Topological Surveillance of Recurrent Mutations in SARS-CoV-2

CoVtRec report as of 25 December 2022

Michael Bleher[2*], Lukas Hahn[2*], Maximilian Neumann[*], Samuel Braun[3], Holger Obermaier[3], Mehmet Soysal[3], René Caspart[3], Andreas Ott[1,2*]

[1]Mathematics Department, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Mathematical Institute, Heidelberg University, Heidelberg, Germany
[3]Steinbuch Centre for Computing, Karlsruhe Institute of Technology, Karlsruhe, Germany

[*]Correspondence: mbleher@mathi.uni-heidelberg.de (M.B.)
lhahn@mathi.uni-heidelberg.de (L.H.)
neumann.mn@icloud.com (M.N.)
andreas.ott@kit.edu (A.O.)

Abstract

The appearance of new variants of the coronavirus SARS-CoV-2 in the current COVID-19 pandemic underlines the importance of being able to quickly identify mutations that could confer some adaptive advantage to the virus, such as immune evasion or higher infectivity. Here we apply CoVtRec, a fast and scalable early warning system based on Topological Data Analysis, for the identification and surveillance of emerging potentially adaptive mutations in the ongoing evolution of SARS-CoV-2. CoVtRec is based on a new topological approach to the study of recurrent mutations in large genomic datasets developed in [1].

Results

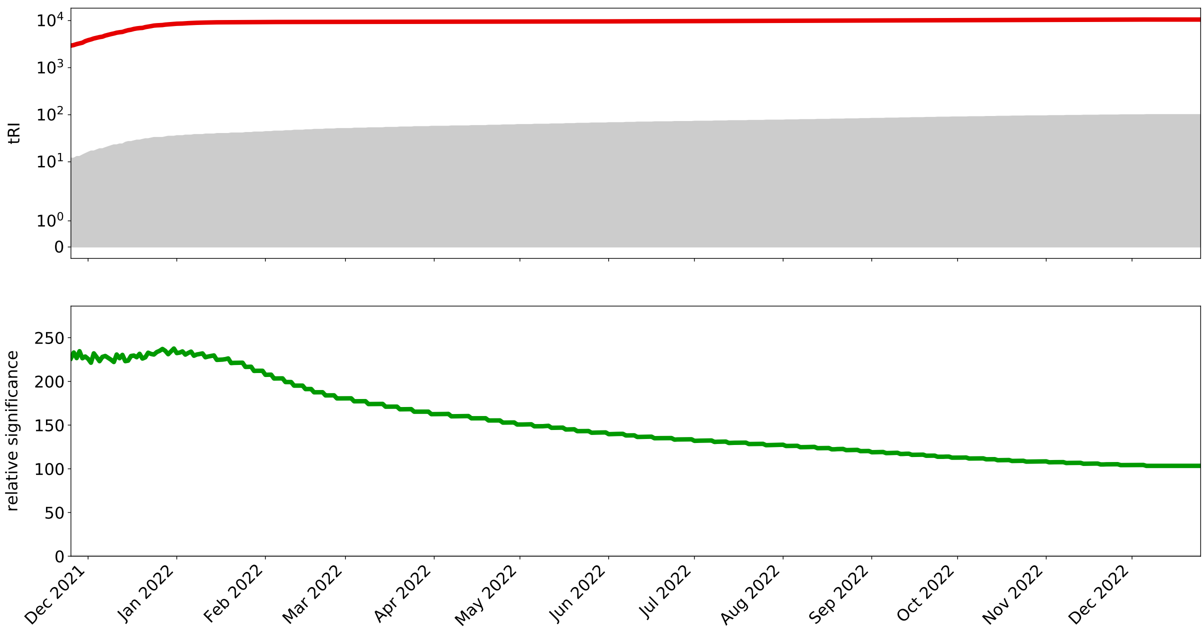
We analyzed topological signals for the ongoing convergent evolution of the coronavirus SARS-CoV-2 on the Spike gene from 25 November 2021 until 25 December 2022. To that end, we performed a topological recurrence analysis for a curated alignment of 12,924,524 high-quality SARS-CoV-2 Spike gene sequences shared via GISAID, the global data science initiative [2, 3]. For each Spike mutation we computed its topological recurrence index (tRI) and the corresponding time series analysis chart. The topological recurrence index is a topological measure for the potential adaptiveness of a given mutation (see [1, 4] for details).

