify mutations relative to the reference sequence that are monomorphic for the samples of interest. For example, a set of known B.1.1.7 samples can be used to define a set of characteristic mutations, which can then be used to identify unknown samples. Once samples are characterized as variants in this way, they can be compared to other variants in terms of geography, patient outcome, and other epidemiological factors.

Conflicts of interest

All authors are employees of bioMérieux, a company designing, developing, and selling diagnostic tests for infectious diseases. For this reason, it is impossible to provide the software used free of charge to all except for BIONUMERICS evaluation licenses and for a limited period of a month only. We do welcome collaborations in order to expand the current type of analyses and look forward to suggestions to that effect. bioMérieux marketing and sales departments had no part in the design and the written documentation of this work.

Acknowledgements

We gratefully acknowledge Dr. Maud Tournoud (bioMérieux, Data Analytics, Grenoble, France) for editing the paper and advising on proper statistical procedures to be used.

References

Asselta, R., Paraboschi, E.M., Mantovani, A., Duga, S., 2020. ACE2 and TMPRSS2 variants and expression as candidates to sex and country differences in COVID-19 severity in Italy. Aging (Albany NY) 12 (11), 10087–10098. https://doi.org/10.18632/aging.103415 (PMID: 32501810).

Bosso, M., Thanaraj, T.A., Abu-Farha, M., Alanbaei, M., Abubaker, J., Al-Mulla, F., 2020 Jun 25. The two faces of ACE2: the role of ACE2 receptor and its polymorphisms in hypertension and COVID-19. Version 2. Mol. Ther. Methods Clin. Dev. 18, 321–327. https://doi.org/10.1016/j.omtm.2020.06.017. 32665962.

Callaway, E., Ledford, H., Mallapaty, S., 2020 Jul. Six months of coronavirus: the mysteries scientists are still racing to solve. Nature. 583 (7815), 178–179. https://doi.org/10.1038/d41586-020-01989-z.

Dearlove, B., Lewitus, E., Bai, H., Li, Y., Reeves, D.B., Joyce, M.G., Scott, P.T., Amare, M. F., Vasan, S., Michael, N.L., Modjarrad, K., Rolland, M., 2020. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. Proc. Natl. Acad. Sci. U. S. A. 202008281. https://doi.org/10.1073/pnas.2008281117. Online ahead of print. Aug 31.

Gallego, O., Martin-Carbonero, L., Aguero, J., de Mendoza, C., Corral, A., Soriano, V., 2004 Oct. Correlation between rules-based interpretation and virtual phenotype interpretation of HIV-1 genotypes for predicting drug resistance in HIV-infected individuals. J. Virol. Methods 121 (1), 115–118. https://doi.org/10.1016/j.jviromet.2004.06.003. 15350741.

Grubaugh, et al., 2020. Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. Cell. https://doi.org/10.1016/j.cell.2020.06.040.

Grundy, E.J., Suddek, T., Filippidis, F.T., Majeed, A., Coronini-Cronberg, S., 2020. Smoking, SARS-CoV-2 and COVID-19: A review of reviews considering implications for public health policy and practice. Tob. Induc. Dis. 18, 58. https://doi.org/10.18332/tid/124788. Jul 3. (eCollection 2020. PMID: 32641924).

Gudbjartsson, D.F., Helgason, A., Jonsson, H., Magnusson, O.T., Melsted, P., Norddahl, G.L., Saemundsdottir, J., Sigurdsson, A., Sulem, P., Agustsdottir, A.B., Eiriksdottir, B., Fridriksdottir, R., Gardarsdottir, E.E., Georgsson, G., Gretarsdottir, O.S., Gudmundsson, K.R., Gunnarsdottir, T.R., Gylfason, A., Holm, H., Jensson, B.O., Jonasdottir, A., Jonsson, F., Josefsdottir, K.S., Kristjansson, T., Magnusdottir, D.N., le Roux, L., Sigmundsdottir, G., Sveinbjornsson, G., Sveinsdottir, K.E., Sveinsdottir, M., Thorarensen, E.A., Thorbjornsson, B., Löve, A., Masson, G., Jonsdottir, I., Möller, A.D., Gudnason, T., Kristinsson, K.G., Thorsteinsdottir, U., Stefansson, K., 2020. Spread of SARS-CoV-2 in the Icelandic Population. N. Engl. J. Med. 382 (24), 2302–2315. https://doi.org/10.1056/NEJMoa2006100. Jun 11. (Epub 2020 Apr 14).

