Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology

Samira M. Hamed
October University for Modern Science and Arts (MSA)
Walid F. Elkhatib (walid-elkhatib@pharma.asu.edu.eg)
Ain Shams University
Ahmed S. Khairallah
Beni Suef University
Ayman M. Noreddin
Galala University

Research Article

Keywords: SARS-CoV-2, COVID-19, clade, age, gender

Posted Date: October 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-89876/v1

Abstract

Expansion of COVID-19 worldwide increases interest in unraveling genomic variations of novel SARS-CoV-2 virus. Metadata of 60,703 SARS-CoV-2 genomes submitted to GISAID database were analyzed with respect to genomic clades and their geographic, age, and gender distributions. Clade GR was the most frequently identified followed by G and GH. Chronological analysis revealed expansion in SARS-CoV-2 clades with D614Gmutations indicating adaptation-driven evolution. Of them, clade GH showed a slight regression. GR, GH and L clades prevail in countries with higher deaths. GR clade showed higher prevalence among severe/deceased patients. Metadata analysis showed higher ($p > 0.05$) prevalence of severe/deceased cases among males than females and predominance of GR clade in female and children patients. Furthermore, severe disease/death was more prevalent ($p < 0.05$) in elderly than in adults/children. These findings uniquely provide an evidence-based evolution of SARS-CoV-2 leading to altered infectivity, virulence, and mortality.

Introduction

Late in December 2019, an outbreak of atypical pneumonia of unknown etiology was described in Wuhan province in China. A novel coronavirus named “Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2)” was then identified as the etiologic agent.[1,2] Later, the disease was designated COrona VIrus Disease – 2019 (COVID-19).[3] The rapid expansion of COVID-19 cases in number and geographic distribution prompted the World Health Organization (WHO) to declare a global health emergency. Containment of the disease was hindered by the lack of antiviral treatment, lack of vaccines and existence of asymptomatic carriers. In March 11, 2020, the virus was officially classified by the WHO as a pandemic.

After declaration of COVID-19 as pandemic, there was a global interest in exploring genomic variations in the novel virus. The first genomic sequence of SARS-CoV-2 was reported by Wu and colleagues.[2] Subsequently, publicly available resources were developed to provide dynamic and updated data on SARS-CoV-2 genome, thus offering an extraordinary opportunity for comparative genomic studies. Among the open access repositories of SARS-Cov-2 genomic sequences are the Global Initiative for Sharing All Influenza Data (GISAID) database (https://www.gisaid.org)[4], National Center for Biotechnology Information database (NCBI) (www.ncbi.nlm.nih.gov), and Virus Pathogen Resource database (ViPR) (www.viprbrc.org). Genome analysis tools were also provided by several platforms such as The China National Center for Bioinformation (https://bigd.big.ac.cn/ncov/tool/annotation)[5] Nextstrain project (https://nextstrain.org)[6], and CoV-GLUE (http://cov-glue.cvr.gla.ac.uk).[7]

According to GISAID nomenclature system, most of the currently sequenced SARS-CoV-2 genomes were clustered into one of six major clades. In their submission order, SARS-CoV-2 clades include L, to which SARS-CoV-2 virus reference strain belongs, S, G, V, GR and GH. They exhibit few changes in relation to the

reference strain (GenBank accession number NC_045512, GISIAD accession ID: EPI_ISL_402124). Such changes include: L84S in NS8 for clade S; coexisting L37F and G251V mutations in NSP6 and NS3, respectively for clade V; D614G mutation in the spike protein (S) for clade G. In addition to D614G, NS3-Q57H and N-G204R mutations characterize the clades GH and GR, respectively. Genomes that don't belong to any of the six major clades had the designation "O clade".

