 doi:10.20944/preprints202004.0405.v1

Some Findings on Genes over SARS-CoV2 Genomes

Sk. Sarif Hassan[a,*], Atanu Moitra[b], Pabitra Pal Choudhury[c], Prasanta Pramanik[d], Siddhartha Jana[e]

[a] *Department of Mathematics, Pingla Thana Mahavidyalaya, Maligram 721140, India*
[b] *CMO, Government of West Bengal, India.*
[c] *Applied Statistics Unit, Indian Statistical Institute, Kolkata 700108, West Bengal, India.*
[d] *Finance Department, Government of West Bengal, India.*
[e] *School of Biological Sciences, Indian Association for the Cultivation of Science, West Bengal, 700032, India.*

Abstract

Coronaviruses are a large family of RNA viruses which cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS), Severe Acute Respiratory Syndrome (SARS) and COVID-19. This article highlights some key findings based on a thorough scanning of genes of 470 SARS-CoV2 genomes, including the co-presence of ORF7a and ORF8 over the 251 SARS-CoV2 genomes and the absence of the gene ORF7b over the 219 SARS-CoV2 genomes collected from various countries including India.

Keywords: SARS-CoV2; ORF7a; ORF7b; ORF8; SARS-CoV2 Genomes.

1. Introduction

The outbreak of the SARS-CoV2, a novel coronavirus becomes now a pandemic [1, 2]. The genome of SARS-CoV2 contains approximately 30kbp nucleotides and each genomes contains around 11 genes of various type such as S,

*Corresponding author
Email addresses: sarimif@gmail.com (Sk. Sarif Hassan), atanu.mli@gmail.com (Atanu Moitra), pabitrapalchoudhury@gmail.com (Pabitra Pal Choudhury), prasanta.pramanik@gmail.com (Prasanta Pramanik), siddhartha.jana@gmail.com (Siddhartha Jana)

E, M, N etc [3]. The SARS-CoV2 genomes consist of structural protein gene S which specifically bind to the receptor of the host cell, and this is the key protein for viruses to invade susceptible cells. The gene M and E are involved in the formation of the virus envelope, while the gene N is involved in the assembly of the virus [4]. The origin of the source of the virus SARS-CoV2 and its intermediate host is still controversial [5]. The phylogenetic analysis implied that the coronavirus was the most similar to Bat coronavirus isolate RaTG13 (GenBank No.: MN996532), with 96.2% nucleotide homology in the whole genome [5]. It is also reported that the SARS-CoV2 was closely related (with 88% identity) to two bat-derived severe acute respiratory syndrome (SARS)-like coronaviruses, bat-SL-CoVZC45 and bat-SL-CoVZXC21, collected in 2018 in Zhoushan, eastern China, but were more distant from SARS-CoV (about 79%) and MERS-CoV (about 50%) [6]. It is reported that the genome sequences MT050493, MT012098 from India are highly similar to the genome of the Wuhan seafood market pneumonia virus (accession number: NC 045512) [7]. The other recent findings and present state of the art including review can be obtained from the various articles [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18].

Replication of Coronavirus is caused by a set of highly conserved viral proteins. Only two ORFs such as ORF3a and ORF7a have been in virus-infected cells among the eight putative accessory proteins encoded by the (SARS-CoV) [19]. The ORF7b gene is expressed in virus-infected cell lysates and from a cDNA encoding the gene 7 coding region, indicating that the sgRNA7 is bicistronic. The ORF7b protein is not only an accessory protein but a structural component of the SARS-CoV virion [19].

It is needless to mention that a deep scanning over these genomes and associated genes is necessary for various reasons including pathogenesis [3]. In this article, an attempt has been made to search out the gene variations among

 doi:10.20944/preprints202004.0405.v1

the 470 SARS-CoV2 genomes. This article reports that a pair of genes viz. ORF7a and ORF8 is present across almost all the SARS-CoV2 genomes except MN988668, MN988669 and MT121215 genomes from China. Also it is found that ORF7b is absent across 219 genomes of SARS-CoV2 such as MT050493 (India), MT012098 (India).

