Worldwide SARS-COV-2 haplotype distribution

Andrea Cairo

Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Angelo Bianchi Bonomi Hemophilia and Thrombosis Center and Fondazione Luigi Villa, Milan, Italy

Marilena V. Iorio

Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

Silvia Spena

Università degli Studi di Milano, Department of Pathophysiology and Transplantation, Milan, Italy

Elda Tagliabue

Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

Flora Peyvandi (✉ flora.peyvandi@unimi.it)

Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Angelo Bianchi Bonomi Hemophilia and Thrombosis Center and Fondazione Luigi Villa, Milan, Italy

Research Article

Keywords: SARS-COV-2, genetic variants, haplotypes

DOI: https://doi.org/10.21203/rs.3.rs-50916/v1

Abstract

The world is experiencing one of the most severe viral outbreaks in the last years, the pandemic infection by SARS-COV-2, causative agent of COVID-19 disease. The virus reached over 120 countries, with a total number of 6.5 million infected, and 320000 deaths. A deeper understanding of its genomic diversity is mandatory.

We analyzed 21296 SARS-COV-2 reported sequences, defining the existence of recurrent haplotypes and their specific geographical distribution.

Main Text

In three months after the declaration of the SARS-COV-2 pandemic (**1**), the scientific community has been struggling to understand the complexity of this novel coronavirus of still debated zoonotic origin (**2**), the clinical symptoms (**3, 4**), the risk factors, the potential treatments, in the urgent effort to contain the infection, to predict potentially serious disease outcomes, to find a cure. Despite the circulation of SARS-COV-2 before the pandemic declaration has been ascertained, little is known about its worldwide spreading and the genetic changes that originated different viral strains.

The first case of COVID-19 was reported in Wuhan, China, in December 2019 and, despite the relatively low mortality (2% on average) and the high percentage of asymptomatic or pauci-symptomatic subjects (over 80%), the viral outbreak has literally caused a dramatic collapse of the health care system in the most hit countries, as in Northern Italy.

In the urgent race to find efficient drugs and decrease the complications, a deeper understanding of the genomic diversity of this virus is crucial. Indeed, the existence of different strains and their temporal and geographical distribution can provide relevant information on: how the virus spread all over the world; the possible acquisition of selective advantages; the most conserved sequences suitable for a vaccine design.

Since the first complete genomic sequence of SARS-COV-2 release on January 5th 2020 (**1, 2**), thousands of additional sequences have been deposited. Different virus strains have been reported **(5)** but, to the best of our knowledge, this manuscript describes the first analysis of all the genomic sequences reported so far. The variant call format (VCF) files, containing 21296 sequences of SARS-COV-2 at the date of manuscript preparation, was downloaded from http://covseq.baidu.com/ (**6**).

The geographical distribution of available sequences is reported in **Figure S1**. The samples were collected from December 2019 to April 2020 at different times across different countries (**Figure S2**). The first sequence was reported in Asia (Wuhan, December 2019), where the virus outbreak originated. It is worth noting that few sequences were early obtained (January 2020) in United States (**Supplementary Table S1).**

Using as reference the first reported sequence of SARS-COV-2 (RefSeq NC_045512.2), we described the geographical distribution of the variants reported in the merged VCF file of all available sequences**.**

China is the country with the highest percentage of unmutated viral genomes (13.8%) followed by USA (2.3%) (**Table 1**). In addition, observing the date of the first collected sample with viral sequence identical to the reference, we can notice how USA, Northern Europe, Australia and South Africa have been involved in the spread of the pandemic immediately after China (**Figure 1**).

To further explore the geographical occurrence, the timing of the virus circulation and to track its mutational evolution, we analyzed the minimum number of variants/sequence in each country (**Figure S3)**. This approach allowed us to define how some countries severely hit by COVID-19, as Italy, Spain and Brazil, are characterized by the occurrence and spread of already mutated forms of the virus. The number of variants/sequence spans mainly from 0 to 12; the majority of sequences carried from 4 to 5, and only few cases (63) had more than 12 variants (**Figure S4a**). A plausible hypothesis is that, whereas few variants might have provided favorable features, as an improved infectivity, a higher number of variants did not result in a selective advantage.

