Mapping Data to Deep Understanding: Making the Most of the Deluge of SARS-CoV-2 Genome Sequences

Bahrad A. Sokhansanj, Gail L. Rosen

Drexel University, Ecological and Evolutionary Signal-Processing and Informatics Laboratory, Department of Electrical & Computer Engineering, College of Engineering, Philadelphia, Pennsylvania, USA

ABSTRACT

Next-generation sequencing has been essential to the global response to the COVID-19 pandemic. As of January 2022, nearly 7 million severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequences are available to researchers in public databases. Sequence databases are an abundant resource from which to extract biologically relevant and clinically actionable information. As the pandemic has gone on, SARS-CoV-2 has rapidly evolved, involving complex genomic changes that challenge current approaches to classifying SARS-CoV-2 variants. Deep sequence learning could be a potentially powerful way to build complex sequence-to-phenotype models. Unfortunately, while they can be predictive, deep learning typically produces “black box” models that cannot directly provide biological and clinical insight. Researchers should therefore consider implementing emerging methods for visualizing and interpreting deep sequence models. Finally, researchers should address important data limitations, including (i) global sequencing disparities, (ii) insufficient sequence metadata, and (iii) screening artifacts due to poor sequence quality control.

KEYWORDS

COVID-19, SARS-CoV-2, viral genomics, bioinformatics, deep learning, explainable AI, genomics, machine learning

COVID-19 has been called the “first pandemic in the post-genomic era” (1). The first severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome was published on 12 January 2020, a week after the WHO first reported on the virus. Only 5 days later, the sequence was used to design the mRNA vaccines that have changed the course of the pandemic (2). Since then, next-generation sequencing technology has enabled an unprecedented view of genetic changes in the virus throughout both the duration of the pandemic and different parts of the world (1, 3). Global data sharing of sequence data has been equally critical, much to the credit of the GISAID EpiCoV database project (4). GISAID’s primary mission has been to share flu genomes, in part to help design the annual flu vaccine (its full name being the Global Initiative on Sharing All Influenza Data) (5). Now, at the beginning of 2022, the GISAID EpiCoV database has accumulated nearly 7 million SARS-CoV-2 genome sequences, and at present, around 800,000 sequences are being added each month. So much data has been generated and made available that it has spurred the development of computational tools for high-frequency sequence variant tracking (6) and even daily updates (7, 8).

Despite the surge in research efforts devoted to COVID-19 (9, 10), laboratory study of the virus remains a more specialized and time-consuming effort than sequencing. Clinical and epidemiological data are often superficial, measuring only a few variables, other than data sets specific to particular facilities or narrow populations. We need to fully capitalize on the abundant data that we do have to (i) anticipate how changes in the virus might affect health before we have time to gather empirical data and (ii) better design and interpret experiments to maximize our use of limited resources. So, how can we translate genome sequence data to as much biological understanding and actionable clinical insight as possible?
**SARS-CoV-2 IS RAPIDLY EVOLVING**

SARS-CoV-2 has spent the first 2 years of the pandemic rapidly evolving in ways that have had a big impact on virulence, transmission, and ability to evade our immune responses (11). SARS-CoV-2 is an RNA virus, so its genome is prone to mutate—albeit at a rate mitigated by its large genome size and the proofreading function of its exoribonuclease (12). The most frequent mutations observed in coronaviruses are generally substitutions, although insertions and deletions are observed as well (13). In some cases, insertions from other viral genomes may occur, and, in fact, it appears as though the SARS-CoV-2 genome includes an insertion from human RNA (14). In other human coronaviruses, the estimated mutation rate is around $3 \times 10^{-4}$ substitutions per site per year (15, 16).