Given that most of the immune-based therapeutics and diagnostics of COVID-19 are based on the protein sequence of Wuhan reference strain spike[8], their efficacy could potentially be affected by genomic variations and the associated altered viral phenotype. Moreover, the influence of genetic mutations on the infectivity and/or virulence of SARS-CoV-2 is yet to be established.[9] The acquisition of mutations imparting higher infectivity, virulence and/or immunological resistance is thus an eminent threat. Accordingly, active genomic surveillance and close monitoring of the genomic sequence dynamics of SARS-CoV-2 is urgently required to: a) trace the pattern of geographic spread of the virus during the ongoing pandemic[10,11]; b) ensure the effectiveness of vaccines and immune-based diagnostic or therapeutic interventions currently in use or under investigation[9]; and c) identify putative therapeutic targets.[12-14]

Geographic, gender, and age discrepancy of COVID-19 disease outcome have been reported by several studies.[15-18] Whether this correlates to SARS-CoV-2 genomic variation is still unclear. In addition to laboratory investigations, statistical approaches correlating the distribution of viral clades in different groups to disease severity might provide a good evidence on this bias. The current study aims to analyze the geographic, gender and age distribution SARS-CoV-2 genomic clades with respect to COVID-19 disease epidemiology.

Results

Geographic distribution of SARS-CoV-2 clades

As of July 7, 2020, WHO reported a total number of 11,669,259 confirmed COVID-19 cases and 539,906 deaths.[19] The calculated world case fatality rate (CFR) was 4.63%.

The median number of cases reported from the countries from which SARS-CoV-2 genomes were submitted to the database was 16,799 while the median number of deaths was 274 and that of the CFR was 2.2%. Contributing countries were grouped into two groups according to the relation between the national values of each of the disease epidemiology parameters to the median. Of all continents, the highest number of COVID-19 cases was reported from North America. This was also associated with the highest median CFR (6.26%). On the other hand, most deaths were reported from Europe. GR was the most common clade (29.4%), followed by G (23.4%) and GH (21.5%). Other less common clades including L, S, and V were identified in 6.1%, 7.0% and 6.7% of the submitted genomes, respectively. About 6.0% of the genomes were not clustered into any of the major clades.

Analysis of the continent distribution of viral clades (Fig 1) showed that clade G was the most common in Africa (52.6%) and Oceania (26.4%). Clade GH was the most common in North America (58.5%). GR was the most common clade in Europe (40.0%) and South America (52.9%), while other clade (O) predominated in Asia (34.7%).

The number of coexisting clades was compared between countries with respect to different disease epidemiology parameters including the number of cases, total number of deaths and CFRs. Viral strains belonging to all known clades coexisted in 27 countries (27.0%). Such countries had also reported above median local values for the studied disease epidemiology parameters. Of them, above median number of cases, number of deaths and CFRs were reported by 70.3%, 66.6%, and 62.9%, respectively.

Mann Whitney test showed a significant difference in the distribution of the number of coexisting clades in the two groups with respect to the total number of cases (P-value = 0.008), total deaths (P-value = 0.038) and CFRs (P-value = 0.039). Higher medians were shown for all parameters in the group of countries where above median cases, deaths and CFRs were recorded.

The impact of the distribution of individual clades on the disease epidemiology was also analyzed. Distribution bias of some clades was noted, as shown in Table 1. This was statistically significant for all clades with all disease epidemiology parameters.

Table 1: Geographical distribution of SARS-Cov-2 clades with respect to disease epidemiology parameters

Geographic Region	G%	GH%	GR%	L%	O%	S%	V%
Countries showing above median number of cases	22.9	20.9	31.3	6.4	4.9	6.8	6.8
Countries showing below median number of cases	26.8	26.9	13.7	3.2	15.8	7.9	5.6
Countries showing above median number of deaths	23.1	21.7	31.3	6.3	4.1	6.7	6.8
Countries showing below median number of deaths	25.4	19.6	13.6	4	22.5	8.9	5.9
Countries showing above median CFRs	23.1	21.8	31	6.4	4.1	6.8	6.9
Countries showing below median CFRs	25	19.5	17	3.7	21.2	8.4	5.2

The distribution bias of different clades among the groups of countries showing above median and below median values for all disease epidemiology parameters was statistically significant (P-value <0.05).

Among all studied cases, patient's clinical status were specified for only 1331. Based on the provided data, such cases were grouped into mild/recovered cases (n=1153) and severe/deceased cases (n=178). Only clade GR was significantly more frequently identified among viral genomes isolated from severe/deceased cases (Pearson Chi Square, P-value= <0.001). In contrast all other clades showed higher prevalence in mild/recovered cases than severe/deceased ones. Of them, this was statistically significant only for clade S (Pearson Chi Square, P-value= 0.003) (Table 2).

