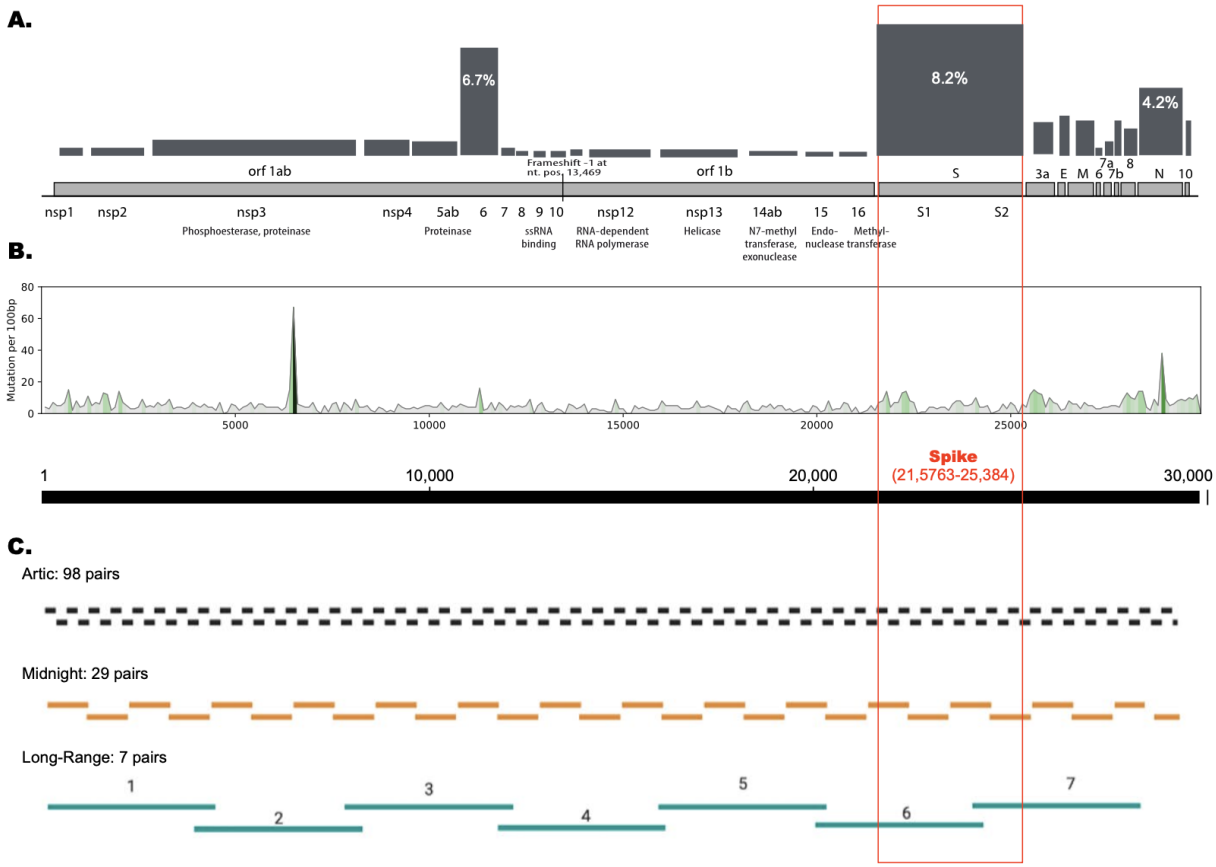
What can we learn from millions of viral genome sequences?

openaccessgovernment.org/article/what-can-we-learn-from-millions-of-viral-genome-sequences/198833

Emily Warrender September 26, 2025



David Ussery and Pratul Agarwal, Professors in the Department of Physiological Sciences at Oklahoma State University, discuss their work using high-performance computing for the analysis of millions of viral genome sequences

Advances in nanotechnology and machine learning methods have made possible disruptive changes in genome sequencing; it is now possible to sequence hundreds of viral genomes in parallel, within a few hours[1], allowing for the routine monitoring of viruses in water[2], and other environments[3]. The dramatic improvements in ease and cost of sequencing have resulted in millions of genome sequences being deposited in public databases[4]. Currently, there are several viral species with more than a million genomes: >9 million SARS-CoV-2 viral genomes available at NCBI[5]; >1.5 million Influenza viral genomes[6];and > 1.1 million human immunodeficiency (AIDS) viral genomes[7].