1.1. Dataset

As on date 14th April, 2020, 470 *complete* genome sequences of SARS-CoV2 were available in the NCBI virus database. From the NCBI Virus Database (*https* : *//www.ncbi.nlm.nih.gov/labs/virus/vssi/*) we fetch all the 470 genomes and their associated genes from each genomes. The accession of the 470 genomes are given in the Table-1. The known gene-organization of a typical SARS-CoV2 genomes is given in the Fig. 1. The first ORF representing approximately 67% of the entire genome encodes 16 non-structural proteins (nsps), while the remaining ORFs encode accessory proteins and structural proteins [20].

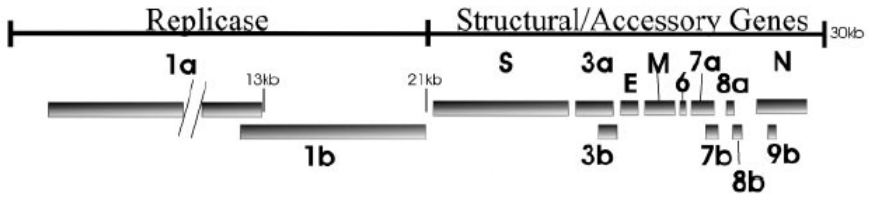

Figure 1: Gene-organizations over a typical SARS-CoV2 genome. [19]

 doi:10.20944/preprints202004.0405.v1

Table 1: List of the 470 SARS-CoV2 Genome Accessions (*Data collected on 14th April, 2020*)