Table 2: Distribution of SARS-CoV-2 clades with respect to patient's clinical status

Patient Status	G%	GH%	GR%	L%	O%	S%	V%
Severe/deceased	21.9%	24.7%	31.5%	5.6%	12.4%	2.8%	1.1%
Mild/recovered	23.0%	25.2%	15.1%	7.4%	18.0%	9.5%	1.9%
P-value	0.751	0.901	<0.001*	0.398	0.062	0.003*	0.761

*P-values < 0.05 are statistically significant.

Analysis of the chronological distribution of SARS-CoV-2 clades was done for 59,425 cases for which the date of collection was available. The analysis showed a gradual regression in the number of genomes that belonged to some clades including L, S, and V as well as those not clustered into any of the major clades (clade O). This was accompanied by an expansion in the number of genomes that belonged to others such as G, GR and GH. A slight recent regression was also noted for clade GH compared to G and GR. The global chronological distribution of SARS-CoV-2 clades is shown in Fig 2.

Gender distribution of SARS-CoV-2 clades

The severity of cases in both genders were compared in 1265 cases for which both gender and patient's clinical status are known. Although severe or deceased cases were more prevalent among male patients than females, gender bias was found to be statistically non-significant (15.0% versus 12.7%, P-value=0.248).

Analysis of 20,939 cases (Males= 11292, Females = 9647) for which patient gender was specified showed gender distribution bias for some clades (Table 3). Some clades were more frequently isolated from males than females such as L and O, while others were more prevalent in females such as GR and V. This was statistically significant for clades GR and O. Clades G, GH and S were nearly equally distributed between the two groups. Deeper analysis into patient's clinical status information showed that GR clades were significantly more prevalent among female patients with severe cases or death than those with mild or recovered disease (31.8% versus 17.2%, P-value = 0.005). Similarly, genomes that belonged to other clade (O) recovered from male patients were more prevalent in mild or recovered cases than others (13.4% versus 11.6%, P-value = 0.603).

Table 3: Gender Distribution of SARS-CoV-2 clades

Gender	G%	GH%	GR%	L%	O%	S%	V%
Male	22.1%	21.5%	21.2%	7.0%	17.7%	7.7%	8.7%
Female	22.8%	21.9%	23.9%	6.0%	8.1%	7.8%	9.5%
P-value	0.217	0.515	<0.001*	0.002	<0.001*	0.990	0.055

*P-values < 0.05 are statistically significant.

Age distribution of SARS-CoV-2 clades

A significant correlation was found between age groups and patient's clinical status. The analysis included 1194 cases for which both patient age and clinical status are known. Age groups were defined as children (up to 18 years), adults (18-64 years) and elderly (65 years or more). Severe/deceased cases were significantly more prevalent in elderly than in adults (38.1 vs 7.9%, Pearson Chi-Square P-value <0.001) or in children (38.1 vs 3.0%, Pearson Chi-Square P-value <0.001). Although Severe/deceased cases were more frequently reported among adults than children, this was not statistically significant (7.9 vs 3.0%, Fisher's Exact test P-value = 0.158).

The distribution of genomes that belonged to different clades in different age groups was analyzed among 20,871 cases for which the patient age was specified. This included 68.0% adults, 28.7% elderly and 3.4% children. The distribution of the genomes that belonged to clade G was nearly the same across the groups. Viral isolates whose genomes belonged to GR were more frequently isolated from children (27.2% versus 22.0% in adults and 22.8% in elderly). Those which belonged to clades GH and O showed higher prevalence in adult patients. The prevalence of GH was 22.1% in genomes from adults, 20.6% in children and 18.6% in elderly. Other clade (O) was identified in 12.4%, 10.9% and 6.4% of viral genomes recovered from adults, children and elderly patients, respectively. Genomes belonged to clades L, S, and V, were isolated with higher percentages from elderly patients (7.5%, 8.6% and 13.7%, respectively). The distribution of SARS-CoV-2 clades in different age groups with respect to patients' clinical status is shown in Fig 3.