Hadfield, James, Megill, Colin, Bell, Sidney M., Huddleston, John, Potter, Barney, Callender, Charlton, Sagulenko, Pavel, Bedford, Trevor, Neher, Richard A., 2018 May. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34 (23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407.

Hodcroft, E.B., Zuber, M., Nadeau, S., Comas, I., Candelas, F.G., SeqCOVID-SPAIN consortium, Stadler, T., Neher, R.A., 2020. Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. medRxiv. https://doi.org/10.1101/2020.10.25.20219063, 10.25.20219063.

Hussain, A., Mahawar, K., Xia, Z., Yang, W., El-Hasani, S., 2020. Obesity and mortality of COVID-19. Meta-analysis. Obes. Res. Clin. Pract. Jul 9. S1871-403X(20)30550-0 https://doi.org/10.1016/j.orcp.2020.07.002 (Online ahead of print. PMID: 32660813).

Jary, A., Leducq, V., Malet, I., Marot, S., Klement-Frutos, E., Teyssou, E., Soulié, C., Abdi, B., Wirden, M., Pourcher, V., Caumes, E., Calvez, V., Burrel, S., Marcelin, A.G., Boutolleau, D., 2020. Evolution of viral quasispecies during SARS-CoV-2 infection. Clin. Microbiol. Infect. 26 (11), P1560.E1–1560.E4. Jul 24. S1198-743X(20)30440–7. https://doi.org/10.1016/j.cmi.2020.07.032 (Online ahead of print).

Ji, X., Tan, W., Zhang, C., Zhai, Y., Hsueh, Y., Zhang, Z., Zhang, C., Lu, Y., Duan, B., Tan, G., Na, R., Deng, G., Niu, G., 2020 Jul 13. TWIRLS, a knowledge-mining technology, suggests a possible mechanism for the pathological changes in the human host after coronavirus infection via ACE2. Drug Dev. Res. https://doi.org/10.1002/ddr.21717. Online ahead of print. 32657473.

Kadkhoda, K., 2020. COVID-19: an Immunopathological View. mSphere 5 (2). https://doi.org/10.1128/mSphere.00344-20 e00344–20. Apr 22. (PMID: 32321823).

Kaushal, N., Gupta, Y., Goyal, M., Khaiboullina, S.F., Baranwal, M., Verma, S.C., 2020. Mutational frequencies of SARS-CoV-2 genome during the beginning months of the outbreak in USA. Pathogens. 9 (7), E565 https://doi.org/10.3390/pathogens9070565 (PMID: 32668692).

Ke, Z., et al., 2020. Structures and distributions of SARS-CoV-2 spike proteins on intact virions. Nature. https://doi.org/10.1038/s41586-020-2665-2.

Korber, et al., 2020. Cell182, 1–16 August 20, 2020. Published by Elsevier Inc. https://doi.org/10.1016/j.cell.2020.06.043ll.

Lamers, M.M., Beumer, J., van der Vaart, J., Knoops, K., Puschhof, J., Breugem, T.I., Ravelli, R.B.G., Paul van Schayck, J., Mykytyn, A.Z., Duimel, H.Q., van Donselaar, E., Riesebosch, S., HJH, Kuijpers, Schipper, D., van de Wetering, W.J., de Graaf, M., Koopmans, M., Cuppen, E., Peters, P.J., Haagmans, B.L., Clevers, H., 2020 Jul 3. SARS-CoV-2 productively infects human gut enterocytes. Science. 369 (6499), 50–54. https://doi.org/10.1126/science.abc1669. Epub 2020 May 1. 32358202.

Leibel, S.L., McVicar, R.N., Winquist, A.M., Niles, W.D., Snyder, E.Y., 2020 Sep. Generation of complete multi-cell type lung organoids from human embryonic and patient-specific induced pluripotent stem cells for infectious disease modeling and therapeutics validation. Curr. Protoc. Stem Cell Biol. 54 (1), e118 https://doi.org/10.1002/cpsc.118.

