JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH | Center for Health Security

SARS-CoV-2 Genetics

Updated April 16, 2020

Key Findings for Public Health

- The outbreak was initiated from either a single introduction into humans or very few animal-to-human transmission events.
- SARS-CoV-1 and SARS-CoV-2 use the same cellular receptor, ACE2, which could be used as a starting point for creating therapeutics for SARS-CoV-2.

Background

Coronaviruses, including the pneumonia-causing novel coronavirus currently known as SARS-CoV-2, are enveloped, nonsegmented, positive-sense RNA viruses. Coronavirus genomes have some of the largest genomes among RNA viruses, with approximately 25-32 kilobases.[1] The typical CoV genome includes a 5'-cap, 5'-untranslated region (UTR), open reading frames, a 3'-UTR, and 3'-poly(A) tail. The first two thirds of the genome typically codes for nonstructural proteins from 2 open reading frames that form the replicase complex. The last third of the genome encodes primarily structural proteins.[2] There are 4 conserved structural proteins across CoVs: the spike (S) protein, membrane (M) protein, envelope (E) protein, and nucleocapsid (N) protein.[1] The S protein is responsible for binding to host cell receptors and viral entry to host cells. The M, E, and N proteins are part of the nucleocapsid of viral particles.

SARS-CoV-2 Naming

In a paper published early in the pandemic,[3] viral sequences collected from the earliest patients were assessed and compared to known viral sequences. Sequence analysis of 11 samples found that SARS-CoV-2 is in the same species as SARS-CoV; the 2 viruses are 94.6% similar in amino acid sequence (80% nucleotide sequence similarity) across the genome.[3] However, other studies from early in the outbreak do not consider the viruses to be the same species, as they differ by more than 10% in the replicase genes.[4] In February, the Coronavirus Study Group (CSG) of the International Committee on Taxonomy of Viruses officially named the novel coronavirus SARS-CoV-2. The CSG analyzed viral genomes from several patients and assessed phylogenetic (evolutionary) relationships between the new virus and known coronaviruses. The committee found that the genome of viruses isolated from patients was similar enough to SARS genomes to be considered a variant of SARS, not an entirely novel virus. While the clinical presentation, epidemiologic patterns, and host range of SARS-CoV-2 may differ from the original SARS-CoV, it is the *genetic* similarity between the 2 viruses that is used to conclude they are the same species. For this reason, the CSG has named SARS-CoV and SARS-CoV-2 as variants of the species known as *Severe acute respiratory syndrome–related coronaviruses*. The name SARS-CoV-2 is distinct from the name of the disease, which the WHO has officially designated COVID-19.

Coronaviruses have a genome made of RNA. Viruses with RNA genomes have an essential gene called the RNA-dependent RNA polymerase (RdRp), which is highly conserved, meaning that there are few changes in the gene from one RNA virus to another. This makes the gene useful for measuring the evolutionary distance and relatedness of one RNA virus to another. The CSG found that there were fewer differences in the RdRp gene when comparing the SARS-CoV and SARS-CoV-2 genomes than between variants of MERS or HCoV-OC43—that is, SARS-CoV and SARS-CoV-2 were more closely related. However, there are relatively few samples of SARS-CoV from the 2003 outbreak available for use in analysis, which could bias calculations of genetic distance between variants within and between species of coronaviruses.

RNA viruses have high mutation rates that result in several slightly different versions of the viral genome being made each time the viral genome is replicated (Figure 1). This creates a viral population with diverse genomes, known as a quasispecies. With each viral replication cycle, the differences accumulate between the original viral genome and the progeny viral genomes. This may contribute to differences in clinical outcomes between patients, as the viral populations that are infecting them are slightly different. The fact that there are multiple versions of the viruses also makes it challenging to categorize viruses as different species. Instead, the genomes will have varying levels of relatedness to one another. This is why the CSG assessed the relative levels of relatedness between and within SARS-CoV and SARS-CoV-2 and compared it to the level of relatedness between other and within other coronavirus species.

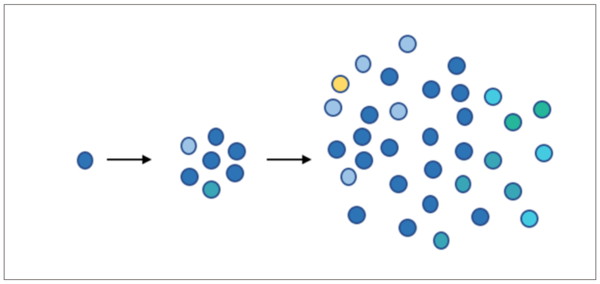

Figure 1. This diagram illustrates the generation of a quasispecies from one initial virus.