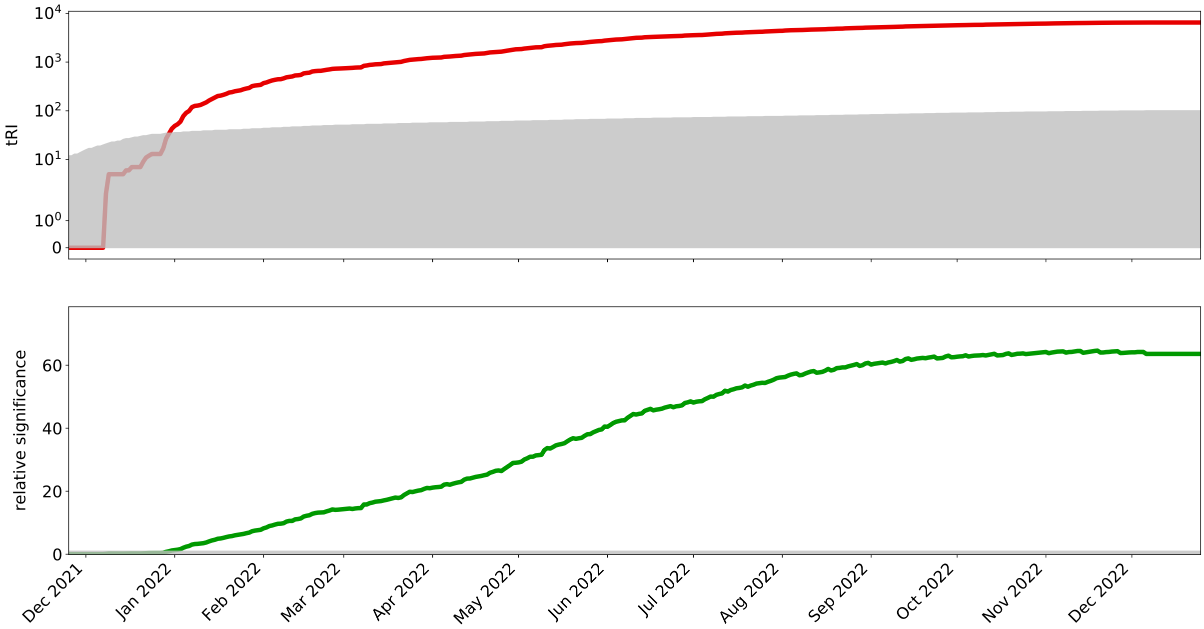
We present a list featuring the top ten amino acid variations on the Spike gene that show strongest topological recurrence index as of 25 December 2022 (see Table 1). Here signals with tRI ≥ 101 are statistically significant ($p < 0.05$). It was demonstrated in [1] that these mutations are potentially adaptive in the current phase of the pandemic and might therefore appear in future variants. We also present time series analysis charts (see Figure 1) showing (i) the development of the topological signal as well as its significance over time, (ii) major lineages containing the mutation, and (iii) the date from which on the topological signal became significant.

SAAV	tRI	relative significance	notable variants
G142D	10545	103.4	Delta, Omicron, Kappa
N440K	6486	63.6	Omicron
L5F	5332	52.3	Iota
T95I	4661	45.7	Delta, Omicron, Iota, Kappa, Mu
R346T	3939	38.6	Omicron
R346K	3815	37.4	Omicron, Mu
R408S	3335	32.7	Omicron
K417N	2440	23.9	Beta, Omicron, Mu
A1020S	2033	19.9	
T76I	1998	19.6	Lambda