Li, Q., Cao, Z., Rahman, P., 2020 Jun. Genetic variability of human angiotensin-converting enzyme 2 (hACE2) among various ethnic populations. Mol. Genet. Genomic Med. 18, e1344 https://doi.org/10.1002/mgg3.1344. 32558308.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., Chen, J., Meng, Y., Wang, J., Lin, Y., Yuan, J., Xie, Z., Ma, J., Liu, W.J., Wang, D., Xu, W., Holmes, E.C., Gao, G.F., Wu, G., Chen, W., Shi, W., Tan, W., 2020 Feb 22. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 395 (10224), 565–574. https://doi.org/10.1016/S0140-6736(20)30251-8. Epub 2020 Jan 30. 32007145.

McCoy, J., Wambier, C.G., Vano-Galvan, S., Shapiro, J., Sinclair, R., Ramos, P.M., Washenik, K., Andrade, M., Herrera, S., Goren, A., 2020 Jul. Racial variations in COVID-19 deaths may be due to androgen receptor genetic variants associated with prostate cancer and androgenetic alopecia. Are anti-androgens a potential treatment for COVID-19? J. Cosmet. Dermatol. 19 (7), 1542–1543. https://doi.org/10.1111/jocd.13455. Epub 2020 Jun 14. 32333494.

Mishra, A.K., Sahu, K.K., George, A.A., Lal, A., 2020. A review of cardiac manifestations and predictors of outcome in patients with COVID-19. Heart Lung. https://doi.org/10.1016/j.hrtlng.2020.04.019. May 3:S0147-9563(20)30157-6. (Online ahead of print. PMID: 32593418).

Nakamichi, K., Shen, J.Z., Lee, C.S., Lee, A.Y., Roberts, E.A., Simonson, P.D., Roychoudhury, P., Andriesen, J.G., Randhawa, A.K., Mathias, P.C., Greninger, A., Jerome, K.R., Van Gelder, R.N., 2020. Outcomes associated with SARS-CoV-2 viral clades in COVID-19. medRxiv. https://doi.org/10.1101/2020.09.24.20201228, 09.24.20201228.

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M., Gallo, R.C., Zella, D., Ippodrino, R., 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. Version 2. J. Transl. Med. 18 (1), 179. https://doi.org/10.1186/s12967-020-02344-6. Apr 22. (PMID: 32321524).

Papadopoulos, V., Li, L., Samplaski, M., 2020. Why does COVID-19 kill more elderly men than women? Is there a role for testosterone? Andrology. https://doi.org/10.1111/andr.12868. Jul 18. (Online ahead of print. PMID: 32681716).

Parlikar, A., Kalia, K., Sinha, S., Patnaik, S., Sharma, N., Vemuri, S.G., Sharma, G., 2020 Jul 17. Understanding genomic diversity, pan-genome, and evolution of SARS-CoV-2. PeerJ. 8, e9576 https://doi.org/10.7717/peerj.9576 (eCollection 2020).

Potere, N., Valeriani, E., Candeloro, M., Tana, M., Porreca, E., Abbate, A., Spoto, S., Rutjes, A.W.S., Di Nisio, M., 2020. Acute complications and mortality in hospitalized patients with coronavirus disease 2019: a systematic review and meta-analysis. Version 2. Crit. Care 24 (1), 389. https://doi.org/10.1186/s13054-020-03022-1. Jul 2. (PMID: 32616077).

Public Health England, 2020. Investigation of Novel SARS-CoV-2 Variant: Variant of Concern 202012/01, Technical Briefing 3. Public Health England, London, United Kingdom.

Rajpal, A., Rahimi, L., Ismail-Beigi, F., 2020. Factors leading to high morbidity and mortality of COVID-19 in patients with type 2 diabetes. J. Diabetes. 14 (4), 295–300. https://doi.org/10.1111/1753-0407.13085. Online ahead of print. Jul 16. (PMID: 32671936).

Saha, I., Ghosh, N., Maity, D., Sharma, N., Sarkar, J.P., Mitra, K., 2020 Jul 10. Genome-wide analysis of Indian SARS-CoV-2 genomes for the identification of genetic mutation and SNP. Infect. Genet. Evol. 85, 104457. https://doi.org/10.1016/j.meegid.2020.104457. 32659347.

Sapoval, N., Mahmoud, M., Jochum, M.D., Liu, Y., Leo Elworth, R.A., Wang, Q., Albin, D., Ogilvie, H., Lee, M.D., Villapol, S., Hernandez, K.M., Berry, I.M., Foox, J., Beheshti, A., Ternus, K., Aagaard, K.M., Posada, D., Mason, C.E., Sedlazeck, F., Treangen, T.J., 2020. Hidden genomic diversity of SARS-CoV-2: implications for

qRT-PCR diagnostics and transmission. bioRxiv. https://doi.org/10.1101/2020.07.02.184481, 2020.07.02.184481. (PMID: 32637955).

