Genetic landscape of SARS-CoV-2

Xueming Zheng[1] and Wen Zhang[1]

[1]Affiliation not available

November 5, 2020

Abstract

The novel coronavirus named SARS-CoV-2 caused human epidemic all over the world at breathtaking speed. It is of great concern for the research community to understand the evolutionary origin and molecular characteristics of this virus. With more and more isolates are sequenced, it is possible to estimate the genomic variation and evolution of SARS-CoV-2. In this study, 17,229 complete genomes of SARS-CoV-2 were analyzed to characterize the genomic diversity. Using Doc2vec algorithm, we got the the genome embeddings of SARS-CoV-2 isolates as well as its related virus species. The results showed that the distance estimated from genome embedding is different from sequence alignment. Additionally, a frequently happened mutations (C to T/U) in -25 upstream of the ORF1ab start codon were identified. On protein level, it seemed that the mutations appeared with unequal distribution among the proteins. ORF1ab, S, ORF3a, ORF8 and N proteins were easier to tolerate mutations while the other proteins showed high conservation among the isolates.

Introduction

SARS-CoV-2 (severe acute respiratory syndrome coronavirus-2) is a novel human infection coronavirus, which is responsible for the outbreak of coronavirus disease in December, 2019, in Wuhan, China and later all over the world [1]. SARS-CoV-2 belongs to beta-coronaviruses which also including bat coronavirus (BCoV) as well as SARS-CoV and MERS-CoV viruses [2]. Although the origin of SARS-CoV-2 is still unclear, a few closely related CoVs with high sequence identity were identified including the BatCoV RaTG13 (MN996532.1, identity: ˜96%) [3]. The similarity analysis between SARS-CoV-2 and the animal-infection CoVs suggests its bat- or pangolin-origin [3,4]. However, the putative inter-species evolution and infection mechanism remains largely unknown.

SARS-CoV-2 contains a positive-sense, single-stranded RNA (ssRNA) genome of about 30 kb in size [1]. Besides the 5'- and 3'-untranslated region (UTR), almost all the genome is occupied by coding regions. The 5'-terminal encodes the largest polyprotein ORF1ab which involves in genome transcription and replication [5,6]. The glycoproteins spike (S) attaches the virus to the cell membrane by interacting with host receptor, initiating the infection [7-9]. The remaining ORFs encoding envelope (E), membrane (M), nucleocapsid (N) proteins as well as a few accessory proteins such as ORF3a, ORF8 [10].

The S protein is processed by host cell furin or another cellular protease to yield the mature S1 and S2 proteins [11,12]. The S1 fragment which contains RBD (receptor-binding domain) is responsible for receptor binding while a second cleavage of S2 leads to the release of a fusion peptide after viral attachment to host cell receptor. It was reported that the SARS-CoV-2 S binds human angiotensin-converting enzyme 2 (ACE2) with higher affinity than SARS-CoV spike protein [13]. Due to the great importance in determining the host specificity and infection efficiency, the coronavirus spike glycoprotein is the key target for vaccines, therapeutic antibodies and diagnostics [14,15].

With the increasing amount of sequencing data of SARS-CoV-2 deposited in public databases, the characteristics of genomic variance of different isolates are emerging [16-19]. Here, more than 17,000 SARS-CoV-2 complete genome sequences were analyzed to provide a landscape of mutations of this novel coronavirus. Furthermore, the genome embeddings method was used to infer the genome similarity, providing novel insights into the phylogenetic origin of SARS-CoV-2.

Materials and Methods

Genomes

SARS-CoV-2 genomes submitted from 2019-12-23 to 2020-08-29, including the reference genome NC_045512.2, were downloaded from NCBI Datasets (https://www.ncbi.nlm.nih.gov/sars-cov-2/) and 17,229 complete genomes (Supplementary S1) were retrieved for downstream analysis. All the other genomes (Supplementary S2) including 267 SARS, 59 MERS, 35 bat coronavirus and one pangolin coronavirus genomes (EPI_ISL_410539), were collected from GenBank or GISAID (https://www.gisaid.org/).

