doi:10.20944/preprints202006.0225.v1

CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation

Joshua B. Singer[1,*], Robert J. Gifford[1], Matthew Cotten[1, 2], David L. Robertson[1]
[1]MRC-University of Glasgow Centre for Virus Research (CVR), Glasgow, Scotland, UK
[2]MRC/UVRI & LSHTM Uganda Research Unit, Entebbe, Uganda
*To whom correspondence should be addressed: email josh.singer@glasgow.ac.uk

Abstract

Summary CoV-GLUE is an online web application for the interpretation and analysis of SARS-CoV-2 virus genome sequences, with a focus on amino acid sequence variation. It is based on the GLUE data-centric bioinformatics environment and provides a browsable database of amino acid replacements and coding region indels that have been observed in sequences from the pandemic. Users may also analyse their own SARS-CoV-2 sequences by submitting them to the web application to receive an interactive report containing visualisations of phylogenetic classification and highlighting genomic variation of potentially high impact, for example linked to primer mismatches.
Availability and implementation Available at http://cov-glue.cvr.gla.ac.uk
Implemented using GLUE, an open source framework for the development of virus sequence data resources.
Contact josh.singer@glasgow.ac.uk
Supplementary information CoV-GLUE User Guide, GISAID EpiCoV™ data contributors

Keywords: SARS-CoV-2; web application; virus genome; lineage assignment; amino acids

Main text

Introduction

In December 2019 a novel virus causing respiratory illness and other symptoms including gastrointestinal problems, cardiovascular dysfunction and thrombosis was identified in Wuhan, China (Zhao et al., 2020). In the following months, this virus (SARS-CoV-2, also known as hCoV-19 (Gorbalenya et al., 2020; Jiang et al., 2020)) has caused a pandemic in which >7 million have been infected and >400,000 have died (Dong et al., 2020). Whole genome sequencing revealed that its species is SARS-related coronavirus, within the *Betacoronavirus* genus of the family *Coronaviridae* (coronaviruses). Coronaviruses are positive-sense single-stranded RNA viruses with large genomes ~30,000 nucleotides (nt) in length (Gorbalenya et al., 2006; Gorbalenya et al., 2020; Zhao et al., 2020). SARS-CoV-2 genome sequences are being collated in the GISAID EpiCoV™ database (Shu et al., 2017), which contains > 30,000 sequences, from over 70 countries, with over 1,000 new genomes being deposited every week, the vast majority spanning almost the whole virus genome.

The analysis of amino acid replacements from virus genome data is important for basic virology. For example a single replacement E627K in protein PB2 of avian influenza A virus (IAV) was found experimentally to be important for efficient replication in mammalian cells (Subbarao et al., 1993) and is amongst a small set of changes associated with the adaptation to humans of highly pathogenic IAV strains (Taubenberger et al., 2005). Genome sequence analysis also has a range of practical applications including real-time genomic epidemiology (Gardy et al., 2015), documenting novel infection routes (Diallo et al., 2016), documenting sources of zoonotic infection (Memish et al., 2014), vaccine design, understanding resistance to antivirals, informing nosocomial-associated transmissions and the design of effective diagnostic PCR primers (Wang et al., 2020).

SARS-CoV-2 is naturally accumulating nucleotide mutations in its RNA genome as the pandemic progresses. Point mutations, specifically nonsynonymous substitutions, will result in amino acid replacements in viral genome sequences, while other mutations will result in insertions or deletions (indels). Without compelling evidence, the observed changes should be expected to have no or minimal consequence for virus biology (MacLean et al., 2020; Grubaugh et al., 2020). However, tracking these changes will help us better understand and control the pandemic as mutations of consequence could arise, for example leading to escape from antiviral drugs and future vaccines or causing attenuation as observed towards the end of the SARS outbreak (Muth et al., 2018).

We created CoV-GLUE, a resource capable of providing high resolution tracking of change within the SARS-CoV-2 genome. To create CoV-GLUE, we exploited GLUE (Singer et al., 2018), a data-centric bioinformatics environment for virus sequence data. In this report we describe CoV-GLUE and its main functionality. A detailed user guide for CoV-GLUE is given in the supplementary materials.

Results

CoV-GLUE (http://cov-glue.cvr.gla.ac.uk) is a publicly accessible web application for tracking and analysing variation within SARS-CoV-2 genome sequences, enabled by data from EpiCoV™ (Shu et al., 2017). Its first aim is to track the key elements of variation as these appear in pandemic sequences, linked to sampling data. Secondly it allows users to submit their own consensus sequence data for analysis, receiving an interactive report detailing genomic variation.