Discussion

A relatively higher genomic stability was reported for SARS-CoV-2 compared to SARS-CoV.[20] Nevertheless, SARS-CoV-2 genomes sequenced so far were clustered into at least six major clades, as defined by GISAID database. Whether the genetic variability in SARS-CoV-2 clades arises due to an ongoing adaptation or merely due to genetic drift is still unknown. Lack of distinct evolutionary patterns or signatures in SARS-CoV-2 genomes was reported[21], while independently emerged recurrent mutations were also identified[22], suggesting an ongoing adaptation. Whether this possible adaptation provides more fitness for transmission and/or virulence is a matter of concern. In the current study, the metadata of 60,703 SARS-CoV-2 genomes submitted to GISAID EpiCoV database as of July 7, 2020 were analyzed with respect to genomic clades and their geographic, age, and gender distribution.

Most of the genomes belonged to one of six major clades namely L, S, V, G, GH, or GR. In addition, genomes that belonged to other clade (O) were also identified. About 74.3% of the genomes belonged to the clades with D614G mutation including the clades G, GH and GR. Of them Clade GR was the most frequently identified followed by G and GH. Earlier in February, clade G characterized by spike D614G mutation was identified and rapidly predominated the pandemic. The mutation was found to be located in a heavily glycosylated residue in the viral spike that is highly conserved in this species.[23] Theoretical evidence strongly suggests that mutations in this region could be coupled to altered capacity for host cell membrane fusion[23-25], an effect that should also lead to higher person to person transmission and

pathogenicity. An experimental evidence was later provided by Korber and colleagues,[9] who could link this mutation to greater infectivity and higher viral loads in COVID-19 patients. Sub-clusters of clade G then started to evolve including the clades GH and GR. Analysis of the chronological distribution of SARS-COV-2 clades in the current study showed that there was much expansion in the number of sequenced genomes that were clustered into the GR clade compared to clade G. A regression in the number of genomes clustered into clade GH was also evident. Together with the predominance of the clades GR followed by G then GH, this suggests higher fitness for transmission by clade GR than genetically related clades. Regression of the newer clade GH in comparison to the ancestral one, clade G, also suggests less fitness for transmission. Based on the mentioned findings, the hypothesis of an adaptation-driven genetic evolution is stronger. However, an experimental evidence, providing comparison between clades, is yet to be established.

Adequate scientific elucidation of the reasons behind the rapid transmission and higher mortality rates of COVID-19 in some geographic regions compared to others is still demanding. Apart from public health issues, intrinsic factors related to viral genome may be implicated. Whether the geographic distribution bias of SARS-CoV-2 clades is related to the discrepancy of COVID-19 disease severity observed worldwide is still unclear.[26] In agreement with others[21,27], a geographic distribution bias of SARS-CoV-2 clades was evident in the current analysis. The predominance of certain clades in different continents with respect to local disease epidemiology parameters was also analyzed. The GR and GH clades predominated the sequenced genomes in the top ranked continents with respect to all disease epidemiology parameters, including Europe and North America, respectively. Coexistence of all clades was evident in 27% of the contributing countries accompanied, in most cases, by relatively higher COVID-19 cases, deaths and CFRs.

Analysis of the geographic distribution of individual clades with respect to disease epidemiology parameters showed higher prevalence of the clades GR and L among the group of countries that showed above median total number of cases than others. In addition, GH, GR and L clades were more frequently identified in the genomes submitted from countries with relatively higher deaths and CFRs. Such findings suggest higher transmission of viral strains whose genome belongs to clades GR and L. Higher virulence of clades GH, GR and L is also suspected. To further examine this hypothesis, the distribution of all clades among viral genomes from patients with mild disease or recovered patients and those from severe disease or deceased patients was analyzed. Only clade GR significantly showed higher prevalence among the group of severe disease or deceased patients. This is in line with the previous finding of higher viral loads in patients infected by SARS-CoV-2 virus strains harboring D614G genomic mutations.[9] In addition, lower prevalence of clade S among mild disease or recovered cases was also statistically significant. In agreement with this finding, clade S was also found to be significantly less prevalent among the group of countries that showed above median values for the studied epidemiologic parameters. Although the reference strain of SARS-CoV-2 belonged to the L clade that also had higher prevalence at the beginning of the pandemic, clade S was found to be evolutionarily more related to animal coronaviruses.[28] In agreement with our findings, this suggests higher fitness for clade L compared

to clade S from which it had rapidly evolved early in the pandemic. Together, our findings support the previous hypothesis of Brufsky about possible ongoing competition between viral clades of varying virulence during the current pandemic.[29]