The amount of mutation observed during the COVID-19 pandemic has been even more substantial than expected (17). An early estimate of SARS-CoV-2 mutation was $6 \times 10^{-4}$ substitutions per site per year (18). But the disease has spread widely around the world since then, and novel variants transmit more quickly—increasing the opportunities for the virus to mutate (19, 20). The SARS-CoV-2 spike protein will continue to change in the future. Studies on another human coronavirus, HCoV-OC43, suggest that genetic drift plays a role in coronavirus adaptive evolution (21). One study estimates that as of July 2021, SARS-CoV-2 had only “explored” 31% of the potential space for spike gene variation, based on comparisons with related sarbecoviruses (22).

**SEQUENCE ANALYSIS HAS STRUGGLED TO KEEP UP**

The first widely used tool for tracking SARS-CoV-2 genomic variation was the Nextstrain project, https://nextstrain.org. Nextstrain, originally developed as a general tool for viruses, was adapted to offer clade definitions for SARS-CoV-2 based on phylogenetic analysis (23). Phylogenetic tree reconstruction has been effective in inferring viral origins and trace transmission changes but not as useful in classifying genomes because the virus can accumulate and drop mutations in parallel across clades and subclades (24). The Pango nomenclature (https://cov-lineages.org/), developed specifically for SARS-CoV-2, has largely supplanted Nextstrain clade definitions (25). New sequences are assigned to Pango classifications, called “lineages,” using the Random Forests classification algorithm. A new Pango lineage is defined when a sufficient number of viral sequences emerges with a phylogenetic dissimilarity from existing sequences above a set threshold (26). Particularly significant Pango lineages have been identified by the World Health Organization (WHO) as variants of concern (VOC), which are given Greek letter designations (27), such as Alpha (Pango lineage B.1.1.7), Beta (B.1.351), Delta (B.1.617.2), and, recently, Omicron (B.1.1.529).

While Pango lineages appear clear and well-defined, the reality is that the genome is much more fluid. If we want to understand how genome affects viral function, we cannot rely on traditional taxonomic categorization. As mutations recur, revert, and proliferate, taxonomy hits its limits of utility (11). As an initial matter, changes to SARS-CoV-2 properties often implicate combinations of multiple mutations that emerge simultaneously—and then sometimes revert in whole or in part as the virus continues to evolve (28, 29). For example, one frequent spike protein amino acid substitution, N501Y, has appeared and reverted contemporaneously in multiple clades and lineages, with no evidence of recombination (30). Simultaneous mutations can also have unpredictable, nonlinear effects, i.e., they can be synergistic, antagonistic, or fully independent (31). This complicates classical and Bayesian logistic regression methods for predicting fitness or protein function from mutations, as they rely on assuming the independence between mutations of individual amino acids or bases (32).

SARS-CoV-2 evolution is also highly nonlinear. Widespread lineages, such as Delta, have spawned complex sublineages with distinct immune evasion and virulence properties, which often genetically share more in common with distantly related lineages than their most recent ancestor (33, 34). The increasingly complex evolutionary history of the virus stymies other proposed methods for genetically subtyping viral variants as well (35–37). Further complicating the picture, some immunocompromised individuals can have chronic infections lasting 6 months to a year (38). During long-term infection,
a spike protein can emerge with multiple variations, which phylogenetic analysis identifies as “long branch” divergence from the phylogenetic tree (39). Some long-term patients may even be treated with convalescent plasma or antibodies, which may select for immune evasive mutations (40). The Omicron variant has such a long branch divergence, indicating that it may have emerged in an immunocompromised host or after incubating in a nonhuman host such as mice (41, 42).

CAN DEEP SEQUENCE LEARNING HELP?