CoV-GLUE maintains a browsable database of amino acid sequence variations within all 26 of the putative viral proteins, as proposed in annotations on NCBI RefSeq NC_045512. CoV-GLUE currently documents 8,778 amino acid replacements detected in pandemic sequences from EpiCoV™. Users may filter these by custom criteria including viral protein, codon position, Grantham/Miyata distance (Grantham, 1974; Miyata et al. 1979) and the number of sequences in which the replacement is found. There are similar capabilities for in-frame insertions (8 in the current version) and in-frame deletions (138). CoV-GLUE additionally provides a web page for each variation, with a table of sequences exhibiting the variation, which may be filtered by criteria including global region, country, collection date and phylogenetic lineage.

The analysis report for submitted FASTA sequences is generated within less than a minute per sequence and contains various sections. Any nucleotide mismatches or indels relative to published sequencing and diagnostic primer designs are reported. Each virus protein coding region of the query sequence may be visualised graphically, displaying nucleotide and amino acid variations relative to the reference sequence, in standardised coordinates. Additionally, each detected amino acid replacement or indel is classified according to whether it is novel or known in pandemic sequences. One or more phylogenetic placement branches are generated for each query sequence within the precomputed reference tree containing representative taxa from the pandemic, and may be visualised graphically. The sequence is assigned to a lineage proposed by Rambaut et al. (2020), based on these placements and their relative likelihoods.

The lineage system release on 19th May 2020 contained hand-curated lineage assignments for 24,688 GISAID sequences, excluding lineage representatives. We found that CoV-GLUE's assignment matched exactly in 95.5% of cases and in most remaining cases assigned to either the parent or a child lineage.

Methods

Throughout CoV-GLUE, Wuhan-Hu-1 (NCBI RefSeq NC_045512) is used as the reference sequence for nucleotide coordinates, codon numbering within viral proteins and as the comparison for detecting variations.

Sequences are regularly downloaded from GISAID EpiCoV™ (Shu et al., 2017) and added to a constrained alignment (Singer et al. 2018), using MAFFT (Katoh and Standley, 2013) as a pairwise aligner. After excluding irrelevant sequences and those with potential quality issues, amino acid replacements and in-frame indels in each sequence are identified. A similar process identifies variations in sequences submitted via the web.

Lineage representative sequences defined by Rambaut et al. (2020) are aligned using MAFFT (Katoh and Standley, 2013) to produce the reference alignment. Excluding the 5' and 3' untranslated regions, the reference tree is then generated from this alignment by RAxML (Stamatakis, 2014) using the GTRGAMMA substitution model, 1000 bootstraps.

For lineage assignment by CoV-GLUE, the sequence is added to the reference alignment using MAFFT (Katoh and Standley, 2013) with the --add and --keeplength options. The RAxML evolutionary placement algorithm (Berger et al., 2011) is then run with the GTRGAMMA substitution model and the --epa-accumulated-threshold option set to 95%, producing zero or more placements with likelihood weight ratios. Each placement branch is examined to determine its ancestor internal nodes, including any node forming a polytomy with the branch. The placement contributes its likelihood weight ratio to each of the lineages represented by these nodes. Each lineage thereby accumulates a total likelihood weight ratio, the sequence is assigned to the most specific lineage with a total of more than 50%.

Discussion

Bioinformatics resources such as CoV-GLUE have an important place in disease outbreaks as they are essential for data curation and analysis (Hufsky et al., 2020).

The GLUE software environment facilitates rapid deployment of novel alignment-centric resources for viruses. It is designed to be scalable and to permit adaptation of existing functionality to construct a bespoke online resource such as CoV-GLUE. Researchers generating sequence data can use CoV-GLUE to obtain a rapid, convenient and visual phylogenetic analysis and classification of new sequences. This will be useful for molecular epidemiology, at least to detect the introduction of lineages that have not been observed locally. CoV-GLUE can also be used to rapidly notify diagnostic laboratories of emerging local lineages for which the implemented primer design may be ineffective.

CoV-GLUE can also help researchers investigate the functional relevance of amino acid sequence variation in SARS-CoV-2. For example, Korber et al. (2020) identified the increasing prevalence of the Spike protein replacement D614G and suggested this may confer a selective transmissibility advantage. So, while it is difficult to show the functional effect of a viral mutation within a pandemic (MacLean et al. 2020), tracking these variants does help formulate hypotheses that could then be tested experimentally.