Analysis of genomes metadata showed higher prevalence of severe or deceased cases among male patients than females but without statistical significance. The worse disease outcome of male patients was also reported by others.[15-18] Several assumptions have been made by scientists to justify this gender bias. Among them are female's superior immune response[30] and higher angiotensin converting enzyme type 2 (ACE2) activity in male or ovariectomized animal models.[31] ACE2 is the main receptor for SARS-CoV-2 spike through which it attaches to target cells.[32] Wambier and colleagues assumed androgen receptor genetic variation as a likely reason.[33] The receptor is thought to regulate transcription of the transmembrane protease serine 2 (TMPRSS2), responsible for S protein priming that allows viral fusion to host cell membranes.[32] To explain the role of genomic variation of SARS-CoV-2, the distribution of SARS-CoV-2 clades in viral genomes from male versus female patients was analyzed. Gender bias was evident for some clades but this was statistically significant only for clade GR that was, strikingly, more prevalent in female patients. Deeper analysis of patients' clinical status showed that such infections were significantly associated with severe or deceased cases. Our hypothesis is that being relatively more resistant to COVID-19 or at least to worse disease outcome, females showing symptomatic disease are more likely get infected by the most virulent clade (clade GR, as assumed in the current study).

Consistent with previous reports[16,34-36], our analysis showed that severe disease or death was significantly more prevalent in elderly than in adults and children. This was previously explained by existence of comorbidities, immune senescence[37] and alterations in ACE2 receptors.[38] Mild disease in children was also reported by many studies.[15,39] Contributing factors may include lower maturity and function of ACE2 receptors[40] and viral co-infection that leads to limited replication of SARS-CoV-2 in the respiratory tract.[41] Similar to the least susceptible gender group, children age group showed the highest prevalence for clade GR. The association between symptomatic disease in the most resistant age group and the most virulent clade was also concluded. Analysis of patients' clinical status showed that only two SARS-COV-2 genomes were recovered from children with severe COVID-19 or deceased cases. The genomes belonged to the clades G and GR. In contrast, the clades L, O and S were only identified among mild or recovered cases.

Conclusion

The current analysis provides a statistical evidence on an ongoing adaptation-driven SARS-CoV-2 evolution whose outcome is higher viral infectivity and virulence. This is suggested by the biased distribution of the newer clades in geographic regions from which higher number of cases and deaths as well as higher CFRs were reported. More frequent isolation of the newer clades from the least susceptible populations including females and children was also noted. Given that the newer clades are thought to have higher virulence, this suggests that further evolution of the virus may put such groups at higher risk

for COVID-19 worse outcome. However, it is worth mentioning that a successful genome-based epidemiologic analysis is limited by the inadequate and imbalanced number of genomes deposited in open access databases. Some constraints in this respect are the lack of whole genome sequencing facilities and data sharing policies by some countries. Accordingly, an experimental evidence is required to confirm or role out our hypothesis.

Materials And Methods

SARS-CoV-2 genomes metadata

Metadata of all SARS-CoV-2 genomes submitted to the GISAID database (https://www.gisaid.org/CoV2020/), were accessed in July 7, 2020 (n = 60,782). Only genomes of viruses isolated from humans (n = 60,703) were selected for analysis. The genomes were submitted by labs from 100 countries around the world. Metadata of genomes included information on collection date, geographic location, patient gender, patient age, patient clinical status and viral genome clade. Genomic clades were defined according to GISAID database nomenclature system at the time of data collection.

Disease epidemiology data

Data of the disease epidemiology including total number of cases and total number of deaths in different countries were obtained from Coronavirus disease (COVID-19) Situation Report – 170, released by the WHO and available at (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports), accessed in July 8, 2020.

Statistical Analyses

Categorical data were expressed as percentages, while the median was used to describe the central tendency of the non-normally distributed numerical data. Group comparisons were done using Mann-Whitney U-test for numerical data and Chi-square (χ^2) or Fisher's exact test for categorical data. All statistical analyses were performed using the Statistical Package for Social Sciences (SPSS) software version 20.0 (IBM Corp., Armonk, NY, USA). P-value of less than 0.05 (two-tailed) was considered to be statistically significant.

Declarations

Author Contributions

S.M.H. and W.F.E planned and designed the research, analyzed the data, wrote and revised the manuscript. A.S.K. contributed to data curation as well as writing, reviewing and editing of the

manuscript. A.M.N contributed to research design, data curation as well as writing, reviewing and editing of the manuscript. All authors approved the final version of the manuscript.

