[Paper Research]

about Insights on early mutational events in SARS CoV-2 virus reveal founder effects across geographical regions

Name: Haidy Essam Elhamady
Prof: Dr/Sara elmetwaly

ABSTRACT

In this study, we characterized the early mutational events across 50 illumina high-quality datasets publicly available on the sequence read archive repository. A total of 30 out of 50 samples (60%) contained at least a single founder variant and most of the variants across samples are missense (over 63%). SARS-CoV-2 founder variants in WA State and USA are dissimilar to Australian SARS-CoV-2 founder variants, which were found to be heterogeneous. However, a mutational signature from USA mutations was found in an Australian sample, suggesting a world-wide spread of this molecular signature consisting of five-point variants. Remarkably, mutations in the helicase and ORF1ab proteins of the virus were found more frequently than others, suggesting that these regions continue to actively evolve. As proof of the latter.

Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus .Most people who fall sick with COVID-19 will experience mild to moderate symptoms and recover without special treatment. positive-stranded RNA virus with a large genome belonging to the family Coronaviridae, order Nidovirales .It is rapidly spreading worldwide, greatly surpassing the 8,000 total cases of the 2002–2004 SARS coronavirus outbreak (SARS-CoV-1) after 1 month of the initially identified case in China. The virus that causes COVID-19 is mainly transmitted through droplets generated when an infected person coughs, sneezes, or exhales.it is a threat to the global population. It is critical to understand SARS-CoV-2 characteristics to deal with this ongoing pandemic and to develop future treatments.

Releted work

The COVID-19 pandemic caused by a novel 2019 SARS coronavirus, known as SARS-CoV-2, is rapidly spreading worldwide, greatly surpassing the 8,000 total cases of the 2002–2004 SARS coronavirus outbreak (SARS-CoV-1) after 1 month of the initially identified case on 31 December 2019, in Wuhan **(Wilder-Smith, Chiew & Lee, 2020)**

A SARS-CoV-2 virus is an enveloped, positive-stranded RNA virus with a large genome (29.9 kb) belonging to the family Coronaviridae, order Nidovirales **(De Wit et al., 2016)**

One of the striking genomic features of this novel virus is the presence of a novel furin-like cleavage site in the S-protein of the virus, which differs from SARS-CoV-1.

(Coutard et al., 2020; Wu et al., 2020a)

(Firstly, it was suggested that SARS-CoV-2 is relative of the RaTG13 bat-derived coronavirus rather than of SARS-CoV-1).(in the early beginning of the outbreak in China, sequencing the virus from nine patients from Wuhan in China revealed 99.9% similarity among samples. That finding suggests 2019-nCoV originated from one source within a very short time, supporting clonality of spreading) **(Lu et al., 2020)**

Due to this association with bat coronaviruses, it was also argued that SARS-CoV-2 virus has the potential to spread into another species, as bat coronaviruses do **(Hu et al., 2018)**

Although bats are likely natural reservoir hosts for SARS-CoV-2, it was recently demonstrated that SARS-CoV-2 is closely related to a pangolin coronavirus (Pangolin-CoV),the closest relationship found so far for SARS-CoV-2 .

(Zhang, Wu & Zhang, 2020)

In that study, genomic analyses revealed that the S1 protein of Pangolin-CoV is related closer to SARS-CoV-2 than to RaTG13 coronavirus. Also, five key amino acid residues involved in the interaction with the human ACE2 receptor are maintained in Pangolin-CoV and SARS-CoV-2, but not in RaTG13 coronavirus. **(Zhang & Holmes, 2020)**

Recently, thousands of GenBank sequences from SARS-CoV-19 available at the NCBI virus database were trackable by region, suggesting that the transmission occurred mainly through clonal events due to clustering of the available sequences.