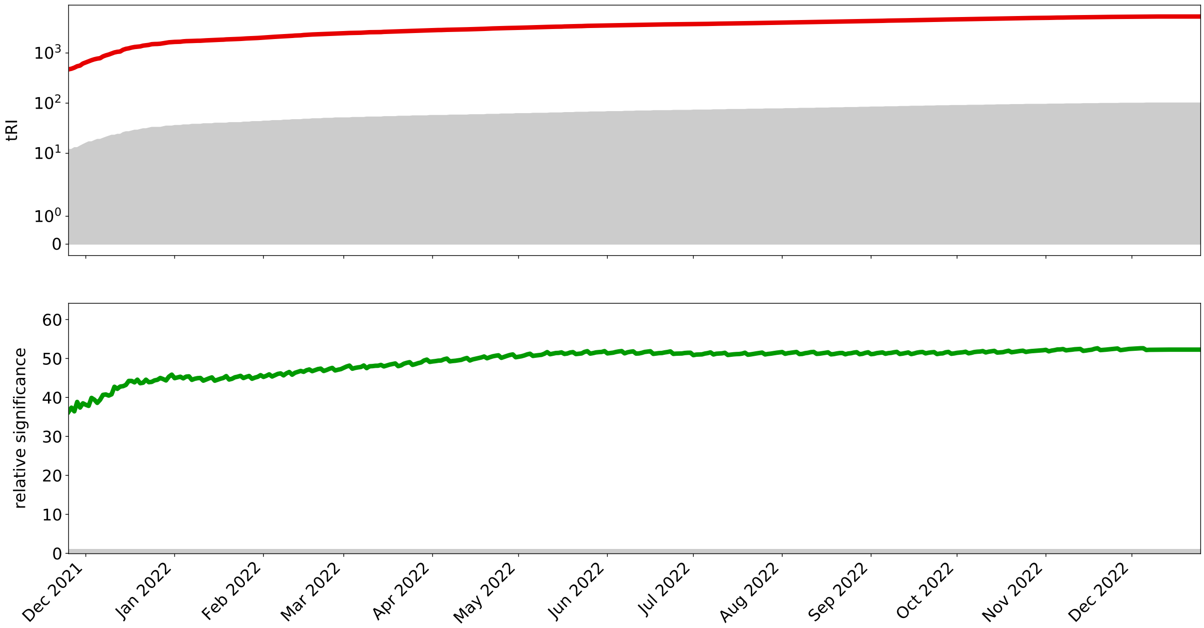
Table 1. The top ten amino acid changes on the Spike gene showing strongest topological recurrence index as of 25 December 2022. For a given mutation, the table displays its topological recurrence index (tRI), its relative significance, and notable variants containing the mutation.

Figure 1. Time series analysis charts for the mutations listed in Table 1. Each chart shows the topological recurrence index (red) and its relative significance (green) from 25 November 2021 until 25 December 2022. In each chart, in the upper diagram the shaded region marks the level of significance.