Genome embedding

To characterize the genomic diversity during time, 1,928 SARS-CoV-2 isolates were sampled from the complete genomes with no more than 10 genomes on the same submitting date (Supplementary S3). Each genomic sequence was broken into a list of k-mers (k = 6) with overlapping and the step of one base from 5' to 3' end. To avoid the effect of gapped sequences, any fragment containing N (any base) was removed from the list. The genome embedding model learns to predict the central k-mer based on the whole genome embedding and the embeddings for a context window (size = 6) of k-mers on either side of the central k-mer. The embedding model was trained on 1,928 SARS-CoV-2 and 362 closely related virus genomes. After unsupervised training, the model was used to infer embeddings of genome sequences. All the genomes were embedded into the same vector space (32 dimensions), allowing comparison and inferring the distance between them. The training and inference of the embedding model was performed in Gensim (*http://radimrehurek.com/gensim/tutorial.html*) using Doc2Vec model (vector_size=32, min_count=3, window=6, epochs=30).

To visualize the distribution of the embeddings, two-dimensional t-distributed stochastic neighbor embedding (t-SNE) was then generated using a perplexity setting of 20, the learning rate of 200 and 5,000 iterations. The t-SNE was calculated and plotted with scikit-learn and matplotlib libraries of python, respectively.

Sequence logo of untranslated region

The 5' and 3'-untranslated region (UTR) were retrieved from each SARS-CoV-2 genome and aligned using MUSCLE v3.8.31 [20]. The conservation of UTR was visualized using WebLogo (*http://weblogo.threeplusone.com/*) and the second structure of 5'UTR was predicted using the RNAfold WebServer (*http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi*) with the default parameters.

Protein mutation analysis

The proteins encoded by each SARS-CoV-2 genomes were retrieved with Biopython [21]. The Multiple sequence alignments (MSA) were done with MUSCLE v3.8.31 for each kind of protein respectively and the the mutations were called based on each reference protein sequence.

The cryo-EM structure of SARS-CoV-2 spike glycoprotein was downloaded from PDB (http://www.rcsb.org/) with the ID of 6vsb. The structure and frequently mutated amino acids of the spike protein were displayed with MOLMOL [22].

Results

Genomic similarity of SARS-CoV-2 to closely related species

To estimate the similarity of SARS-CoV-2 and its related species, we performed genome embedding using Doc2vec model with Gensim. All the genomes were embedded into 32-dimensional space and the distribution was shown with t-SNE (Figure 1). Overall, the genome embeddings were clustered by viral species except a few outliers of SARS coronaviruses. All the SARS-CoV-2 genome embeddings were clustered with no pattern of submitting date. Of note, although the BatCoV RaTG13 (MN996532.1) had high identity (96%) to SARS-CoV-2 reference genome (NC_045512.2) as some SARS-CoV-2 strains, it was not found in SARS-CoV-2 cluster or nearby. The result of genome embedings indicated that SARS-CoV-2 had the same distance to the related species, which was different from alignment-based inference and needed to be carefully investigated in the future.

Figure1

Mutations in 5'UTR and 3'UTR

Although there are high frequencies of deletion and insertion found in the 3'-UTR (reference genome NC_045512.2: 29675 to 29903) and 5'-UTR (reference genome MN908947.3: 1 to 265) of SARS-CoV-2 genomes, the sequences are high conservative (Supplementary S4 & Figure 2A). It is worth noting that a mutation frequently appeared in -25 upstream of the ORF1ab start codon (Figure 2A). The most frequent mutation in this site is transition (C241T/U). Predicted second structure of 5'UTR showed that the mutation site is located in a little loop (Figure 2 B&C). Still, the effect of the mutation should be carefully investigated in the future. Investigating the source of the mutants, we found that most mutations were discovered in USA with later collection time.