Data availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest.

References

1. Gorbalenya, A. E. *et al.* Severe acute respiratory syndrome-related coronavirus: The species and its viruses – a statement of the Coronavirus Study Group. *bioRxiv*, 2020.2002.2007.937862, doi:10.1101/2020.02.07.937862 (2020).
2. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).
3. World Health Organization. Novel Coronavirus (2019-nCoV): situation report, 22 https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf?sfvrsn=fb6d49b1_2.
4. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494, doi:doi:https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 (2017).
5. Zhao, W. M. *et al.* The 2019 novel coronavirus resource. *Yi Chuan* **42**, 212-221, doi:10.16288/j.yczz.20-030 (2020).
6. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123, doi:10.1093/bioinformatics/bty407 (2018).
7. Singer, J., Gifford, R., Cotten, M. & Robertson, D. CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. *Preprints*, doi:10.20944/preprints202006.0225.v1 (2020).
8. Wang, C. *et al.* The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J Med Virol* **92**, 667-674, doi:10.1002/jmv.25762 (2020).
9. Korber, B. *et al.* Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*, doi:https://doi.org/10.1016/j.cell.2020.06.043 (2020).
10. Giovanetti, M., Angeletti, S., Benvenuto, D. & Ciccozzi, M. A doubt of multiple introduction of SARS-CoV-2 in Italy: A preliminary overview. *J Med Virol*, doi:10.1002/jmv.25773 (2020).
11. Castillo, A. E. *et al.* Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. *J Med Virol*, doi:10.1002/jmv.25797 (2020).

12. Zhou, Y. *et al.* Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* **6**, 14, doi:10.1038/s41421-020-0153-3 (2020).
13. Chen, Y. W., Yiu, C. B. & Wong, K. Y. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL (pro)) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Res* **9**, 129, doi:10.12688/f1000research.22457.2 (2020).
14. Robson, B. Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus. *Comput Biol Med* **119**, 103670, doi:10.1016/j.compbiomed.2020.103670 (2020).
15. Guan, W. J. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med* **382**, 1708-1720, doi:10.1056/NEJMoa2002032 (2020).
16. Jin, J. M. *et al.* Gender Differences in Patients With COVID-19: Focus on Severity and Mortality. *Front Public Health* **8**, 152, doi:10.3389/fpubh.2020.00152 (2020).
17. Richardson, S. *et al.* Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA*, doi:10.1001/jama.2020.6775 (2020).
18. Shi, Y. *et al.* Host susceptibility to severe COVID-19 and establishment of a host risk score: findings of 487 cases outside Wuhan. *Crit Care* **24**, 108, doi:10.1186/s13054-020-2833-7 (2020).
19. World Health Organization. Coronavirus disease (COVID-19): situation report, 170 https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200708-covid-19-sitrep-170.pdf?sfvrsn=bca86036_2
20. Jia, Y. *et al.* Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity. *bioRxiv*, 2020.2004.2009.034942, doi:10.1101/2020.04.09.034942 (2020).
21. Chiara, M., Horner, D. S., Gissi, C. & Pesole, G. Comparative genomics suggests limited variability and similar evolutionary patterns between major clades of SARS-CoV-2. *bioRxiv*, 2020.2003.2030.016790, doi:10.1101/2020.03.30.016790 (2020).
22. van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol* **83**, 104351, doi:10.1016/j.meegid.2020.104351 (2020).
23. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292 e286, doi:10.1016/j.cell.2020.02.058 (2020).
24. Brufsky, A. Hyperglycemia, hydroxychloroquine, and the COVID-19 pandemic. *J Med Virol* **92**, 770-775, doi:10.1002/jmv.25887 (2020).
25. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260-1263, doi:10.1126/science.abb2507 (2020).
26. Baud, D. *et al.* Real estimates of mortality following COVID-19 infection. *Lancet Infect Dis* **20**, 773, doi:10.1016/S1473-3099(20)30195-X (2020).