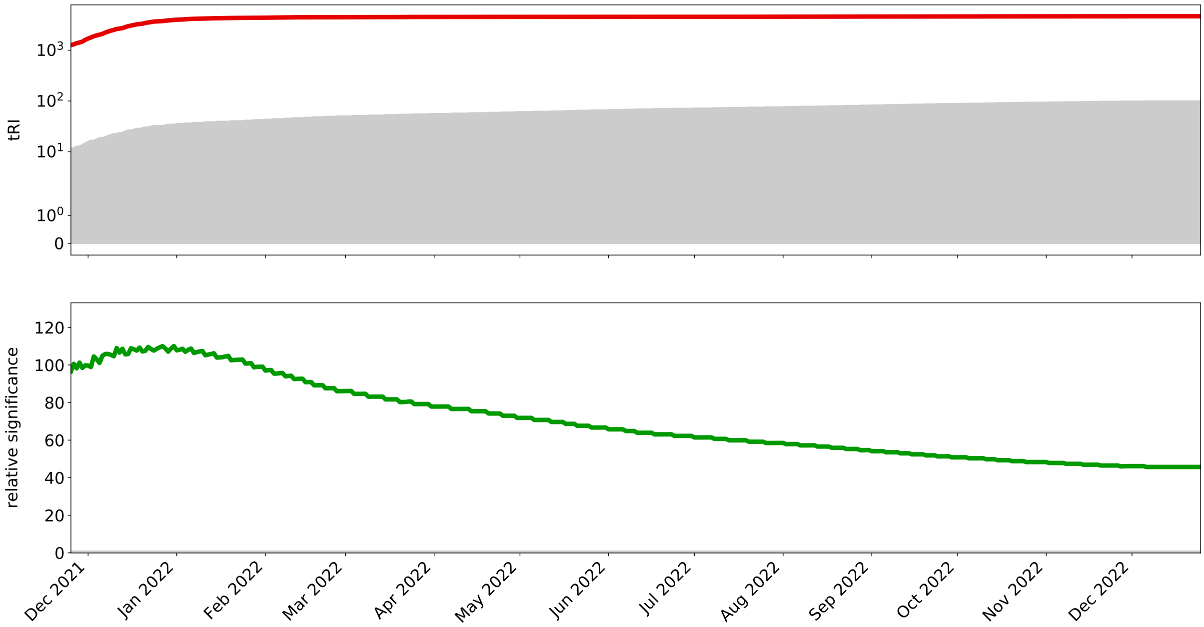

S:G142D | contained in Delta, Omicron, Kappa | significant since 2021-11-25


S:N440K | contained in Omicron | significant since 2021-12-31

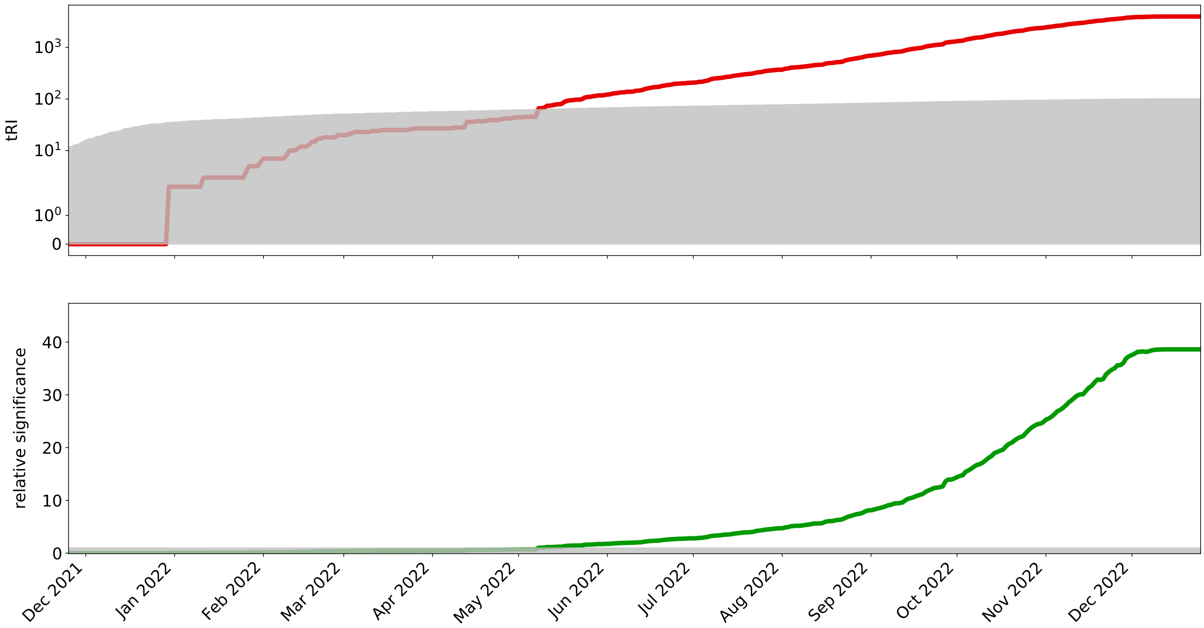
S:L5F | contained in Iota | significant since 2021-11-25



