2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

Discovering SARS-CoV-2 genes and mutations adapted for humans in 2594 genomes

Weitao Sun[1,2]
[1]*School of Aerospace Engineering, Tsinghua University*
[2]*Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University*
Beijing, China
sunwt@tsinghua.edu.cn

2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669705

***Abstract*—Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a positive-sense single-stranded virus approximately 30 kb in length, is the cause of the ongoing global life-threatening novel coronavirus disease-2019 (COVID-19) outbreak. Studies confirmed significant genome differences between SARS-CoV-2 and SARS-CoV, suggesting that the distinctions in pathogenicity and virulence might be related to genomic diversity. However, the relationship between genomic differences and SARS-CoV-2 fitness has not been fully explained, especially for open reading frame (ORF)-encoded accessory proteins. RNA viruses have a high mutation rate, but how SARS-CoV-2 mutations accelerate host adaptation is not clear. This study shows that the host-genome similarity (HGS) of SARS-CoV-2 is significantly higher than that of SARS-CoV, especially in the ORF6 and ORF8 genes that encode proteins antagonizing innate immunity *in vivo*. A power law relationship was discovered between the HGS of ORF3b, ORF6, and N and the expression of interferon (IFN)-sensitive response element (ISRE)-containing promoters. This finding implies that the increase in HGS in the SARS-CoV-2 genome may further inhibit IFN I synthesis and cause delayed host innate immunity. An ORF1ab mutation, 10818G>T, which occurred in virus populations with high HGS but rarely in low-HGS populations, was identified in 2594 genomes with geolocations of China, the USA and Europe. The genomic mutation caused the amino acid mutation M37F in the transmembrane protein nsp6. The results suggest that the ORF6 and ORF8 genes and the residue mutation M37F may play important roles in SARS-CoV-2 adaptation to humans. However, the underlying basis by which the mutations mediate adaptation to humans is still unknown. The findings demonstrate that HGS analysis is a reliable way to identify important genes and mutations in adaptive strains, which may help in the search for potential targets for pharmaceutical agents.**

Keywords*—SARS-CoV-2, open reading frame (ORF)-encoded proteins, host-genome similarity, genes mutations

I. Introduction

In December 2019, a novel coronavirus SARS-CoV-2 was reported as the cause of COVID-19. SARS-CoV-2 has a positive-sense single-stranded RNA with a length of approximately 30 kb[1]. Studies have shown that considerable genetic diversity exists between SARS-CoV-2 and SARS-CoV[2]. Compared with SARS-CoV, SARS-CoV-2 appears to be more contagious and more adapted to humans[3]. The distinctions in pathogenicity and virulence might be related to genomic diversity.

RNA viruses are susceptible to genetic recombination, and viral populations may evolve improved adaptability in the process of infecting hosts. By comparing the genome similarity of the virus to the host, the adaptability of the virus to the host can be inferred. Although the genomes of viruses and hosts are quite different in general, nucleotide sequence similarities do exist. Such similarities may have three biological significances. (1) These similar fragments come from a common ancestor and remain stable over long-term evolution due to their biological significance. (2) Similar genomic fragments are coincidentally preserved in both viruses and hosts over time because of the biological benefits of the gene products. (3) When the virus interacts with the hosts, mutants are created by virus-host gene exchanges, causing genome similarities.

A growing number of studies on virus-host gene similarity have been reported. Simian virus 40 (SV40), the first animal virus to undergo complete full-sequence DNA analysis, can infect monkeys and humans and cause tumors[4]. Rosenberg et al.[5] found that some mutant SV40 viruses contained nucleic acid sequences from their host monkeys. This finding suggests that viruses can recombine with host genes to complete their own physiological processes, which makes up for a lack of function or increases virulence. Genes similar to specific fragments of the human genome in molluscum contagiosum virus (MCV) have been reported[6]. MCV is a human poxvirus and lacks the genes associated with virus-host interactions in other poxvirus species (variola virus). However, genes in MCV with high similarity to specific fragments of the human genome are also hard to find in other poxviruses. These host-like genes may provide MCV-specific strategies for coexistence with the host[6]. In other words, it is very likely that viruses use host-specific genes to perform activities related to virus-host interactions, such as evasion of the host innate immune system. When human peripheral blood DNA was used as a template for polymerase chain reaction (PCR), 5 of 6 samples could be amplified by Epstein-Barr virus (EBV)- or hepatitis C virus (HCV)-specific primers[7]. Therefore, it is speculated that some genes of the two viruses may also exist in the human genome or that the viruses may have homology with human genes. This hypothesis implies that not only can the virus have the host's genes but also the host itself may have genes from the virus.

Selection pressure exerted by the host immune system plays an important role in shaping virus mutations. Homology between virus and host proteins indicates the presence of host gene capture. Evolution of viral genes may involve intergenome gene transfer and intragenome gene duplication[8]. By acquiring immune modulation genes from cells, viruses have evolved proteins that can regulate or inhibit the host's immune system[9, 10]. A recent study showed that human genome evolution was shaped by viral infections[11]. In mammals, nearly 30% of the adaptive amino acid changes in the human proteome are caused by viruses, suggesting that viruses are one of the major driving factors for the evolution of mammalian and human proteomes[12]. These findings support the possibility that SARS-CoV-2 may exchange

genetic information with host cells. It can be inferred that most of the traits and mechanisms retained in "coevolution" between viruses and their hosts, including genetic and mutational mechanisms, benefit at least one or both. At the molecular level of evolution, the exchange of genetic information is necessary for virus-host mutual adaptation, leading to the similarity of nucleotide sequences.

It is interesting to study the relationship between gene similarities and viral transmission/pathological ability. The single-stranded RNA of coronavirus generally encodes three categories of proteins: (1) the replication proteins open reading frame (ORF)1a and ORF1ab; (2) the structural proteins S (spike), E (envelope), M (membrane) and N (nucleocapsid); and (3) accessory proteins with unknown homologues. The structural protein genes are organized as '-S-E-M-N-' in the SARS-CoV-2 genome, and accessory protein genes are distributed between S and E, M and N.