Figure 2

Mutation landscape of SARS-CoV-2 proteome

To provide a mutation landscape of SARS-CoV-2 proteome, all the proteins were translated from the complete genomes and then aligned respectively. After alignments, all the mutated sites of respective proteins were analyzed by python script. It seemed that E, M, ORF6, ORF7a, ORF7b and ORF10 had high conservation while the other proteins showed more divergent. Beside the change of amino acid, lots of of deletions and insertions were found in ORF1ab and the spike protein.

All the mutations of SARS-CoV-2 proteome were showed in Supplementary S5 and some most frequent mutations was showed in Figure 3. Seven frequent point mutations were found in the large replicase polyproteins ORF1ab (T265I, L1599F, F3071Y, L3606F, P4715L, P5828L and Y5865C). One frequent mutations happened in S1 domain of the spike (S) protein (D614G) and three in the nucleocapsid (N) protein (S194L, R203K, G204R). But the other two structural proteins Envelope (E) and Membrane (M) protein were less prone to tolerate mutations. For the accessory proteins, three frequent mutations appeared in ORF3a (Q57H, G196V, G251V) and ORF8 (S24L, V62L, L84S), respectively while ORF6, ORF7a, ORF7b, ORF10 were more conservative. Of note, the R203K mutation of N protein was caused by three nucleotide mutations, which indicated strong positive selection and the significance should be investigated.

Figure 3

Spike protein mutation sites

The spike (S) glycoprotein, which mediates entry into host cells and therefore determines the specificity, is the mostly intensively investigated protein of coronavirus. The S protein is composed of the putative N-terminal signal peptide, S1 which contains receptor-binding domain (RBD) and S2. Because of many Sporadic mutations, we only showed some representative mutations frequently happened in early submitted genomes. Thanks to the cryo-EM structure of SARS-CoV-2 S proteins (PDB ID: 6vsb), all these mutated sites were analyzed from the view of 3D structure. Twelve mutations were mapped onto the structure (Figure 4) and six more mutations (L5F, N74K, Y144del., G181V, S247R, G476S) were not shown in the structure because of the resolution and sequence length. In addition to one mutation (L5F) in the signal peptide and three in S2 fragment (F797C, A930V, D936Y), fourteen mutations appeared in the S1 fragment. To be specific, four mutations (A348T, R408I, D428E, G476S) were discovered in the RBD domain (left upper corner) and ten mutations (Y28N, H49Y, L54F, N74K, Y144del., F157L, G181V, S221W, S247R, D614G) were found in other part of S1.

Figure 4

Discussion

With the increasing number of sequenced SARS-CoV-2 genomes, more and more mutations will be discovered. In this study, more than 17,000 complete SARS-CoV-2 genomes collected all over the world were analyzed to characterize the mutations on both nucleotide and protein levels. Except the deletion/insertion in the two ends of genome, a few frequent mutations were discovered. These mutations may result from the positive selection which should be carefully studied in the future. Also, the mutations may be used as marker to track the origin of different isolates and the conservative regions provide useful information to develop robust molecular diagnostics methods.

To investigate the phylogenetic of SARS-CoV-2, a Doc2vec model was used for embedding genome sequences. Doc2vec is an unsupervised learning algorithm, which is used to predict vectors to represent different documents and hence infer the similarity between them. It seems that the distance estimated from genome embedding is different from sequence alignment. Because of interspecies exchange of genetic fragments, the overall similarity of whole genomes may not sufficient to reveal the evolutionary relationships. The result of genome embedding should not be neglected, but need to be carefully investigated in the future.

Besides the genomic variation, the mutations of the encoded proteins of SARS-CoV-2 were also analyzed. Obviously, some proteins, including spike protein, showed less evolutional constrain and some frequent mutations were identified. Whether these mutations result from positive selection and the biological significance should be investigated in the future. The conservation and diversity of SARS-CoV-2 proteome will benefit discovering the infection mechanism and developing therapeutic methods.