A total of 7197 variants were identified, the majority are missense (60%) and synonymous (32%) **(Figure S4b).** As expected, putative loss of function mutations (nonsense, deletion or insertion) were not found among the most frequent variants. Moreover, we observed 7733 different haplotypes, classified the 20 most frequent according to the number of cases (**Table 3**) and evaluated their frequencies, geographical distribution, temporal occurrence, and potential connection (**Figure S6-S9**). 5161 haplotypes are unique, whereas the remaining occurred in more than one patient.

Tracing mutations in the virus' genome is crucial to evaluate potential functional consequences and in the attempt to obtain an efficient vaccine. Indeed, more evolutionary conserved regions should be preferential target for the production of a vaccine. Our analysis underlines that, despite the high number of variants, the regions coding for the polyprotein ORF1ab, the spike (S) and the membrane (M) proteins are the most conserved (**Figure S4c**). This is not surprising, since most evolutionary conserved regions usually encode for essential proteins. However, we identified the missense variant p.Asp614Gly in the gene encoding for the spike protein as the most frequent (found in 13451 sequences, 63% of the total) (**Table 2)**. This variant was previously reported in smaller cohorts of samples (**7, 8**) and preliminary studies suggest that it might improve the binding affinity of the S protein to the human ACE2 receptor, reported as main entry site of the virus into human cells (**9**), through the cleavage of the S1/S2 domain (**10**). Indeed, it was previously shown that SARS-CoV infection can be enhanced by exogenous proteases (**11**). In addition, a very recent report (**12**) indicates a positive correlation between the frequency of this variant and a higher case-fatality rate, although further studies should be performed to demonstrate a causal role of this variant in a more severe disease outcome. Moreover, it is extremely relevant to notice that we found the p.Asp614Gly variant as a single mutation only in two patients, suggesting that this variant alone might not provide a selective advantage to the virus. Indeed, in the large majority of cases p.Asp614Gly occurs in concomitance with other variants, in particular with two variants located in the

ORF1ab gene, the missense mutation p.Pro4715Leu and the c.-25C>T in the 5' untranslated region, all characterized by similar frequency.

These three variants are indeed in strong linkage and constitute the most common haplotype, identified in 1523 sequences (haplotype 1, **Table 3**). This haplotype, first reported in Northern Italy (Lombardia) in February and representing 40% of Italian cases, has mainly spread to the rest of the European countries (1059 cases) and to North America (213 cases). Moreover, this triplet of variants represents the "common core" of 4400 different haplotypes, affecting 12949 patients (61% of total), of which the most frequent are reported in **Table 3.** Interestingly, 90% of the sequences reported in Italy (73 out of 81) has these three variants (*data not shown*).

A deeper evaluation of the haplotypes sharing the same variants and their geographical occurrence revealed the existence of specific clusters, likely reflecting a temporal and spatial spread of virus strains. Whereas haplotype 1 originated different haplotypes more common in North America (haplotypes 2, 8, 13) or in Europe (haplotype 4) (**Figure S5**), others seem to be peculiar for specific areas, as haplotype 3, first reported in USA, mainly present in North America, and characterized by 4 completely different variants (**Table 3**).

It would be extremely interesting and relevant to associate the different haplotypes to a specific outcome of the disease, and to understand whether the acquired mutations have functional consequences in terms of infectiveness, clinical severity, and potential responsiveness to specific treatments.

Declarations

Acknowledgments.

This work was supported by the Italian Ministry of Health - Bando Ricerca Corrente.

Dr. Iorio MV is supported by the Young Researcher Grant from the Italian Ministry of Health (grant N.GR-2016-02361750) and by a Career Foundation Grant from Berlucchi Foundation.

Author contributions.

A.C. conceived the work, performed the analyses, evaluated the results and edited the text; M.V.I. discussed the results and wrote the text; S.S. contributed to revise the text; E.T. helped in data discussion; F.P. supervised the work, discussed the data and revised the text.

Competing interests.

We have no competing interest to declare.