The accessory protein genes play a key role in inhibiting the innate immune response *in vivo* and are more susceptible than the other genes to species-specific mutations under the pressure of evolutionary selection. Once inside the cell, the virus immediately confronts other critical proteins known as host-restriction factors (HRFs)[13]. HRFs are proteins that recognize and block viral replication. Virus-host interactions control species specificity and viral infection ability. Under pressure from the host immune system, viruses must be able to overcome a range of constraints associated with the host species and often show evolutionary mutation selections. It is hypothesized that accessory ORFs may retain beneficial mutations to increase host-genome similarity (HGS). Identifying emerging genetic mutations in virus populations with high HGS may aid the understanding of how SARS-CoV-2 evolved adaptation to humans. To the best of our knowledge, studies on the genetic similarity between SARS-CoV-2 and the human genome have not been reported.

This study investigated the HGS of SARS-CoV-2 genes and elucidated the links between HGS and virus adaptation to humans. A power law relationship was discovered between the expression of genes with interferon (IFN)-stimulated response elements (ISREs) and HGS. ORFs with higher HGS suppressed the gene expression of ISRE-regulated genes to a greater extent. Applying HGS analysis to 2594 SARS-CoV-2 genomes from China, the USA and Europe, it was found that the ORF6 and ORF8 genes of SARS-CoV-2 had more significant HGS increments than SARS-CoV. In addition, three different sets of surviving mutations were identified in SARS-CoV-2 genomes for China, the USA and Europe. Interestingly, an ORF1ab mutation, 10818G>T, which resulted in the residue mutation M37F in the transmembrane protein nsp6, was observed in virus populations of all three regions. This mutation did not occur in strain populations with low HGS but gradually appeared in populations with high HGS. This finding provides strong evidence that SARS-CoV-2 may accelerate adaptation in humans through increasing HGS of the ORF6 and ORF8 genes and selecting the M37F mutation. However, the underlying mechanism by which these genes and mutations make SARS-CoV-2 more adapted to humans remains unclear.

II. Results

A. *SARS-CoV-2 have higher HGS than those of SARS-CoV*

The SARS-CoV-2 (GenBank: MN908947.3) and SARS-CoV (GenBank: AY394850.2) RNA sequences were used as references to establish the genome organization. SARS-CoV-2 has 14 5'-ORFs, while SARS-CoV has 19 5'-ORFs. The length of each ORF is no less than 75 nt. A quantitative definition of HGS was proposed to investigate the similarity between viral coding sequences (CDSs) and the human genome (*Homo sapiens* GRCh38.p12 chromosomes). The CDS alignment scores were determined by using NCBI Blastn[14], and HGS was calculated by the formulas described in the Methods for each ORF in the coronavirus genome. The overall HGS of a full-length virus genome was obtained by the weighted sum of ORF HGSs. The weighting factor was the ratio of ORF length to the full-genome length.

The HGS of ORFs was calculated for 2594 SARS-CoV-2 genomes with geolocation from China, the USA and Europe. Phylogenetic trees representing the HGS relationship among virus strains are shown in Extended Data Fig. E1, E2, and E3 for all three regions. The tree clusters were formed based on the distance between vectors containing ORF HGS values. Most of the genomes had moderate HGS values. Genomes with similar HGS values were usually in the same cluster and shared a common ancestor. The genomes with high HGS were not all concentrated in the same cluster but may form several separate populations in the tree.

The full-length genome data were obtained from the Global Initiative on Sharing All Influenza Data (GISAID) database[15]. The sequence requirements were full-length sequences only, sequences with definite collection dates and locations, and no nucleotide names other than A, G, C and T. The number of genomes that met such requirements was 200 for China, 1538 for the USA and 856 for Europe at the time of article preparation. The HGS of human SARS-CoV genomes was also calculated. In NCBI GenBank[16], a total of 25 SARS-CoV CDSs met the above sequence requirements.

Fig. 1 shows that ORF 7b of SARS-CoV had the highest similarity with human genome, followed by ORF6, ORF7a, ORF3a and ORF 8. For SARS-CoV-2, ORF 7b, ORF 6 and ORF 8 were the top 3 genes with the highest HGSs. The mean HGS values of ORF6 and ORF8 in SARS-CoV-2 increased significantly, reaching 122% and 148% of those of SARS-CoV ORF6 and ORF8, respectively (Fig. 1). The roles of the HGS changes are not clear. However, by investigating the function of the SARS-CoV genes and proteins, the mechanism of the rapid spread of the new emerged COVID-19 may be inferred from the HGS changes in SARS-CoV-2 genomes.

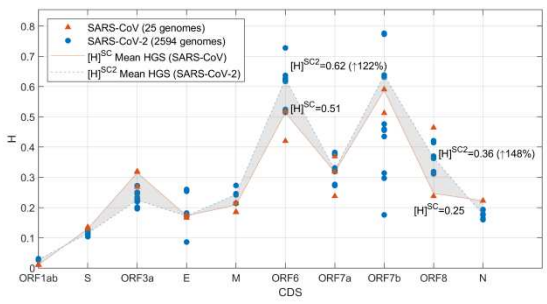

Fig. 1. The HGS values of SARS-CoV-2 and SARS-CoV genes. ORF6 and ORF8 of SARS-CoV-2 have apparently higher mean HGS values than those of SARS-CoV, reaching 122% and 148% of that of SARS-CoV ORF6 and ORF8, respectively.

Studies have shown that ORF6 suppresses the induction of IFN and signaling pathways[17]. A membrane protein with

63 amino acids, ORF 6 blocked the IFNAR-STAT signaling pathway by limiting the mobility of the importin subunit KPNB1 and preventing the STAT1 complex from moving into the nucleus for ISRE activation[18]. Laboratory studies confirmed that the expression of ORF 6 transformed a sublethal infection into lethal encephalitis and enhanced the growth of the virus in cells[19]. In addition, ORF 6 circumvented IFN production by inhibiting IRF-3 phosphorylation in the (TRAF3)-(TBK1+IKKε)-(IRF3)-(IFNβ) signaling pathway (Extended Data Fig. E4), which is an essential signaling pathway triggered by the viral sensors RIG-1/MDA5 and TLRs[20].

