The discovery of gene mutations making SARS-CoV-2 well adapted for humans: host-genome similarity analysis of 2594 genomes from China, the USA and Europe

Weitao Sun[1,2*&]

[1]School of Aerospace Engineering, Tsinghua University, Beijing, 100084, China

[2]Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, 100084, China

*** Corresponding author**

Email: sunw@tsinghua.edu.cn

Short title

Discovery of SARS-CoV-2 gene mutations by host-genome similarity analysis

[&]The author contributed equally to this work.

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a positive-sense single-stranded virus approximately 30 kb in length, causes the ongoing novel coronavirus disease-2019 (COVID-19). Studies confirmed significant genome differences between SARS-CoV-2 and SARS-CoV, suggesting that the distinctions in pathogenicity 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 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 encoding 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 high HGS of 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 10818G>T caused the amino acid mutation M37F in the transmembrane protein nsp6. The results suggest that the ORF6 and ORF8 genes and the mutation M37F may play important roles in causing COVID-19. The findings demonstrate that HGS analysis is a promising way to identify important genes and mutations in adaptive strains, which may help in searching potential targets for pharmaceutical agents.

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[1, 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.

Materials and Methods

Viral genome data

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

The SARS-CoV-2 genomes were obtained from the GISAID database[15]. By May 20, 2020, the GISAID database (https://www.gisaid.org/) had 416 SARS-CoV-2 genomes from China, 5184 genomes from the USA and 10954 genomes from Europe. Complete and high-coverage genomes were used to ensure accurate HGS calculations. The sequences containing nucleotide names other than A, G, C and T were removed from the dataset. In total, 2594 SARS-CoV-2 genomes were used in the current study, including 200 from China, 1538 from the USA and 856 from Europe. The CDSs of the SARS-CoV-2 genome were identified by using MATLAB (https://www.mathworks.com/help/bioinfo/ref/seqshoworfs.html). The accession IDs of the genomes used in the article can be found in the Supplemental Information.

Human SARS-CoV genomes were collected from NCBI GenBank[16]. There were 25 CDSs of SARS-CoVs (full-length sequences only, with all ORF sequences, no nucleotide names other than A, G, C and T) at the time of article preparation. The accession IDs of these viral sequences can be found in the Supplemental Information.

Host-genome similarity (HGS)

The target CDSs were aligned with the human genome (*Homo sapiens* GRCh38.p12 chromosomes) by Blastn[17] to obtain matching fragments. Blastn sequence alignment gives an original score of S. To facilitate the comparison of Blast results among different subgenomic groups, the original score is standardized to S' by Blastn:

$$S' = \frac{\lambda S - \ln K}{\ln 2}, \quad (1)$$

$$E = mn2^{-S'}. \quad (2)$$

Here, the 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[18]. Assuming that the fragment of length a matches perfectly in the two random sequences, one has the following formula:

$$E = (m-a)(n-a)4^{-a}. \quad (3)$$

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, HGS is defined as

$$H = \frac{\sum a}{n} = \frac{\sum S'}{2n}, \quad (4)$$

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.

Data availability

The SARS-CoV-2 genomes used in this study can be obtained at GISAID website (https://www.gisaid.org/). The SARS-CoV genomes can be obtained at NCBI database (https://www.ncbi.nlm.nih.gov/). The accession number and corresponding HGS of 2594 SARS-CoV-2 genomes and those of 25 SARS-CoV genomes are in Supplemental Information. The code for HGS calculation is available in GitHub (https://github.com/WeitaoNSun/HGS).

Results

SARS-CoV-2 ORFs 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 (**Table 1**).

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[17], 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 ORF lengths of SARS-CoV and SARS-CoV-2 genomes are given in **Table 1**.

Table 1. Location, length and residue number of each ORF of SARS-CoV-2 and SARS-CoV genomes. The ORF names defined in different papers are listed in the first three columns(9, 18, 19).

ORF names			SARS-CoV				SARS-CoV-2			
Narayanan	Marra	Rota	start	stop	length	residues	start	stop	length	residues
ORF1a	ORF1a	1a	3361	13413	10053	3350	266	13483	13218	4405
N/R	N/R	N/R	13685	13759	75	24	13685	13759	75	24

ORF1b	ORF1b	1b	13398	21485	8088	2628	13768	21555	7788	2595
S	Sprotein	S	21492	25259	3768	1282	21536	25384	3849	1282
N/R	N/R	N/R	25207	25329	123	40	25332	25448	117	38
ORF3a	ORF3	X1	25268	26092	825	274	25393	26220	828	275
E	Eprotein	E	26117	26347	231	76	26245	26472	228	75
M	Mprotein	M	26398	27063	666	221	26523	27191	669	222
ORF6	ORF7	X3	27074	27265	192	63	27202	27387	186	61
ORF7a	ORF8	X4	27273	27641	369	122	27394	27759	366	121
ORF7b	ORF9	N/R	27638	27772	135	44	27756	27887	132	43
ORF8a	ORF10	N/R	27779	27853	75	24	27894	28259	366	121
ORF8b	ORF11	X5	27862	28116	255	84	N/R	N/R	N/R	N/R
N	Nprotein	N	28118	29386	1269	422	28274	29533	1260	419
N/R	N/R	N/R	29413	29490	78	25	29558	29674	117	38

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 Fig 1, Fig 2 and Fig 3 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.

Fig 1. 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 color bar represents the overall HGS value of each genome (weighted sum of ORF HGS). 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 HGS.

Fig 2. 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 color

bar represents the overall HGS value of each genome (weighted sum of ORF HGS). 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 HGS.

Fig 3. 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 color bar represents the overall HGS value of each genome (weighted sum of ORF HGS). 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 HGS.

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 4 shows that ORF 7b of SARS-CoV had the highest similarity with the 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 4**. The roles of such HGS changes are not clear. However, by investigating the function of the SARS-CoV viral genes and proteins, the mechanism of the rapid spread of the newly emerged COVID-19 may be inferred from the HGS changes in SARS-CoV-2 genomes.

Fig 4. 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[19]. 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[20]. 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[21]. In addition, ORF 6 circumvented IFN production by inhibiting IRF-3 phosphorylation in the (TRAF3)-(TBK1+IKKε)-(IRF3)-(IFNβ) signaling pathway (**Fig 5**), which is an essential signaling pathway triggered by the viral sensors RIG-1/MDA5 and TLRs[22].

Fig 5. 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.

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[23]. ORF8a and 8b have been observed in most SARS-

CoV-infected cells[24]. Wong et al.[25] 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[26]. 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%[27]. 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[25]. The 8b region of SARS-CoV protein ORF8 functions in ubiquitination binding, ubiquitination and glycosylation, which may interact with IRF3[28]. The expression of ORF8b and 8ab enhanced IRF3 degradation, thus regulating the immune functions of IRF3 (**Fig 5**). 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.

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[29]. 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[29]. 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 6**), 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.[29]. For 293T cells transfected with the SARS-CoV proteins and infected by Sendai virus[29], 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.

Fig 6. 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.

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[30]. 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.

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[31, 32]. 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 inFig 6, Fig 7 and Fig 8. SARS-CoV-2 in different regions developed its own conserved mutations independently (**Table 2**). 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.

Fig 7. 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 8. 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 9. 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.

Table 2. Conserved mutations identified in SARS-CoV-2 genomes with geolocations of China, the USA and Europe.

mutation	Location(ORF)	protein	residue mutation	residue type	residue property	Geolocation
TTG>TTT	10818	nsp6	M37F	Phenylalanine	Hydrophobic	China
GTA>ATA	1132	nsp2	V198I	Isoleucine	Hydrophobic	China
AGC>AGT	8517	nsp4	S76S	Serine	Polar	China
TTA>TCA	251	ORF8	L84S	Serine	Polar	China
TTG>CTG	415	N	M139M	Methionine	Hydrophobic	China
GAT>GGT	1868	S	D623G	Glycine	Special	China
GGT>GTT	752	ORF3a	G251V	Valine	Hydrophobic	China
ATC>ACC	794	nsp2	I85T	Threonine	Polar	USA
CAT>CAG	171	ORF3a	H57Q	Glutamine	Polar	USA
TTT>TTC	2772	nsp3	F106F	Phenylalanine	Hydrophobic	USA
CTT>CCT	13859	nsp12	L228P	Proline	Special	USA
GGT>GAT	1868	S	G623D	Aspartic acid	Negative	USA
TTG>TTT	10818	nsp6	M37F	Phenylalanine	Hydrophobic	USA
AGC>AGT	8517	nsp4	S76S	Serine	Polar	USA
TTT>TTC	2772	nsp3	F106F	Phenylalanine	Hydrophobic	Europe
CTT>CCT	13859	nsp12	L228P	Proline	Special	Europe
TTG>TTT	10818	nsp6	M37F	Phenylalanine	Hydrophobic	Europe
GGT>GAT	1868	S	G623D	Aspartic acid	Negative	Europe
GGT>GTT	752	ORF3a	G251V	Valine	Hydrophobic	Europe
TAC>TAT	14256	nsp12	Y360Y	Tyrosine	Hydrophobic	Europe
CAG>CAT	171	ORF3a	Q57H	Histidine	Positive	Europe

Of all the gene mutations, the ORF1ab 10818G>T(TTG>TTT) mutation is the most interesting. This mutation survived in all three regions (**Fig 10**). In addition, this mutation occurred only in the high HGS population rather than in that with a lower HGS (**Table 3**). 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)[33], which are organelle-like structures for viral genome replication and protect against host cell defenses.

Fig 10. 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.

Table 3. The mutation 10818G>T in ORF1ab (codon TTG>TTT) of SARS-CoV-2 mostly occurs in high-HGS (the first four columns) and rarely occurs in low-HGS population (the last four columns). The top 15 genomes with high HGS are chosen as high-HGS population. The last 15 genomes with low HGS are chosen as low-HGS population. GISAID accession ID and locations are given for genomes from China, the USA and Europe.