How can we predict the virulence, fitness, antibody evasion, and other key properties of novel SARS-CoV-2 variants from complex, nonlinear changes in genetic sequence? Machine learning can tackle complex pattern recognition problems by training a model that can classify the organisms or genes by phylogeny or phenotype based on features of their genetic sequences. For example, we can extract k-mer (short subsequence) frequencies or other combinations of bases/amino acids and use them as features to train classifiers using naive Bayes classifier (NBC), support vector machines (SVM), decision tree-based methods, and neural networks (43–49). Machine learning with k-mer features has been used for SARS-CoV-2 to identify genetic fingerprints of specific infections (50), classify variants (51, 52), and train a model to predict the pathogenicity of unknown viruses (53). Another approach is to build profile hidden Markov models (HMMs), which can identify taxonomic lineages and variants of viruses. HMMs have been used to align SARS-CoV-2 sequences and compare its spike protein to that of other coronaviruses (22, 54, 55).

Deep learning has emerged as an even more powerful and flexible tool to find patterns in large and complicated data sets (56–59). Deep learning models use multiple layers of neural networks to automatically extract and transform features during training (56–58). We can borrow deep learning methods developed for natural language processing (NLP) to find patterns in sequence data, where bases and amino acids that make up genome and protein sequences are analogous to semantic relationships between the words that make up sentences (60–63). For example, one group of researchers has used concepts from semantic processing, e.g., the frequency of correlated words, to identify potential mutagenic sites in viruses including SARS-CoV-2 (64). An emerging approach to deep sequencing learning is to transform protein sequences to embeddings that reflect their semantic structure, using the BERT (bidirectional encoder representations from transformers) neural network architecture, which Google developed to handle natural language search (65–68). An example of this approach is k-means clustering of “ProtBERT” SARS-CoV-2 protein embeddings generated by pretraining a BERT model on millions of UniProt sequences, which can be used to identify mutational hot spots within the genome that may give rise to future variants (69).

A key goal for modeling is to predict the health risk of emerging variants before empirical data are available. To this end, our group has developed a deep learning model to predict patient outcomes for emerging sequence variants that takes into account patient demographics (70). Others are working to integrate sequence learning with computational protein structure models. For example, one project combines models of cell receptor binding and immune epitope alteration with transformer-based deep learning models to predict the fitness advantage of mutations (71). Deep learning has also been used to identify the relationship between protein sequence and function using data from deep mutational scanning, an experimental technique for massively parallel functional analysis of protein sequence site mutations (72, 73). Using this approach, another project predicts the risk for emerging variants by using a neural network to predict infectivity and vaccine breakthrough in combination with protein structure and binding prediction to model antibody resistance (74).

LOOKING INSIDE THE DEEP LEARNING BLACK BOX

Deep learning methods excel at identifying complex features within data that allow classification. But they have a major weakness. Deep learning relies on neural networks, and it is very hard to determine why a neural network makes a particular classification or prediction. Interpretable, or explainable, machine learning can fill this important gap (75, 76). Interpretable machine learning is particularly important in bioinformatics, since explaining
a model’s predictions is critical to justify making high-stakes clinical or research decisions based on machine learning predictions (77, 78). Accordingly, developers of deep learning approaches to SARS-CoV-2 should consider providing some functionality to interpret or explain predictions.

Analytical tools for interpretability in deep learning include examining neural network structure through relevance propagation, activation difference propagation, sensitivity analysis, and saliency map methods (79–81). Integrated gradients have been used to analyze RNA splicing models (82). An increasingly popular approach is the “attention” mechanism originally developed for NLP (83, 84). Attention can highlight important features in text processed by deep learning models (85–87). The amount of “attention” at a position in a sequence correlates with the weight put on that position in a trained model, where high attention at a position implies potential significance. Architectures combining convolutional neural networks (CNNs) with attention have been used to identify sequence motifs for functional genomics, e.g., transcription factor binding site detection (88, 89). Another group generated predictive models of adverse drug reactions based on chemical structures by combining attention with a CNN for each chemical property and structural feature in the model (90). Our group has shown that attention in combination with a recurrent neural network-based sequence model can provide insight into taxonomic and phenotypic classification of microbial 16s rRNA sequences (91), as well as gene ontology classifications of protein sequences (92).

