On December 31, 2019 the Chinese Center for Disease Control (CDC) reported a cluster of severe pneumonia cases of unknown etiology in the city of Wuhan, China. Shortly thereafter, public health professionals identified the likely causative agent to be a novel betacoronavirus (SARS-CoV-2). The virus has been spreading rapidly worldwide. As of May 3, 2020, the WHO reports over 3.3 million confirmed cases with almost 240,000 deaths from 213 countries, areas or territories. These numbers are changing rapidly. For up-to-date information about the COVID-19 pandemic, see the WHO website - https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

Comparative genomics

In order to understand the origins and evolution of SARS-CoV-2, the ViPR team has been performing comparative genomics analyses on this new coronavirus. In this report, we present an updated phylogenetic analysis of the SARS-CoV-2 genomes.

Genome sequence source and quality control

A total of 3500 sequence records were obtained from the ViPR SARS-CoV-2 portal and GISAID on April 1, 2020 for this analysis:

Data from both sources were processed using the in-house SequenceQC program which trims off terminal Ns, removes gaps (spaces and dashes), filters genomes based on the minimum length requirement (29,000 nt was used in this study), and filters out genomes with excessive consecutive Ns (10 nt was used here). The data sanitization process cleaned 1416 sequences with terminal Ns and 220 sequences with gaps from GISAID. After this data sanitization and filtering, 2275 sequences were selected for comparative analysis.

Metadata for ViPR sequences was downloaded from the ViPR website. Metadata for GISAID sequences was manually curated from the acknowledgement table downloaded from GISAID.

Sequence clustering

In order to limit the impact of rare random mutations on the phylogenetic analysis, the QC-ed dataset was clustered with the cd-hit program to remove highly similar sequences. Various
sequence identity thresholds were tried. The results reported here were obtained from cd-hit clustering based on 0.9999 sequence identity, which resulted in the selection of 725 representative sequences.

**Sequence alignment**

The selected sequences were aligned with MAFFT FFT-NS-1 mode. The computed alignment was further trimmed at both the 5’ and 3’ ends to remove terminal gaps in the alignment.

**Phylogenetic tree**

The curated sequence metadata and sequence alignment were uploaded to the ViPR workbench and subsequently used to compute the phylogenetic tree with the FastME program in ViPR. MT233522 had an unexpectedly long branch length that may be attributed to many intermittent Ns in the sequence and was deleted from the tree.

The phylogenetic tree indicates that the SARS-CoV-2 sequences are highly similar to each other, shown by the very short branch lengths in the SARS-CoV-2 subtree (Figure 1A, in red). The SARS-CoV-2 subtree is most similar to a coronavirus isolated from bats in China in 2013, indicating that they share a common ancestor. Although 6 betacoronaviruses from pangolin are also closely related to human SARS-CoV-2 isolates, they are isolated from different provinces in different years and appear to have two distinct origins than SARS-CoV-2.

The SARS-CoV-2 subtree colored by region (Figure 1B) suggests some level of clustering by geographic region, probably reflecting founder effects.

**Sequence variations**

The extent of SARS-CoV-2 genome sequence variation was evaluated by polymorphism analysis. 2268 SARS-CoV-2 human isolates out of the selected 2275 genomes were included in the analysis. The genomes were first aligned with the MAFFT FFT-NS-1 mode. The resulting alignment was trimmed to remove gap regions at both ends, and then analyzed using an in-house SNP analysis tool. The SNP program calculates an entropy score for each aligned position using the following formula:

\[
S = -100 \times \text{Sum} (\Pi_i \times \log_2 \Pi_i)
\]

where \( \Pi_i \) is the frequency of the \( i \)-th allele. Among the ambiguous residues, \( \text{N} \) is excluded from entropy calculation while the others (e.g., \( \text{R, W, Y} \)) are treated as separate valid states.

The SNP positions were mapped to Wuhan-Hu-1 (MN908947) to facilitate comparison. A total of 3020 positions showed some level of sequence variations, with 60% (1817) having an entropy score of < 1.0 which corresponds to a single sequence with the substitution (Figure 2A). The single occurrence of substitutions suggests that these are either sequencing errors or random substitutions experiencing either neutral or purifying selection.
The highest entropy score 100.18 is found at position 23403 (23480 in the alignment used to calculate the tree), with the residue distribution of 1313 A, 950 G, 5 R (A or G) (Figure 1C). Examination of the phylogenetic tree by isolation year and month shows that this substitution first appeared in a strain from Shanghai, February 6, 2020, and another strain from Wuhan, February 7, 2020, shown in red branch and green label in Figure 1D. By the last week of February, this substitution was found in another 10 strains from Belgium, France, Italy, Switzerland and Mexico (Figure 1E). The Switzerland and Mexico strains belong to the same cluster, while other strains are dispersed in the tree. In March, an entire lineage of 253 isolates had gained the A23403G substitution. This substitution was also observed in 4 isolates in a different lineage. Further analysis of this substitution at the amino acid level shows that it leads to a D614G substitution in the middle of the S protein. Our previous study had identified potential targets for immune responses to SARS-CoV-2 (Grifoni et al, 2020). The D614G substitution that emerged in February 2020 and quickly gained prevalence afterwards is located within one of the predicted B-cell epitopes (Figure 2B). The fact that the substitution appeared a month after the onset of the outbreak and is positioned within a predicted B-cell epitope suggests that this position may have experienced immune selection pressure or that it confers some other selective advantage to S glycoprotein function. The occurrence of the same substitution in multiple lineages in the phylogenetic tree could be the evidence for convergent evolution.

Among the other positions with high polymorphism scores, positions 8782 (orf1ab, synonymous, Figure 1G, 8810 in the alignment used to calculate the tree) and 28144 (ORF8 S84L, Figure 1H, 28768 in the alignment used to calculate the tree) appear to be correlated with each other and have been used to define the S and L types in a previous study (Tang et al, 2020).

References


ViPR SARS-CoV-2 portal: https://www.viprbrc.org/brc/home.sp?decorator=corona_ncov
Figure 1. Phylogenetic analysis of SARS-CoV-2. Full length genome sequences of SARS-CoV-2 after sequence QC and clustering were selected for phylogenetic analysis using FastME following MAFFT sequence alignment and manual editing. A. Tree labels are colored by isolation host. B-H. SARS-CoV-2 subtree from A. B. Tree labels are colored by isolation region. C-F. Tree branches and labels are colored by residue at position 23480 (23403 in Wuhan-Hu-1). Isolates from January 2020, February 2020 and March 2020 are overlaid in green in D, E and F, respectively. The first occurrences of A23403G are marked with arrows in E. G-H. Tree branches and labels are colored by residue at position 8810 (8782 in Wuhan-Hu-1) and 28768 (28144 in Wuhan-Hu-1), respectively.
Figure 2. Genetic variations of SARS-CoV-2 human strains. A. Full length genome sequences of SARS-CoV-2 from human host after sequence QC were selected for sequence variation analysis using an in-house SNP program following MAFFT sequence alignment and manual editing. Sequence variation scores are plotted on the Wuhan-Hu-1 (MN908947) coordinates. B. A new substitution of D614G in the S protein (A23403G in genome) was found in SARS-CoV-2 human strains from February 2020 and onward. This substitution is located within a predicted B-cell epitope defined by our previous study (Grifoni et al, 2020) as shown in this alignment.