Genome ID	HGS	Seq. (10813-10823)	Mut.	Genome ID	HGS	Seq. (10813-10823)	Mut.
EPI-ISL-416331|China	0.07085	TTTTTTTATGA	T	EPI-ISL-406801|China	0.06999	TTTTTGTATGA	G
EPI-ISL-431783|China	0.07081	TTTTTTTATGA	T	EPI-ISL-412982|China	0.06986	TTTTTGTATGA	G
EPI-ISL-431180|China	0.07080	TTTTTGTATGA	G	EPI-ISL-421262|China	0.06943	TTTTTTTATGA	T
EPI-ISL-416373|China	0.07078	TTTTTTTATGA	T	EPI-ISL-406534|China	0.06942	TTTTTGTATGA	G
EPI-ISL-424360|China	0.07074	TTTTTGTATGA	G	EPI-ISL-413520|China	0.06936	TTTTTGTATGA	G
EPI-ISL-405839|China	0.07072	TTTTTGTATGA	G	EPI-ISL-416397|China	0.06935	TTTTTGTATGA	G
EPI-ISL-416325|China	0.07071	TTTTTGTATGA	G	EPI-ISL-421259|China	0.06935	TTTTTGTATGA	G
EPI-ISL-402127|China	0.07071	TTTTTGTATGA	G	EPI-ISL-406533|China	0.06934	TTTTTGTATGA	G
EPI-ISL-406595|China	0.07069	TTTTTGTATGA	G	EPI-ISL-421250|China	0.06931	TTTTTGTATGA	G
EPI-ISL-408515|China	0.07069	TTTTTGTATGA	G	EPI-ISL-416330|China	0.06930	TTTTTGTATGA	G
EPI-ISL-421253|China	0.07069	TTTTTGTATGA	G	EPI-ISL-418991|China	0.06924	TTTTTGTATGA	G
EPI-ISL-412978|China	0.07067	TTTTTGTATGA	G	EPI-ISL-406798|China	0.06922	TTTTTGTATGA	G
EPI-ISL-421256|China	0.07066	TTTTTGTATGA	G	EPI-ISL-416399|China	0.06915	TTTTTGTATGA	G
EPI-ISL-412459|China	0.07066	TTTTTGTATGA	G	EPI-ISL-413749|China	0.06914	TTTTTGTATGA	G
EPI-ISL-403932|China	0.07065	TTTTTGTATGA	G	EPI-ISL-421261|China	0.06866	TTTTTGTATGA	G
EPI-ISL-427288|USA	0.07205	TTTTTTTATGA	T	EPI-ISL-437866|USA	0.06931	TTTTTGTATGA	G
EPI-ISL-427190|USA	0.07195	TTTTTGTATGA	G	EPI-ISL-437384|USA	0.06928	TTTTTGTATGA	G
EPI-ISL-435558|USA	0.07153	TTTTTTTATGA	T	EPI-ISL-424901|USA	0.06928	TTTTTGTATGA	G
EPI-ISL-436064|USA	0.07148	TTTTTTTATGA	T	EPI-ISL-417353|USA	0.06927	TTTTTGTATGA	G
EPI-ISL-430939|USA	0.07144	TTTTTTTATGA	T	EPI-ISL-445078|USA	0.06923	TTTTTGTATGA	G
EPI-ISL-430404|USA	0.07144	TTTTTGTATGA	G	EPI-ISL-444068|USA	0.06922	TTTTTGTATGA	G
EPI-ISL-406223|USA	0.07142	TTTTTTTATGA	T	EPI-ISL-429641|USA	0.06921	TTTTTGTATGA	G
EPI-ISL-437857|USA	0.07138	TTTTTGTATGA	G	EPI-ISL-413458|USA	0.06921	TTTTTGTATGA	G
EPI-ISL-444760|USA	0.07138	TTTTTGTATGA	G	EPI-ISL-427209|USA	0.06915	TTTTTGTATGA	G
EPI-ISL-435444|USA	0.07137	TTTTTTTATGA	T	EPI-ISL-417453|USA	0.06914	TTTTTGTATGA	G
EPI-ISL-428776|USA	0.07133	TTTTTGTATGA	G	EPI-ISL-416711|USA	0.06914	TTTTTGTATGA	G
EPI-ISL-411956|USA	0.07132	TTTTTGTATGA	G	EPI-ISL-416677|USA	0.06911	TTTTTGTATGA	G
EPI-ISL-418897|USA	0.07128	TTTTTGTATGA	G	EPI-ISL-429645|USA	0.06910	TTTTTGTATGA	G
EPI-ISL-437799|USA	0.07085	TTTTTTTATGA	T	EPI-ISL-435562|USA	0.06864	TTTTTGTATGA	G
EPI-ISL-422966|USA	0.07081	TTTTTTTATGA	T	EPI-ISL-427247|USA	0.06856	TTTTTGTATGA	G
EPI-ISL-437322|Europe	0.07226	TTTTTTTATGA	T	EPI-ISL-445227|Europe	0.06929	TTTTTGTATGA	G
EPI-ISL-437303|Europe	0.07155	TTTTTTTATGA	T	EPI-ISL-418264|Europe	0.06925	TTTTTGTATGA	G
EPI-ISL-448774|Europe	0.07151	TTTTTGTATGA	G	EPI-ISL-447679|Europe	0.06925	TTTTTGTATGA	G
EPI-ISL-447665|Europe	0.07151	TTTTTTTATGA	T	EPI-ISL-447510|Europe	0.06925	TTTTTGTATGA	G
EPI-ISL-418243|Europe	0.07149	TTTTTTTATGA	T	EPI-ISL-413489|Europe	0.06922	TTTTTGTATGA	G
EPI-ISL-430852|Europe	0.07148	TTTTTTTATGA	T	EPI-ISL-428688|Europe	0.06922	TTTTTGTATGA	G
EPI-ISL-447654|Europe	0.07148	TTTTTTTATGA	T	EPI-ISL-428691|Europe	0.06922	TTTTTGTATGA	G
EPI-ISL-420423|Europe	0.07139	TTTTTGTATGA	G	EPI-ISL-434619|Europe	0.06921	TTTTTGTATGA	G
EPI-ISL-445241|Europe	0.07139	TTTTTGTATGA	G	EPI-ISL-445243|Europe	0.06913	TTTTTGTATGA	G
EPI-ISL-434662|Europe	0.07135	TTTTTGTATGA	G	EPI-ISL-448468|Europe	0.06910	TTTTTGTATGA	G
EPI-ISL-437896|Europe	0.07133	TTTTTGTATGA	G	EPI-ISL-434665|Europe	0.06908	TTTTTGTATGA	G
EPI-ISL-447634|Europe	0.07133	TTTTTGTATGA	G	EPI-ISL-430859|Europe	0.06867	TTTTTGTATGA	G
EPI-ISL-426379|Europe	0.07081	TTTTTTTATGA	T	EPI-ISL-418265|Europe	0.06856	TTTTTGTATGA	G
EPI-ISL-448502|Europe	0.07079	TTTTTGTATGA	G	EPI-ISL-435144|Europe	0.06848	TTTTTGTATGA	G
EPI-ISL-408430|Europe	0.07078	TTTTTTTATGA	T	EPI-ISL-437096|Europe	0.06795	TTTTTGTATGA	G

Studies on the nsp6 protein showed that the protein is a transmembrane protein with 6 transmembrane regions[34]. 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 11**). 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 11. 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. Viral genome data

Discussion

The HGS differences between SARS-CoV-2 and SARS-CoV genomes are critical to understanding clinical manifestations of the ongoing pandemic. ORF 6, 7b, 8 are the top 3 genes with significant HGS in SARS-CoV-2. What's more, ORF 6 and ORF 8 of the SARS-CoV-2 have clear increments in HGS, up to about

122% and 148% of that of SARS-CoV. Such apparent HGS changes suggest that these ORFs are important in defining the difference between SARS-CoV-2 and SARS-CoV. In the ongoing SARS-CoV-2 pandemic, the number of infected people is growing much faster than SARS and the total number of diagnosed cases exceeded that of SARS. But the mortality rate (about 3 %) was lower than SARS (about 11%)[35]. It is known that the primary targets of SARS-CoV are lung and small intestine[36, 37]. Recent studies have found that the SARS-CoV-2 may impair kidney function[38], infect the digestive system[39] and heart[40], and cause liver damage[41]. Recent study showed that SARS-CoV-2 can cause thromboembolic complications[42]. It has been reported that the SARS-CoV-2 virus can be found in stools and urine[40, 43]. In addition, an unusually long incubation period has been reported, during which more than half of the patients had no signs of disease and the virus carriers may be highly contagious [43]. Why the COVID-19 is so different from SARS is still not clear. But the mutations in virus genome and encoded proteins (such as spike protein S) are believed as an important factor.

The knowledge on SARS-CoV-2 accessory proteins by now is quite limited. However, the viral genome and proteins of SARS-CoV have been studied in depth in the past decade. Coronavirus has evolved to escape the innate immune (especially IFN-I expression and signaling) through suppression of IFN induction and singling pathways by non-structural proteins (nsps), structural proteins (S, E, M, N), and accessory proteins (ORF 3a, 6, 7a, 7b, 8a, 8b) [20, 44-53]. By comparing the SARS-CoV gene HGS with that of SARS-CoV-2, the obvious host-genome similarity changes shed light on the cause of rapid spread of COVID-19.

A power-law relationship is recognized between HGS and the expression of ISRE promoter, which is a direct indicator of the virus to inhibit interferon synthesis. The HGS of ORF6 and ORF8 increase greatly in SARS-CoV-2, which represents enhanced ability in suppressing innate immune. Although the functions of accessory proteins of SARS-CoV-2 have not been well studied, the secondary structure prediction reveals that ORF 6 and 8 are transmembrane proteins and may have related functions as in SARS-CoV. In fact, the SARS-CoV-2 contains a full-length ORF 8, which in SARS-CoV this reading frame is divided into ORF 8a and ORF 8b. Linking of ORF 8a and ORF 8b into a single continuous gene fragment had no significant effect on virus growth and RNA replication *in vitro*[54], which indicates that there are ORFs of SARS-CoV-2 may be similar to ORFs of SARS-CoV in function.

The discovery of increased HGS of ORF 6 and ORF 8 provide a strong evidence that SARS-COV-2 evolved to be more adaptable to humans than SARS-CoV. Based on these findings, following conjecture is proposed that the SARS-CoV-2 genes involved in suppressing the host's innate immunity are more powerful. Therefore, SARS-CoV-2 causes the delayed response of host innate immunity, which results in rapid transmission, low mortality and asymptomatic infection. These inferences are based on bioinformatics data, but offer a valuable picture of how SARS-CoV-2 could become different from SARS-CoV. In addition, the HGS method can also identify genetic mutations that help the virus adapt to humans.

It took the coronavirus 17 years to update from SARS-CoV to SARS-CoV-2. The significant increase in host-genome similarity distinguishes SARS-COV-2 from SARS-COV. SARS-CoV-2 found out a way to improve host adaptation. It is reasonable that high HGS may explain the quite different epidemiological characteristics of SARS-CoV-2, such as mild symptoms, rapid spread and low mortality. But the mechanism behind the impairment remains poorly understood and calls for future laboratory investigations. The COVID-19 appears to be less able to cause deaths than SARS and MERS during the ongoing pandemic. However, there is still a serious warning sign about viral mutation. The threat of another coronavirus outbreak with high infectiousness and mortality remains an alarming possibility.

Funding

This work was supported by the C.C. Lin specific fund.

Acknowledgments

I thank the laboratories for sharing SARS-CoV-2 and SARS-CoV genomes through GISAID and NCBI database.

References

1.Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-9. Epub 2020/02/06. PubMed PMID: 32015508; PubMed Central PMCID: PMCPMC7094943.

2.Xu X, Chen P, Wang J, Feng J, Zhou H, Li X, 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.He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nature Medicine. 2020;26(5):672-5.

4.Fiers W, Contreras R, Haegeman G, Rogiers R, Van de Voorde A, Van Heuverswyn H, et al. Complete nucleotide sequence of SV40 DNA. Nature. 1978;273(5658):113-20.

5.Rosenberg M, Segal S, Kuff EL, Singer MF. The nucleotide sequence of repetitive monkey DNA found in defective simian virus 40. Cell. 1977;11(4):845-57.

6.Senkevich TG, Bugert JJ, Sisler JR, Koonin EV, Darai G, Moss B. Genome sequence of a human tumorigenic poxvirus: prediction of specific host response-evasion genes. Science (New York, NY). 1996;273(5276):813-6. Epub 1996/08/09. PubMed PMID: 8670425.

7.Chang Y, Ma J, Zhang M, Yu Y. Prel iminary study on genome homology of viruses and human. J N BETHUNE UNIV MED SCI. 1997;23(3):242-4.

8.Shackelton LA, Holmes EC. The evolution of large DNA viruses: combining genomic information of viruses and their hosts. Trends in Microbiology. 2004;12(10):458-65.

9.Rouse BT, Sehrawat S. Immunity and immunopathology to viruses: what decides the outcome? Nature Reviews Immunology. 2010;10(7):514-26.

10.Van Kaer L, Joyce S. Viral evasion of antigen presentation: not just for peptides anymore. Nature Immunology. 2006;7(8):795-7.

11.Enard D, Petrov DA. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. Cell. 2018;175(2):360-71.e13. PubMed PMID: 30290142.

12.Enard D, Cai L, Gwennap C, Petrov DA. Viruses are a dominant driver of protein adaptation in mammals. Elife. 2016;5:e12469. PubMed PMID: 27187613.