Recently, transformer-based architectures have emerged, like the aforementioned BERT (93). Transformers are built on multiple attention modules (“heads”), which could be used for interpretability (94). For example, one recent paper demonstrated how different attention heads attended to different aspects of a learning task to identify nucleotide motifs for promoter sequences (95). However, attention cannot be inherently drawn out of transformers. Further processing steps are generally required to connect attention to specific linguistic features (96). Our group recently applied a self-attention layer after a transformer as a way to more readily extract and visualize attention across the sequence and applied it to SARS-CoV-2 (70). An important caveat is that, based on comparing attention to empirical evidence, attention does not necessarily imply explanation—at least in the sense of explaining precisely why a prediction took place (97). Attention can only highlight features that the attention layer of the deep learning model weighted most heavily during training, so it may only weakly indicate the complete set of important features for a classification problem.

SEQUENCING DISPARITIES AND DATA CHALLENGES

Finally, we highlight three important data limitations that researchers should address. First, as Fig. 1 shows, there are serious global inequities in sequencing data, with the overwhelming majority of sequences coming from Europe and North America. GISAID has encouraged data sharing from developing countries by trading restrictions on republishing sequence information for access to that information (98). But global sequencing resources are disparately available (99). Even within Europe and the United States, racial and regional disparities in sequencing found in other surveys (100) hamper SARS-CoV-2 sequencing as well. Second, the task of interpreting sequencing data is complicated by insufficient sample metadata, making it difficult to understand how SARS-CoV-2 sequences affect patient outcomes, for example. In GISAID, most sequences only have information about a patient’s age or gender (if available) and the location where the sample was collected. As of 7 January 2022, a little over 270,000 sequences (4%) of the nearly 6.9 million have any metadata for patient outcomes, and many metadata entries are unintelligible. Sequencing projects should be encouraged to collect and curate as much information as possible about the sample and meet minimum information standards for sequence metadata (101). Third, sequencing errors can lead to spurious results. Quality control is critical to make sure that low-frequency sequence variants are real (102). Sequences can pick up contaminants from other variants in the amplification process, leading to what appear to be recombinant variants but which are in fact simply artifacts (103).
ACKNOWLEDGMENTS

G.L.R. received National Science Foundation (NSF) grants no. 1919691 and no. 2107108. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

We gratefully acknowledge all data contributors, i.e., the authors and their originating laboratories responsible for obtaining the specimens and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, from which Fig. 1 was generated.

REFERENCES

1. van Dorp L, Houldcroft CJ, Richard D, Balloux F. 2021. COVID-19, the first pandemic in the post-genomic era. Curr Opin Virol 50:40–48. https://doi.org/10.1016/j.coviro.2021.07.002.

2. Corbett KS, Edwards DK, Leist SR, Abiona OM, Boyoglu-Barnum S, Gillespie RA, Himansu S, Schafer A, Ziwaow CT, DiPiazza AT, Dinnon KH, Elbashir SM, Shaw CA, Woods A, Fritch EJ, Martinez DR, Bock KW, Minai M, Nagata BM, Hutchinson GB, Wu K, Henry C, Bahl K, Garcia-Dominguez D, Ma L, Renzi I, Kong WP, Schmidt SD, Wang L, Zhang Y, Phung E, Chang LA, Loomis RJ, Altaras NE, Narayanen E, Metkar M, Presnyak V, Liu C, Louder MK, Shi W, Leung K, Yang ES, West A, Gully KL, Stevens LJ, Wang N, Wrapp D, Doria-Rose NA, Stewart-Jones G, Bennett H, et al. 2020. SARS-CoV-2 mRNA vaccine design enabled by prototype pathogen preparedness. Nature 586:567–571. https://doi.org/10.1038/s41586-020-2622-0.

3. Chiara M, D’Erchia AM, Gissi C, Manzari C, Parisi A, Resta N, Zambelli F, Picardi E, Pavesi G, Horner DS, Pesole G. 2021. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. Brief Bioinform 22:616–630. https://doi.org/10.1093/bib/bbaa297.

4. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, Akite N, Ho J, Lee RT, Yeo W, Team GCC, Maurer-Stroh S. 2021. GISAID’s role in pandemic response. China CDC Wkly 3:1049–1051. https://doi.org/10.46234/ccdcw2021.255.

5. Shu Y, McCauley J. 2017. GISAID: global initiative on sharing all influenza data – from vision to reality. Eurosurveill 22:30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

6. Bernasconi A, Mari L, Casagrandi R, Ceri S. 2021. Data-driven analysis of amino acid change dynamics timely reveals SARS-CoV-2 variant emergence. Sci Rep 11:21068. https://doi.org/10.1038/s41598-021-00496-z.

7. McBroome J, Thornlow B, Hinrichs AS, Kramer A, De Maio N, Goldman N, Haussler D, Corbett-Detig R, Turakhia Y. 2021. A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. Mol Biol Evol 12:5819–5824. https://doi.org/10.1093/molbev/msab264.

8. Chen C, Nadeau S, Yared M, Voinov P, Xie R, Roemer C, Stadler T. 2022. CoV-spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Bioinformatics 38:1735–1737. https://doi.org/10.1093/bioinformatics/btaa856.

9. Rando HM, MacLean AL, Lee AJ, Lordan R, Ray S, Bansal V, Skelly AN, Sell E, Dziak JJ, Shinholster L, McGowan LD, Guebila MB, Wellhausen N, Knyazev S, Bocca SM, Capone S, Qi Y, Park Y, Mai D, Sun Y, Boerckel JD, Brueffer C, Byrd JB, Kamil JP, Wang J, Velazquez R, Szeto GL, Barton JP, Goel RR, Mangul S, Lubiana T, Gitter A, Greene CS, COVID-19 Review Consortium. 2021. Pathogenesis, symptomatology, and transmission of SARS-CoV-2 through analysis of viral genomics and structure. mSystems 6:e00095-21. https://doi.org/10.1128/mSystems.00095-21.

10. Rando HM, Wellhausen N, Ghosh S, Lee AJ, Dattoli AA, Hu F, Byrd JB, Rafzadeh DN, Lordan R, Qi Y, Sun Y, Brueffer C, Field JM, Guebila MB, Jadavji NM, Skelly AN, Ramsundar B, Wang J, Goel RR, Park Y, Bocca SM, Gitter A, Greene CS, COVID-19 Review Consortium. 2021. Identification and development of therapeutics for COVID-19. mSystems 6:e00233-21. https://doi.org/10.1128/mSystems.00233-21.

FIG 1 Total number of sequences submitted to GISAID (4) (left axis and bars, blue) and ratio of the number of submitted sequences to the total number of reported cases (104) as of 7 January 2022 for the 30 countries that have submitted the most sequences to GISAID. The overwhelming majority of sequences in GISAID come from North America and Europe. Over half of all sequences are from the United States and United Kingdom alone. Sequencing rates show even greater disparities.

REFERENCES

1. van Dorp L, Houldcroft CJ, Richard D, Balloux F. 2021. COVID-19, the first pandemic in the post-genomic era. Curr Opin Virol 50:40–48. https://doi.org/10.1016/j.coviro.2021.07.002.

2. Corbett KS, Edwards DK, Leist SR, Abiona OM, Boyoglu-Barnum S, Gillespie RA, Himansu S, Schafer A, Ziwaow CT, DiPiazza AT, Dinnon KH, Elbashir SM, Shaw CA, Woods A, Fritch EJ, Martinez DR, Bock KW, Minai M, Nagata BM, Hutchinson GB, Wu K, Henry C, Bahl K, Garcia-Dominguez D, Ma L, Renzi I, Kong WP, Schmidt SD, Wang L, Zhang Y, Phung E, Chang LA, Loomis RJ, Altaras NE, Narayanen E, Metkar M, Presnyak V, Liu C, Louder MK, Shi W, Leung K, Yang ES, West A, Gully KL, Stevens LJ, Wang N, Wrapp D, Doria-Rose NA, Stewart-Jones G, Bennett H, et al. 2020. SARS-CoV-2 mRNA vaccine design enabled by prototype pathogen preparedness. Nature 586:567–571. https://doi.org/10.1038/s41586-020-2622-0.