(Chen, Allot & Lu,2020; Kupferschmidt, 2020)(https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Protein&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049)

Results

Inspection of variants reveals well-defined signatures with founder effects across sequenced samples We aimed to call variants of SARS-CoV-2 datasets sequenced with the Illumina technology, due to its depth and sequencing quality, in terms of error rate (Nielsen et al., 2011). As of 27 March 2020 we obtained 282 accession numbers for SARS-CoV-2, from the sequence read archive, containing 27 illumina datasets. By searching in Sequence Read Archive repository (SRA) we added 24 more datasets yielding in total 51 illumina raw datasets to analyse (see Table S1, sheets 1 and 2). From this list, we excluded the Chinese Sample nCoV5 (SRR11059943) due to large gaps in genome coverage, as explained in (Shen et al., 2020) and we subsequently worked with these illumina datasets. We aligned each fastq reads against the SARS-CoV-2 reference genome NC_045512.2, corresponding to the initial isolate Wuhan-Hu-1. We checked coverage of each sample by using the Integrative Genomics Viewer tool (Robinson et al., 2011) and samtools (Table S1, sheet 2). Variant calling in each sample by using Strelka2 reveals a diverse number of variants per sample, yielding 137 single nucleotide polymorphisms (SNVs) and nine indels (see Fig. 1A; Table S1, sheet 3). Founder variants were obtained by doing variant calling with bcftools after strict filtering (see Table S1, sheet 2). Remarkably, thirteen out of fourteen datasets from the USA-WA State study (hereafter referred as USA-WA) displayed variants presenting a defined variant signature, consisting in a core of five founder variants at positions 8,782, 17,747, 17,858, 18,060 and 28,144 in the SARS-CoV-2 reference genome, also detected within the 137 SNVs from the next generation sequencing datasets (see Fig. 1B; Table S1, sheets 2 and 3). Mutational landscape analysis of SARS-CoV-2 samples in Australia (Australia-VIC samples, hereafter Australia-VIC) demonstrated that these samples were clearly heterogeneous, displaying a variety of founder mutations per sample but also shared variants were observed within samples. One variant (position 26,144) is present in 5/11 Australia-VIC samples and variants 8,782 and 28,144 from USA-WA signature are also present in 3/11 Australian samples (see Fig. 1C; Table S1, sheet 5). Notably, one Australian-VIC sample displayed the same five-point variant signature of USA-WA samples, two samples contain the same variant signature presenting one deletion and one novel signature presents a SNP that creates a stop codon (see Table S1, sheet 5). All of these called variants present mutant allele frequencies near or equal to 100%, evidenced in the number of mutant_alleles/reference_alleles (see mutant allele frequency in Table S1, sheets 4 and 5, respectively) easily visualized in the aligned bam files (see Fig. 1C from USA-WA and Fig. 1D for Australian-VIC samples, respectively). These analyses suggest that these variants were already spread in the infected population in the early days of the outbreak, they are not restricted by country and that they will continue to spread along with the growing cases. To support the latter, as of 22 April 2020 we downloaded 1,599 GenBank sequences of SARS-CoV-2 from Asia, Europe and North America origin, respectively and we aligned them against the SARS-CoV-2 reference genome. A phylogenetic tree was constructed from all genbank sequences and depict a mixed clustering of sequences between Asia, Europe and North America, supporting the existance of different viral signatures. (Fig. S1). Variant calling from these alignment reveals the substantial presence of USA-WA signature in North America sequences, with allele

frequencies (AF) ranging 33–39% (see Fig. 1E; Table S1, sheet 6). Variants 8,782 and 28,144 from USA-WA signature are also present in Asia and Europe, suggesting these two variants arises in the beginning of the pandemic and spread worldwide. Thus, the USA-WA signature is likely widespread among SARS-CoV-2 infections in USA due to founder effect. Also, in North America GenBank sequences, six more variants were detected with similar allele frequencies as reported for USA-WA variants. A summary of the USA-WA mutational signature is depicted in Table 1. Less founder variants with lower allele frequencies are present in Asian samples suggesting high clonality of the original strain of SARS-CoV-2. Conversely, in Europe 16 founder variants with higher allele frequencies were found, supporting SARS-CoV-2 evolves as the pandemic spread.