References

1. Wu, F. *et al*. *Nature* 579, 265–269 (2020).
2. Zhou, P. et al. *Nature* 579, 270-273 (2020).
3. Huang, C. et al. *Lancet* 395, 497-506 (2020).
4. Panigada, M. et al. *J Thromb Haemost.* https://doi.org/10.1111/jth.14850 (2020).
5. Van Dorp, L. et al. *Infect Genet Evol.* 83, 104351 (2020).
6. Liu, B. et al. Preprint https://doi.org/10.1101/2020.05.01.071050 (2020)
7. Gudbjartsson, D. F. et al. *Engl. J. Med.* https://doi.org/10.1056/NEJMoa2006100 (2020).
8. Biswas, N. K., Majumder, P. P. *Indian J. Med. Res.* (ACCEPTED) (2020).
9. Hoffmann, M. et al. 181(2):271-280 (2020).
10. Bhattacharyya, C. et al. Preprint https://doi.org/10.1101/2020.05.04.075911 (2020).
11. Matsuyama, S., Ujike, M., Morikawa, S., Tashiro, M., Taguchi, F. *Natl. Acad. Sci. U. S. A*. 102, 12543–12547 (2005).
12. Becerra-Flores, M., Cardozo, T. *Int J Clin Pract*. https://doi.org/10.1111/ijcp.13525 (2020)

Tables

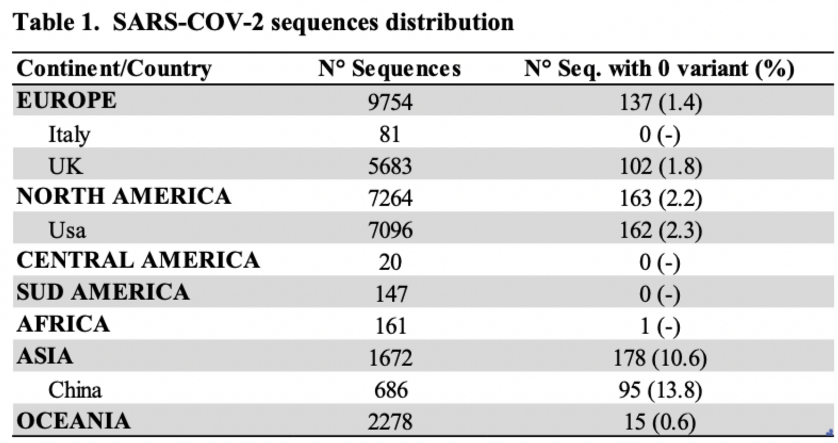
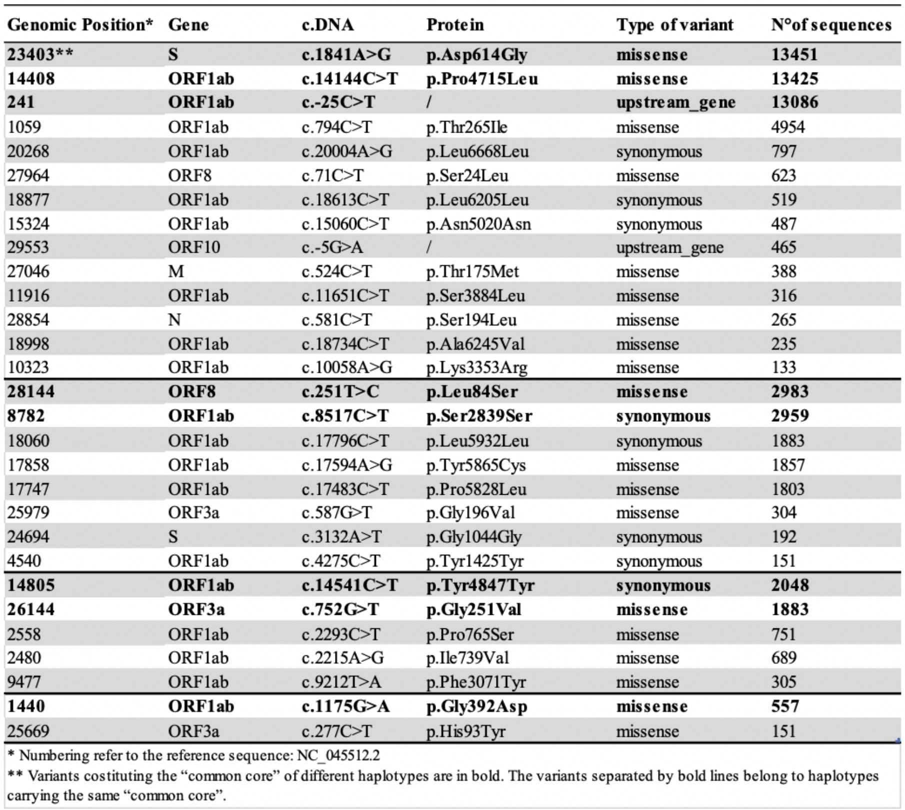