An intact gene, ORF8 encodes a single accessory protein at the early stage of SARS-CoV infection and splits into two fragments, ORF8a and ORF8b, at later stages[21]. ORF8a and 8b have been observed in most SARS-CoV-infected cells[22]. Wong et al.[23] found that the proteins ORF8b and ORF8ab in SARS-CoV inhibited the IFN response during viral infection. It was also reported that ORF8b formed insoluble intracellular aggregates and triggered cell death[24]. Amazingly, studies showed that SARS-CoV-related CoVs in horseshoe bats had 95% genome identities to human and civet SARS-CoVs, but the ORF8 protein amino acid similarities varied from 32% to 81%[25]. These findings indicate that the ORF8 gene is more prone than other CoV genes to mutations in virus-host interactions. Overexpression of ORF 8b and ORF 8ab had a significant effect on IRF3 dimerization rather than IRF3 phosphorylation[23]. The 8b region of SARS-CoV protein ORF8 functions in ubiquitination binding, ubiquitination and glycosylation, which may interact with IRF3[26]. The expression of ORF8b and 8ab enhanced IRF3 degradation, thus regulating the immune functions of IRF3 (Extended Data Fig. E4). Interestingly, ORF8 is an IFN antagonist expressed in the later stage of SARS-CoV infection. Studies showed that activation of IRF3 was blocked in the late stage of SARS-CoV infection, which was consistent with the late expression of ORF8b. Therefore, the expression of ORF8 may help to suppress the innate immune response that occurs in the later stages of infection and delay IFNβ signaling. This may explain why the virus expresses a late-stage IFN antagonist, such as ORF8.

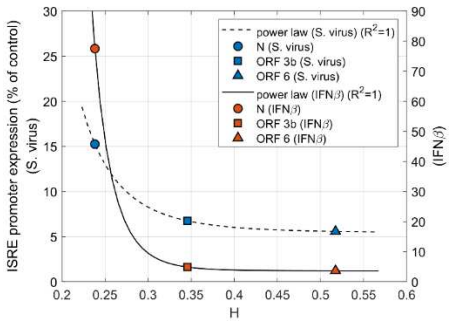

Fig. 2. Inhibition of a promoter containing an ISRE by SARS-CoV proteins with different genome HGS values. Cells were cotransfected with the SARS-CoV proteins and either infected with Sendai virus (S. virus) or treated with IFNβ after 24 hours. The expression of the promoter decays rapidly with the increasing HGS of ORF 3b, ORF 6 and N, conforming to a power law.

This work found that genes with high HGS were critical in suppressing innate immunity. Studies have shown that the ORF3b, ORF6 and N proteins of SARS-CoV enhance suppression of IFNβ expression in host innate immunity[27]. When IFN binds to the cell receptor IFNAR, the JAK/STAT signaling pathway is activated, leading to activation of IFN-stimulated genes (ISGs) containing an ISRE in their promoter. Expression of genes with an ISRE will trigger the production of hundreds of antiviral proteins inhibiting viral infections. Therefore, a reduction in expression from ISRE-containing promoters is a direct indicator of the enhanced ability to inhibit IFN synthesis.

ISRE-containing promoter expression after Sendai virus infection needs both IFN synthesis and signaling. However, ISRE-containing promoter expression after IFNβ treatment requires only IFN signaling. In cells treated with IFNβ, it was found that N did not significantly inhibit the expression of the ISRE promoter[17]. The expression level was approximately 78% of the value for the empty control. However, ORF3b and ORF6 still inhibited the expression of the ISRE promoter. We calculated the HGSs of ORF3b, ORF6 and N for SARS-CoV. Amazingly, the results clearly demonstrated that the ISRE-containing promoter expression decreased rapidly with increasing HGS (Fig. 2), which provided evidence that there was a power law dependence of IFN synthesis inhibition based on HGS. The ISRE-containing promoter expression data followed the work of Kopecky-Bromberg et al.[17]. For 293T cells transfected with the SARS-CoV proteins and infected by Sendai virus[17], IFN inhibition obeys the following power law equation: $P = 0.004H^{-0.539} + 5.421$, where H is the HGS value of the viral genes ORF3b, ORF6 and N, and P is the expression of genes with an ISRE as a percentage of the value for the empty control. The power law equation for cells treated with IFNβ is $P = 0.00001H^{-11.007} + 3.633$. The coefficient of determination R^2 reaches 1 for both data sets, indicating a perfect fit for the power law dependence on HGS.

The findings suggested that HGS, i.e., similarity between the virus and host genome, is a reliable indicator of the suppression of innate immunity by viral proteins. Channappanavar et al. found that rapid SARS-CoV replication and a relative delay in IFN I signaling resulted in immune dysregulation and severe disease in infected mice[28]. Considering the significant HGS increments of ORF6 and ORF8 and their roles in suppressing innate immunity, it could be speculated that SARS-CoV-2 would further suppress IFN I synthesis and delay host innate immunity as HGS increases. This hypothesis may explain the delayed immune response and uncontrolled inflammatory response that lead to the epidemiological manifestations of SARS-CoV-2, such as long incubation periods, mild symptoms, rapid spread and low mortality. However, the mechanism of how viral proteins cause further delay of immune signaling and how it leads to new immunopathological features remain largely unknown.

The discovery of increased HGS of ORF 6 and ORF 8 provides strong evidence that SARS-CoV-2 evolved to be more adapted to humans than SARS-CoV. These inferences offer a valuable picture of how SARS-CoV-2 could have become different from SARS-CoV. In addition, genetic mutations making the virus genome adapted to humans can also be identified through HGS analysis.

B. The SARS-CoV-2 mutation 10818G>T is adapted to humans

Recent studies have shown that SARS-CoV-2 had a high mutation rate, and new mutations have emerged in ORF1ab, S, ORF3a and ORF8[29]. However, the types of mutations that contribute to viral adaptations in humans are not clear.

To understand how mutations aid survival of SARS-CoV-2 populations under selective pressure, the accumulated nucleotide variants in consensus sequences were identified in 2594 genomes from China, the USA and Europe. The virus genome was identified by its HGS values of ten ORFs (ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, and N). The percentages of virus strains with unique ORF HGSs were 18% (36 out of 200), 9% (140 out of 1538) and 11% (98 out of 856) for genomes with geolocations of China, the USA and Europe, respectively. A total of 74 mutations, 162 mutations and 145 mutations were identified in genomes for these three regions, respectively. Gene mutation profiles of SARS-CoV-2 genomes with different HGSs are shown in Extended Data Fig. E5, E6 and E7. SARS-CoV-2 in different regions developed its own conserved mutations independently. For example, the mutations in genomes with a geolocation of China included the ORF1ab mutations 10818G>T (TTG>TTT), 1132G>A (GTA>ATA), and 8517C>T (AGC>AGT); ORF8 mutation 251T>C (TTA>TCA); N mutation 415T>C (TTG>CTG); S mutation 1868A>G (GAT>GGT); and ORF3a mutation 752G>T (GGT>GTT). Here, the number before the mutated nucleotide represents the sequence position relative to the starting point of the ORF where the mutation is located.