13.Rothenburg S, Brennan G. Species-Specific Host–Virus Interactions: Implications for Viral Host Range and Virulence. Trends in Microbiology. 2020;28(1):46-56.

14.NCBI RC. Database resources of the National Center for Biotechnology Information. Nucleic acids research. 2018;46(D1):D8-d13. Epub 2017/11/16. PubMed PMID: 29140470; PubMed Central PMCID: PMCPMC5753372.

15.Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges. 2017;1(1):33-46.

16.Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic acids research. 2013;41(Database issue):D36-42. Epub 2012/11/30. PubMed PMID: 23193287; PubMed Central PMCID: PMCPMC3531190.

17.Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403-10.

18.Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A. 1990;87(6):2264-8. PubMed PMID: 2315319.

19.Frieman M, Yount B, Heise M, Kopecky-Bromberg SA, Palese P, Baric RS. Severe Acute Respiratory Syndrome Coronavirus ORF6 Antagonizes STAT1 Function by Sequestering Nuclear

Import Factors on the Rough Endoplasmic Reticulum/Golgi Membrane. Journal of Virology. 2007;81(18):9812.

20.Totura AL, Baric RS. SARS coronavirus pathogenesis: host innate immune responses and viral antagonism of interferon. Current Opinion in Virology. 2012;2(3):264-75.

21.Pewe L, Zhou H, Netland J, Tangadu C, Olivares H, Shi L, et al. A SARS-CoV-specific protein enhances virulence of an attenuated strain of mouse hepatitis virus. Advances in experimental medicine and biology. 2006;581:493-8. Epub 2006/10/14. PubMed PMID: 17037583.

22.Chau TL, Gioia R, Gatot JS, Patrascu F, Carpentier I, Chapelle JP, et al. Are the IKKs and IKK-related kinases TBK1 and IKK-epsilon similarly activated? Trends in biochemical sciences. 2008;33(4):171-80. Epub 2008/03/21. PubMed PMID: 18353649.

23.Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, et al. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science (New York, NY). 2003;302(5643):276-8. Epub 2003/09/06. PubMed PMID: 12958366.

24.Keng C-T, Choi Y-W, Welkers MRA, Chan DZL, Shen S, Gee Lim S, 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. Virology. 2006;354(1):132 - 42.

25.Wong HH, Fung TS, Fang S, Huang M, Le MT, Liu DX. 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. 2018;515:165-75.

26.Shi C-S, Nabar NR, Huang N-N, Kehrl JH. SARS-Coronavirus Open Reading Frame-8b triggers intracellular stress pathways and activates NLRP3 inflammasomes. Cell Death Discov. 2019;5:101-. PubMed PMID: 31231549.

27.Lau SKP, Feng Y, Chen H, Luk HKH, Yang W-H, Li KSM, et al. Severe Acute Respiratory Syndrome (SARS) Coronavirus ORF8 Protein Is Acquired from SARS-Related Coronavirus from Greater Horseshoe Bats through Recombination. Journal of virology. 2015;89(20):10532-47. Epub 08/12. PubMed PMID: 26269185.

28.Le TM, Wong HH, Tay FPL, Fang S, Keng C-T, Tan YJ, 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. 2007;274(16):4211-22.

29.Kopecky-Bromberg SA, Martínez-Sobrido L, Frieman M, Baric RA, Palese P. Severe Acute Respiratory Syndrome Coronavirus Open Reading Frame (ORF) 3b, ORF 6, and Nucleocapsid Proteins Function as Interferon Antagonists. Journal of Virology. 2007;81(2):548.

30.Channappanavar R, Fehr Anthony R, Vijay R, Mack M, Zhao J, Meyerholz David K, et al. Dysregulated Type I Interferon and Inflammatory Monocyte-Macrophage Responses Cause Lethal Pneumonia in SARS-CoV-Infected Mice. Cell Host & Microbe. 2016;19(2):181-93.

31.Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. Journal of Medical Virology. 2020;92(6):667-74.

32.Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. Journal of translational medicine. 2020;18(1):179. Epub 2020/04/24. PubMed PMID: 32321524; PubMed Central PMCID: PMCPMC7174922.

33.Baliji S, Cammer SA, Sobral B, Baker SC. Detection of nonstructural protein 6 in murine coronavirus-infected cells and analysis of the transmembrane topology by using bioinformatics and molecular approaches. Journal of virology. 2009;83(13):6957-62. Epub 04/22. PubMed PMID: 19386712.

34.Oostra M, Hagemeijer MC, van Gent M, Bekker CPJ, te Lintelo EG, Rottier PJM, 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. 2008;82(24):12392.

35.World Health Organization (2003) Consensus document on the epidemiology of severe acute respiratory syndrome (SARS). Department of Communicable Disease Surveillance and Response, WHO: 2003.

36.Lee N, Hui D, Wu A, Chan P, Cameron P, Joynt GM, et al. A major outbreak of severe acute respiratory syndrome in Hong Kong. The New England journal of medicine. 2003;348(20):1986-94. Epub 2003/04/16. doi: 10.1056/NEJMoa030685. PubMed PMID: 12682352.

37.Chan WS, Wu C, Chow SCS, Cheung T, To K-F, Leung W-K, et al. Coronaviral hypothetical and structural proteins were found in the intestinal surface enterocytes and pneumocytes of severe acute respiratory syndrome (SARS). Modern Pathology. 2005;18(11):1432-9. doi: 10.1038/modpathol.3800439.

38.Li Z, Wu M, Guo J, Yao J, Liao X, Song S, et al. Caution on Kidney Dysfunctions of 2019-nCoV Patients. medRxiv. 2020:2020.02.08.20021212. doi: 10.1101/2020.02.08.20021212.

39.Zhang H, Kang Z, Gong H, Xu D, Wang J, Li Z, et al. The digestive system is a potential route of 2019-nCov infection: a bioinformatics analysis based on single-cell transcriptomes. bioRxiv. 2020:2020.01.30.927806. doi: 10.1101/2020.01.30.927806.

40.Zou X, Chen K, Zou J, Han P, Hao J, Han Z. The single-cell RNA-seq data analysis on the receptor ACE2 expression reveals the potential risk of different human organs vulnerable to Wuhan 2019-nCoV infection Frontiers of Medicine 2020;online.

41.Chai X, Hu L, Zhang Y, Han W, Lu Z, Ke A, et al. Specific ACE2 Expression in Cholangiocytes May Cause Liver Damage After 2019-nCoV Infection. bioRxiv. 2020:2020.02.03.931766. doi: 10.1101/2020.02.03.931766.

42.Lamamri M, Chebbi A, Mamane J, Abbad S, Munuzzolini M, Sarfati F, et al. Priapism in a patient with coronavirus disease 2019 (COVID-19): A case report. Am J Emerg Med. 2020. doi: 10.1016/j.ajem.2020.06.027. PubMed PMID: PMC7301054.

43.Guan W-j, Ni Z-y, Hu Y, Liang W-h, Ou C-q, He J-x, et al. Clinical characteristics of 2019 novel coronavirus infection in China. medRxiv. 2020:2020.02.06.20020974. doi: 10.1101/2020.02.06.20020974.

44.Dosch SF, Mahajan SD, Collins AR. SARS coronavirus spike protein-induced innate immune response occurs via activation of the NF-κB pathway in human monocyte macrophages in vitro. Virus Research. 2009;142(1):19-27. doi: https://doi.org/10.1016/j.virusres.2009.01.005.

45.Fung TS, Liu DX. Human Coronavirus: Host-Pathogen Interaction. Annual review of microbiology. 2019;73:529-57. Epub 2019/06/22. doi: 10.1146/annurev-micro-020518-115759. PubMed PMID: 31226023.

46.Kikkert M. Innate Immune Evasion by Human Respiratory RNA Viruses. J Innate Immun. 2020;12(1):4-20. Epub 10/14. doi: 10.1159/000503030. PubMed PMID: 31610541.

47.Wong L-YR, Lui P-Y, Jin D-Y. A molecular arms race between host innate antiviral response and emerging human coronaviruses. Virologica Sinica. 2016;31(1):12-23. doi: 10.1007/s12250-015-3683-3.

48.Zhang Q, Yoo D. Immune evasion of porcine enteric coronaviruses and viral modulation of antiviral innate signaling. Virus Research. 2016;226:128-41. doi: https://doi.org/10.1016/j.virusres.2016.05.015.

49.DeDiego ML, Nieto-Torres JL, Jimenez-Guardeño JM, Regla-Nava JA, Castaño-Rodriguez C, Fernandez-Delgado R, et al. Coronavirus virulence genes with main focus on SARS-CoV envelope gene. Virus Research. 2014;194:124-37. doi: https://doi.org/10.1016/j.virusres.2014.07.024.

50.Saitoh T, Akira S. Regulation of innate immune responses by autophagy-related proteins. J Cell Biol. 2010;189(6):925-35. doi: 10.1083/jcb.201002021. PubMed PMID: 20548099.

51.Lim YX, Ng YL, Tam JP, Liu DX. Human Coronaviruses: A Review of Virus-Host Interactions. Diseases. 2016;4(3):26. doi: 10.3390/diseases4030026. PubMed PMID: 28933406.

52.Nelemans T, Kikkert M. Viral Innate Immune Evasion and the Pathogenesis of Emerging RNA Virus Infections. Viruses. 2019;11(10):961. doi: 10.3390/v11100961. PubMed PMID: 31635238.

53.MC F, GR S. Evasion of Host Innate Immunity by Emerging Viruses: Antagonizing Host RIG-I Pathways. J Emerg Dis Virol. 2017;3(3):1-8.

54.Yount B, Roberts RS, Sims AC, Deming D, Frieman MB, Sparks J, et al. Severe Acute Respiratory Syndrome Coronavirus Group-Specific Open Reading Frames Encode Nonessential Functions for Replication in Cell Cultures and Mice. Journal of Virology. 2005;79(23):14909. doi: 10.1128/JVI.79.23.14909-14922.2005.

Figure captions

Fig 1. 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 color bar represents the overall HGS value of each genome (weighted sum of ORF HGS). 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 HGS.

Fig 2. 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 color bar represents the overall HGS value of each genome (weighted sum of ORF HGS). 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 HGS.

Fig 3. 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 color bar represents the overall HGS value of each genome (weighted sum of ORF HGS). 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 HGS.

Fig 4. 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.

Fig 5. 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 6. 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.

Fig 7. 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 8. 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 9. 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.

Fig 10. 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.

Fig 11. 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.