Table 1. SARS-COV-2 sequences distribution

Continent/Country	N° Sequences	N° Seq. with 0 variant (%)
EUROPE	9754	137 (1.4)
Italy	81	0 (-)
UK	5683	102 (1.8)
NORTH AMERICA	7264	163 (2.2)
Usa	7096	162 (2.3)
CENTRAL AMERICA	20	0 (-)
SUD AMERICA	147	0 (-)
AFRICA	161	1 (-)
ASIA	1672	178 (10.6)
China	686	95 (13.8)
OCEANIA	2278	15 (0.6)

Table 2. Most frequent SARS-COV-2 variants.

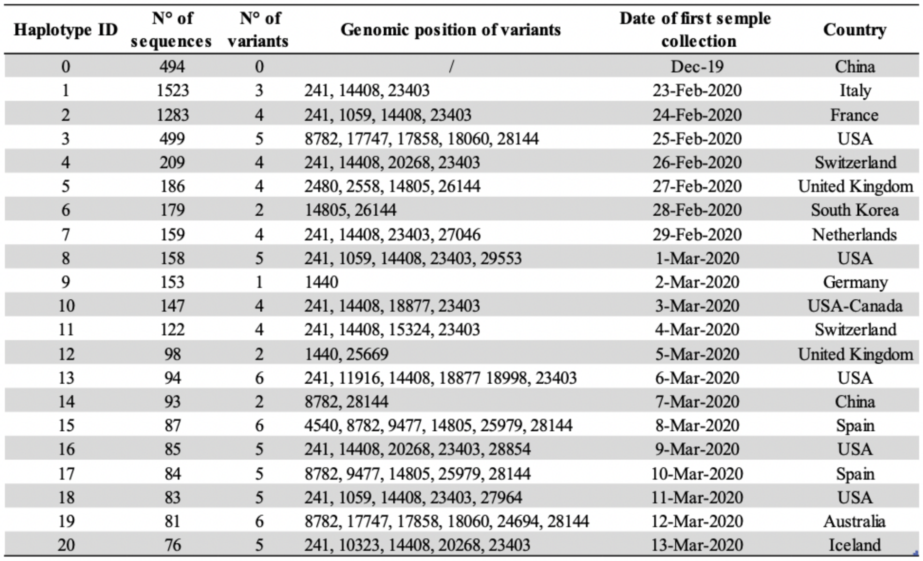


Genomic Position*	Gene	c.DNA	Protein	Type of variant	N°of sequences
23403**	**S**	**c.1841A>G**	**p.Asp614Gly**	**missense**	**13451**
14408	**ORF1ab**	**c.14144C>T**	**p.Pro4715Leu**	**missense**	**13425**
241	**ORF1ab**	**c.-25C>T**	**/**	**upstream_gene**	**13086**
1059	ORF1ab	c.794C>T	p.Thr265Ile	missense	4954
20268	ORF1ab	c.20004A>G	p.Leu6668Leu	synonymous	797
27964	ORF8	c.71C>T	p.Ser24Leu	missense	623
18877	ORF1ab	c.18613C>T	p.Leu6205Leu	synonymous	519
15324	ORF1ab	c.15060C>T	p.Asn5020Asn	synonymous	487
29553	ORF10	c.-5G>A	/	upstream_gene	465
27046	M	c.524C>T	p.Thr175Met	missense	388
11916	ORF1ab	c.11651C>T	p.Ser3884Leu	missense	316
28854	N	c.581C>T	p.Ser194Leu	missense	265
18998	ORF1ab	c.18734C>T	p.Ala6245Val	missense	235
10323	ORF1ab	c.10058A>G	p.Lys3353Arg	missense	133
28144	**ORF8**	**c.251T>C**	**p.Leu84Ser**	**missense**	**2983**
8782	**ORF1ab**	**c.8517C>T**	**p.Ser2839Ser**	**synonymous**	**2959**
18060	ORF1ab	c.17796C>T	p.Leu5932Leu	synonymous	1883
17858	ORF1ab	c.17594A>G	p.Tyr5865Cys	missense	1857
17747	ORF1ab	c.17483C>T	p.Pro5828Leu	missense	1803
25979	ORF3a	c.587G>T	p.Gly196Val	missense	304
24694	S	c.3132A>T	p.Gly1044Gly	synonymous	192
4540	ORF1ab	c.4275C>T	p.Tyr1425Tyr	synonymous	151
14805	**ORF1ab**	**c.14541C>T**	**p.Tyr4847Tyr**	**synonymous**	**2048**
26144	**ORF3a**	**c.752G>T**	**p.Gly251Val**	**missense**	**1883**