3. Chiara M, D’Erchia AM, Gissi C, Manzari C, Parisi A, Resta N, Zambelli F, Picardi E, Pavesi G, Horner DS, Pesole G. 2021. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. Brief Bioinform 22:616–630. https://doi.org/10.1093/bib/bbaa297.

4. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, Akite N, Ho J, Lee RT, Yeo W, Team GCC, Maurer-Stroh S. 2021. GISAID’s role in pandemic response. China CDC Wkly 3:1049–1051. https://doi.org/10.46234/ccdcw2021.255.

5. Shu Y, McCauley J. 2017. GISAID: global initiative on sharing all influenza data – from vision to reality. Eurosurveill 22:30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

6. Bernasconi A, Mari L, Casagrandi R, Ceri S. 2021. Data-driven analysis of amino acid change dynamics timely reveals SARS-CoV-2 variant emergence. Sci Rep 11:21068. https://doi.org/10.1038/s41598-021-00496-z.

7. McBroome J, Thornlow B, Hinrichs AS, Kramer A, De Maio N, Goldman N, Haussler D, Corbett-Detig R, Turakhia Y. 2021. A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. Mol Biol Evol 12:5819–5824. https://doi.org/10.1093/molbev/msab264.

8. Chen C, Nadeau S, Yared M, Voinov P, Xie R, Roemer C, Stadler T. 2022. CoV-spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Bioinformatics 38:1735–1737. https://doi.org/10.1093/bioinformatics/btaa856.

9. Rando HM, MacLean AL, Lee AJ, Lordan R, Ray S, Bansal V, Skelly AN, Sell E, Dziak JJ, Shinholster L, McGowan LD, Guebila MB, Wellhausen N, Knyazev S, Bocca SM, Capone S, Qi Y, Park Y, Mai D, Sun Y, Boerckel JD, Brueffer C, Byrd JB, Kamil JP, Wang J, Velazquez R, Szeto GL, Barton JP, Goel RR, Mangul S, Lubiana T, Gitter A, Greene CS, COVID-19 Review Consortium. 2021. Pathogenesis, symptomatology, and transmission of SARS-CoV-2 through analysis of viral genomics and structure. mSystems 6:e00095-21. https://doi.org/10.1128/mSystems.00095-21.

10. Rando HM, Wellhausen N, Ghosh S, Lee AJ, Dattoli AA, Hu F, Byrd JB, Rafzadeh DN, Lordan R, Qi Y, Sun Y, Brueffer C, Field JM, Guebila MB, Jadavji NM, Skelly AN, Ramsundar B, Wang J, Goel RR, Park Y, Bocca SM, Gitter A, Greene CS, COVID-19 Review Consortium. 2021. Identification and development of therapeutics for COVID-19. mSystems 6:e00233-21. https://doi.org/10.1128/mSystems.00233-21.
Perspective

11. Tao K, Tzou PL, Nouhin J, Gupta RK, de Oliveira T, Kosakovski PS, Fera D, Shafer RW. 2021. The biological and clinical significance of emerging SARS-CoV-2 variants. Nat Rev Genet 22:1–17. https://doi.org/10.1038/s41576-021-00408-x.

12. Plante JA, Mitchell BM, Plante KS, Debinkl K, Weaver SC, Menachery VD. 2021. The variant gambit: COVID-19’s next move. Cell Host Microbe 29:508–515. https://doi.org/10.1016/j.chom.2021.02.020.