SARS-CoV-2 Evolution

Phylogenetic analysis of 30 publicly available SARS-CoV-2 samples concluded that emergence of SARS-CoV-2 in the human population likely occurred in mid-November 2019.[5] The sequences have limited variability in consensus sequences, suggesting the outbreak was initiated from either a single introduction into humans or from a very few animal-to-human transmission events.[6] The mutation rate has been estimated in various groups, ranging from about $1.05x10^{-3}$ to$1.26x10^{-3}$ substitutions per site per year, which is similar to some estimates of MERS-CoV mutation rates.[5,7-9] As more viral genomes are made publicly available, scientists will better be able to track viral evolution and mutation rates, so the exact estimates will vary.

Selection analysis of the genome suggests that 2 genes in the SARS-CoV-2, the S and N genes, are under episodic selection as the virus is transmitted between humans.[10] This is normal for emerging viruses and means that parts of the genome are undergoing positive selection.[11,12] Mutations and adaptation in the S and N genes could affect virus stability and pathogenicity.[9] As more genomes are made publicly available, analysis of the genome sequence diversity across samples has revealed the highest diversity occurring in the structural genes, especially the S protein, ORF3a, and ORF8.

SARS-CoV-2 is evolving over the course of the pandemic. However, this evolution is not occurring faster than expected compared to other viruses during an outbreak. There are different clades of SARS-CoV-2 developing as COVID-19 spread across the globe.[13] Different clades emerge as viruses evolve. This is entirely normal and does not mean there are new strains of SARS-CoV-2 that are more pathogenic than others circulating right now.

Scientists have done an incredible job sequencing samples of SARS-CoV-2 and sharing results during this pandemic. These sequences are allowing public health officials to estimate several important parameters of the epidemiology of COVID-19, such as the reproductive number and introduction of the virus into new regions.

References

1. Masters PS. Coronavirus genomic RNA packaging. *Virology* 2019;537:198-207. doi:10.1016/j.virol.2019.08.031
2. Lo C-Y, Tsai T-L, Lin C-N, Lin C-H, Wu H-Y. Interaction of coronavirus nucleocapsid protein with the 5'- and 3'-ends of the coronavirus genome is involved in genome circularization and negative-strand RNA synthesis. *FEBS J* 2019;286(16):3222-3239. doi:10.1111/febs.14863
3. Zhou P, Yang X-L, Wang X-G, et al. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. *Microbiology* 2020. doi:10.1101/2020.01.22.914952
4. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020. doi:10.1056/NEJMoa2001017
5. Rambaut A. Phylodynamic analysis of SARS-CoV-2 genomes- 27-Jan-2020. *Virological* January 27, 2020. http://virological.org/c/novel-2019-coronavirus/33/l/latest. Accessed January 28, 2020.
6. Bedford T, Richard N, Hadfield J, Hodcroft E, Muller N, Llcisin M. Narrative: Genomic analysis of nCoV spread. Situation report 2020-01-23. nextstrain. https://nextstrain.org/narratives/ncov/sit-rep/2020-01-23?n=1. Published January 23, 2020. Accessed January 24, 2020.
7. Cotten M, Watson SJ, Zumla AI, et al. Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. *mBio* 2014;5(1). doi:10.1128/mBio.01062-13
8. Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. *eLife* 2018;7:e31257. doi:10.7554/eLife.31257
9. Baric RS, Yount B, Hensley L, Peel SA, Chen W. Episodic evolution mediates interspecies transfer of a murine coronavirus. *J Virol* 1997;71(3):1946-1955. doi:10.1128/ JVI.71.3.1946-1955.1997
10. Benvenuto D, Giovanetti M, Ciccozzi A, Spoto S, Angeletti S, Ciccozzi M. The 2019-new Coronavirus epidemic: evidence for virus evolution. *bioRxiv* January 2020. doi:https://doi.org/10.1101/2020.01.24.915157
11. Sironi M, Cagliani R, Forni D, Clerici M. Evolutionary insights into host–pathogen interactions from mammalian sequence data. *Nat Rev Genet* 2015;16(4):224-236. doi:10.1038/nrg3905
12. Nextstrain / ncov (2/3/2020). nextstrain. https://nextstrain.org/ncov?m=num_date. Accessed February 3, 2020. February 3, 2020.
13. Bell SM, Müller N, Wagner C, et al. Narrative: genomic analysis of COVID-19 spread. Situation report 2020-04-10. nextstrain April 10, 2020. https://nextstrain.org/narratives/ncov/sit-rep/2020-04-10. Accessed April 11, 2020.