Of all the gene mutations, the ORF1ab 10818G>T(TTG>TTT) mutation is the most interesting. This mutation survived in all three regions (Fig. 3). In addition, this mutation occurred only in the high HGS population rather than in that with a lower HGS. The SARS-CoV-2 ORF1ab gene encodes the precursor polyprotein pp1ab, which is then cleaved into 16 nonstructural proteins (nsp1 to nsp16) by virus-encoded proteinases. nsp6 plays a critical role in membrane anchoring of the RNA replication/transcription complex. The expression of the nonstructural protein nsp6 along with nsp3 and nsp4 mediates the formation of double-membrane vesicles (DMVs)[30], which are organelle-like structures for viral genome replication and protect against host cell defenses.

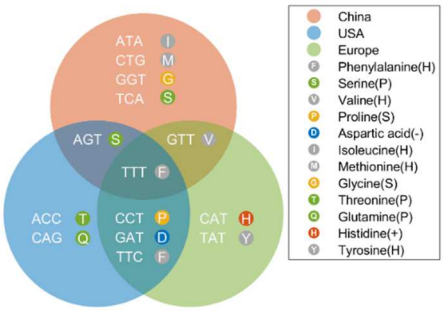

Fig. 3. Highly conserved mutations identified in SARS-CoV-2 genomes with geolocations of China, the USA and Europe. The three regions have different sets of mutations. The TTT (F, Phenylalanine) mutation occurred in all three regions. TTT represents the mutation 10818G>T(TTG>TTT) in ORF1ab. The F in the circle represents the amino acid mutation M37F (Methionine to Phenylalanine) in nonstructural protein nsp6. The P, H, +, - and S in brackets in the legend represent polar, hydrophobic, positively charged, negatively charged and special residues, respectively.

Studies on the nsp6 protein showed that the protein is a transmembrane protein with 6 transmembrane regions[31]. This 10818G>T ORF1ab mutation caused an amino acid mutation, M37F, in the nonstructural protein nsp6, which is located in a loop between the first and second transmembrane domains on the N-terminal side (Fig. 4). This finding strongly suggested that the 10818G>T (M37F) mutation survived a selection event and resulted in a new population of SARS-CoV-2 with high HGS, which could be more adapted to humans. In addition, the simultaneous occurrence of ORF1ab 10818G>T in all three regions demonstrated that the mutation was highly stable in human-adapted strains. Although mutations in the nonstructural proteins nsp4 and nsp6 may affect the assembly of DMVs and viral autophagy, the underlying basis of how the M37F mutation results in SARS-CoV-2 adaptation in humans is not clear.

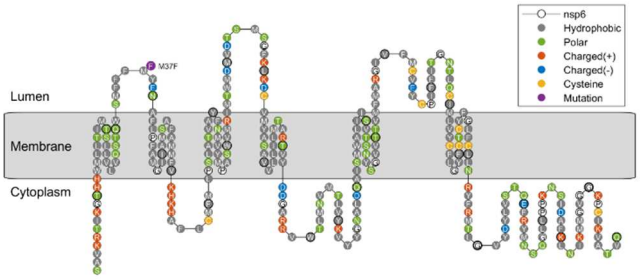

Fig. 4. The topology of transmembrane protein nsp6 and the identified M37F mutation located in a loop between the first and second transmembrane domains on the N-terminal side.

The identification of conserved mutations demonstrates that SARS-CoV-2 can improve host adaptation. It is reasonable to hypothesize that high HGS in SARS-CoV-2 genomes and conserved mutations may explain the epidemiological characteristics of COVID-19, such as mild symptoms, rapid spread and low mortality. However, the mechanism behind the impairment remains poorly understood and calls for future laboratory investigations.

III. Methods

By using BLAST ORFfinder[32], 31 open reading frames (ORF) were detected in RNA genome sequence (29903 nt) of SARS-CoV-2 (GenBank: MN908947.3). Only ATG was used as ORF start codon and nested ORFs were ignored. Among all the ORFs, we selected the top 14 longest ones as targets, whose lengths were no less than 75 nt. For genome comparison, ORFs in SARS-CoV genome with a length of 29728 nt (GenBank: AY394850.2) were also identified. There were totally 19 ORFs with length no less than 75 nt in SARS-nCoV sequence.

The SARS-CoV-2 genomes were obtained from Global Initiative on Sharing All Influenza Data (GISAID) database [15]. By May 20, 2020, the GISAID database (https://www.gisaid.org/) had 416 SARS-CoV-2 genomes with location as China, 5184 genomes with location as USA and 10954 genomes with location as Europe. Complete and high coverage genome were used to ensures accurate HGS calculations. The sequences contain nucleotide other than A, G, C and T were removed from the dataset. Totally 2594 SARS-CoV-2 genomes were used in current study, including 200 from China, 1538 from USA and 856 from Europe. The coding sequences (CDS) of SARS-CoV-2 genome were identified by using Matlab (www.mathworks.com/help/bioinfo/ref/seqshoworfs.html).

Human SARS-CoV genomes were collected from NCBI Genebank[16]. There were 25 CDS sequences of SARS-CoV isolated in mainland China (full-length only, sequences with definite collection date and location, no nucleotide names other than A, G, C and T) by the time of preparing this article.

The accession id of these viral sequences can be found in Supplemental Information.

The target CDS sequences are aligned with the human genome (Homo sapiens GRCh38.p12 chromosomes) by Blastn[14] to obtain the matching fragment. Blastn sequence alignment gives an original score of S. In order to facilitate the comparison of Blast results among different subgenomic groups, the original score is standardized to S' by Blastn $S' = (\lambda S - \ln K)/\ln 2$, $E = mn2^{-S'}$ .Here E value represents the expected number of times when two random sequences of length m and n are matched and the score is not lower than S'. Parameters K and λ describe the statistical significance of the results[33]. Assuming that the fragment of length a matches perfectly in the two random sequences, one has following formula $E = (m-a)(n-a)4^{-a}$ .Since the viral genome is quite different from the human genome, matching fragments are usually very short. When a is particularly small compared to m and n , $a = S'/2$ is obtained by combining Equation (3) and Equation (4). Thus, host-genome-similarity (HGS) is defined as $H = (\sum a)/n = (\sum S')/2n$, where n represents the length of the target sequence. The meaning of H is the ratio of the number of matched base pairs to the total length of the sequence when the matched sequences are converted into sequences of the same length.