Accession	Accession	Accession	Accession	Accession	Accession
NC_045512	MT291836	MT262912	MT263440	MT253707	MT163716
MT077125	MT291831	MT262913	MT263447	MT251972	MT159706
MT039874	MT293170	MT262914	MT263449	MT251974	MT159716
MT322402	MT292574	MT262993	MT263455	MT251975	MT159719
MT322407	MT293178	MT263074	MT263444	MT251973	MT159707
MT322396	MT293181	MT263381	MT263445	MT251976	MT159717
MT322394	MT293183	MT263391	MT263451	MT251979	MT159709
MT322397	MT293195	MT262901	MT263420	MT253697	MT159715
MT322398	MT293196	MT262903	MT263441	MT253699	MT159718
MT322400	MT293197	MT262907	MT263454	MT253696	MT159722
MT322401	MT293204	MT262909	MT263464	MT253698	MT159708
MT322403	MT293212	MT262911	MT263465	MT253700	MT159705
MT322404	MT293215	MT262899	MT263468	MT251977	MT159710
MT322405	MT293216	MT262904	MT263394	MT251978	MT159711
MT322406	MT293219	MT262915	MT263395	MT251980	MT159712
MT322408	MT293224	MT262916	MT263396	MT233526	MT159713
MT322409	MT293225	MT262897	MT259226	MT246667	MT159714
MT322410	MT293206	MT262898	MT259275	MT246451	MT159720
MT322411	MT293208	MT262905	MT259247	MT246461	MT159721
MT322412	MT293209	MT262910	MT258377	MT246471	MT121215
MT322413	MT295464	MT263400	MT258378	MT246472	MT066156
MT322414	MT293160	MT263382	MT258379	MT246474	MT050493
MT322415	MT293166	MT263383	MT259231	MT246450	MT012098
MT322416	MT293171	MT263384	MT259228	MT246453	MT152824
MT322417	MT293190	MT262896	MT259248	MT246454	MT135044
MT322418	MT293161	MT263415	MT259227	MT246462	MT135042
MT322419	MT293167	MT263406	MT259236	MT246464	MT135041
MT322420	MT293168	MT263408	MT258380	MT246470	MT135043
MT322421	MT293174	MT263422	MT259235	MT246473	MT126808
MT322422	MT293175	MT263469	MT259237	MT246480	MT123293
MT322423	MT293182	MT263439	MT259239	MT246484	MT123291
MT322424	MT293191	MT263457	MT259281	MT246449	MT123290
MT322395	MT293158	MT263459	MT259282	MT246455	MT123292
MT320538	MT293162	MT263432	MT259243	MT246456	MT118835
MT320891	MT293163	MT263450	MT259249	MT246478	MT106052
MT308702	MT293164	MT263458	MT259250	MT246485	MT106053
MT308703	MT281577	MT263467	MT259251	MT246488	MT106054
MT308704	MT293156	MT263411	MT259256	MT246452	MT093571
MT304476	MT293159	MT263413	MT259258	MT246460	MT093631
MT304474	MT291834	MT263426	MT259266	MT246481	MT072688
MT304475	MT291829	MT263421	MT259267	MT246482	MT066176
MT304478	MT291827	MT263443	MT259274	MT246490	MT066175
MT304482	MT291830	MT263412	MT259286	MT246459	MT044258
MT304488	MT291828	MT263416	MT259287	MT246468	MT044257
MT304479	MT293169	MT263417	MT259241	MT246475	MT049951
MT304480	MT293200	MT263423	MT258381	MT246477	MT039887
MT304481	MT293210	MT263431	MT259257	MT246479	MT039888
MT304486	MT293211	MT263410	MT259261	MT246457	MT039890
MT304487	MT293218	MT263424	MT259263	MT246458	MT039873
MT304490	MT295465	MT263425	MT259264	MT246466	MT027062
MT304491	MT293198	MT263442	MT259268	MT246467	MT027063
MT304483	MT293205	MT263402	MT259269	MT246469	MT027064
MT304484	MT293207	MT263405	MT259271	MT246476	MT020881
MT304485	MT293213	MT263418	MT259273	MT246486	MT019530
MT304489	MT293220	MT263419	MT259277	MT246487	MT019531
MT300186	MT293222	MT263398	MT259278	MT246489	MT019533
MT304477	MT292569	MT263399	MT259280	MT240479	MT020880
MT292571	MT293172	MT263403	MT258383	MT233523	MT019532
MT293186	MT293177	MT263404	MT258382	MT233519	MT019529
MT292570	MT293176	MT263414	MT259246	MT233522	MT007544
MT292573	MT293199	MT263430	MT259244	MT226610	MN996531
MT293173	MT293165	MT263390	MT259245	MT198652	MN996530
MT292575	MT276597	MT263434	MT259252	MT192773	MN996527
MT293179	MT276598	MT263436	MT259253	MT192772	MN996528
MT293180	MT276323	MT263446	MT259254	MT192765	MN996529
MT293184	MT276328	MT263448	MT259284	MT192759	MN997409
MT293189	MT276331	MT263452	MT259229	MT188341	MN988668
MT293192	MT276329	MT263453	MT259230	MT188340	MN988669
MT293194	MT276330	MT263456	MT259260	MT188339	MN994467
MT293201	MT276324	MT263462	MT259285	MT184909	MN994468
MT293202	MT276325	MT263463	MT253710	MT184911	MN988713
MT292572	MT276327	MT263386	MT253709	MT184912	MN938384
MT293185	MT276326	MT263387	MT253705	MT184913	MN975262
MT293187	MT263388	MT263428	MT253708	MT184910	MN985325
MT293188	MT263392	MT263429	MT253701	MT184907	MN908947
MT291826	MT262900	MT263433	MT253702	MT184908	
MT291832	MT262902	MT263435	MT253703	MT163718	
MT291833	MT262906	MT263437	MT253704	MT163719	
MT291835	MT262908	MT263438	MT253706	MT163717	

 doi:10.20944/preprints202004.0405.v1

2. Findings

Based on a thorough scanning over the set of genes of all the 470 genomes, the following observations are made:

- A pair of genes ORF7a and ORF8 of length 366 is present among almost all the SARS-CoV2 genomes except three genomes viz. MN988668 (China), MN988669 (China) and MT121215 (China). The loci of these two genes across the all the 467 genomes is gapped by 135bp.