2558	ORF1ab	c.2293C>T	p.Pro765Ser	missense	751
2480	ORF1ab	c.2215A>G	p.Ile739Val	missense	689
9477	ORF1ab	c.9212T>A	p.Phe3071Tyr	missense	305
1440	**ORF1ab**	**c.1175G>A**	**p.Gly392Asp**	**missense**	**557**
25669	ORF3a	c.277C>T	p.His93Tyr	missense	151

* Numbering refer to the reference sequence: NC_045512.2

** Variants costituting the "common core" of different haplotypes are in bold. The variants separated by bold lines belong to haplotypes carrying the same "common core".

Table 3. Most frequent SARS-COV-2 haplotypes.



Haplotype ID	N° of sequences	N° of variants	Genomic position of variants	Date of first semple collection	Country
0	494	0	/	Dec-19	China
1	1523	3	241, 14408, 23403	23-Feb-2020	Italy
2	1283	4	241, 1059, 14408, 23403	24-Feb-2020	France
3	499	5	8782, 17747, 17858, 18060, 28144	25-Feb-2020	USA
4	209	4	241, 14408, 20268, 23403	26-Feb-2020	Switzerland
5	186	4	2480, 2558, 14805, 26144	27-Feb-2020	United Kingdom
6	179	2	14805, 26144	28-Feb-2020	South Korea
7	159	4	241, 14408, 23403, 27046	29-Feb-2020	Netherlands
8	158	5	241, 1059, 14408, 23403, 29553	1-Mar-2020	USA
9	153	1	1440	2-Mar-2020	Germany
10	147	4	241, 14408, 18877, 23403	3-Mar-2020	USA-Canada
11	122	4	241, 14408, 15324, 23403	4-Mar-2020	Switzerland
12	98	2	1440, 25669	5-Mar-2020	United Kingdom
13	94	6	241, 11916, 14408, 18877 18998, 23403	6-Mar-2020	USA
14	93	2	8782, 28144	7-Mar-2020	China
15	87	6	4540, 8782, 9477, 14805, 25979, 28144	8-Mar-2020	Spain
16	85	5	241, 14408, 20268, 23403, 28854	9-Mar-2020	USA
17	84	5	8782, 9477, 14805, 25979, 28144	10-Mar-2020	Spain
18	83	5	241, 1059, 14408, 23403, 27964	11-Mar-2020	USA
19	81	6	8782, 17747, 17858, 18060, 24694, 28144	12-Mar-2020	Australia
20	76	5	241, 10323, 14408, 20268, 23403	13-Mar-2020	Iceland