For the most intensively studied S protein of coronavirus, all the mutations were analyzed. It seems that S protein is under fast evolution and the SBD domain is most susceptible to mutation. How these mutations determine the receptor specificity and affinity need further research. Identification of the S protein's mutations will provide the basis for optimizing the design of diagnostic, antiviral and vaccination strategies for this emerging infection.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by National Key Research and Development Programs of China No. 2017YFC1200201, Jiangsu Provincial Key Research and Development Projects No. BE2017693 and Independent Project of Chengdu Research Base of Giant Panda Breeding No. 2020CPB-C11.

Figure legends

Figure 1 t-SNE plot of genome embeddings. Genomes of different species were shown with different colors and the submitting date of different SARS-CoV-2 isolates were indicated with different gray intensity. SARS2: SARS-CoV-2; SARS: SARS-CoV; MERS: MERS-CoV; BCov: bat coronavirus; PCov: pangolin coronavirus.

Figure 2 Mutations in the 5'UTR of SARS-CoV-2 genomes. A. Sequence logo of part of 5'UTR. B&C. Predicted secondary structure of two alleles of 5'UTR. The negative number showed the distance to ORF1ab start codon and the mutated site was showed as indicated.

Figure 3 Mutations of SARS-CoV-2 proteins. For all the analyzed genomes, the encoded proteins were deduced and most frequently mutated sites were shown as indicated in each protein.

Figure 4 Mutation mapping of the spike protein. The mutation sites on the ribbon representation of SARS-CoV-2 S proteins (PDB ID: 6vsb) were shown as indicated.

All these mutations of the S protein were frequently happeded in early submitted SARS-CoV-2 genomes.

References

1. Lu R, Zhao X, Li J, Niu P, Yang B, et al. (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395: 565-574.

2. Chan JF, To KK, Tse H, Jin DY, Yuen KY (2013) Interspecies transmission and emergence of novel viruses: lessons from bats and birds. Trends Microbiol 21: 544-555.

3. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, et al. (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579: 270-273.

4. Lam TT, Shum MH, Zhu HC, Tong YG, Ni XB, et al. (2020) Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature.

5. Azzi A, Lin SX (2004) Human SARS-coronavirus RNA-dependent RNA polymerase: activity determinants and nucleoside analogue inhibitors. Proteins 57: 12-14.

6. Ma Y, Wu L, Shaw N, Gao Y, Wang J, et al. (2015) Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. Proc Natl Acad Sci U S A 112: 9436-9441.

7. Yuan Y, Cao D, Zhang Y, Ma J, Qi J, et al. (2017) Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. Nat Commun 8: 15092.

8. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, et al. (2020) Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 181: 281-292 e286.

9. Walls AC, Xiong X, Park YJ, Tortorici MA, Snijder J, et al. (2019) Unexpected Receptor Functional Mimicry Elucidates Activation of Coronavirus Fusion. Cell 176: 1026-1039 e1015.

10. Forni D, Cagliani R, Clerici M, Sironi M (2017) Molecular Evolution of Human Coronavirus Genomes. Trends Microbiol 25: 35-48.

11. Follis KE, York J, Nunberg JH (2006) Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. Virology 350: 358-369.

12. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, et al. (2020) The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res 176: 104742.

13. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, et al. (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 367: 1260-1263.

14. Du L, He Y, Zhou Y, Liu S, Zheng BJ, et al. (2009) The spike protein of SARS-CoV–a target for vaccine and therapeutic development. Nat Rev Microbiol 7: 226-236.

15. Du L, Yang Y, Zhou Y, Lu L, Li F, et al. (2017) MERS-CoV spike protein: a key target for antivirals. Expert Opin Ther Targets 21: 131-143.

16. Ceraolo C, Giorgi FM (2020) Genomic variance of the 2019-nCoV coronavirus. J Med Virol 92: 522-528.

17. Chan JF, Kok KH, Zhu Z, Chu H, To KK, et al. (2020) Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect 9: 221-236.