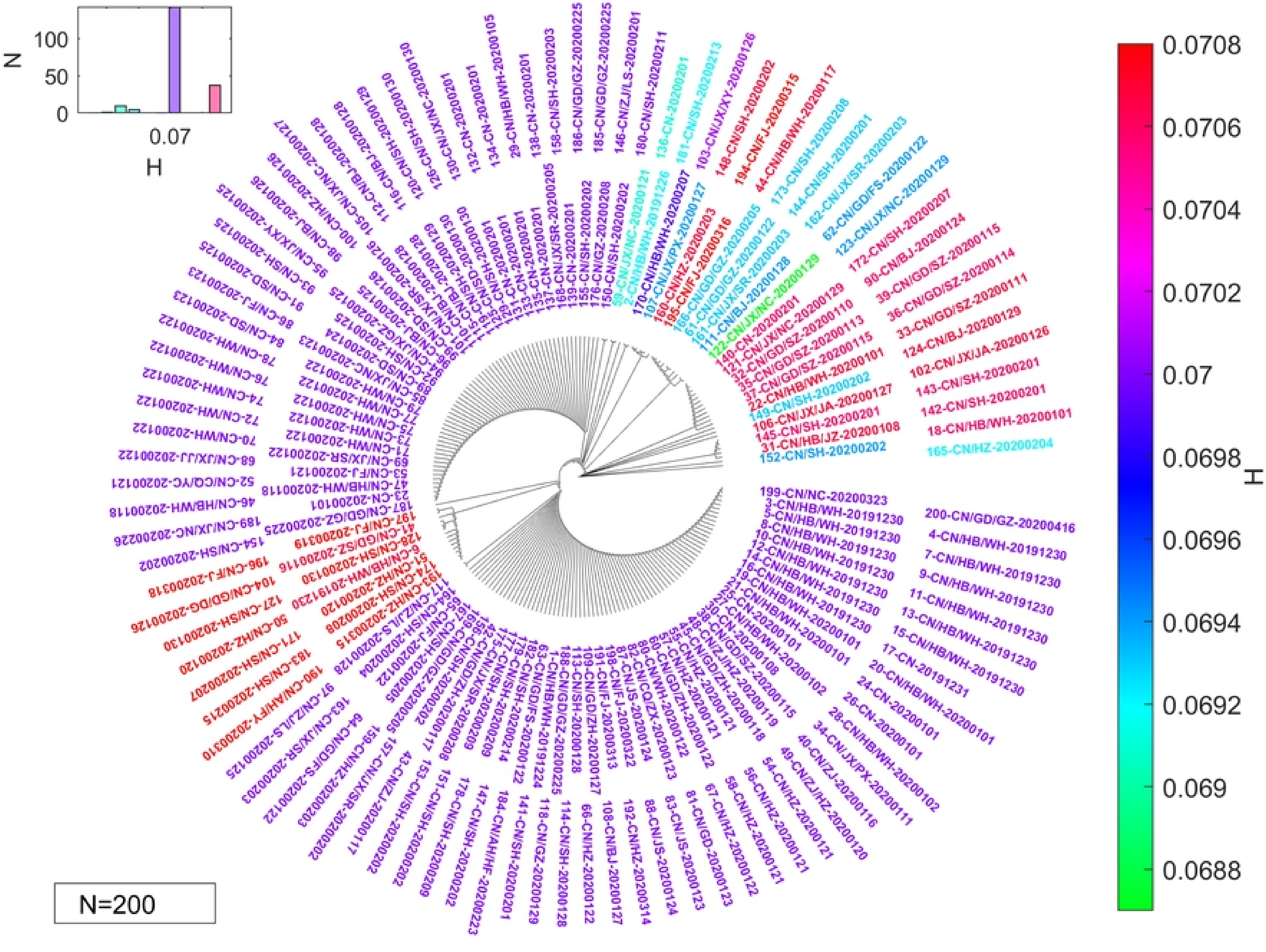
Supporting information

Dataset S1 (separate file). The accession number and corresponding HGS of 200 SARS-CoV-2 genomes with geolocation of China. Filename is DatasetS1_China_SARS-CoV-2_nstrain200_ORFHGS_allinone.xls. The file contains accession ID, collection date, location, HGS values for 10 ORFs (ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, and N) and the weighted HGS of the whole genome.

Dataset S2 (separate file). The accession number and corresponding HGS of 1538 SARS-CoV-2 genomes with geolocation of the USA. Filename is DatasetS2_USA_SARS-CoV-2_nstrain1538_ORFHGS_allinone.xls. The file contains accession ID, collection date, location, HGS values for 10 ORFs (ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, and N) and the weighted HGS of the whole genome.

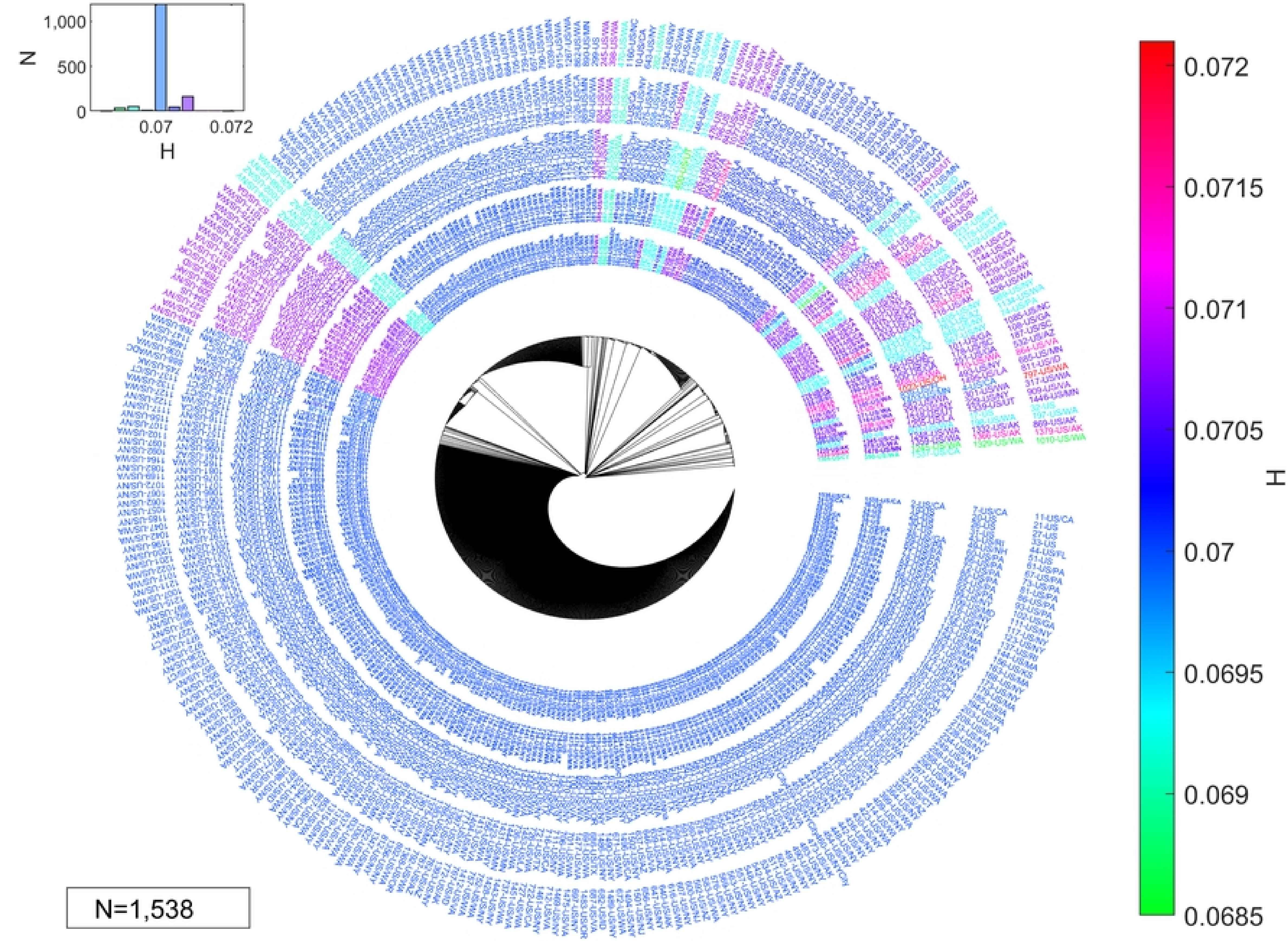
Dataset S3 (separate file). The accession number and corresponding HGS of 856 SARS-CoV-2 genomes with geolocation of Europe. Filename is DatasetS3_Europe_SARS-CoV-2_nstrain856_ORFHGS_allinone.xls. The file contains accession ID, collection date, location, HGS values for 10 ORFs (ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, and N) and the weighted HGS of the whole genome.

Dataset S4 (separate file). The accession number and corresponding HGS of 25 SARS-CoV genomes. Filename is DatasetS4_SARS-CoV_nstrain25_ORFHGS_allinone.xls. The file contains accession ID, HGS values for 10 ORFs (ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, and N) and the weighted HGS of the whole genome.