REFERENCES

[1] F. Wu *et al.*, "A new coronavirus associated with human respiratory disease in China," (in eng), *Nature,* vol. 579, no. 7798, pp. 265-269, Mar 2020.

[2] X. Xu *et al.*, "Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission," *SCIENCE CHINA Life Sciences,* 2020.

[3] X. He *et al.*, "Temporal dynamics in viral shedding and transmissibility of COVID-19," *Nature Medicine,* vol. 26, no. 5, pp. 672-675, 2020/05/01 2020.

[4] W. Fiers *et al.*, "Complete nucleotide sequence of SV40 DNA," *Nature,* vol. 273, no. 5658, pp. 113-120, 1978/05/01 1978.

[5] M. Rosenberg, S. Segal, E. L. Kuff, and M. F. Singer, "The nucleotide sequence of repetitive monkey DNA found in defective simian virus 40," *Cell,* vol. 11, no. 4, pp. 845-857, 1977/08/01/ 1977.

[6] T. G. Senkevich, J. J. Bugert, J. R. Sisler, E. V. Koonin, G. Darai, and B. Moss, "Genome sequence of a human tumorigenic poxvirus: prediction of specific host response-evasion genes," (in eng), *Science,* vol. 273, no. 5276, pp. 813-6, Aug 9 1996.

[7] Y. Chang, J. Ma, M. Zhang, and Y. Yu, "Prel iminary study on genome homology of viruses and human," *J. N. BETHUNE UNIV. MED. SCI.,* vol. 23, no. 3, pp. 242-244, 1997.

[8] L. A. Shackelton and E. C. Holmes, "The evolution of large DNA viruses: combining genomic information of viruses and their hosts," *Trends in Microbiology,* vol. 12, no. 10, pp. 458-465, 2004/10/01/ 2004.

[9] B. T. Rouse and S. Sehrawat, "Immunity and immunopathology to viruses: what decides the outcome?," *Nature Reviews Immunology,* vol. 10, no. 7, pp. 514-526, 2010/07/01 2010.

[10] L. Van Kaer and S. Joyce, "Viral evasion of antigen presentation: not just for peptides anymore," *Nature Immunology,* vol. 7, no. 8, pp. 795-797, 2006/08/01 2006.

[11] D. Enard and D. A. Petrov, "Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans," (in eng), *Cell,* vol. 175, no. 2, pp. 360-371.e13, 2018.

[12] D. Enard, L. Cai, C. Gwennap, and D. A. Petrov, "Viruses are a dominant driver of protein adaptation in mammals," (in eng), *eLife,* vol. 5, p. e12469, 2016.

[13] S. Rothenburg and G. Brennan, "Species-Specific Host–Virus Interactions: Implications for Viral Host Range and Virulence," *Trends in Microbiology,* vol. 28, no. 1, pp. 46-56, 2020/01/01/ 2020.

[14] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology,* vol. 215, no. 3, pp. 403-410, 1990/10/05/ 1990.

[15] S. Elbe and G. Buckland-Merrett, "Data, disease and diplomacy: GISAID's innovative contribution to global health," *Global Challenges,* vol. 1, no. 1, pp. 33-46, 2017/01/01 2017.

[16] D. A. Benson *et al.*, "GenBank," (in eng), *Nucleic Acids Res,* vol. 41, no. Database issue, pp. D36-42, Jan 2013.

[17] M. Frieman, B. Yount, M. Heise, S. A. Kopecky-Bromberg, P. Palese, and R. S. Baric, "Severe Acute Respiratory Syndrome Coronavirus ORF6 Antagonizes STAT1 Function by Sequestering Nuclear Import Factors on the Rough Endoplasmic Reticulum/Golgi Membrane," *Journal of Virology,* vol. 81, no. 18, p. 9812, 2007.

[18] A. L. Totura and R. S. Baric, "SARS coronavirus pathogenesis: host innate immune responses and viral antagonism of interferon," *Current Opinion in Virology,* vol. 2, no. 3, pp. 264-275, 2012/06/01/ 2012.

[19] L. Pewe *et al.*, "A SARS-CoV-specific protein enhances virulence of an attenuated strain of mouse hepatitis virus," (in eng), *Adv Exp Med Biol,* vol. 581, pp. 493-8, 2006.

[20] T. L. Chau *et al.*, "Are the IKKs and IKK-related kinases TBK1 and IKK-epsilon similarly activated?," (in eng), *Trends Biochem Sci,* vol. 33, no. 4, pp. 171-80, Apr 2008.

[21] Y. Guan *et al.*, "Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China," (in eng), *Science,* vol. 302, no. 5643, pp. 276-8, Oct 10 2003.

[22] C.-T. Keng *et al.*, "The human severe acute respiratory syndrome coronavirus (SARS-CoV) 8b protein is distinct from its counterpart in animal SARS-CoV and down-regulates the expression of the envelope protein in infected cells," (in en), *Virology,* vol. 354, no. 1, pp. 132 - 142, 2006 2006.

[23] H. H. Wong, T. S. Fung, S. Fang, M. Huang, M. T. Le, and D. X. Liu, "Accessory proteins 8b and 8ab of severe acute respiratory syndrome coronavirus suppress the interferon signaling pathway by mediating ubiquitin-dependent rapid degradation of interferon regulatory factor 3," *Virology,* vol. 515, pp. 165-175, 2018/02/01/ 2018.

[24] C.-S. Shi, N. R. Nabar, N.-N. Huang, and J. H. Kehrl, "SARS-Coronavirus Open Reading Frame-8b triggers intracellular stress pathways and activates NLRP3 inflammasomes," (in eng), *Cell death discovery,* vol. 5, pp. 101-101, 2019.

[25] S. K. P. Lau *et al.*, "Severe Acute Respiratory Syndrome (SARS) Coronavirus ORF8 Protein Is Acquired from SARS-Related Coronavirus from Greater Horseshoe Bats through Recombination," (in eng), *Journal of virology,* vol. 89, no. 20, pp. 10532-10547, 2015.

[26] T. M. Le *et al.*, "Expression, post-translational modification and biochemical characterization of proteins encoded by subgenomic mRNA8 of the severe acute respiratory syndrome coronavirus," *The FEBS Journal,* vol. 274, no. 16, pp. 4211-4222, 2007/08/01 2007.

