Genomic characterization of SARS-CoV-2

1st Atanas Krstev
Faculty of engineering and computer science
Skopje, Republic of North Macedonia
atanas.krstev@students.finki.ukim.mk

2nd Vasil Manavski
Faculty of engineering and computer science
Skopje, Republic of North Macedonia
vasil.manavski@students.finki.ukim.mk

***Abstract*—A new severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) associated with human to human transmission and extreme human sickness has been as of late announced from the city of Wuhan in China. Our objectives were to mutation analysis between recently reported genomes at various times and locations and to characterize the genomic structure of SARS-CoV-2 using bioinformatics programs.**

***Index Terms*—SARS-CoV-2, Mutation, COVID-19**

I. Introduction

Due to the ubiquity of the virus, we have an abundance of data that can be studied. The aim of this project is to analyze all the genomes that have been sequenced so far, in order to better understand the mutations that occur in SARS-CoV-2. Our main goal is to find mutations, where those mutations occur, in which gene / non-encoding region they are located and what significance they have for the virus itself.

The database that we are using is the NCBI Virus database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/).

Importantly, the genome size of the SARS-CoV-2 varies from 29.8 kb to 29.9 kb and its genome structure followed the specific gene characteristics to known CoVs; the 5' more than two-thirds of the genome comprises orf1ab encoding orf1abpolyproteins, while the 3' one third consists of genes encoding structural proteins including surface (S), envelope (E), membrane (M), and nucleocapsid N proteins ("Fig. 1"). Additionally, the SARS-CoV-2 contains 6 accessory proteins, encoded by ORF3a, ORF6, ORF7a, ORF7b, and ORF8 genes.

II. Methodology

We took 3815 complete genomes from the aforementioned database. Of all the genomes, some were not used because they contained "N" nucleotides, leaving 2769 sequences. NC_045512 genome sequence was used for reference and the genomic coordinate in this study is based on this reference genome. We store the data locally/on google drive because this method was more time efficient as opposed to connecting to the database via Entrez. After pre-processing, we saved the data in a file and thus the data is ready for deeper analysis. Our reference sequence is in a separate file, while all other sequences (1998) were in another file, and we used the online tool EMBOSS [3] and got the result of the alignment in a single file.

In the next step, we separated all the alignments from the single file because the AlignIO.read() method does not allow reading multiple alignments from the same file. We go through each file individually and find the differences, looking for which gene they are in terms of the reference sequence. Then in the gene found we record which codon has changed and which type of change has occurred. For each of the mutations, we count how many times it appeared in the alignments. From the obtained mutations we identify those that have the highest frequency of occurrence, and from them we find the pairs that coincide and we also find the triplets of mutations that coincide. The definition of coincide that we are using is that of an implication i.e. if mutation A occurs then mutation B also occurs, but the opposite doesn't have to hold.

III. Results

A total of 2327 mutations have been found, of which 1600 occur only once i.e. they are unique, as some of them are shown in Table IV. Among the 1999 genomes we analyzed, 27 samples did not exhibit any variants except for missing parts in the non-coding regions of the genome.

Some of the most common mutations are shown below along with the gene they are found in and the amount of times the appear in the sequences, Table I:

TABLE I
Most common mutations detected in SARS-CoV-2 genomes

Mutation	Gene	Count
3036C > T	ORF1ab	1436
23402A > G	S	1434
14407C > T	ORF1ab	1420
25562G > T	ORF3a	1029
1058C > T	ORF1ab	724
28143T > C	ORF8	322
8781C > T	ORF1ab	313
18059C > T	ORF1ab	206
17857A > G	ORF1ab	202
17746C > T	ORF1ab	193
28880G > A	N	185
28881G > A	N	185
28882G > C	N	185
18876C > T	ORF1ab	159
2415C > T	ORF1ab	140
11082G > T	ORF1ab	113
26734C > T	M	92

No special technique was used for the threshold value, but according to the obtained data, we arbitrarily selected it.