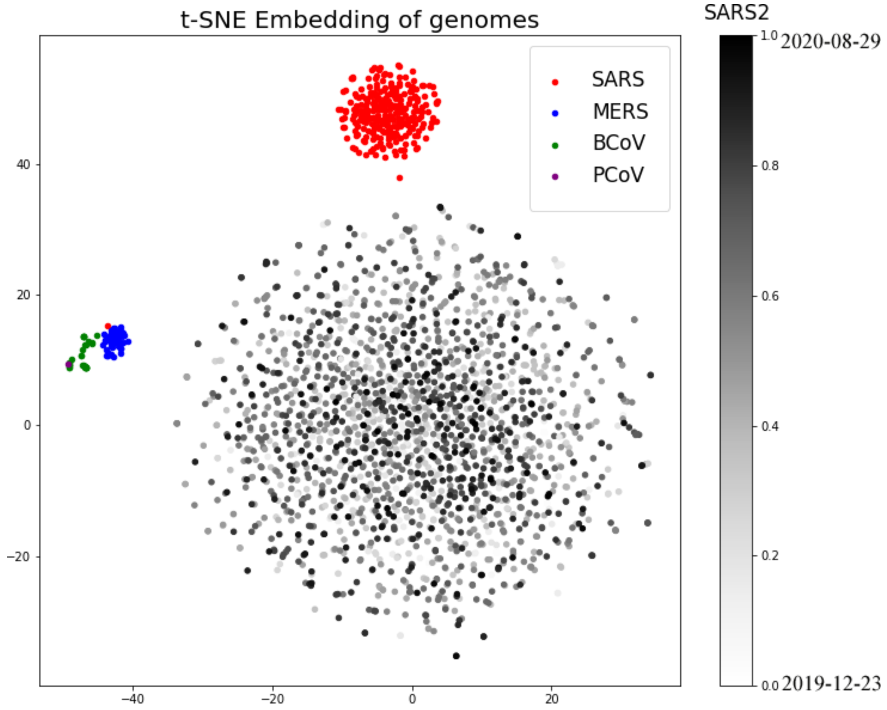
18. Shen Z, Xiao Y, Kang L, Ma W, Shi L, et al. (2020) Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. Clin Infect Dis.

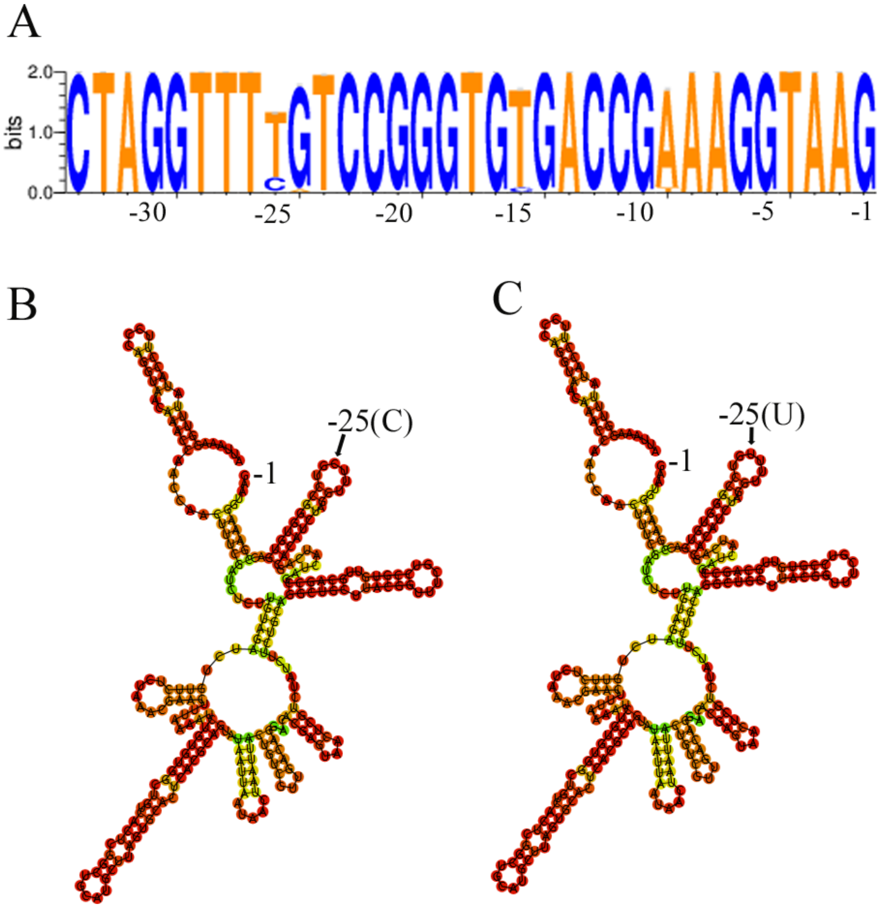
19. Khailany RA, Safdar M, Ozaslan M (2020) Genomic characterization of a novel SARS-CoV-2. Gene Rep 19: 100682.