[27] S. A. Kopecky-Bromberg, L. Martínez-Sobrido, M. Frieman, R. A. Baric, and P. Palese, "Severe Acute Respiratory Syndrome Coronavirus Open Reading Frame (ORF) 3b, ORF 6, and Nucleocapsid Proteins Function as Interferon Antagonists," *Journal of Virology,* vol. 81, no. 2, p. 548, 2007.

[28] R. Channappanavar *et al.*, "Dysregulated Type I Interferon and Inflammatory Monocyte-Macrophage Responses Cause Lethal Pneumonia in SARS-CoV-Infected Mice," *Cell Host & Microbe,* vol. 19, no. 2, pp. 181-193, 2016/02/10/ 2016.

[29] C. Wang *et al.*, "The establishment of reference sequence for SARS-CoV-2 and variation analysis," *Journal of Medical Virology,* vol. 92, no. 6, pp. 667-674, 2020/06/01 2020.

[30] M. Pachetti *et al.*, "Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant," (in eng), *J Transl Med,* vol. 18, no. 1, p. 179, Apr 22 2020.

[31] S. Baliji, S. A. Cammer, B. Sobral, and S. C. Baker, "Detection of nonstructural protein 6 in murine coronavirus-infected cells and analysis of the transmembrane topology by using bioinformatics and molecular approaches," (in eng), *Journal of virology,* vol. 83, no. 13, pp. 6957-6962, 2009.

[32] M. Oostra *et al.*, "Topology and Membrane Anchoring of the Coronavirus Replication Complex: Not All Hydrophobic Domains of nsp3 and nsp6 Are Membrane Spanning," *Journal of Virology,* vol. 82, no. 24, p. 12392, 2008.

IV. Extended Data

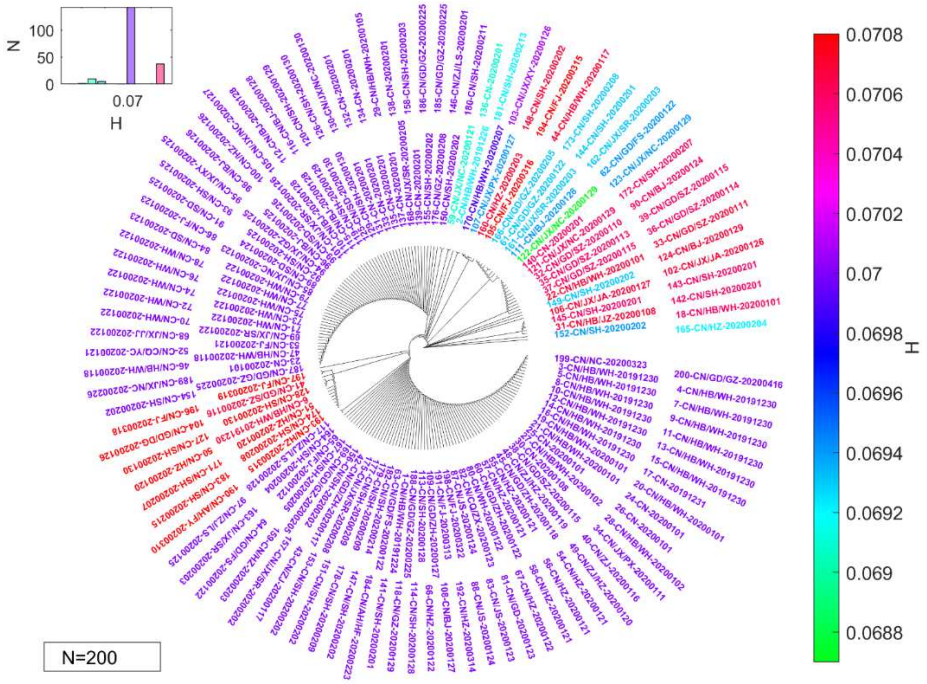

Fig. E5. The HGS tree contains 200 SARS-CoV-2 genomes from China. Distance between leaves is the unweighted pair distance between the 10-ORF-HGS vector of genomes. The colorbar represents the overall HGS value of each genome (weighted sum of ORF HGSs). Out of a total of 200 viral genomes, 36 have unique ORF HGS values. The histogram at the top left shows the distribution of all genome HGSs.