- There are 251 genomes out of the 470 genomes that contain a gene ORF7b of length 132bp, located in between the gap-frame of 135bp as mentioned above. The presence of ORF7b over 251 genomes across different geo-locations is given in Table 2. It is worth observing from the reference [19] that the presence of ORF7b gene makes a CoV protein more virulent.

- There are 219 SARS-CoV2 genomes out of 470, which do not contain the ORF7b gene. The list of the genomes with their respective geographic locations, which do not contain ORF7b is given in the Table-3. The absence of ORF7b over 251 genomes across different countries is given in Table 2. It is to be noted that 55.25% of the SARS-CoV2 genomes from USA contain ORF7b gene which may be an indication of the strong pandemic situation in USA.

- The gene E of length 228 is present over all the 470 SARS-CoV2 genomes.

- The gene M of length 669 is present over all the 470 SARS-CoV2 genomes.

- The gene N of length 1260 is present over all the 470 SARS-CoV2 genomes.

- The spike gene S is a necessary structural gene and is present over all 470 genomes consisting of gene length 3822bp for 468 genomes and 3819bp for

only two genomes, one from India MT012098 and other one from France MT320538.

- There are only three genomes MN988668 (China), MN988669 (China), MT121215 (China) have only five genes (ORFs) viz. ORF1ab, S, E, M and N.
- There are 245 genomes having exactly 11 genes viz. ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, N and ORF10.
- There are 221 genomes having exactly 10 genes viz. ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF8, N and ORF10. It is noted that ORF7b is absent in this list of ten genes.
- There is only one genome MN938384 (China) consisting exactly nine genes viz. ORF1ab, S, ORF3, E, M, ORF6, ORF7, ORF8 and N.

Table 2: Presence and Absence of ORF7b over 470 SARS-CoV2 Genomes

Country	Absent of ORF7b	Present of ORF7b
USA	166	205
China	36	28
Italy	1	1
France	0	1
Iran	0	1
South Korea	1	3
Spain	4	7
Israel	0	2
Pakistan	0	2
Peru	0	1
Viet Nam	2	0
Taiwan	3	0
India	2	0
Brazil	1	0
Sweden	1	0
Nepal	1	0
Australia	1	0

From the Table-2, it is inferred that screening of ORF7b specific test may provide the rate of SARS-CoV2's infectivity and mortality in a country.

 doi:10.20944/preprints202004.0405.v1

Table 3: List of the genome accessions with their respective geographic location, which do not contain the gene ORF7b