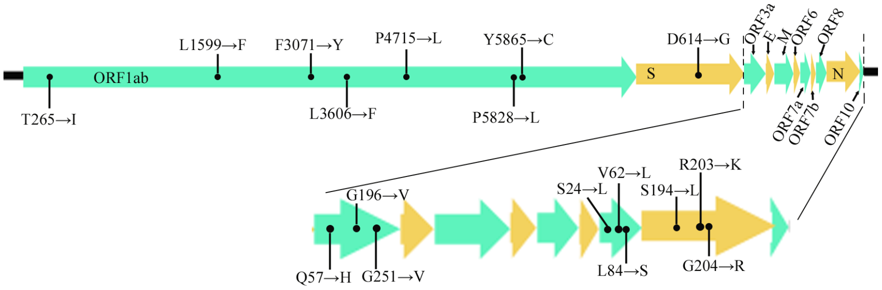
20. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

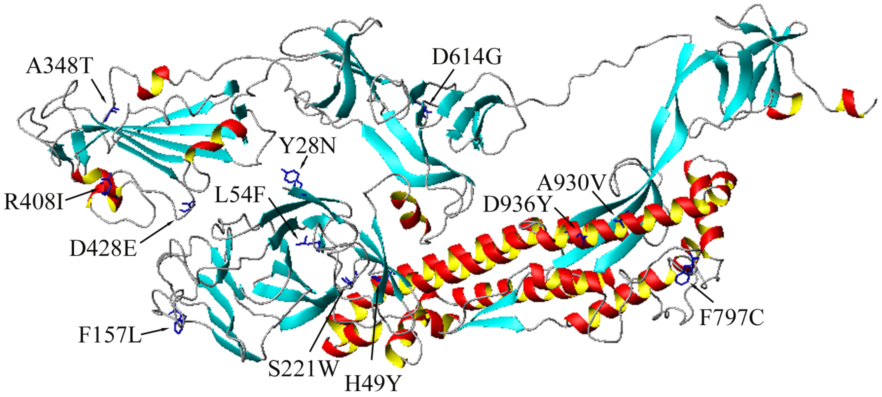
21. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422-1423.

22. Koradi R, Billeter M, Wuthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 14: 51-55, 29-32.




A
2.0
1.0
0.0
bits
CTAGGTTTTGTCCGGGTGTGACCGAAAGGTAAG
C
-30
-25
-20
-15
-10
-5
-1
B
-25(C)
-1
C
-25(U)
-1


ORF1ab
L1599→F
F3071→Y
P4715→L
Y5865→C
D614→G
T265→I
L3606→F
P5828→L
S
ORF3a
E
M
ORF6
ORF8
N
ORF7a
ORF7b
ORF10
G196→V
S24→L
V62→L
S194→L
R203→K
Q57→H
G251→V
L84→S
G204→R


A348T
D614G
Y28N
R408I
L54F
A930V
D936Y
D428E
F157L
F797C
S221W
H49Y