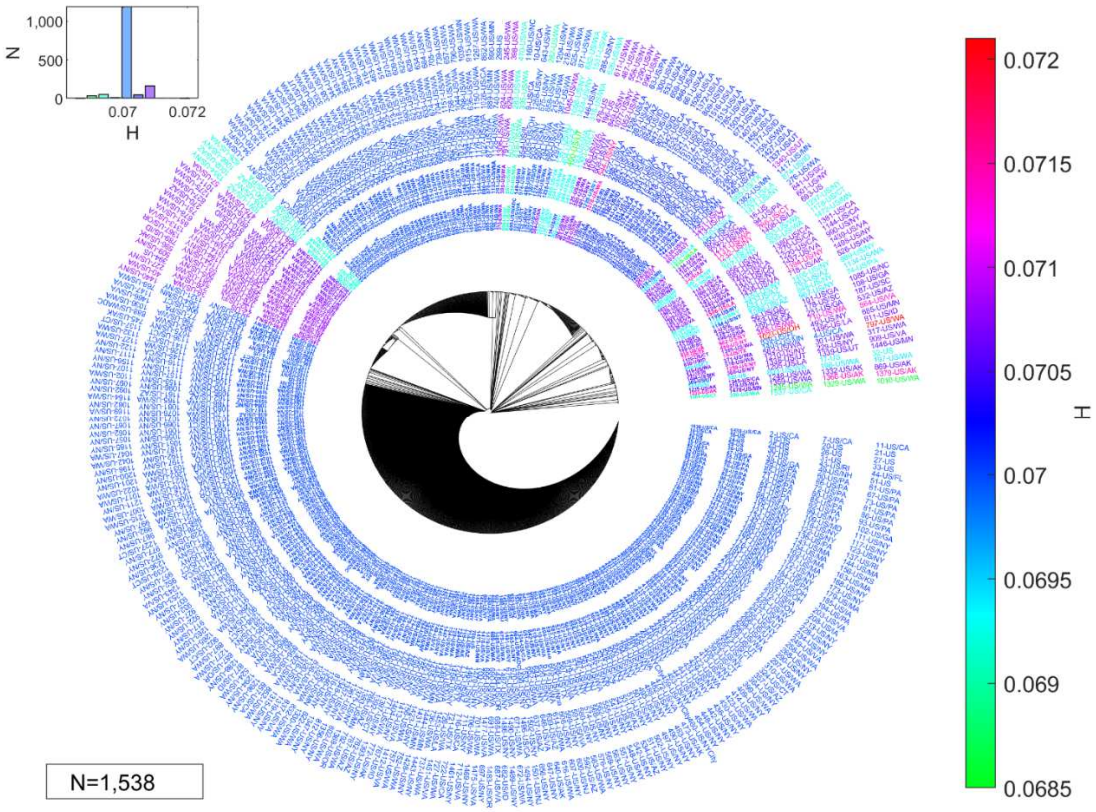

Fig. E6. The HGS tree contains 1538 SARS-CoV-2 genomes from the USA. Distance between leaves is the unweighted pair distance between the 10-ORF-HGS vector of genomes. The colorbar represents the overall HGS value of each genome (weighted sum of ORF HGSs). Out of a total of 1538 viral genomes, 140 have unique ORF HGS values. The histogram at the top left shows the distribution of all genome HGSs.

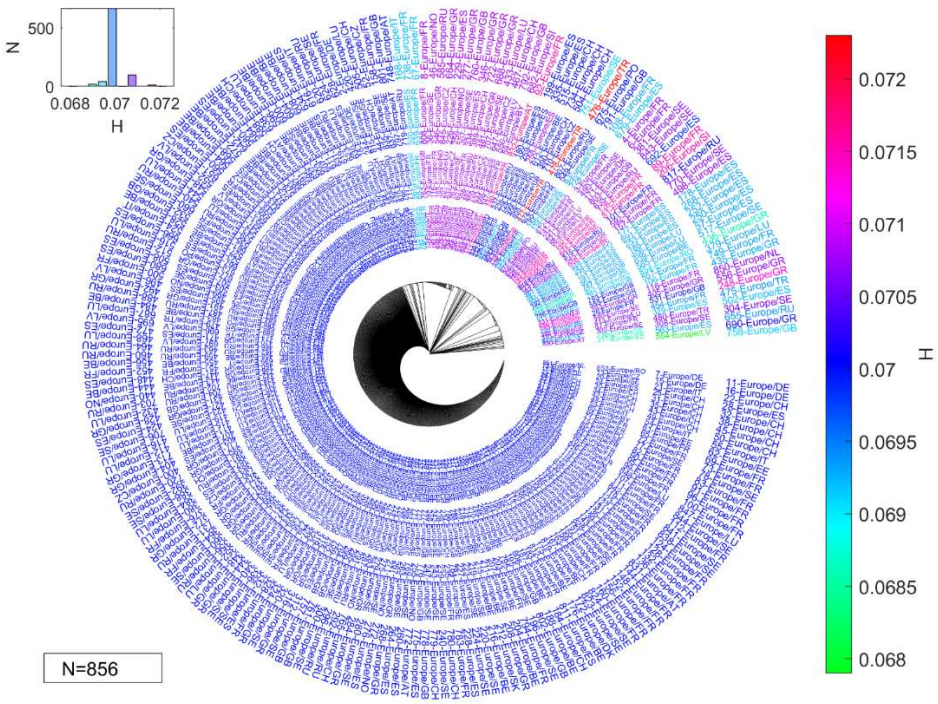

Fig. E7 The HGS tree contains 856 SARS-CoV-2 genomes from Europe. Distance between leaves is the unweighted pair distance between the 10-ORF-HGS vector of genomes. The colorbar represents the overall HGS value of each genome (weighted sum of ORF HGSs). Out of a total of 856 viral genomes, 98 have unique ORF HGS values. The histogram at the top left shows the distribution of all genome HGSs.

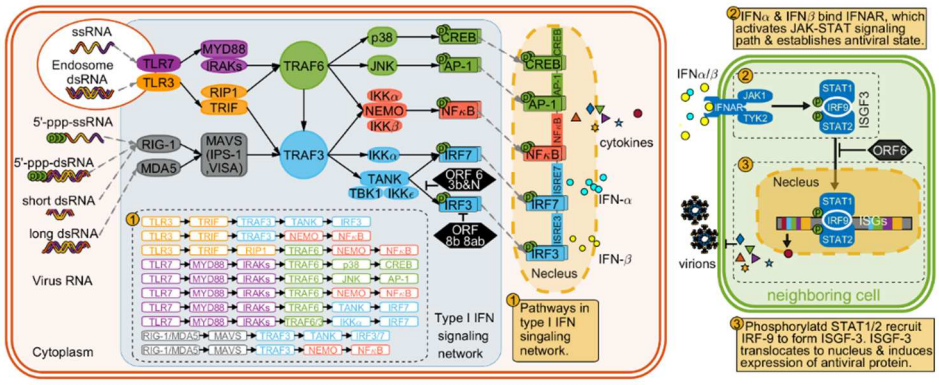

Fig. E8 SARS-CoV induced immune response in host cells. Host cell detect virus invasion mainly by TLPs and RIG1/MDA5 and lead to type I IFN signaling pathway. The receptor IFNAR senses type I IFN and leads to the JAK1-STAT signaling pathway, which expresses antiviral proteins and bring neighboring cell into anti-virus state. The ORF6 suppresses type I IFN expression by inhibiting translocation of STAT1+STAT2+IRF9 complex into nucleus. ORF 6 also circumvent IFN production by inhibit IRF-3 phosphorylation in signaling pathway (TRAF3)-(TBK1+IKKε)-(IRF3)-(IFNβ). The expression of ORF8b and 8ab enhance the IRF3 degradation, thus regulating immune functions of IRF3.

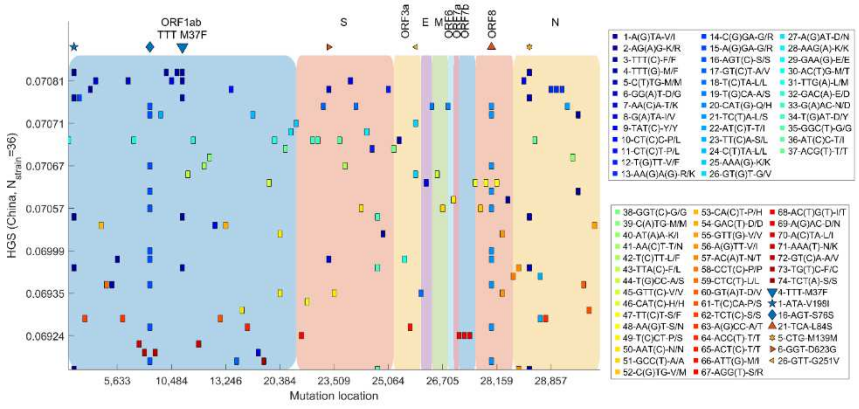

Fig. E9 Mutation profile for SARS-CoV-2 genomes (geolocation of China) with different HGS. Out of a total of 200 viral genomes, 36 genomes have unique HGS values. A total of 74 mutations were identified in all the genomes. The top 7 conserved mutations with were shown with special markers at the top of colored blocks representing ORFs. Mutation 10818G>T in ORF1ab (codon TTG>TTT) occurred in populations with high HGS, which results in amino acid M37F mutation in transmembrane protein nsp6. The mutation rarely occurred in populations with low/moderate HGS.