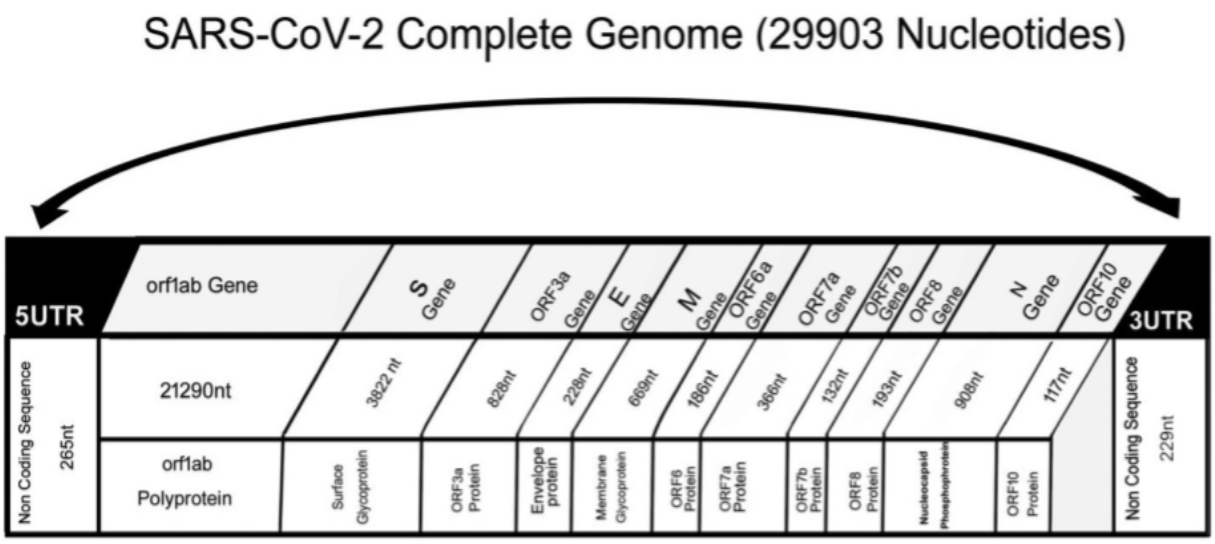

Fig. 1. Structure of the SARS-CoV-2 genome. [2]

The following Table II also shows some of those mutations that coincide in pairs, according to the previously defined definition:

TABLE II
COINCIDENTAL PAIRS OF MUTATIONS DETECTED IN SARS-COV-2 GENOMES

Mutation A	Mutation B
17857A > G	18059C > T
17746C > T	18059C > T
2415C > T	3036C > T
3891G > T	3036C > T
28880G > A	3036C > T
28881G > A	3036C > T
28882G > C	3036C > T
26734C > T	3036C > T
28853C > T	3036C > T
3891G > T	23402A > G
26734C > T	23402A > G
2415C > T	23402A > G
28880G > A	23402A > G
28881G > A	23402A > G
28882G > C	23402A > G

We also found those triplets that coincide. The previous definition is easily expandable and in this case it would mean that mutation A implies two other mutations B and C, but not the other way around. From this it is also known that A coincides with each of the other two mutations separately, but nothing can be claimed about the relation of mutations B and C from this relationship. We choose not to show them here as doing so would take even more space, and we certainly aren't short on tables.

We also analyzed the mutations that are present in ORF1ab, which is the longest ORF occupying 2/3 of the entire genome. ORF1ab is cleaved into many nonstructural proteins (NSP1-NSP16).

The distribution of mutations across this gene is given on the following Table III:

TABLE III
NON-STRUCTURAL PROTEINS IN THE ORF1AB GENE AND THE AMOUNT OF MUTATIONS IN EACH ONE.

Nonstructural protein	Count
nsp1	148
nsp2	1291
nsp3	2286
nsp4	496
nsp5	161
nsp6	178
nsp7	74
nsp8	73
nsp9	43
nsp10	23
nsp11	1
nsp12	1810
nsp13	604
nsp14	562
nsp15	162
nsp16	84

Nonstructural protein 3 is a multifunctional protein comprising up to 16 different domains and regions. Nsp3 binds to viral RNA, nucleocapsid protein, as well as other viral proteins,