Sekizuka, T., Itokawa, K., Kageyama, T., Saito, S., Takayama, I., Asanuma, H., Nao, N., Tanaka, R., Hashino, M., Takahashi, T., Kamiya, H., Yamagishi, T., Kakimoto, K., Suzuki, M., Hasegawa, H., Wakita, T., Kuroda, M., 2020 Jul. Haplotype networks of SARS-CoV-2 infections in the diamond princess cruise ship outbreak. Proc. Natl. Acad. Sci. U. S. A. 28, 202006824. https://doi.org/10.1073/pnas.2006824117. Online ahead of print.

Sneath, H.A., Sokal, R.R., 1973. Numerical Taxonomy: The Principles and Practice of Numerical Classification. Freeman, San Francisco, 573.

Thompson, J.V., Meghani, N., Powell, B.M., Newell, I., Craven, R., Skilton, G., Bagg, L.J., Yaqoob, I., Dixon, M.J., Evans, E.J., Kambele, B., Rehman, A., Kwong, G.N.M., 2020. Patient characteristics and predictors of mortality in 470 adults admitted to a district general hospital in England with Covid-19. medRxiv. https://doi.org/10.1101/2020.07.21.20153650 preprint.

Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y., Kiyotani, K., 2020 Jul. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. J. Hum. Genet. 22, 1–8. https://doi.org/10.1038/s10038-020-0808-9. Online ahead of print.

Williamson, E.J., Walker, A.J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C.E., Curtis, H.J., Mehrkar, A., Evans, D., Inglesby, P., Cockburn, J., McDonald, H.I., MacKenna, B., Tomlinson, L., Douglas, I.J., Rentsch, C.T., Mathur, R., Wong, A.Y.S., Grieve, R., Harrison, D., Forbes, H., Schultze, A., Croker, R., Parry, J., Hester, F., Harper, S., Perera, R., Evans, S.J.W., Smeeth, L., Goldacre, B., 2020 Aug. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 584 (7821), 430–436. https://doi.org/10.1038/s41586-020-2521-4.

Yang, H.C., Chen, C., Wang, J.H., Liao, H.C., Yang, C.T., Chen, C.W., Lin, Y.C., Kao, C.H., Lu, M.Y.J., Liao, J.C., 2020. Analysis of genomic disctributions of SARS.CoV-2 reveals a dominant strain type with strong allelic associations. Proc. Natl. Acad. Sci. https://doi.org/10.1073/2007840117.

Young, B.E., Fong, S.W., Chan, Y.H., Mak, T.M., Ang, L.W., Anderson, D.E., Lee, C.Y., Amrun, S.N., Lee, B., Goh, Y.S., Su, Y.C.F., Wei, W.E., Kalimuddin, S., Chai, L.Y.A., Pada, S., Tan, S.Y., Sun, L., Parthasarathy, P., Chen, Y.Y.C., Barkham, T., Lin, R.T.P., Maurer-Stroh, S., Leo, Y.S., Wang, L.F., Renia, L., Lee, V.J., Smith, G.J.D., Lye, D.C., Ng, L.F.P., 2020 Aug 29. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. Lancet. 396 (10251), 603–611. https://doi.org/10.1016/S0140-6736(20)31757-8.

Zaki, N., Alashwal, H., Ibrahim, S., 2020 Jul 8. Association of hypertension, diabetes, stroke, cancer, kidney disease, and high-cholesterol with COVID-19 disease severity and fatality: a systematic review. Diabetes Metab. Syndr. 14 (5), 1133–1142. https://doi.org/10.1016/j.dsx.2020.07.005. Online ahead of print. 32663789.

Zhang, X., Tan, Y., Ling, Y., Lu, G., Liu, F., Yi, Z., Jia, X., Wu, M., Shi, B., Xu, S., Chen, J., Wang, W., Chen, B., Jiang, L., Yu, S., Lu, J., Wang, J., Xu, M., Yuan, Z., Zhang, Q., Zhang, X., Zhao, G., Wang, S., Chen, S., Lu, H., 2020 Jul. Viral and host factors related to the clinical outcome of COVID-19. Nature. 583 (7816), 437–440. https://doi.org/10.1038/s41586-020-2355-0 (Epub 2020 May 20).