Acknowledgements

We would like to thank Joseph Hughes, Emma Thomson, Oscar MacLean, Richard Orton, Ana da Silva Filipe, Massimo Palmarini, Andrew Rambaut, Áine O’Toole, Nuno Faria and Sarah Hill for valuable feedback and input. CoV-GLUE is enabled by data from the GISAID EpiCoV™ database. The full table of data contributors is given in the supplementary materials.

Funding

The authors are funded by the UK Medical Research Council award number MC_UU_1201412. MC is also funded by the Wellcome Trust and the UK Department for International Development under award number 220977/Z/20/Z. CoV-GLUE also receives support from the UK Department of Health and Social Care and UK Research and Innovation as part of the COVID-19 UK Genomics Consortium, award number MC_PC_19027.

References

Berger, S. A., Krompass, D., & Stamatakis, A. (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood.

Systematic Biology, 60(3), 291–302.

Diallo, B., Sissoko, D., Loman, N. J., Bah, H. A., Bah, H., Worrell, M. C., … Duraffour, S. (2016). Resurgence of Ebola Virus Disease in Guinea Linked to a Survivor With Virus Persistence in Seminal Fluid for More Than 500 Days. Clinical Infectious Diseases, 63(10), 1353–1356.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases, 20(5), 533–534.

Drosten, C., Lauber, C., Penzar, D., Perlman, S., Sola, I., Ziebuhr, J., … Sidorov, I. A. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nature Microbiology, 5(4), 536–544.

Gardy, J., Loman, N. J., & Rambaut, A. (2015). Real-time digital pathogen surveillance — the time is now. Genome Biology, 16(1).

Gorbalenya, A. E., Enjuanes, L., Ziebuhr, J., & Snijder, E. J. (2006). Nidovirales: Evolving the largest RNA virus genome. Virus Research, 117(1), 17–37.

Grantham, R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. Science, 185(4154), 862–864.

Grubaugh, N. D., Petrone, M. E., & Holmes, E. C. (2020). We shouldn’t worry when a virus mutates during disease outbreaks. Nature Microbiology, 5(4), 529–530.

Hufsky, F., Lamkiewicz, K., Almeida, A., Aouacheria, A., Arighi, C., Bateman, A., … Marz, M. (2020). Computational Strategies to Combat COVID-19: Useful Tools to Accelerate SARS-CoV-2 and Coronavirus Research. Preprints. https://doi.org/10.20944/preprints202005.0376.v1

Jiang, S., Shi, Z., Shu, Y., Song, J., Tan, W., Guo, D., & Gao, G. F. (2020). A distinct name is needed for the new coronavirus. The Lancet, 395(10228), 949.

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution, 30(4), 772–780.

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., … Montefiori, D. C. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. https://doi.org/10.1101/2020.04.29.069054

Maclean, O. A., Orton, R. J., Singer, J. B., & Robertson, D. L. (2020). No evidence for distinct types in the evolution of SARS-CoV-2. Virus Evolution, 6(1).

Memish, Z. A., Cotten, M., Meyer, B., Watson, S. J., Alsahafi, A. J., Rabeeah, A. A. A., … Drosten, C. (2014). Human Infection with MERS Coronavirus after Exposure to Infected Camels, Saudi Arabia, 2013. Emerging Infectious Diseases, 20(6), 1012–1015.

Miyata, T., Miyazawa, S., & Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. Journal of Molecular Evolution, 12(3), 219–236.

Muth, D., Corman, V. M., Roth, H., Binger, T., Dijkman, R., Gottula, L. T., … Drosten, C. (2018). Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. Scientific Reports, 8(1).

Rambaut, A., Holmes, E. C., Hill, V., O'Toole, Á., Ruis, C., Pybus, O. G., … du Plessis, L. (2020). A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. bioRxiv. https://doi.org/10.1101/2020.04.17.046086

Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance, 22(13).

Singer, J. B., Thomson, E. C., Mclauchlan, J., Hughes, J., & Gifford, R. J. (2018). GLUE: a flexible software system for virus sequence data. BMC Bioinformatics, 19(1).

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), 1312–1313.

Subbarao, E. K., London, W., & Murphy, B. R. (1993). A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. Journal of Virology, 67(4), 1761–1764.

Taubenberger, J. K., Reid, A. H., Lourens, R. M., Wang, R., Jin, G., & Fanning, T. G. (2005).

Characterization of the 1918 influenza virus polymerase genes. Nature, 437(7060), 889–893.

Wang, C., Liu, Z., Chen, Z., Huang, X., Xu, M., He, T., & Zhang, Z. (2020). The establishment of reference sequence for SARS‐CoV‐2 and variation analysis. Journal of Medical Virology, 92(6), 667–674.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., … Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. Nature, 579(7798), 265–269.