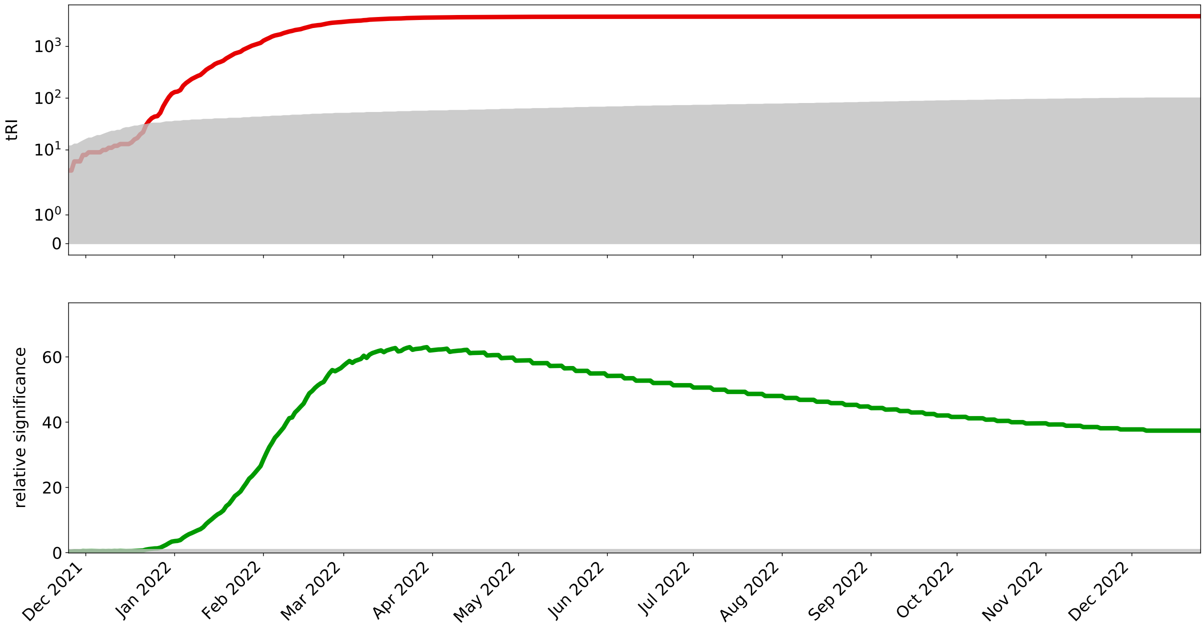
S:T95I | contained in Delta, Omicron, Iota, Kappa, Mu | significant since 2021-11-25



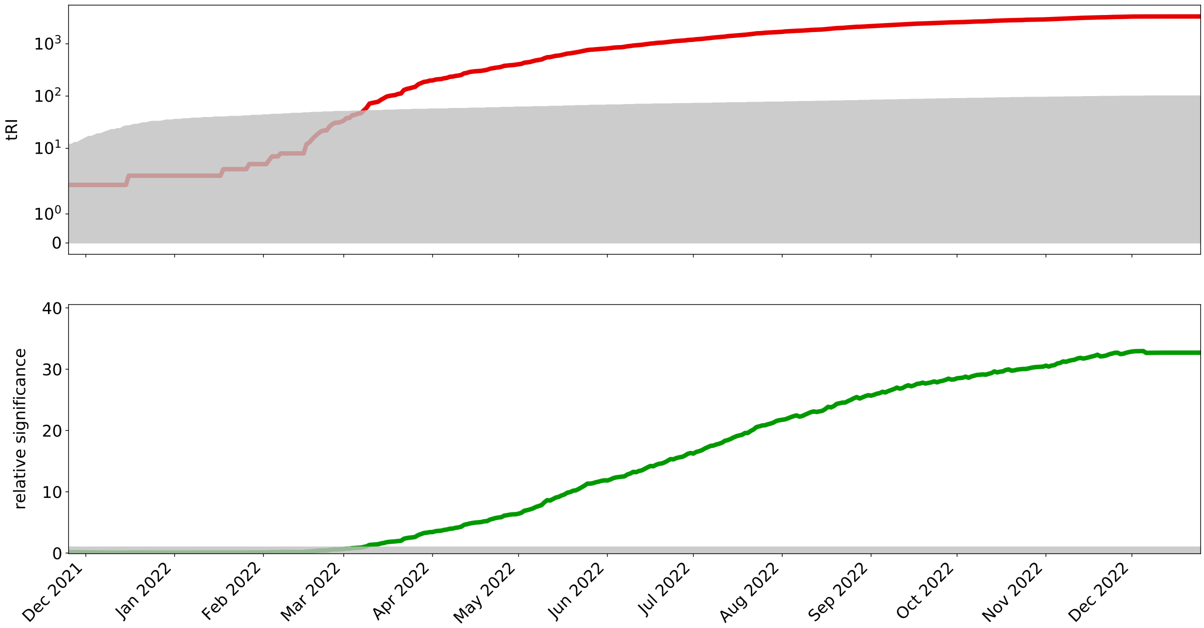
S:R346T | contained in Omicron | significant since 2022-05-08