13. Sanjuan R, Netot MR, Chirico N, Mansky LM, Belshaw R. 2021. Viral mutational diversity and evolution rates. J Virol 84:9733–9748. https://doi.org/10.1128/JVI.00694-20.

14. Peacock TP, Bauer DLV, Barclay WS. 2021. Putative host origins of RNA virus Delta to antibody neutralization. Nature 596:276–280. https://doi.org/10.1038/s41586-021-03777-9.

15. Mohammadi E, Shafer C, Imai O. 2021. The biological and clinical significance of emerging SARS-CoV-2 variants. Nat Rev Genet 22:1–17. https://doi.org/10.1038/s41576-021-00408-x.

16. Pyrc K, Dijkman R, Deng L, Jubbink MF, Shafer C. 2021. Novel and emerging mutations of SARS-CoV-2: biomedical implications. Biomed Pharmacother 139:111599. https://doi.org/10.1016/j.biopha.2021.111599.

17. Vervier K, Mahé P, Tournoud M, Veyrieras JB, Vert JP. 2016. Large-scale sequencing reveals the complete mouse origin of the SARS-CoV-2 Omicron variant. J Genet Genom 48:1111. https://doi.org/10.1016/j.jgg.2021.12.003.

18. Baj A, Novazzi F, Pasciuti R, Genoni A, Ferrante FD, Valli M, Partenope M, Trippicino R, Cersichia A, Catano G, Focosi D, Maggi F. 2021. Breakthrough infections of E484K-harbouring SARS-CoV-2 Delta variant. Lombardy, Italy. Emerg Infect Dis 27:3180–3182. https://doi.org/10.3201/eid2711.211792.

19. Baj A, Novazzi F, Pasciuti R, Genoni A, Ferrante FD, Valli M, Partenope M. 2021. Pitfalls of barcodes in the study of worldwide SARS-CoV-2 variation and genomics. Nat Rev Genet 16:321. https://doi.org/10.1038/s41576-020-00412-3.

20. Kjusten B, Omer SB, Grubaugh ND, Iwasaki A, Yale SARS-CoV-2 Genomic Surveillance Initiative. 2021. Impact of circulating SARS-CoV-2 variants on mRNA vaccine immunity. Nature 600:523–529. https://doi.org/10.1038/s41586-021-04085-y.

21. Obermeyer F, Schaffner SF, Jankowiak M, Barkas N, Pyle JD, Park DJ, Macniris BL, Luban J, Sabeti PC, Lemieux AE. 2021. Analysis of 2.1 million SARS-CoV-2 genomes identifies mutations associated with transmissibility. medRxiv https://doi.org/10.1101/2021.09.21.2126323.

22. Vervier K, Mahé P, Tournoud M, Veyrieras JB, Vert JP. 2016. Large-scale sequencing reveals the complete mouse origin of the SARS-CoV-2 Omicron variant. J Genet Genom 48:1111. https://doi.org/10.1016/j.jgg.2021.12.003.

23. Nusselblatt V, Roder AE, Das S, de Wit E, Youn JH, Banakis S, Mushgian A, Mederos C, Wang C, Chung M, Perez-Perez L, Palmore T, Brudno JN, Kong CC, Chen YY. 2021. Year-long SARS-CoV-2 lineage inference from within-host evolution of SARS-CoV-2 in a patient with B cell depletion. medRxiv https://doi.org/10.1101/2021.02.12.2124627.

24. Wolch CM, da Silva Francisco R, de Almeida LGP, Brustolini OJ, Cardoso TF, Nussenblatt V, Roder AE, Das S, de Wit E, Youn JH, Banakis S, Mushgian A, Mederos C, Wang C, Chung M, Perez-Perez L, Palmore T, Brudno JN, Kong CC, Chen YY. 2021. Year-long SARS-CoV-2 lineage inference from within-host evolution of SARS-CoV-2 in a patient with B cell depletion. medRxiv https://doi.org/10.1101/2021.02.12.2124627.