N
100
50
0
0.07
H
0.0708
0.0706
0.0704
0.0702
0.07
0.0698
0.0696
0.0694
0.0692
0.069
0.0688
H
N=200

Fig1



N
1,000
500
0
0.07
0.072
H
N=1,538
0.072
0.0715
0.071
0.0705
0.07
0.0695
0.069
0.0685
H

Fig2

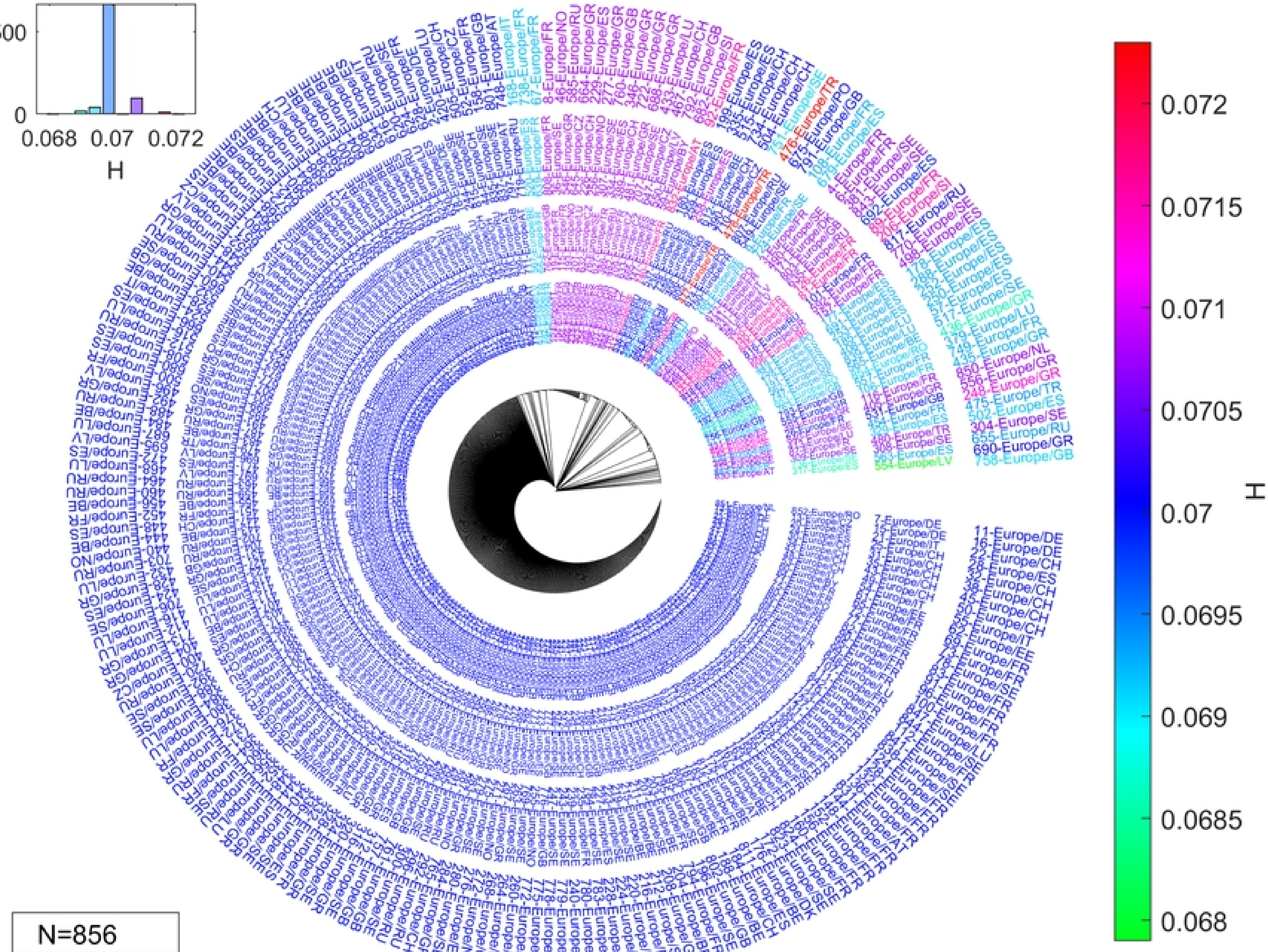
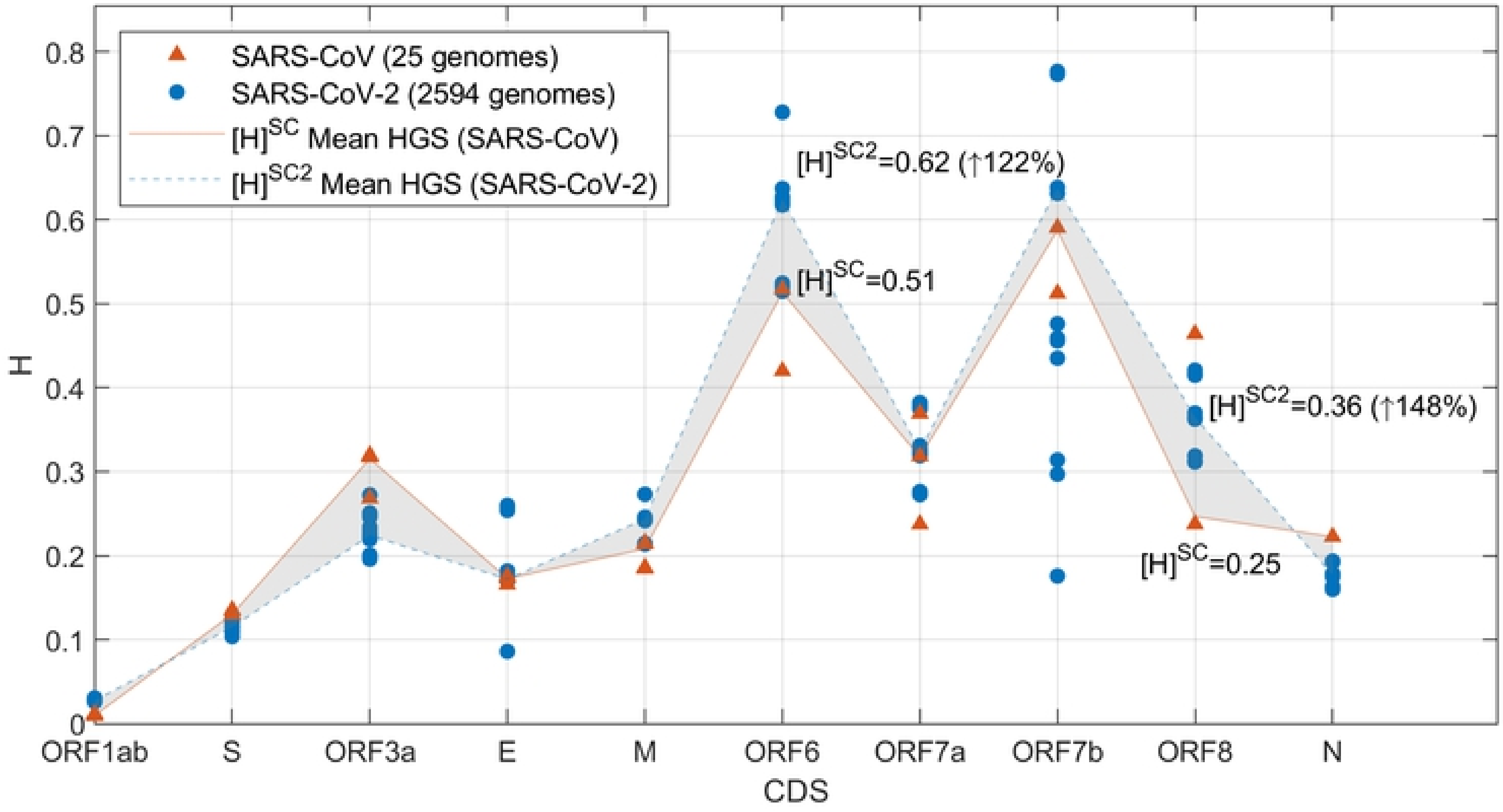

N
500
0
0.068
0.07
0.072
H
0.072
0.0715
0.071
0.0705
0.07
0.0695
0.069
0.0685
0.068
H
N=856

Fig3



SARS-CoV (25 genomes)
SARS-CoV-2 (2594 genomes)
[H]SC Mean HGS (SARS-CoV)
[H]SC2 Mean HGS (SARS-CoV-2)
[H]SC2=0.62 (↑122%)
[H]SC=0.51
[H]SC2=0.36 (↑148%)
[H]SC=0.25
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
H
ORF1ab
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
CDS

Fig4

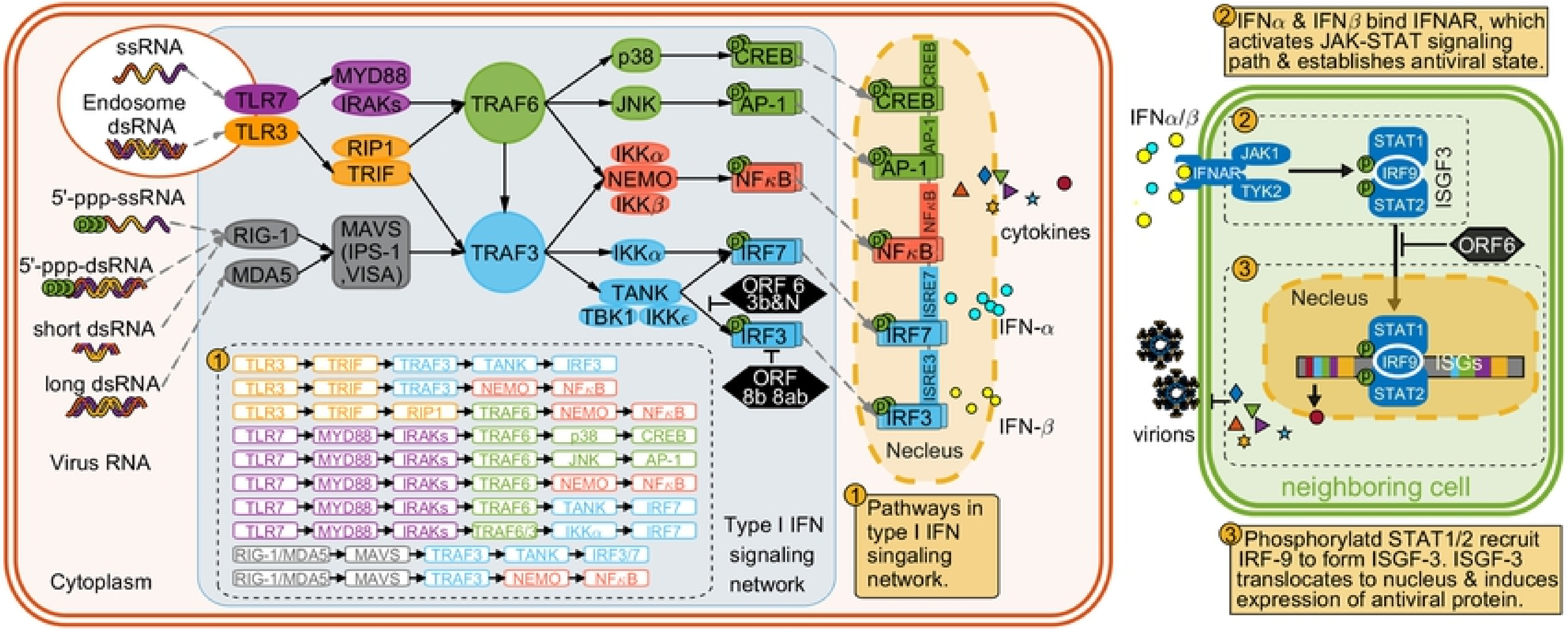

ssRNA
Endosome
dsRNA
TLR7
TLR3
MYD88
IRAKs
RIP1
TRIF
TRAF6
TRAF3
p38
JNK
IKKα
NEMO
IKKβ
CREB
AP-1
NFκB
IRF7
IRF3
TANK
TBK1 IKKε
ORF 6 3b&N
ORF 8b 8ab
5'-ppp-ssRNA
5'-ppp-dsRNA
short dsRNA
long dsRNA
Virus RNA
RIG-1
MDA5
MAVS (IPS-1 ,VISA)
Cytoplasm
Type I IFN signaling network
TLR3 → TRIF → TRAF3 → TANK → IRF3
TLR3 → TRIF → TRAF3 → NEMO → NFκB
TLR3 → TRIF → RIP1 → TRAF6 → NEMO → NFκB
TLR7 → MYD88 → IRAKs → TRAF6 → p38 → CREB
TLR7 → MYD88 → IRAKs → TRAF6 → JNK → AP-1
TLR7 → MYD88 → IRAKs → TRAF6 → NEMO → NFκB
TLR7 → MYD88 → IRAKs → TRAF6 → TANK → IRF7
TLR7 → MYD88 → IRAKs → TRAF6/3 → IKKα → IRF7
RIG-1/MDA5 → MAVS → TRAF3 → TANK → IRF3/7
RIG-1/MDA5 → MAVS → TRAF3 → NEMO → NFκB
Necleus
cytokines
IFN-α
IFN-β
1 Pathways in type I IFN singaling network.
2 IFNα & IFNβ bind IFNAR, which activates JAK-STAT signaling path & establishes antiviral state.
IFNα/β
JAK1
IFNAR
TYK2
STAT1
IRF9
STAT2
ISGF3
ORF6
Necleus
ISGs
virions
neighboring cell
3 Phosphorylatd STAT1/2 recruit IRF-9 to form ISGF-3. ISGF-3 translocates to nucleus & induces expression of antiviral protein.