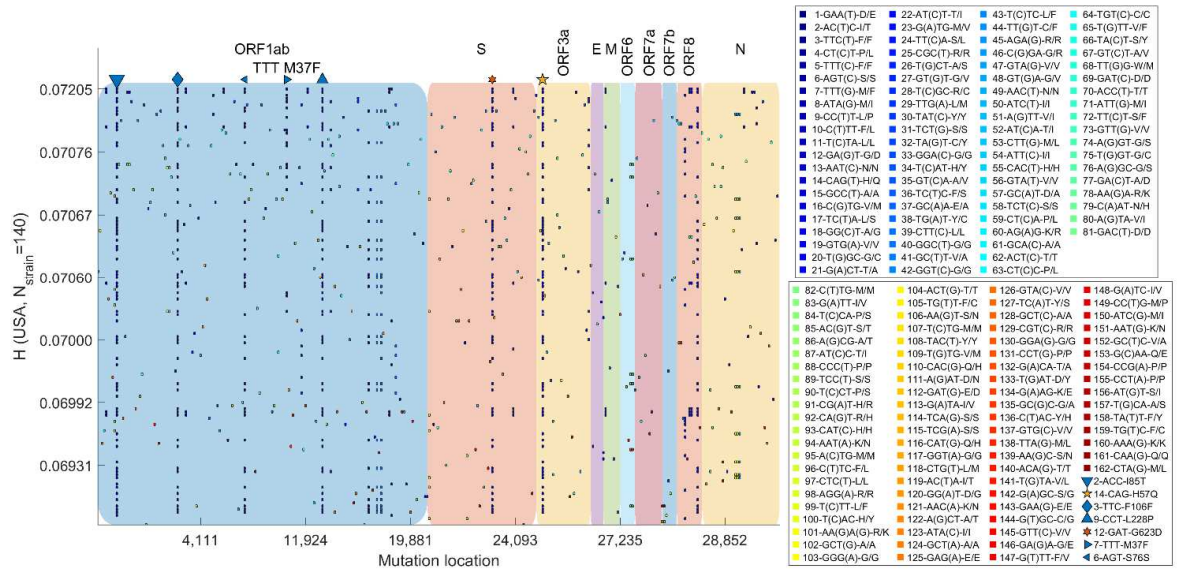

Fig. E10 Mutation profile for SARS-CoV-2 genomes (geolocation of the USA) with different HGS. Out of a total of 1538 viral genomes, 140 genomes have unique HGS values. A total of 162 mutations were identified in all the genomes. The top 7 conserved mutations with were shown with special markers at the top of colored blocks representing ORFs. Mutation 10818G>T in ORF1ab (codon TTG>TTT) occurred in populations with high HGS, which results in amino acid M37F mutation in transmembrane protein nsp6. The mutation rarely occurred in populations with low/moderate HGS.

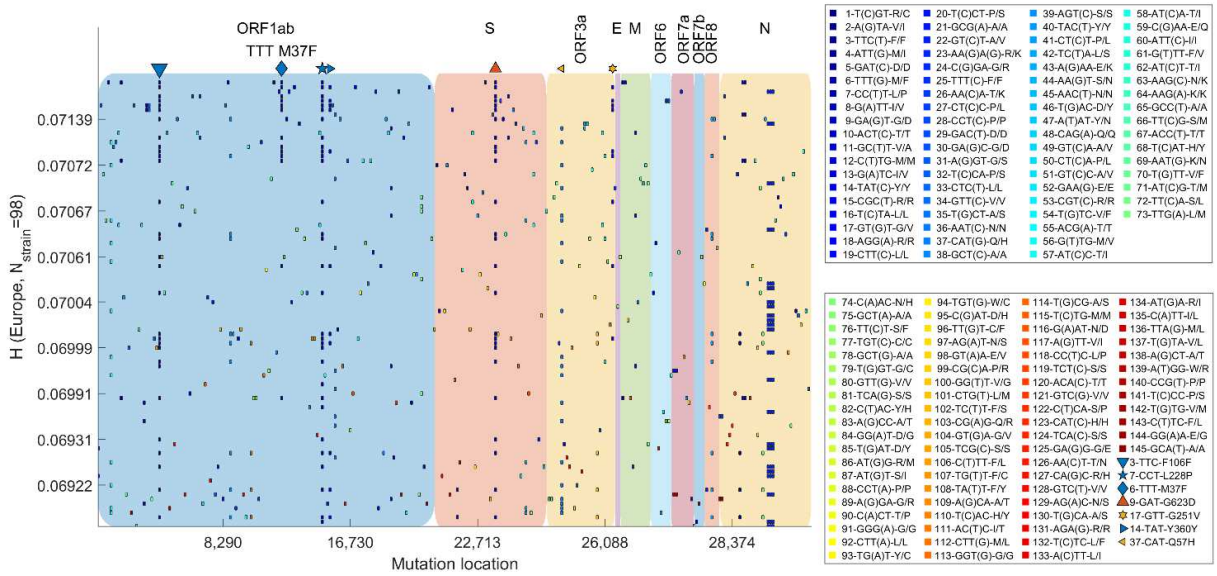

Fig. E11 Mutation profile for SARS-CoV-2 genomes (geolocation of Europe) with different HGS. Out of a total of 856 viral genomes, 98 genomes have unique HGS values. A total of 145 mutations were identified in all the genomes. The top 7 conserved mutations with were shown with special markers at the top of colored blocks representing ORFs. Mutation 10818G>T in ORF1ab (codon TTG>TTT) occurred in populations with high HGS, which results in amino acid M37F mutation in transmembrane protein nsp6. The mutation rarely occurred in populations with low/moderate HGS.