27. Joshi, M. *et al.* Genomic variations in SARS-CoV-2 genomes from Gujarat: Underlying role of variants in disease epidemiology. *bioRxiv*, 2020.2007.2010.197095, doi:10.1101/2020.07.10.197095 (2020).
28. Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *National Science Review* **7**, 1012-1023, doi:10.1093/nsr/nwaa036 (2020).
29. Brufsky, A. Distinct viral clades of SARS-CoV-2: Implications for modeling of viral spread. *J Med Virol*, doi:10.1002/jmv.25902 (2020).
30. Schurz, H. *et al.* The X chromosome and sex-specific effects in infectious disease susceptibility. *Hum Genomics* **13**, 2, doi:10.1186/s40246-018-0185-z (2019).
31. Liu, J. *et al.* Sex differences in renal angiotensin converting enzyme 2 (ACE2) activity are 17beta-oestradiol-dependent and sex chromosome-independent. *Biol Sex Differ* **1**, 6, doi:10.1186/2042-6410-1-6 (2010).
32. Hoffmann, M. *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, doi:10.1016/j.cell.2020.02.052 (2020).
33. Wambier, C. G. *et al.* Androgen sensitivity gateway to COVID-19 disease severity. *Drug Dev Res*, doi:10.1002/ddr.21688 (2020).
34. Chen, N. *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* **395**, 507-513, doi:10.1016/S0140-6736(20)30211-7 (2020).
35. Zhang, J. J. *et al.* Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China. *Allergy* **75**, 1730-1741, doi:10.1111/all.14238 (2020).
36. Wang, D. *et al.* Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA*, doi:10.1001/jama.2020.1585 (2020).
37. Alpert, A. *et al.* A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. *Nat Med* **25**, 487-495, doi:10.1038/s41591-019-0381-y (2019).
38. Koff, W. C. & Williams, M. A. Covid-19 and Immunity in Aging Populations - A New Research Agenda. *N Engl J Med*, doi:10.1056/NEJMp2006761 (2020).
39. Wu, Z. & McGoogan, J. M. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*, doi:10.1001/jama.2020.2648 (2020).
40. Dong, Y. *et al.* Epidemiology of COVID-19 Among Children in China. *Pediatrics* **145**, doi:10.1542/peds.2020-0702 (2020).
41. Brodin, P. Why is COVID-19 so mild in children? *Acta Paediatr* **109**, 1082-1083, doi:10.1111/apa.15271 (2020).

Figures

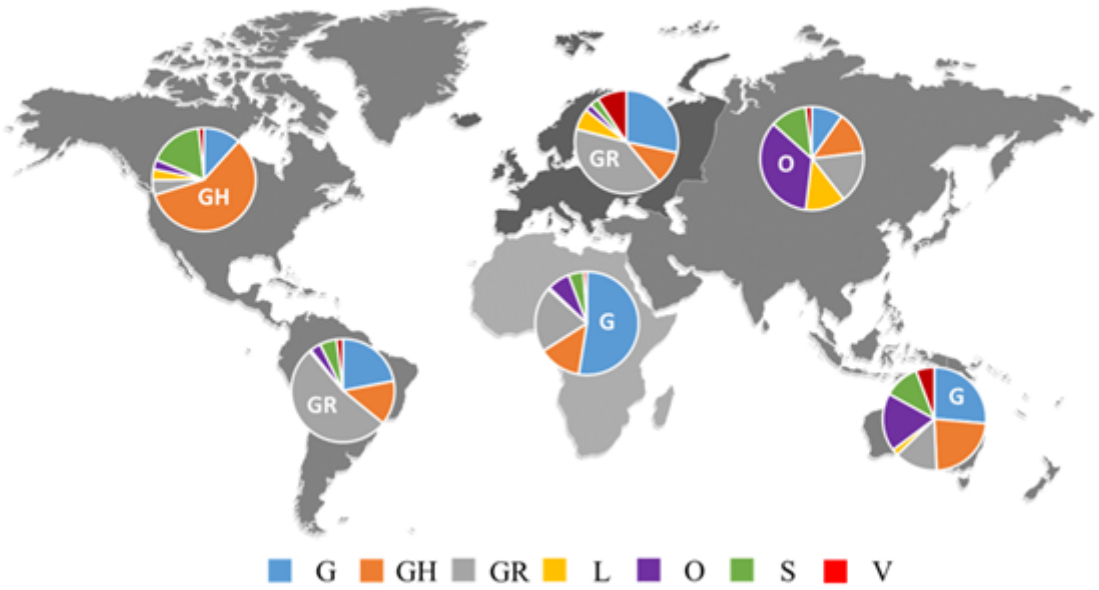

Figure 1

Continent distribution of various SARS-CoV-2 clades. The figure shows the predominance of G, O, GR, GH, G, and GR clades in Africa, Asia, Europe, North America, Oceania, and South America, respectively.