Fig5

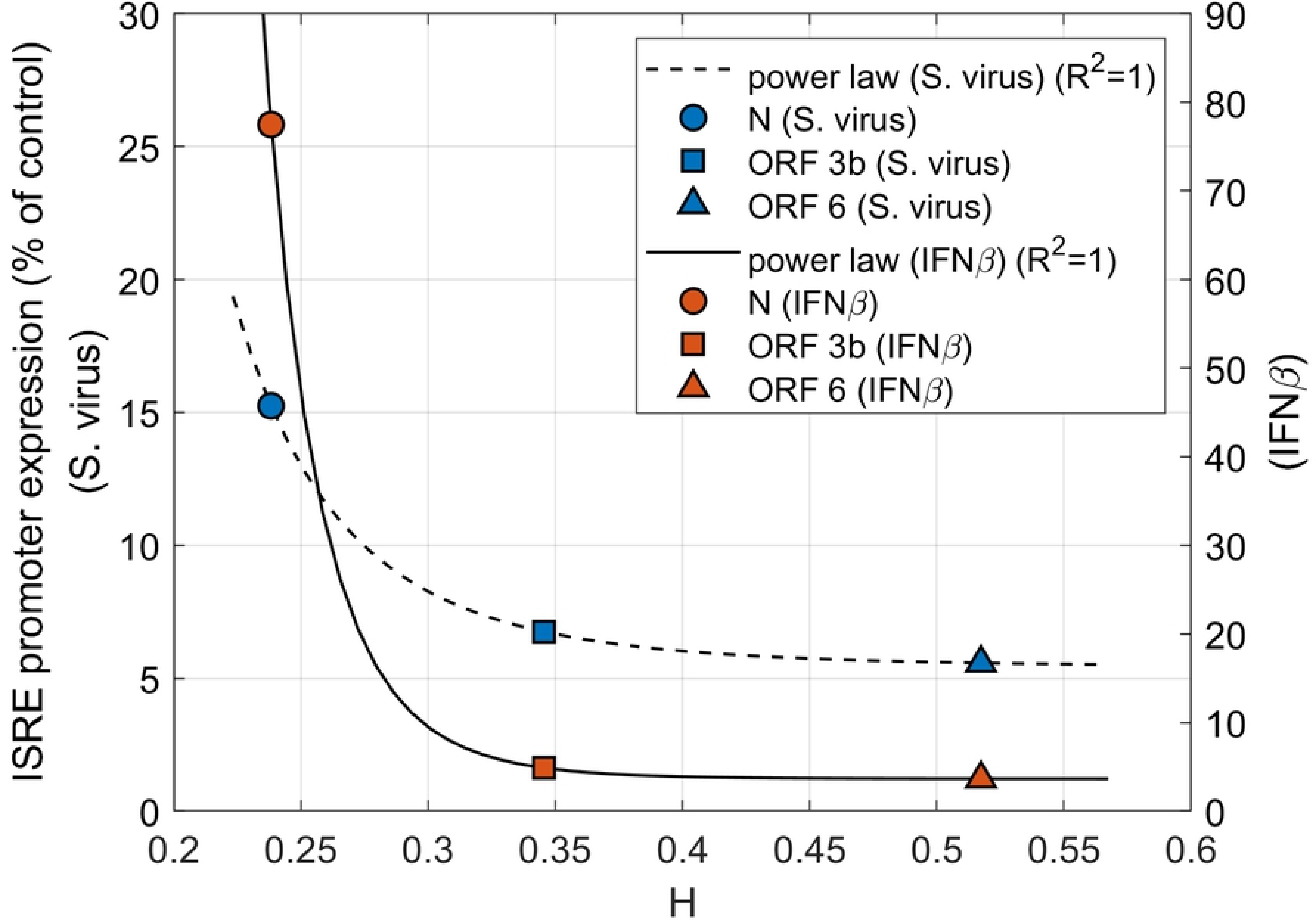

ISRE promoter expression (% of control)
(S. virus)
(IFNβ)
H
power law (S. virus) (R2=1)
N (S. virus)
ORF 3b (S. virus)
ORF 6 (S. virus)
power law (IFNβ) (R2=1)
N (IFNβ)
ORF 3b (IFNβ)
ORF 6 (IFNβ)
0
5
10
15
20
25
30
0
10
20
30
40
50
60
70
80
90
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6

Fig6

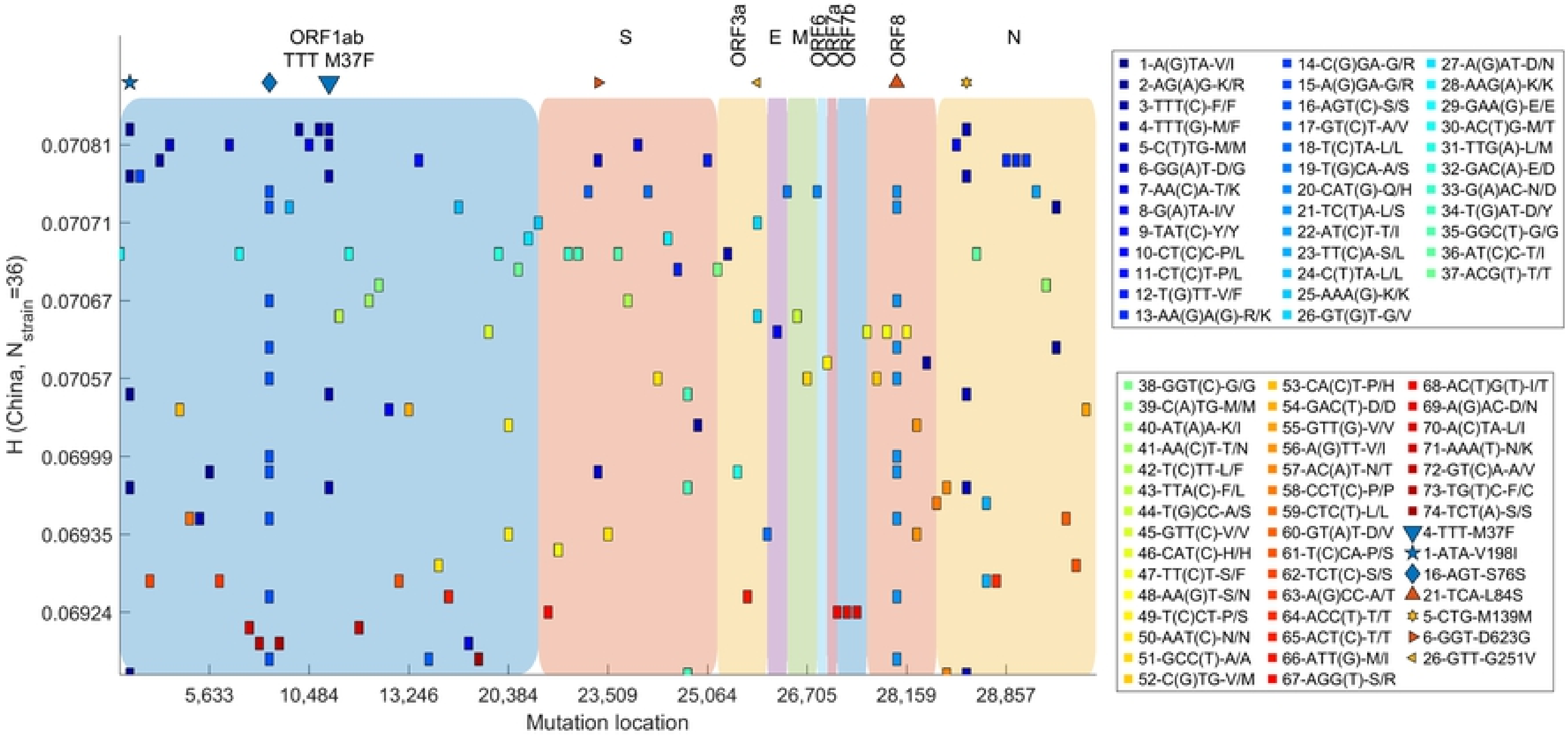

ORF1ab
TTT M37F
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
H (China, N_strain=36)
0.07081
0.07071
0.07067
0.07057
0.06999
0.06935
0.06924
5,633
10,484
13,246
20,384
23,509
25,064
26,705
28,159
28,857
Mutation location
1-A(G)TA-V/I
2-AG(A)G-K/R
3-TTT(C)-F/F
4-TTT(G)-M/F
5-C(T)TG-M/M
6-GG(A)T-D/G
7-AA(C)A-T/K
8-G(A)TA-I/V
9-TAT(C)-Y/Y
10-CT(C)C-P/L
11-CT(C)T-P/L
12-T(G)TT-V/F
13-AA(G)A(G)-R/K
14-C(G)GA-G/R
15-A(G)GA-G/R
16-AGT(C)-S/S
17-GT(C)T-A/V
18-T(C)TA-L/L
19-T(G)CA-A/S
20-CAT(G)-Q/H
21-TC(T)A-L/S
22-AT(C)T-T/I
23-TT(C)A-S/L
24-C(T)TA-L/L
25-AAA(G)-K/K
26-GT(G)T-G/V
27-A(G)AT-D/N
28-AAG(A)-K/K
29-GAA(G)-E/E
30-AC(T)G-M/T
31-TTG(A)-L/M
32-GAC(A)-E/D
33-G(A)AC-N/D
34-T(G)AT-D/Y
35-GGC(T)-G/G
36-AT(C)C-T/I
37-ACG(T)-T/T
38-GGT(C)-G/G
39-C(A)TG-M/M
40-AT(A)A-K/I
41-AA(C)T-T/N
42-T(C)TT-L/F
43-TTA(C)-F/L
44-T(G)CC-A/S
45-GTT(C)-V/V
46-CAT(C)-H/H
47-TT(C)T-S/F
48-AA(G)T-S/N
49-T(C)CT-P/S
50-AAT(C)-N/N
51-GCC(T)-A/A
52-C(G)TG-V/M
53-CA(C)T-P/H
54-GAC(T)-D/D
55-GTT(G)-V/V
56-A(G)TT-V/I
57-AC(A)T-N/T
58-CCT(C)-P/P
59-CTC(T)-L/L
60-GT(A)T-D/V
61-T(C)CA-P/S
62-TCT(C)-S/S
63-A(G)CC-A/T
64-ACC(T)-T/T
65-ACT(C)-T/T
66-ATT(G)-M/I
67-AGG(T)-S/R
68-AC(T)G(T)-I/T
69-A(G)AC-D/N
70-A(C)TA-L/I
71-AAA(T)-N/K
72-GT(C)A-A/V
73-TG(T)C-F/C
74-TCT(A)-S/S
4-TTT-M37F
1-ATA-V198I
16-AGT-S76S
21-TCA-L84S
5-CTG-M139M
6-GGT-D623G
26-GTT-G251V