Figures

Figure 1

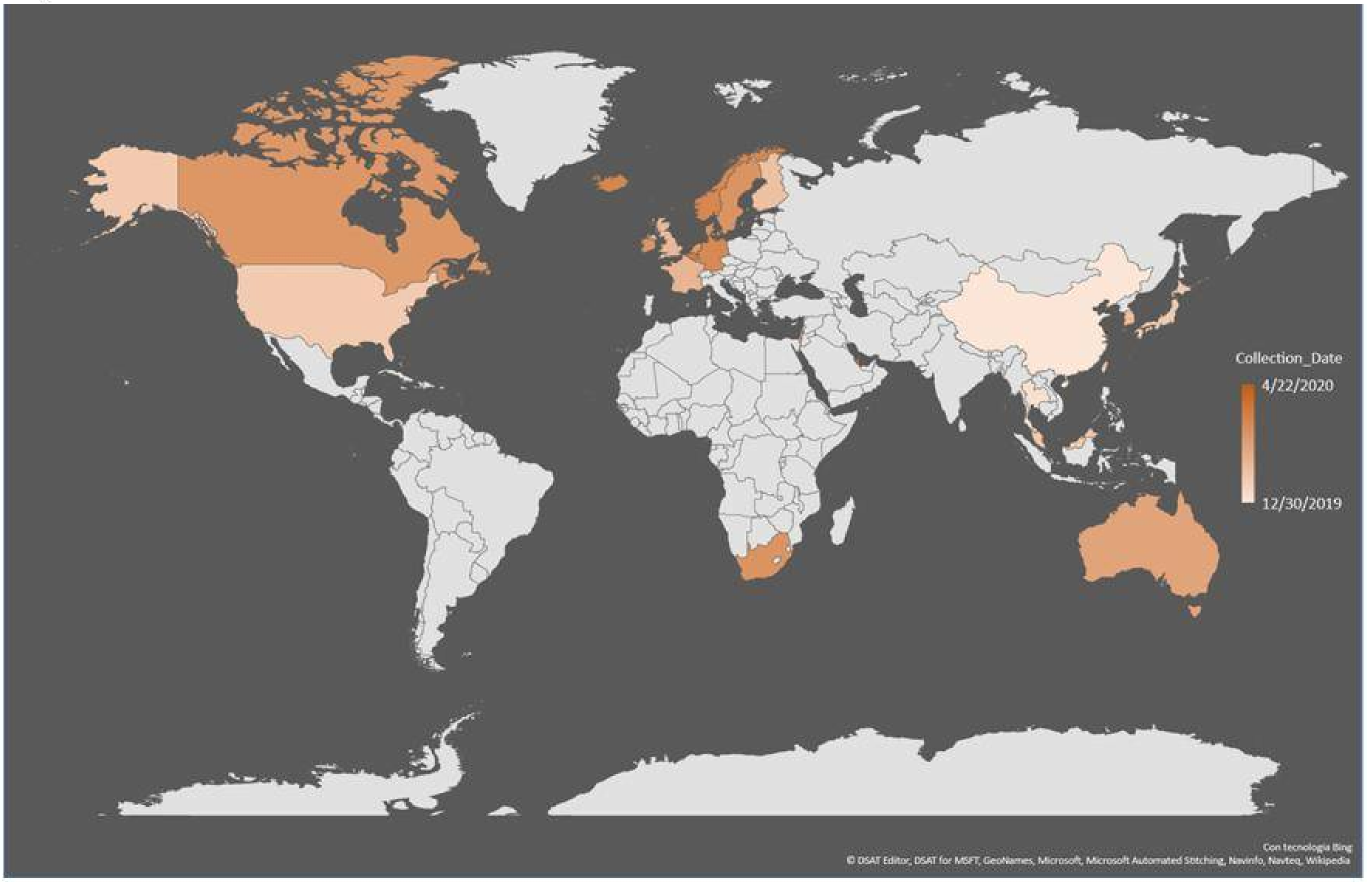

Collection_Date
4/22/2020
12/30/2019
Con tecnologia Bing
© DSAT Editor, DSAT for MSFT, GeoNames, Microsoft, Microsoft Automated Stitching, Navinfo, Navteq, Wikipedia

Figure 1

Geographical distribution of the reference sequence based on the date of the first reported unmutated sequence. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinfo.docx
- SupplementaryTable1datasetsarscov2sequencesinformations.xlsx
- SupplementaryFigures.pdf