TABLE IV
CODING MUTATION LIST DETECTED IN SARS-COV-2 GENOMES

Accession	Location	Nucleotide variation	Gene	Amino acid change	Mutation type
MT374101.1	Taiwan	8781C > T	ORF1ab		Synonymous mutation
MT374101.1	Taiwan	27893_28228del	ORF8		Deletion
MT374101.1	Taiwan	27847_27886del	ORF8		Deletion
MT374101.1	Taiwan	16840A > G	ORF1ab	K > E	Missense
MT374101.1	Taiwan	21706C > T	S	H > Y	Missense
MT374101.1	Taiwan	24212C > T	S	S > F	Missense
MT259240.1	USA	1058C > T	ORF1ab	T > I	Missense
MT259240.1	USA	3036C > T	ORF1ab		Synonymous mutation
MT259240.1	USA	14407C > T	ORF1ab		Synonymous mutation
MT259240.1	USA	23402A > G	S	D > G	Missense
MT259240.1	USA	25562G > T	ORF3a	Q > H	Missense
MT259239.1	USA	1058C > T	ORF1ab	T > I	Missense
MT259239.1	USA	3036C > T	ORF1ab		Synonymous mutation
MT259239.1	USA	14407C > T	ORF1ab		Synonymous mutation
MT259239.1	USA	23402A > G	S	D > G	Missense
MT259239.1	USA	25562G > T	ORF3a	Q > H	Missense
MT419820.1	Puerto Rico	1058C > T	ORF1ab	T > I	Missense
MT419820.1	Puerto Rico	1631A > T	ORF1ab	K > I	Missense
MT419820.1	Puerto Rico	3036C > T	ORF1ab		Synonymous mutation
MT419820.1	Puerto Rico	14407C > T	ORF1ab		Synonymous mutation
MT419820.1	Puerto Rico	23402A > G	S	D > G	Missense
MT419820.1	Puerto Rico	25562G > T	ORF3a	Q > H	Missense
MT459847.1	Greece	2479A > G	ORF1ab	I > V	Missense
MT459847.1	Greece	514_519del	ORF1ab		Deletion
MT459847.1	Greece	2557C > T	ORF1ab	P > S	Missense
MT459847.1	Greece	9490C > T	ORF1ab	H > Y	Missense
MT459847.1	Greece	11082G > T	ORF1ab	L > F	Missense
MT459847.1	Greece	14804C > T	ORF1ab	T > I	Missense
MT459847.1	Greece	26143G > T	ORF3a	G > V	Missense
MT459862.1	Greece	2479A > G	ORF1ab	I > V	Missense
MT459862.1	Greece	514_519del	ORF1ab		Deletion
MT459862.1	Greece	2557C > T	ORF1ab	P > S	Missense
MT459862.1	Greece	11082G > T	ORF1ab	L > F	Missense
MT459862.1	Greece	14804C > T	ORF1ab	T > I	Missense
MT459862.1	Greece	26143G > T	ORF3a	G > V	Missense
MT450936.1	Australia	2479A > G	ORF1ab	I > V	Missense
MT450936.1	Australia	2557C > T	ORF1ab	P > S	Missense
MT450936.1	Australia	6970T > C	ORF1ab		Synonymous mutation

and participates in polyprotein processing. The papain-like protease of Nsp3 is an established target for new antivirals. Through its de-ADP-ribosylating, de-ubiquitinating, and de-ISGylating activities, Nsp3 counteracts host innate immunity. Structural data are available for the N-terminal two thirds of Nsp3, but domains in the remainder are poorly characterized [4].

Among NSP's, as we can see, NSP3 has more variants in the analyzed samples, while some NSP's, such as nsp11, have very little to no mutations at all. The function of NSP11 seems to be unknown [6].

Nsp12 also exhibits a high number of mutations. Nsp12 itself is capable of conducting the polymerase reaction with extremely low efficiency, whereas the presence of nsp7 and nsp8 cofactors remarkably stimulates its polymerase activity.

The nsp12-nsp7-nsp8 subcomplex is thus defined as the minimal core component for mediating coronavirus RNA synthesis. To achieve complete transcription and replication

TABLE V
MUTATIONS AND FREQUENCY IN WHICH THEY OCCUR IN DIFFERENT COUNTRIES.

Mutation	USA	Australia	Bangladesh	Spain	Italy	China	Germany	Czech Republic	Greece	Jamaica	Tunisia	Puerto Rico	India	Serbia
3036C > T	961	132	15	7	5	3	37	20	61	3	4	4	161	7
240C > T	958	132	15	7	5	3	37	20	61	3	4	4	161	7
23402A > G	960	132	15	7	5	3	37	20	61	3	4	4	160	7
14407C > T	964	132	14	7	5	3	23	20	60	3	4	4	158	7
25562G > T	847	49	0	0	0	0	16	4	6	3	3	3	97	1
1058C > T	653	42	0	0	0	0	16	4	1	3	1	3	1	0
28143T > C	219	57	1	3	0	20	1	0	2	0	1	1	3	0
8781C > T	209	57	1	3	0	20	1	0	2	0	1	1	3	0
18059C > T	177	27	0	0	0	1	0	0	0	0	0	1	0	0
17857A > G	174	27	0	0	0	0	0	0	0	0	0	1	0	0
17746C > T	171	21	0	0	0	0	0	0	0	0	0	1	0	0
28880G > A	45	42	14	0	1	1	4	15	48	0	0	1	2	3
28881G > A	45	42	14	0	1	1	4	15	48	0	0	1	2	3
28882G > C	45	42	14	0	1	1	4	15	48	0	0	1	2	3
18876C > T	56	1	0	0	0	0	0	0	5	0	1	0	95	1
2415C > T	133	6	0	0	0	0	0	0	0	0	1	0	0	0
35C > T	134	0	0	0	0	0	0	0	0	0	0	0	0	0
11082G > T	42	30	0	1	1	3	0	4	12	4	0	0	9	0
26734C > T	28	34	0	3	0	1	0	0	13	3	0	0	0	0
14804C > T	25	20	0	1	1	2	0	0	12	3	0	0	0	0
3891G > T	71	0	0	0	0	0	0	0	0	0	0	0	0	0
26143G > T	4	3	0	0	0	0	0	0	0	0	0	0	65	0
28853C > T	72	0	0	0	0	0	0	0	0	0	0	0	0	0
34A > T	0	0	0	0	0	0	0	0	0	0	0	0	92	0

of the viral genome, several other nsp subunits are required to assemble into a holoenzyme complex, including nsp10, nsp13, nsp14 and nsp16, for which the precise functions in RNA synthesis are not well understood [7].