Lessons from millions of SARS-CoV-2 genomes

In principle, genomic epidemiology can follow viral outbreaks in near-real time, as they occur, and make predictions, kind of like the weather forecast for the next few days – here's where the virus is today, and where it's likely to spread tomorrow.

The SARS-CoV-2 outbreak was the first pandemic with large-scale viral genome sequencing, allowing the tracking of novel variants as they appeared. There are several useful and important observations from analysis of millions of SARS-CoV-2 genome sequences. Here we will discuss three lessons we've learned from looking at millions of viral genome sequences, as shown in the Figure.

1. At the protein level, only a few viral proteins have significant mutational frequencies; changes in non-synonymous mutations can predict possible surges in new variant strains. Based on more than 8 million SARS-CoV-2 genome sequences, a comprehensive overview of all 26 proteins encoded in the SARS-CoV-2 genome has been mapped[8], and regularly updated[9] (https://pandemics.okstate.edu/covid19/). Panel A in the figure shows the relative mutational frequency of the proteins encoded along the SARS-CoV-2 genome. The height of the bar represents the number of mutations in amino acids for each protein, normalized to the total length of the protein. This is a modification of Figure 4 from Najar et al. (2023)[9]. Note that the spike protein has the highest mutational frequency (8.2%), followed by the nucleocapsid protein (6.7%), and then the envelope protein(4.2%); the 23 remaining proteins all have mutational frequencies of less that 3%.
2. At the genomic RNA level, SARS-CoV-2 has the few mutational hot-spots, but most of the genome is relatively stable, with a limited mutational repertoire. This is represented by Panel B of the figure, which shows the relative mutational frequency of the genome; this is a modification from Figure 3 of our work describing a comparison of about 6 million Covid-19 genomes[10]. At the time, we saw changes in about 5% of the genome, but as can be seen from the figure, the frequency of these changes varies throughout the genome, with a few spikes localized at specific regions along the chromosome, representing mutational 'hot-spots'.

3. It is possible to design optimal primers in stable regions of the genome, which will be resistant to mutations in the viral genome over time. One of the major problems with large-scale sequencing of the SARS-CoV-2 genomes during the Covid-19 outbreak was the failure of primers, due to mutations in the primer-binding regions. This caused problems with the detection of specific variants if the primers did not bind to that region.The most commonly used procedure (Arctic11) requires 98 different pairs of primers (196 primers in total!), which allows tiling across the entire genome of about 30,000 nt. Fewer primers means fewer chances of mutations in the primer binding sites. The “Midnight” protocol12 takes advantage of longer reads using third-generation sequencing, and only requires 29 sets of primers. We have designed a set of 7 primers that can be used for long-range sequencing of the entire genome, avoiding primer failure in variable regions of the genome13. This is summarized in Panel C of the figure, which is modified from Figure 1 in Kandel et al., 2024[13].

What does the future look like in terms of genomic epidemiology?

The future of viral genomics is headed towards ‘big data’, in terms of the vast amount of viral genomes becoming available. It might be possible to predict new viral outbreaks, just like weather reports, on a daily basis. In addition to the explosion of (DNA) viral genomes, third-generation sequencing technologies allow for the direct sequencing of RNA viruses (dRNAseq), including detection of modified bases in the RNA genomes[14]. Direct RNA sequencing of SARS-CoV-2 genomes continues to improve[15,16], and is becoming a fast and economical method of choice. Third-generation sequencing technologies will allow quantification of viral genome length distributions, as well as alternative splicing and epigenetic modifications.