Fig7

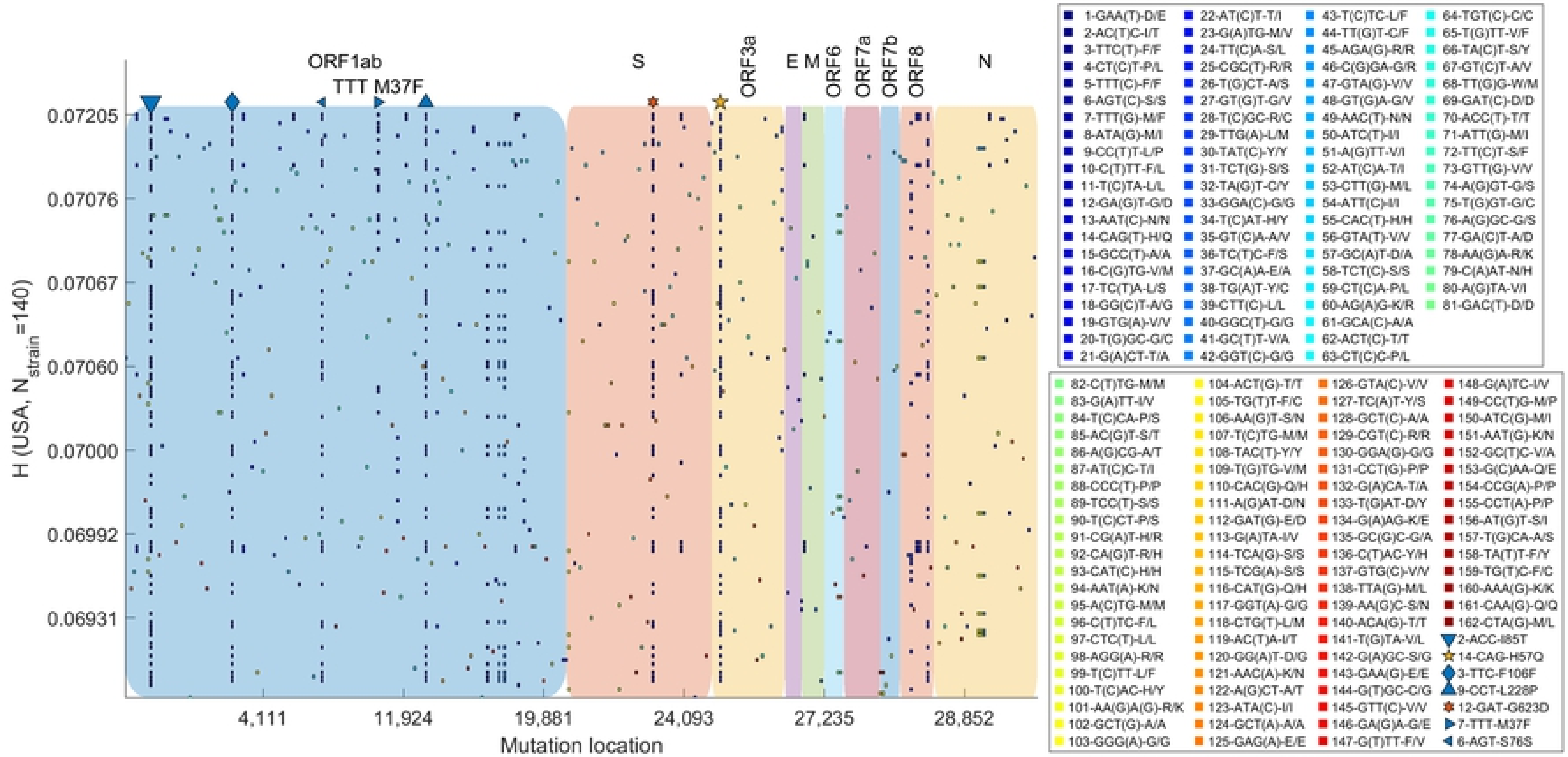

ORF1ab
TTT M37F
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
0.07205
0.07076
0.07067
0.07060
0.07000
0.06992
0.06931
H (USA, N_{strain}=140)
4,111
11,924
19,881
24,093
27,235
28,852
Mutation location
1-GAA(T)-D/E
2-AC(T)C-I/T
3-TTC(T)-F/F
4-CT(C)T-P/L
5-TTT(C)-F/F
6-AGT(C)-S/S
7-TTT(G)-M/F
8-ATA(G)-M/I
9-CC(T)T-L/P
10-C(T)TT-F/L
11-T(C)TA-L/L
12-GA(G)T-G/D
13-AAT(C)-N/N
14-CAG(T)-H/Q
15-GCC(T)-A/A
16-C(G)TG-V/M
17-TC(T)A-L/S
18-GG(C)T-A/G
19-GTG(A)-V/V
20-T(G)GC-G/C
21-G(A)CT-T/A
22-AT(C)T-T/I
23-G(A)TG-M/V
24-TT(C)A-S/L
25-CGC(T)-R/R
26-T(G)CT-A/S
27-GT(G)T-G/V
28-T(C)GC-R/C
29-TTG(A)-L/M
30-TAT(C)-Y/Y
31-TCT(G)-S/S
32-TA(G)T-C/Y
33-GGA(C)-G/G
34-T(C)AT-H/Y
35-GT(C)A-A/V
36-TC(T)C-F/S
37-GC(A)A-E/A
38-TG(A)T-Y/C
39-CTT(C)-L/L
40-GGC(T)-G/G
41-GC(T)T-V/A
42-GGT(C)-G/G
43-T(C)TC-L/F
44-TT(G)T-C/F
45-AGA(G)-R/R
46-C(G)GA-G/R
47-GTA(G)-V/V
48-GT(G)A-G/V
49-AAC(T)-N/N
50-ATC(T)-I/I
51-A(G)TT-V/I
52-AT(C)A-T/I
53-CTT(G)-M/L
54-ATT(C)-I/I
55-CAC(T)-H/H
56-GTA(T)-V/V
57-GC(A)T-D/A
58-TCT(C)-S/S
59-CT(C)A-P/L
60-AG(A)G-K/R
61-GCA(C)-A/A
62-ACT(C)-T/T
63-CT(C)C-P/L
64-TGT(C)-C/C
65-T(G)TT-V/F
66-TA(C)T-S/Y
67-GT(C)T-A/V
68-TT(G)G-W/M
69-GAT(C)-D/D
70-ACC(T)-T/T
71-ATT(G)-M/I
72-TT(C)T-S/F
73-GTT(G)-V/V
74-A(G)GT-G/S
75-T(G)GT-G/C
76-A(G)GC-G/S
77-GA(C)T-A/D
78-AA(G)A-R/K
79-C(A)AT-N/H
80-A(G)TA-V/I
81-GAC(T)-D/D
82-C(T)TG-M/M
83-G(A)TT-I/V
84-T(C)CA-P/S
85-AC(G)T-S/T
86-A(G)CG-A/T
87-AT(C)C-T/I
88-CCC(T)-P/P
89-TCC(T)-S/S
90-T(C)CT-P/S
91-CG(A)T-H/R
92-CA(G)T-R/H
93-CAT(C)-H/H
94-AAT(A)-K/N
95-A(C)TG-M/M
96-C(T)TC-F/L
97-CTC(T)-L/L
98-AGG(A)-R/R
99-T(C)TT-L/F
100-T(C)AC-H/Y
101-AA(G)A(G)-R/K
102-GCT(G)-A/A
103-GGG(A)-G/G
104-ACT(G)-T/T
105-TG(T)T-F/C
106-AA(G)T-S/N
107-T(C)TG-M/M
108-TAC(T)-Y/Y
109-T(G)TG-V/M
110-CAC(G)-Q/H
111-A(G)AT-D/N
112-GAT(G)-E/D
113-G(A)TA-I/V
114-TCA(G)-S/S
115-TCG(A)-S/S
116-CAT(G)-Q/H
117-GGT(A)-G/G
118-CTG(T)-L/M
119-AC(T)A-I/T
120-GG(A)T-D/G
121-AAC(A)-K/N
122-A(G)CT-A/T
123-ATA(C)-I/I
124-GCT(A)-A/A
125-GAG(A)-E/E
126-GTA(C)-V/V
127-TC(A)T-Y/S
128-GCT(C)-A/A
129-CGT(C)-R/R
130-GGA(G)-G/G
131-CCT(G)-P/P
132-G(A)CA-T/A
133-T(G)AT-D/Y
134-G(A)AG-K/E
135-GC(G)C-G/A
136-C(T)AC-Y/H
137-GTG(C)-V/V
138-TTA(G)-M/L
139-AA(G)C-S/N
140-ACA(G)-T/T
141-T(G)TA-V/L
142-G(A)GC-S/G
143-GAA(G)-E/E
144-G(T)GC-C/G
145-GTT(C)-V/V
146-GA(G)A-G/E
147-G(T)TT-F/V
148-G(A)TC-I/V
149-CC(T)G-M/P
150-ATC(G)-M/I
151-AAT(G)-K/N
152-GC(T)C-V/A
153-G(C)AA-Q/E
154-CCG(A)-P/P
155-CCT(A)-P/P
156-AT(G)T-S/I
157-T(G)CA-A/S
158-TA(T)T-F/Y
159-TG(T)C-F/C
160-AAA(G)-K/K
161-CAA(G)-Q/Q
162-CTA(G)-M/L
2-ACC-I85T
14-CAG-H57Q
3-TTC-F106F
9-CCT-L228P
12-GAT-G623D
7-TTT-M37F
6-AGT-S76S

Fig8

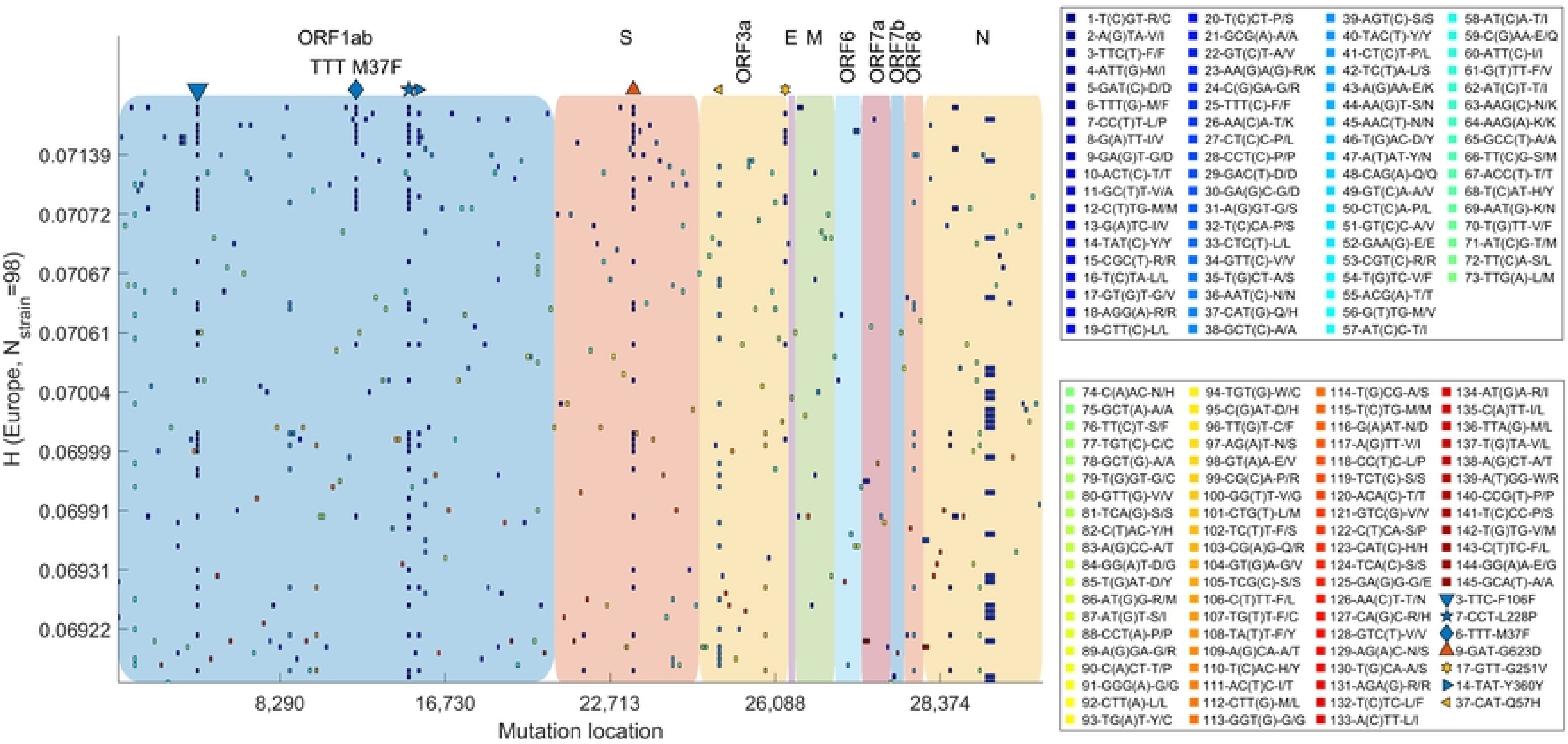
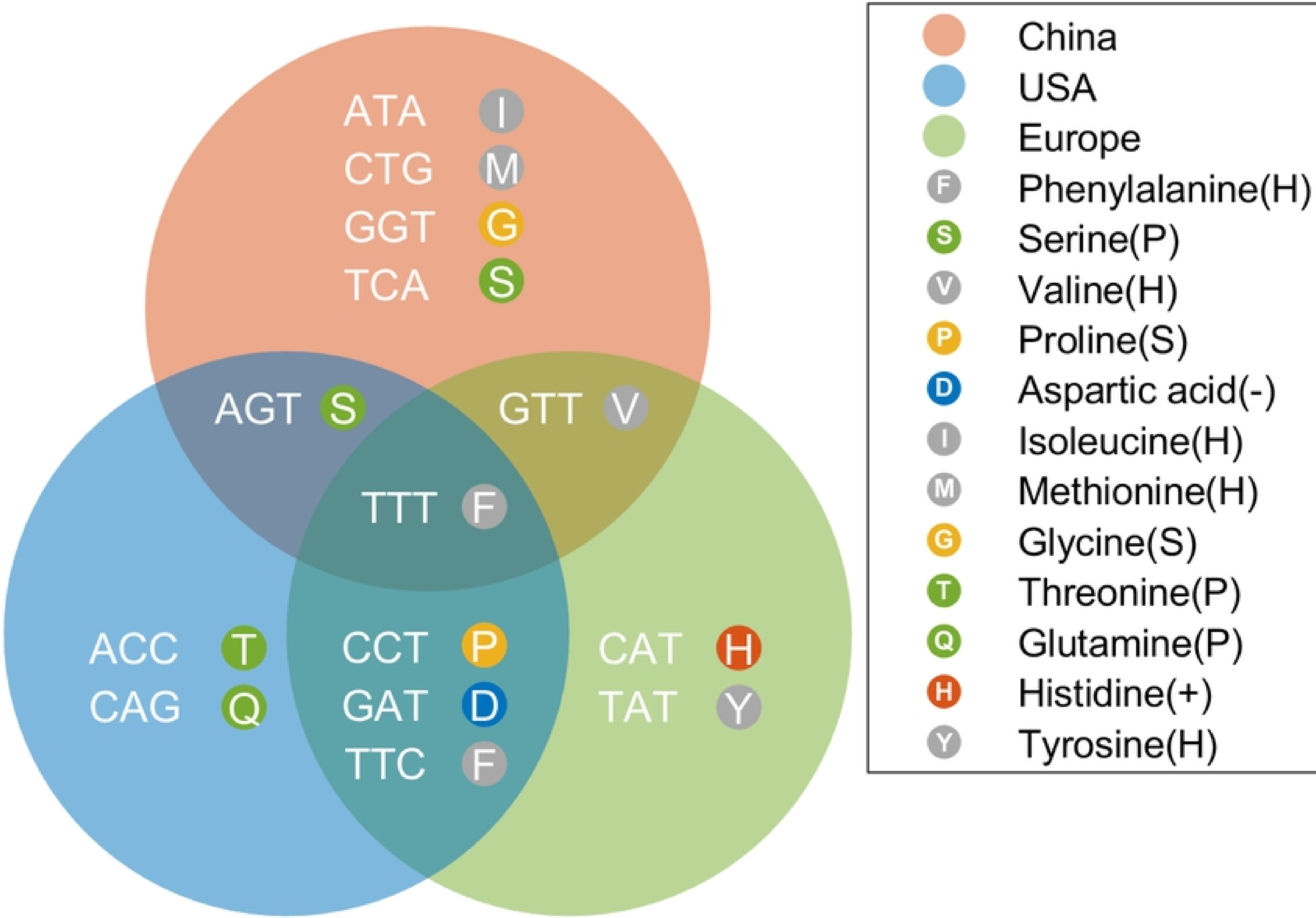