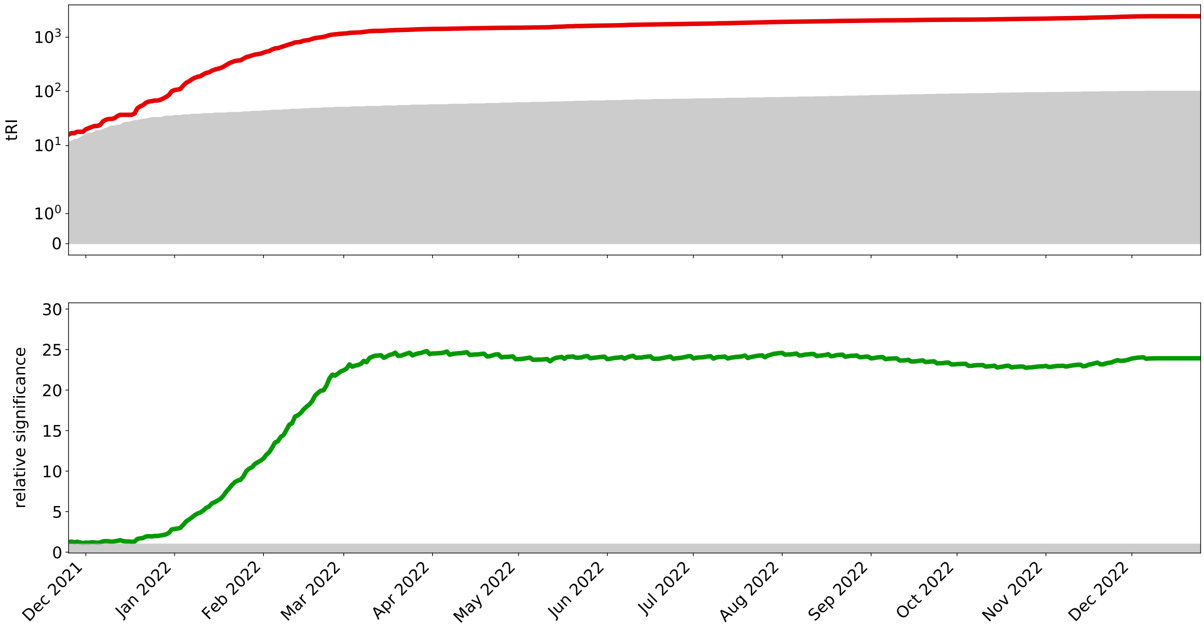
S:R346K | contained in Omicron, Mu | significant since 2021-12-23



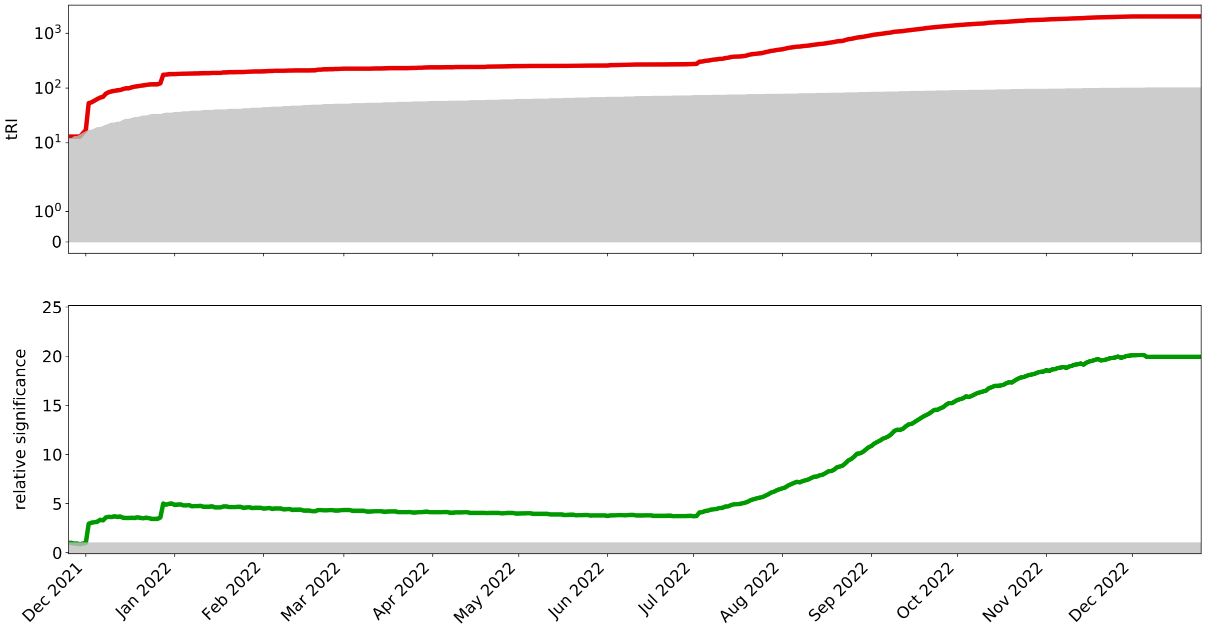
S:R408S | contained in Omicron | significant since 2022-03-08



