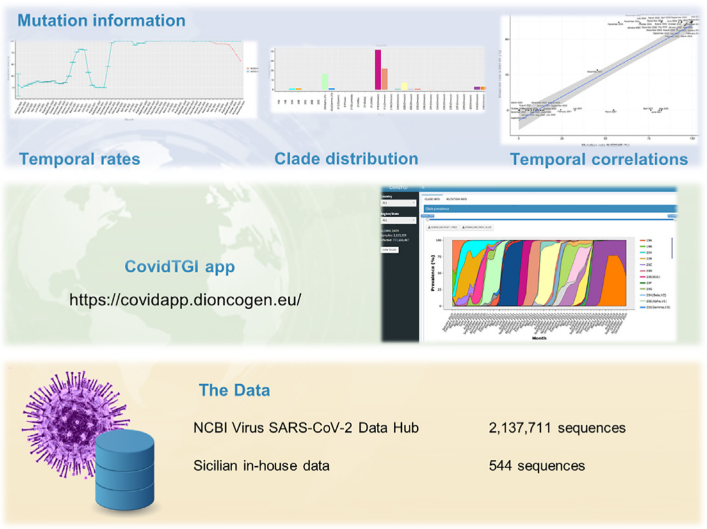
iScience

Article

CovidTGI: A tool to investigate the temporal genetic instability of SARS-CoV-2 variants

Graphical abstract



Highlights

- CovidTGI tracks SARS-CoV-2 mutations across time and geography
- Bayesian analysis reveals key co-occurring mutations in Omicron
- CovidTGI enables rapid, region-specific mutation analysis and comparison
- CovidTGI identifies mutation patterns shaping SARS-CoV-2 evolution

Authors

Grete Francesca Privitera, Nicolò Musso, Giovanni Micale, ..., Guido Scalia, Stefania Stefani, Alfredo Pulvirenti

Correspondence

alfredo.pulvirenti@unict.it

In brief

Classification Description: Bioinformatics; Genetics

Privitera et al., 2025, iScience *28*, 112315
April 18, 2025
https://doi.org/10.1016/j.isci.2025.112315

CellPress

CellPress
OPEN ACCESS

Article

CovidTGI: A tool to investigate the temporal genetic instability of SARS-CoV-2 variants

Grete Francesca Privitera,[1,4] Nicolò Musso,[2,4] Giovanni Micale,[1] Carmelo Bonomo,[2] Salvatore Alaimo,[1] Dalida Bivona,[2] Paolo Giuseppe Bonacci,[2] Guido Scalia,[3] Stefania Stefani,[2] and Alfredo Pulvirenti[1,5,*]

[1]Department of Clinical and Experimental Medicine, Bioinformatic Unit, University of Catania, Via Santa Sofia, 95125 Catania, Italy
[2]Department of Biomedical and Biotechnological Sciences (BIOMETEC), Medical Molecular Microbiology and Antibiotic Resistance Laboratory (MMAR Lab), University of Catania, Via Santa Sofia, 95125 Catania, Italy
[3]U.O.C. Laboratory Analysis Unit, A.O.U. 'Policlinico-Vittorio Emanuele', University of Catania, Via Santa Sofia, 95125 Catania, Italy
[4]These authors contributed equally
[5]Lead contact
*Correspondence: alfredo.pulvirenti@unict.it

https://doi.org/10.1016/j.isci.2025.112315

SUMMARY

The COVID-19 pandemic has underscored the need for fast and accurate epidemiology, particularly due to the high observed mutation frequency in SARS-CoV-2. This study aims to explore the evolution of SARS-CoV-2 through a global analysis. To facilitate a comparative analysis of temporal mutation data, we developed CovidTGI, a Shiny web application. CovidTGI provides insights into observed mutation frequencies and the temporal relationships among mutations across various clades in different geographical regions. Our tool relies on a database that includes 2 million samples obtained from the National Center for Biotechnology Information (NCBI), along with 500 in-house Sicilian samples collected between May 2021 and June 2022. From this smaller group of samples, we identified key variants that are prevalent within a specific clade. Our tool is designed to study the evolution of SARS-CoV-2, which clearly follows a complex trajectory. This complexity highlights the necessity for sophisticated tools like CovidTGI to understand and track the evolution of this virus.

INTRODUCTION

In March 2020, the World Health Organization declared COVID-19 disease as a worldwide pandemic after the global spread of SARS-CoV-2 starting from the city of Wuhan in December 2019.[1] At the beginning of 2020, a global epidemiological surveillance program began, leading to the sequencing of millions of SARS-CoV-2-positive samples using next-generation sequencing (NGS).[2] Three years later, this pandemic highlighted the importance of promptness in fighting emerging infectious diseases as well as the importance of collaboration and data sharing among scientists. Thanks to its efforts, the scientific community has made significant progress in developing effective methods to trace SARS-CoV-2 evolution. In this context, NGS has been shown as an essential tool allowing researchers to rapidly sequence large amounts of viral samples to analyze in detail the changes of their mutational profiles.[2,3]

Even if the observed mutation frequency of SARS-CoV-2 is relatively lower than other RNA viruses,[4,5] its spread across the globe as well as its propensity to recombination led to the rapid occurrence of new variants.[6] After the initial prevalence of the Alpha variant, the coexistence of three main variants of concern (VOCs), Alpha, Beta, and Gamma, was reported starting in May 2021. In June 2021, the Delta variant emerged and represented the prevalent variant worldwide until November 2021, when the Omicron variant (the dominant VOC at the moment of writing) emerged, showing a unique mutational pattern that combines SNPs already described in previous variants[4,7] jointly with new mutations.