25. Chen L, Zody MC, Di Gerlando C, Martínez-Torres F, Salas A. 2021. Pitfalls of barcodes in the study of worldwide SARS-CoV-2 variation and pheOnodynamics. Zool Res 42:87–93. https://doi.org/10.24272/zissn.2020-0364.
92. Zhao Z, Rosen G. 2020. Visualizing and annotating protein sequences using a deep neural network, p 506–510. In 2020 54th Asilomar Conference on Signals, Systems, and Computers.

93. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need, p 6000–6010. In Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS’17, Curran Associates Inc., Red Hook, NY, USA.

94. Vig J. 2019. BertViz: a tool for visualizing multilayer self-attention in the BERT model. In ICLR Workshop: Debugging Machine Learning Models.

95. Clauwaert J, Menschaert G, Waegeman W. 2021. Explainability in transformer models for functional genomics. Briefings Bioinform 22:bbab060. https://doi.org/10.1093/bib/bbab060.

96. Kobayashi G, Kuribayashi T, Yokoi S, Inui K. 2020. Attention is not only a weight: analyzing transformers with vector norms. arXiv 2004.10102 [cs].

97. Jain S, Wallace BC. 2019. Attention is not explanation. arXiv 1902.10186.

98. Van Noorden R. 2021. Scientists call for fully open sharing of coronavirus genome data. Nature 590:195–196. https://doi.org/10.1038/d41586-021-00305-7.

99. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, Hill SC, Sabino EC, Pybus OG, Dye C, Bhatt S, Flaxman S, Suchard MA, Grubaugh ND, Baele G, Faria NR, Danish Covid-19 Genome Consortium. 2021. Global disparities in SARS-CoV-2 genomic surveillance. medRxiv https://doi.org/10.1101/2021.08.21.21262393.

100. Spratt DE, Chan T, Waldron L, Speers C, Feng FY, Ogunwobi OO, Osborne JR. 2016. Racial/ethnic disparities in genomic sequencing. JAMA Oncol 2:1070–1074. https://doi.org/10.1001/jamaoncol.2016.1854.

101. Schirin LM, Chuvchina M, Davies N, Elve-Fadrosch EA, Finn RD, Hugenholtz P, Hunter CI, Hurwitz BL, Kyprides NC, Meyer F, Mizrachi IK, Sansone SA, Sutton G, Tighe S, Walls R. 2020. COVID-19 pandemic reveals the peril of ignoring metadata standards. Sci Data 7:188. https://doi.org/10.1038/s41597-020-0524-5.

102. Jacot D, Pillonel T, Greub G, Bertelli C. 2021. Assessment of SARS-CoV-2 genome sequencing: quality criteria and low-frequency variants. J Clin Microbiol 59:e00944-21. https://doi.org/10.1128/JCM.00944-21.

103. Lagerborg KA, Normandin E, Bauer MR, Adams G, Figueroa K, Loreth C, Gladden-Young A, Shaw BM, Pearlman LR, Benenzy D, Dewey HB, Kales S, Dobbins ST, Shenoy ES, Hooper D, Pierce VM, Zachary KC, Park DJ, Macinnis BL, Tewhey R, Lemieux JE, Sabeti PC, Reilly SK, Siddle KJ. 2022. Synthetic DNA spike-ins (SDSIs) enable sample tracking and detection of inter-sample contamination in SARS-CoV-2 sequencing workflows. Nat Microbiol 7:108–119. https://doi.org/10.1038/s41564-021-01019-2.

104. Hasell J, Mathieu E, Beltekian D, Macdonald B, Giattino C, Ortiz-Ospina E, Roser M, Ritchie H. 2020. A cross-country database of COVID-19 testing. Sci Data 7:345. https://doi.org/10.1038/s41597-020-00688-8.