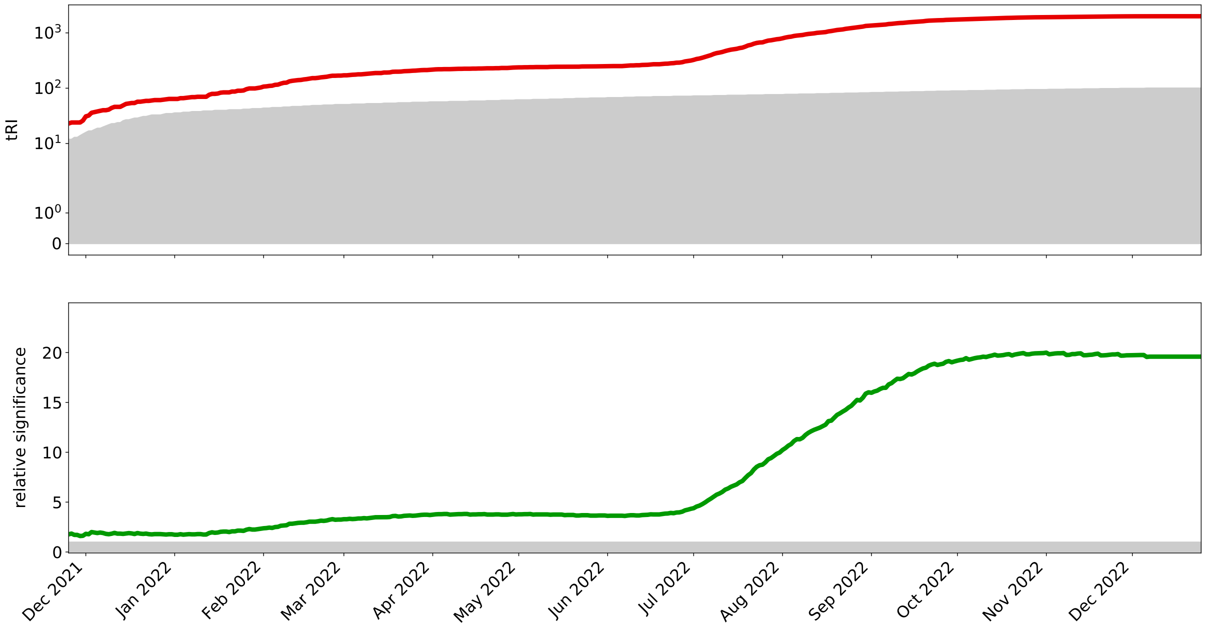
S:K417N | contained in Beta, Omicron, Mu | significant since 2021-11-25



S:A1020S | significant since 2021-12-01



S:T76I | contained in Lambda | significant since 2021-11-25



Methods

Data acquisition and data preparation

Our analysis is based on the alignment `msa_1225.fasta` downloaded from the GISAID EpiCoV Database [2, 3] on 26 December 2022. This alignment comprises 12,924,524 SARS-CoV-2 whole genome sequences that have been aligned to the reference sequence Wuhan/WIV04 with GISAID accession number `EPI_ISL_402124` using `MAFFT` [5]. Sequences in this alignment were truncated to the Spike gene (reference site positions 21,563 to 25,384), and subsequently sequences containing any characters other than A, C, T, G or - were removed. This resulted in an alignment comprising 4,966,409 complete SARS-CoV-2 Spike genes of length 7,592nt. A list of accession numbers of all sequences in this alignment, along with an acknowledgement of the contributions of both the submitting and the originating laboratories, is accessible at https://doi.org/10.55876/gis8.221228od.