The viral nonstructural protein 1 (nsP1) is the only membrane-associated protein that anchors the replication complex to the cellular membranes. NSP1 inhibits host translation by interacting with the 40S ribosomal subunit.

The nsp1-40S ribosome complex further induces an endonucleolytic cleavage near the 5'UTR of host mRNAs, targeting them for degradation. Viral mRNAs are not susceptible to nsp1-mediated endonucleolytic RNA cleavage thanks to the presence of a 5'-end leader sequence and are therefore protected from degradation. By suppressing host gene expression, nsp1 facilitates efficient viral gene expression in infected cells and evasion from host immune response. [5]

Given the vast amount of data that we have on our hands, it would be a pity not to analyze it in terms of the localization of mutations in different geographic regions. Shown on Table V is just that i.e. for a given mutation we have the number of occurrences in each country, though there may be a slight bias given that most of our data is from the USA.

IV. DISCUSSION

When comparing our analysis with [2], we have seen several differences in the results. Namely, the most common mutation in [2] is 8782C > T(ORF1ab), while according to our research this mutation occurs in the middle of the spectrum of frequency of mutations, although the number of occurrences in our case (313) is fairly larger that in the research aforementioned (13).

The other difference is in the coincident mutations,in their case the occurrences of 8782C > T and 28144T > C coincide, while we have found our coincident mutations in Table II and we haven't detected that type of coincidental pair of mutation.

All non-coding mutations are located in 5'UTR or 3'UTR regions. In terms of base changes, the most frequently observed mutation is C > T.

These differences are due to differences in data volume, and mutation positions are slightly different due to different ways of indexing(we use zero based indexing).

TABLE VI
MOST COMMON MUTATIONS IN NON-CODING REGIONS.

Nucleotide variation	Count
240C > T	1435
35C > T	134
34A > T	71
29699A > G	51
33A > T	49

V. CONCLUSION

Viruses may seem like cunning villains, purposefully mutating to increasingly deadlier forms to outwit their human hosts. Over the last century alone, several global epidemics have claimed millions of lives, including the 1957/58 influenza A (H2N2) pandemic, the sixth (1899–1923) and seventh cholera pandemic (1961–1975), as well as the HIV/AIDS pandemic (1981-today) [1]. COVID-19 acts as an unwelcome reminder of the major threat that infectious diseases represent in terms of deaths and disruption. One of the positive aspects of the current pandemic is the availability of an enormous amount of data for analysis and processing related to SARS-CoV-2.

The mobilisation to address the COVID-19 pandemic by scientists worldwide has been remarkable. This includes the

feat of the global scientific community who has already produced and publicly shared above 11,000 complete SARS-CoV-2 genome sequences at the time of writing (July 2, 2020).

In this study, we focused on mutations in the genomes discovered, the frequency of a particular mutation per state, and the coinciding between the mutations. As a community we should continue gathering and sharing information about variants, we believe that this kind of approach will be a step forward in the fight against the pandemic.

REFERENCES

[1] Lucy van Dorp, Mislav Acman, Damien Richard, Liam P. Shaw, Charlotte E. Ford, Louise Ormond, Christopher J. Owen, Juanita Pang, Cedric C.S. Tan, Florencia A.T. Boshier, Arturo Torres Ortiz, François Balloux. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2.

[2] Khailany, Rozhgar & Safdar, Muhammad & Ozaslan, Mehmet. (2020). Genomic characterization of a novel SARS-CoV-2. Gene.

[3] "Emboss Needle Tool"

[4] Jian Lei, Yuri Kusov, Rolf Hilgenfeld. (2018). Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein

[5] "SARS-CoV-2 (COVID-19) NSP1 Protein, His Tag"

[6] Yoshimoto, F.K. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. Protein J 39, 198–216 (2020)

[7] Qi Peng, Ruchao Peng, Bin Yuan, Jingru Zhao, Min Wang, Xixi Wang, Qian Wang, Yan Sun, Zheng Fan, Jianxun Qi, George F. Gao, Yi Shi. (2020). Structural and Biochemical Characterization of the nsp12-nsp7-nsp8 Core Polymerase Complex from SARS-CoV-2