Accession	Geo Location	Accession	Geo Location	Accession	Geo Location
MT262896	USA	MT251976	USA	MT159709	USA
MT262900	USA	MT251977	USA	MT159710	USA
MT262901	USA	MT251978	USA	MT159711	USA
MT262902	USA	MT251979	USA	MT159712	USA
MT262903	USA	MT251980	USA	MT159713	USA
MT262906	USA	MT253696	China	MT159714	USA
MT262907	USA	MT253697	China	MT159705	USA
MT262908	USA	MT253698	China	MT159706	USA
MT262909	USA	MT253699	China	MT159716	USA
MT262911	USA	MT251972	USA	MT159717	USA
MT262912	USA	MT251973	USA	MT159719	USA
MT262913	USA	MT251974	USA	MT159707	USA
MT262914	USA	MT251975	USA	MT121215	China
MT262897	USA	MT253700	China	MT066156	Italy
MT262898	USA	MT198652	Spain	MT012098	India
MT262899	USA	MT233526	USA	MT050493	India
MT262904	USA	MT246667	USA	MT152824	USA
MT262905	USA	MT246450	USA	MT135042	China
MT262910	USA	MT246452	USA	MT135041	China
MT262915	USA	MT246453	USA	MT135043	China
MT262916	USA	MT246454	USA	MT135044	China
MT259281	USA	MT246460	USA	MT126808	Brazil
MT259282	USA	MT246461	USA	MT123293	China
MT258377	USA	MT246462	USA	MT123290	China
MT258378	USA	MT246464	USA	MT123291	China
MT258379	USA	MT246470	USA	MT123292	China
MT259257	USA	MT246471	USA	MT118835	USA
MT259261	USA	MT246449	USA	MT106052	USA
MT259263	USA	MT246473	USA	MT106053	USA
MT259264	USA	MT246474	USA	MT106054	USA
MT259268	USA	MT246455	USA	MT093571	Sweden
MT259269	USA	MT246456	USA	MT072688	Nepal
MT259271	USA	MT246480	USA	MT066175	Taiwan
MT259273	USA	MT246457	USA	MT066176	Taiwan
MT259275	USA	MT246458	USA	MT044258	USA
MT259277	USA	MT246481	USA	MT044257	USA
MT259278	USA	MT246451	USA	MT039890	South Korea
MT259280	USA	MT246472	USA	MT039887	USA
MT258383	USA	MT246486	USA	MT039888	USA
MT259247	USA	MT246487	USA	MT039873	China
MT259248	USA	MT246488	USA	MT027062	USA
MT259236	USA	MT246489	USA	MT027063	USA
MT258380	USA	MT246485	USA	MT027064	USA
MT259235	USA	MT246459	USA	MT019529	China
MT259237	USA	MT246482	USA	MT020881	USA
MT259239	USA	MT246466	USA	MT019530	China
MT258381	USA	MT246467	USA	MT019531	China
MT259243	USA	MT246484	USA	MT019533	China
MT259249	USA	MT246468	USA	MT020880	USA
MT259250	USA	MT246490	USA	MT019532	China
MT259251	USA	MT246469	USA	MT007544	Australia
MT259256	USA	MT246475	USA	MN988668	China
MT259258	USA	MT246477	USA	MN988669	China
MT259266	USA	MT246478	USA	MN994467	USA
MT259267	USA	MT246479	USA	MN994468	USA
MT259274	USA	MT246476	USA	MN997409	USA
MT259286	USA	MT233519	Spain	MN988713	USA
MT259287	USA	MT233522	Spain	MN985325	USA
MT259241	USA	MT233523	Spain	MN975262	China
MT259260	USA	MT226610	China	MN938384	China
MT259284	USA	MT192773	Viet Nam	MN908947	China
MT259285	USA	MT192759	Taiwan		
MT258382	USA	MT192772	Viet Nam		
MT259246	USA	MT188341	USA		
MT259244	USA	MT188339	USA		
MT259245	USA	MT188340	USA		
MT259252	USA	MT184909	USA		
MT259253	USA	MT184911	USA		
MT259254	USA	MT184912	USA		
MT253701	China	MT184913	USA		
MT253702	China	MT184910	USA		
MT253703	China	MT184908	USA		
MT253704	China	MT184907	USA		
MT253706	China	MT159708	USA		
MT253707	China	MT159715	USA		
MT253710	China	MT159718	USA		
MT253705	China	MT159720	USA		
MT253708	China	MT159721	USA		
MT253709	China	MT159722	USA		

 doi:10.20944/preprints202004.0405.v1

References

[1] L. van der Hoek, K. Pyrc, M. F. Jebbink, W. Vermeulen-Oost, R. J. Berkhout, K. C. Wolthers, P. M. Wertheim-van Dillen, J. Kaandorp, J. Spaargaren, B. Berkhout, Identification of a new human coronavirus, Nature medicine 10 (4) (2004) 368–373.

[2] K. V. Holmes, Sars-associated coronavirus, New England Journal of Medicine 348 (20) (2003) 1948–1951.