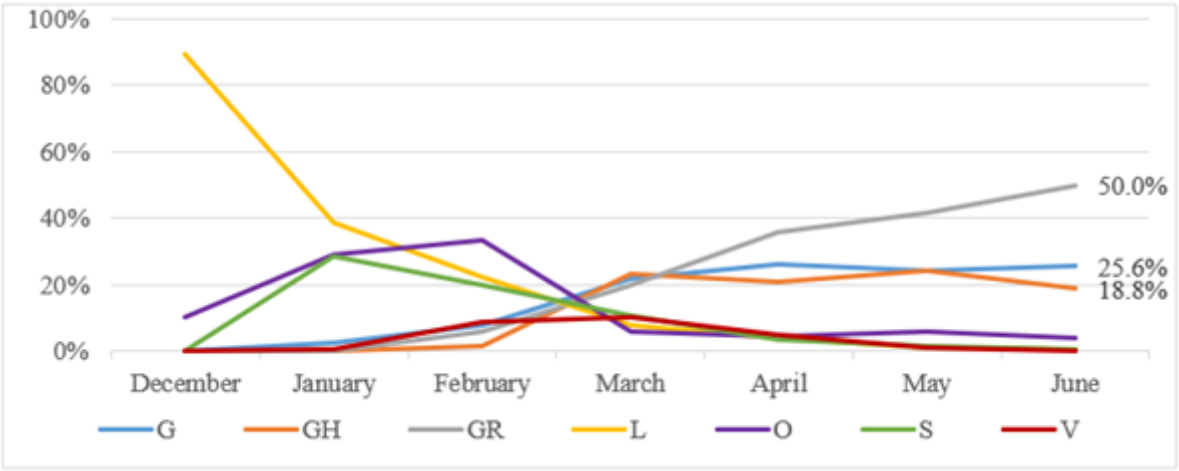

Figure 2

Global chronological distribution of SARS-CoV-2 clades in the period from December 2019 till June 2020. The figure shows a gradual regression in the number of genomes that belonged to clades L, S, O and V and expansion in the number of genomes that belonged to clades G, GR and GH.

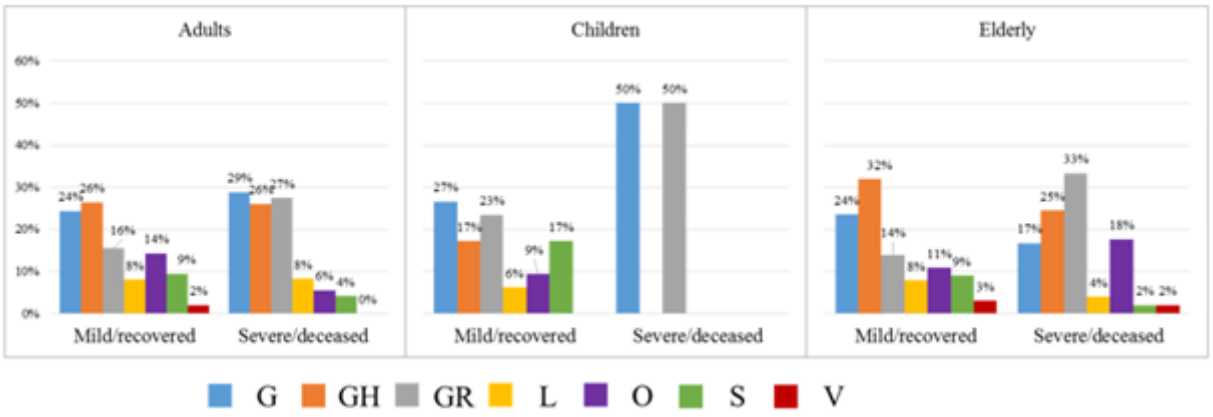

Figure 3

Distribution of SARS-CoV-2 clades in patients with mild/recovered cases versus severe/deceased cases in different age groups.