References:

1. Park SY, Faraci G, Ward P, Lee HY.2024.Utilizing cost-effective portable equipment to enhance COVID-19 variant tracking both on-site and at a large scale. J Clin Microbiol 62:e01558-23. https://doi.org/10.1128/jcm.01558-23
2. Werner, D., Acharya, K., Blackburn, A., Zan, R., Plaimart, J., Allen, B., Mgana, S. M., Sabai, S. M., Halla, F. F., Massawa, S. M., Haile, A. T., Hiruy, A. M., Mohammed, J., Vinitnantharat, S., Thongsamer, T., Pantha, K., Mota Filho, C. R., & Lopes, B. C. (2022). MinION Nanopore Sequencing Accelerates Progress towards Ubiquitous Genetics in Water Research. Water, 14(16), 2491. https://doi.org/10.3390/w14162491
3. Buddle, S., Forrest, L., Akinsuyi, N. et al. Evaluating metagenomics and targeted approaches for diagnosis and surveillance of viruses. Genome Med 16, 111 (2024). https://doi.org/10.1186/s13073-024-01380-x
4. see for example https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/
5. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2697049

6. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:197911&VirusLineage_ss=taxid:197912&VirusLineage_ss=taxid:197913&VirusLineage_ss=taxid:1511083
7. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Human%20immunodeficiency%20virus%201,%20taxid:11676
8. Fares Z Najar Evan Linde Chelsea L Murphy Veniamin A Borin Huan Wang Shozeb Haider Pratul K Agarwal (2023) Future COVID19 surges prediction based on SARS-CoV-2 mutations surveillance eLife 12:e82980. https://elifesciences.org/articles/82980
9. Fares Z. Najar, Chelsea L. Murphy, Evan Linde, Veniamin A. Borin, Huan Wang, Shozeb Haider, Pratul K. Agarwal. Pandemic preparedness through genomic surveillance: Overview of mutations in SARS-CoV-2 over the course of COVID-19 outbreak. bioRxiv 2023.08.12.553079; doi: https://doi.org/10.1101/2023.08.12.553079
10. Trudy M Wassenaar, Visanu Wanchai, Gregory Buzard, David W Ussery, The first three waves of the Covid-19 pandemic hint at a limited genetic repertoire for SARS-CoV-2, FEMS Microbiology Reviews, Volume 46, Issue 3, May 2022, fuac003, https://doi.org/10.1093/femsre/fuac003
11. Itokawa K, Sekizuka T, Hashino M, Tanaka R, Kuroda M (2020) Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. PLoS ONE 15(9): e0239403. https://doi.org/10.1371/journal.pone.0239403
12. Pembaur, A., Sallard, E., Weil, P. P., Ortelt, J., Ahmad-Nejad, P., & Postberg, J. (2021). Simplified Point-of-Care Full SARS-CoV-2 Genome Sequencing Using Nanopore Technology. Microorganisms, 9(12), 2598. https://doi.org/10.3390/microorganisms9122598
13. Kandel S, Hartzell SL, Ingold AK, Turner GA, Kennedy JL and Ussery DW (2024) Genomic surveillance of SARS-CoV-2 using long-range PCR primers. Front. Microbiol. 15:1272972. doi: https://doi.org/10.3389/fmicb.2024.1272972
14. Kim, D., Lee, J.Y., Yang, J.S., Kim, J.W., Kim, V.N., Chang, H., The Architecture of SARS-CoV-2 Transcriptome, Cell, 181:914-920, (2020). https://doi.org/10.1016/j.cell.2020.04.011
15. van der Toorn, W., Bohn, P., Liu-Wei, W. et al. Demultiplexing and barcode-specific adaptive sampling for nanopore direct RNA sequencing. Nat Commun 16, 3742 (2025). https://doi.org/10.1038/s41467-025-59102-9
16. Hottel W, Reeb V, Twait E, Zanon K, Hwang M, Choi H, Chatterjee P, Xiang J, Meier J, Pentella M, McIndoo E, Ammons MCB, Jinadatha C, Stapleton JT. 2025. Direct comparison of clear DX Nanopore and Illumina sequencing of SARS-CoV-2. Microbiol Spectr 13:e00427- 25. https://doi.org/10.1128/spectrum.00427-25

Primary Contributor

David Ussery
Oklahoma State University

Additional Contributor(s)

Pratul Agarwal
Oklahoma State University