[3] L. Mousavizadeh, S. Ghasemi, Genotype and phenotype of covid-19: Their roles in pathogenesis, Journal of Microbiology, Immunology and Infection (2020).

[4] P. Yang, X. Wang, Covid-19: a new challenge for human beings, Cellular & Molecular Immunology (2020) 1–3.

[5] C. Li, Y. Yang, L. Ren, Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species, Infection, Genetics and Evolution (2020) 104285.

[6] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, et al., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, The Lancet 395 (10224) (2020) 565–574.

[7] P. D. Yadav, V. A. Potdar, M. L. Choudhary, D. A. Nyayanit, M. Agrawal, S. M. Jadhav, T. D. Majumdar, A. Shete-Aich, A. Basu, P. Abraham, et al., Full-genome sequences of the first two sars-cov-2 viruses from india., The Indian Journal of Medical Research (2020).

[8] M. Lipsitch, D. L. Swerdlow, L. Finelli, Defining the epidemiology of covid-19—studies needed, New England Journal of Medicine (2020).

[9] A. S. Fauci, H. C. Lane, R. R. Redfield, Covid-19—navigating the uncharted (2020).

[10] W. Liu, Q. Zhang, J. Chen, R. Xiang, H. Song, S. Shu, L. Chen, L. Liang, J. Zhou, L. You, et al., Detection of covid-19 in children in early january 2020 in wuhan, china, New England Journal of Medicine (2020).

[11] F. Jiang, L. Deng, L. Zhang, Y. Cai, C. W. Cheung, Z. Xia, Review of the clinical characteristics of coronavirus disease 2019 (covid-19), Journal of General Internal Medicine (2020) 1–5.

[12] J. Stebbing, A. Phelan, I. Griffin, C. Tucker, O. Oechsle, D. Smith, P. Richardson, Covid-19: combining antiviral and anti-inflammatory treatments, The Lancet Infectious Diseases (2020).

[13] J. F.-W. Chan, C. C.-Y. Yip, K. K.-W. To, T. H.-C. Tang, S. C.-Y. Wong, K.-H. Leung, A. Y.-F. Fung, A. C.-K. Ng, Z. Zou, H.-W. Tsoi, et al., Improved molecular diagnosis of covid-19 by the novel, highly sensitive and specific covid-19-rdrp/hel real-time reverse transcription-polymerase chain reaction assay validated in vitro and with clinical specimens, Journal of Clinical Microbiology (2020).

[14] S. Zhang, M. Y. Diao, L. Duan, Z. Lin, D. Chen, The novel coronavirus (sars-cov-2) infections in china: prevention, control and challenges, Intensive Care Medicine (2020) 1–3.

[15] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, R. Agha, World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19), International Journal of Surgery (2020).

[16] P. Sun, X. Lu, C. Xu, W. Sun, B. Pan, Understanding of covid-19 based on current evidence, Journal of Medical Virology (2020).

[17] J. K. Das, P. P. Choudhury, A. Chaudhuri, S. S. Hassan, P. Basu, Analysis of purines and pyrimidines distribution over mirnas of human, gorilla, chimpanzee, mouse and rat, Scientific reports 8 (1) (2018) 1–19.

[18] J. P. Banerjee, J. K. Das, P. P. Choudhury, S. Mukherjee, S. S. Hassan, P. Basu, The variations of human mirnas and ising like base pairing models, BioRxiv (2018) 319301.

[19] S. R. Schaecher, J. M. Mackenzie, A. Pekosz, The orf7b protein of severe acute respiratory syndrome coronavirus (sars-cov) is expressed in virus-infected cells and incorporated into sars-cov particles, Journal of virology 81 (2) (2007) 718–731.

[20] A. Wu, Y. Peng, B. Huang, X. Ding, X. Wang, P. Niu, J. Meng, Z. Zhu, Z. Zhang, J. Wang, et al., Genome composition and divergence of the novel coronavirus (2019-ncov) originating in china, Cell host & microbe (2020).