ORF1ab
TTT M37F
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
H (Europe, N_{strain}=98)
0.07139
0.07072
0.07067
0.07061
0.07004
0.06999
0.06991
0.06931
0.06922
8,290
16,730
22,713
26,088
28,374
Mutation location
1-T(C)GT-R/C
2-A(G)TA-V/I
3-TTC(T)-F/F
4-ATT(G)-M/I
5-GAT(C)-D/D
6-TTT(G)-M/F
7-CC(T)T-L/P
8-G(A)TT-I/V
9-GA(G)T-G/D
10-ACT(C)-T/T
11-GC(T)T-V/A
12-C(T)TG-M/M
13-G(A)TC-I/V
14-TAT(C)-Y/Y
15-CGC(T)-R/R
16-T(C)TA-L/L
17-GT(G)T-G/V
18-AGG(A)-R/R
19-CTT(C)-L/L
20-T(C)CT-P/S
21-GCG(A)-A/A
22-GT(C)T-A/V
23-AA(G)A(G)-R/K
24-C(G)GA-G/R
25-TTT(C)-F/F
26-AA(C)A-T/K
27-CT(C)C-P/L
28-CCT(C)-P/P
29-GAC(T)-D/D
30-GA(G)C-G/D
31-A(G)GT-G/S
32-T(C)CA-P/S
33-CTC(T)-L/L
34-GTT(C)-V/V
35-T(G)CT-A/S
36-AAT(C)-N/N
37-CAT(G)-Q/H
38-GCT(C)-A/A
39-AGT(C)-S/S
40-TAC(T)-Y/Y
41-CT(C)T-P/L
42-TC(T)A-L/S
43-A(G)AA-E/K
44-AA(G)T-S/N
45-AAC(T)-N/N
46-T(G)AC-D/Y
47-A(T)AT-Y/N
48-CAG(A)-Q/Q
49-GT(C)A-A/V
50-CT(C)A-P/L
51-GT(C)C-A/V
52-GAA(G)-E/E
53-CGT(C)-R/R
54-T(G)TC-V/F
55-ACG(A)-T/T
56-G(T)TG-M/V
57-AT(C)C-T/I
58-AT(C)A-T/I
59-C(G)AA-E/Q
60-ATT(C)-I/I
61-G(T)TT-F/V
62-AT(C)T-T/I
63-AAG(C)-N/K
64-AAG(A)-K/K
65-GCC(T)-A/A
66-TT(C)G-S/M
67-ACC(T)-T/T
68-T(C)AT-H/Y
69-AAT(G)-K/N
70-T(G)TT-V/F
71-AT(C)G-T/M
72-TT(C)A-S/L
73-TTG(A)-L/M
74-C(A)AC-N/H
75-GCT(A)-A/A
76-TT(C)T-S/F
77-TGT(C)-C/C
78-GCT(G)-A/A
79-T(G)GT-G/C
80-GTT(G)-V/V
81-TCA(G)-S/S
82-C(T)AC-Y/H
83-A(G)CC-A/T
84-GG(A)T-D/G
85-T(G)AT-D/Y
86-AT(G)G-R/M
87-AT(G)T-S/I
88-CCT(A)-P/P
89-A(G)GA-G/R
90-C(A)CT-T/P
91-GGG(A)-G/G
92-CTT(A)-L/L
93-TG(A)T-Y/C
94-TGT(G)-W/C
95-C(G)AT-D/H
96-TT(G)T-C/F
97-AG(A)T-N/S
98-GT(A)A-E/V
99-CG(C)A-P/R
100-GG(T)T-V/G
101-CTG(T)-L/M
102-TC(T)T-F/S
103-CG(A)G-Q/R
104-GT(G)A-G/V
105-TCG(C)-S/S
106-C(T)TT-F/L
107-TG(T)T-F/C
108-TA(T)T-F/Y
109-A(G)CA-A/T
110-T(C)AC-H/Y
111-AC(T)C-I/T
112-CTT(G)-M/L
113-GGT(G)-G/G
114-T(G)CG-A/S
115-T(C)TG-M/M
116-G(A)AT-N/D
117-A(G)TT-V/I
118-CC(T)C-L/P
119-TCT(C)-S/S
120-ACA(C)-T/T
121-GTC(G)-V/V
122-C(T)CA-S/P
123-CAT(C)-H/H
124-TCA(C)-S/S
125-GA(G)G-G/E
126-AA(C)T-T/N
127-CA(G)C-R/H
128-GTC(T)-V/V
129-AG(A)C-N/S
130-T(G)CA-A/S
131-AGA(G)-R/R
132-T(C)TC-L/F
133-A(C)TT-L/I
134-AT(G)A-R/I
135-C(A)TT-I/L
136-TTA(G)-M/L
137-T(G)TA-V/L
138-A(G)CT-A/T
139-A(T)GG-W/R
140-CCG(T)-P/P
141-T(C)CC-P/S
142-T(G)TG-V/M
143-C(T)TC-F/L
144-GG(A)A-E/G
145-GCA(T)-A/A
3-TTC-F106F
7-CCT-L228P
6-TTT-M37F
9-GAT-G623D
17-GTT-G251V
14-TAT-Y360Y
37-CAT-Q57H

Fig9


ATA I
CTG M
GGT G
TCA S
AGT S
GTT V
TTT F
ACC T
CAG Q
CCT P
GAT D
TTC F
CAT H
TAT Y
China
USA
Europe
Phenylalanine(H)
Serine(P)
Valine(H)
Proline(S)
Aspartic acid(-)
Isoleucine(H)
Methionine(H)
Glycine(S)
Threonine(P)
Glutamine(P)
Histidine(+)
Tyrosine(H)

Fi10

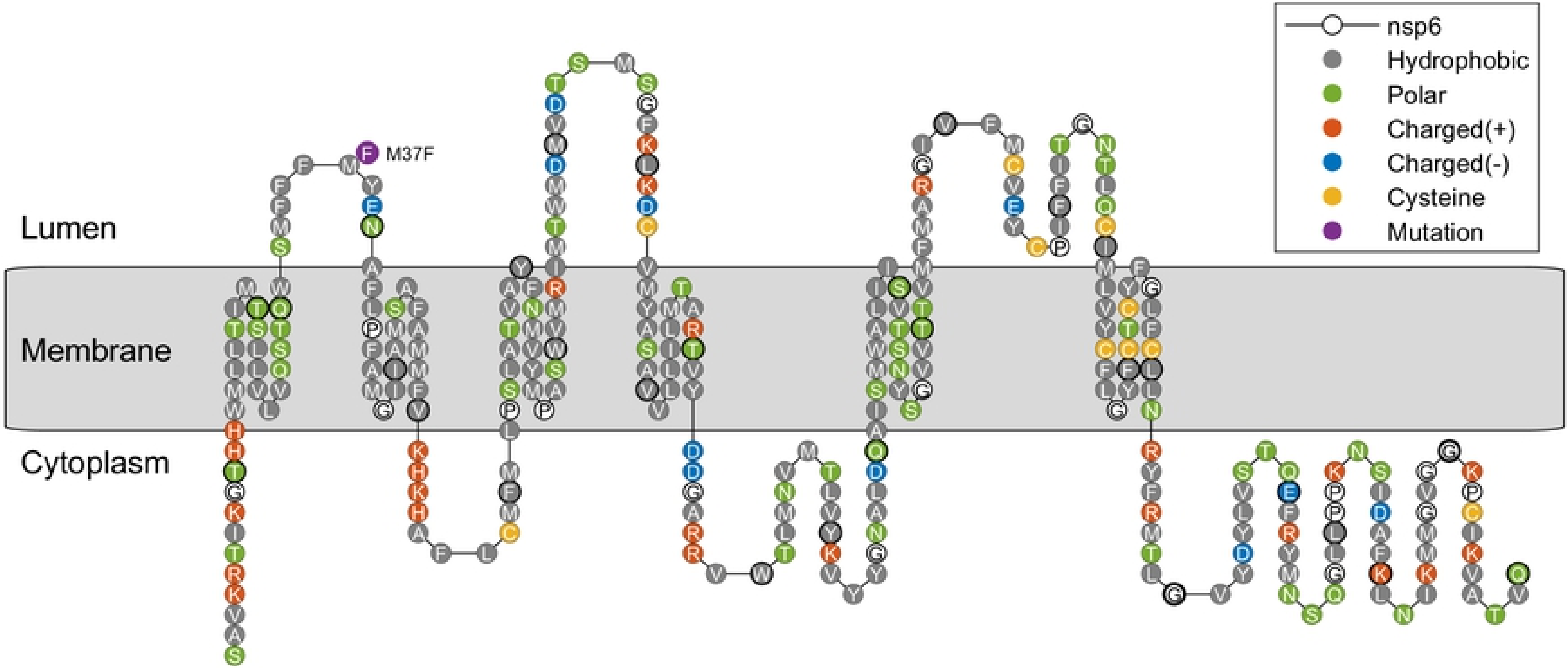

nsp6
Hydrophobic
Polar
Charged(+)
Charged(-)
Cysteine
Mutation
M37F
Lumen
Membrane
Cytoplasm

Fig11