Topological recurrence analysis

The Spike gene alignment contains 312,528 genetically distinct sequences. We used `Hammingdist (Version 0.18.0)` [6] to compute the genetic distance matrix of this alignment. Subsequently we used `Ripser` [7] to compute the representative cycles for the persistent homology of the Vietoris–Rips filtration associated to the genetic distance matrix. The computation of persistence barcodes was restricted to small genetic distance scales (`Ripser` scale parameter threshold set to 2). Next a complete list of SNV cycles (topological cycles all of whose edges correspond to single nucleotide variations) in the given alignment was generated from the corresponding `Ripser` output. Then we used custom code implemented in Python to compute the *topological recurrence index (tRI)* for each such SNV. Summing over all SNVs determining an SAAV (single amino acid variation), we computed the tRI for each SAAV. Lastly, from the distribution of the tRI measurements over the whole Spike gene we inferred the level of significance for the tRI per SAAV. For each SAAV, its *relative significance* is then defined as the quotient of its tRI by the level of significance. Using Vietoris-Rips transformations in multipersistent homology, we computed tRI time series analysis charts at daily resolution from the natural stratification by time of genomic data. For a more detailed description of the topological recurrence analysis see [1, 4].

Data availability

All SARS-CoV-2 genome data used in this work are available from the GISAID EpiCov Database [2, 3] at https://www.gisaid.org and are accessible at https://doi.org/10.55876/gis8.221228od.

Code availability

Code used for the analyses is available at https://github.com/ssciwr/hammingdist and https://github.com/Ripser/ripser/tree/tight-representative-cycles. All other code is available from the corresponding authors upon request.

Acknowledgements

The authors 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 [2, 3], on which this research is based. An acknowledgement table is accessible at https://doi.org/10.55876/gis8.221228od.

The authors acknowledge the use of de.NBI Cloud and the support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen and the German Federal Ministry of Education and Research (BMBF) through grant no 031 A535A. They thank M. Hanussek for IT support and early access to VALET [8]. The authors further acknowledge support from the Interdisciplinary Center for Scientific Computing at Heidelberg University and the development work of the Scientific Software Center of Heidelberg University carried out by L. Keegan and D. Kempf [6]. This research was funded by the Federal Ministry of Education and Research (BMBF) and the Baden-Württemberg Ministry of Science as part of the Excellence Strategy of the German Federal and State Governments (KIT Centers, "Topological Genomics"), and by the Vector Foundation ("Topological Genomics"). M.B. and L.H. were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster). L.H. thanks the Evangelisches Studienwerk Villigst for their support.

Author contributions

M.B., L.H., M.N., A.O. designed the study; M.B., L.H., A.O. curated data; M.B., L.H., M.N., A.O. performed computational analyses; M.B., L.H., M.N., A.O., S.B., H.O., M.S., R.C. developed and implemented software; M.B., L.H., A.O. acquired computing resources; M.B., L.H., A.O., M.N. drafted the manuscript; all authors contributed to the final version of the report.

Competing interests

The authors declare no competing interests.

References

[1] M. Bleher, L. Hahn, J. Á. Patiño-Galindo, et al. "Topological data analysis identifies emerging adaptive mutations in SARS-CoV-2". *medRxiv* (2021). DOI: 10.1101/2021.06.10.21258550.

[2] Yuelong Shu and John McCauley. "GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality". *Eurosurveillance* 22.13 (2017). DOI: 10.2807/1560-7917.ES.2017.22.13.30494.

[3] Shruti Khare, Céline Gurry, Lucas Freitas, et al. "GISAID's Role in Pandemic Response". *China CDC Weekly* 3.49 (2021), pp. 1049–1051. DOI: 10.46234/ccdcw2021.255.

[4] M. Bleher, L. Hahn, M. Neumann, et al. "MuRiT: efficient computation of pathwise persistence barcodes in multi-filtered flag complexes via Vietoris-Rips transformations". *arXiv* (2022). DOI: 10.48550/arXiv.2207.03394.

[5] K. Katoh. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform". *Nucleic Acids Research* 30.14 (2002), pp. 3059–3066. DOI: 10.1093/nar/gkf436.

[6] Liam Keegan and Dominic Kempf. *Hammingdist: A Fast Tool to Calculate Hamming Distances.* Version 0.15.0. 2021. URL: https://github.com/ssciwr/hammingdist.

[7] Ulrich Bauer. "Ripser: efficient computation of Vietoris-Rips persistence barcodes". *Journal of Applied and Computational Topology* (2021). DOI: 10.1007/s41468-021-00071-5.

[8] Maximilian Hanussek. *VALET.* 2021. URL: https://github.com/MaximilianHanussek/VALET.