Despite the efficiency of NGS technology and the significant efforts made by public health authorities, the scientific community lags the rapid emergence of new variants, leading to a significant delay in implementing crucial public health interventions.[8] Our goal is to provide the scientific community with convenient access to pre-analyzed, annotated data, enabling deeper exploration of emerging SARS-CoV-2 mutation and co-mutations patterns. By making annotated data readily available, we aim to significantly reduce analysis time and reliance on high-performance workstations. Moreover, an accurate retrospective analysis of NGS data may help improve the knowledge about the real evolutionary path of the virus.

Starting from the emergence of SARS-CoV-2 in 2020, several web-based systems have been developed to explore the evolution of SARS-CoV-2 variants and their mutations in different locations.

COVID-19 CG[9] lets the user track the evolution of mutations in different lineages and retrieve information about the distributions of sequences sampled in different countries. The user can also obtain the co-occurrence counts of amino acid mutations and compare observed mutation frequencies. Though COVID-19

CG is constantly updated, the system has some important limitations: it provides no statistical testing for these analyses, and data about mutation co-occurrence counts and rates are shown only in US and Canada and for the last 90 days.

Outbreak.info[10,11] gives insights into the temporal evolution of SARS-CoV-2 and its variants in several countries and geographical areas. The system provides detailed reports about the diffusion of a specific lineage in each country and frequencies of characteristic mutations. Outbreak.info also lets the user search for publications and clinical trials related to SARS-CoV-2. However, the system provides no comparative analysis between mutations characterizing different lineages.

ViruClust[12] can be used to investigate the prevalence of lineages in specific countries or regions and compare the evolution of SARS-CoV-2 genome in different regions and time intervals. More precisely, based on specific spatiotemporal or lineage-based filters, the user can build two custom sets of SARS-CoV-2 genomic sequences, denoted as target and background, respectively, and statistically compare the prevalence of characteristic amino acid changes in the target and in the background. VariantHunter[13] extends the ViruClust tool, by introducing a statistical analysis of the increasing (or decreasing) trend of a specific amino acid change in space or time. However, this kind of analysis can be performed in a user-specified limited time window of 4 weeks, as the database used by VariantHunter is designed to pre-compute intermediate and aggregate results and is optimized for 4-week observations.

Overall, most of the existing web-based systems for studying the spatial and temporal evolution of SARS-CoV-2 variants and their mutations miss a statistical correlation analysis, or the proposed analyses are limited to specific time intervals or countries.

In this paper, we present the CovidTGI (Covid Temporal Genetic Instability), an innovative web application designed to track and analyze the evolutionary paths of SARS-CoV-2 mutations. This tool provides researchers and public health officials with a comprehensive platform to visualize and compare mutational trajectories of the virus across different geographic regions and time periods. By integrating sequencing data from various sources, including the National Center for Biotechnology Information (NCBI) database, CovidTGI enables users to explore the temporal dynamics and geographic prevalence of specific SARS-CoV-2 variants. This approach enables us to track the sequence of mutations and their relationships, helping us to predict the virus's behavior and understand its dynamics at specific points in time. The app's user-friendly interface allows for detailed correlation analyses of observed mutation frequencies and patterns, offering valuable insights into the virus's evolution and potential future mutations. CovidTGI offers a comprehensive global perspective on SARS-CoV-2 mutations cataloged in the NCBI database. By combining temporal data with observed frequencies specific to both clades and time periods, this tool surpasses state-of-the-art alternatives. It delivers a richer and more detailed correlation analysis of COVID-19 mutations, enabling users to explore the interactions and co-evolution of mutations within various clades. CovidTGI facilitates a deeper understanding of the virus's mutational dynamics, providing an invaluable resource for researchers and public health officials worldwide. Ultimately, CovidTGI serves as a crucial resource for understanding the ongoing evolution of SARS-CoV-2, aiding in the global effort to monitor and combat the COVID-19 pandemic. CovidTGI is available at https://covidapp.dioncogen.eu/.

RESULTS

Statistical analysis from whole world

CovidTGI app provides access to statistical analysis results. Users can explore data by clades or mutations and filter by region, country, or US state (Figures 1 and 2, see STAR Methods section for more details).

The analysis of the set of more than 2.1 million samples confirms that all over the world, the variants' distribution followed mostly the same trend. Through our analysis, we help the user to track the incidence of mutations through the months. Moreover, our app allows studying pairs of mutations, following their co-occurrence. We can highlight in the app some noteworthy mutations that we have extracted using Bayesian networks.

Omicron Bayesian network

By employing a Bayesian network inference model, we constructed two graphs illustrating ten mutations that frequently co-occur within the Omicron variant. Notably, as shown in Figure 3, several of these mutations have been previously identified as enhancing viral replication and significantly contributing to increased infectivity. The first network (Figure 3A) is based on samples collected in-house from Sicily, while the second (Figure 3B) includes all globally collected Omicron samples. A comparison of the two networks reveals that, over time, these mutations tend to co-occur more consistently, suggesting that they may play a role in stabilizing the virus. This co-occurrence pattern leads to the hypothesis that some mutations could confer a selective advantage, promoting viral stability and persistence.

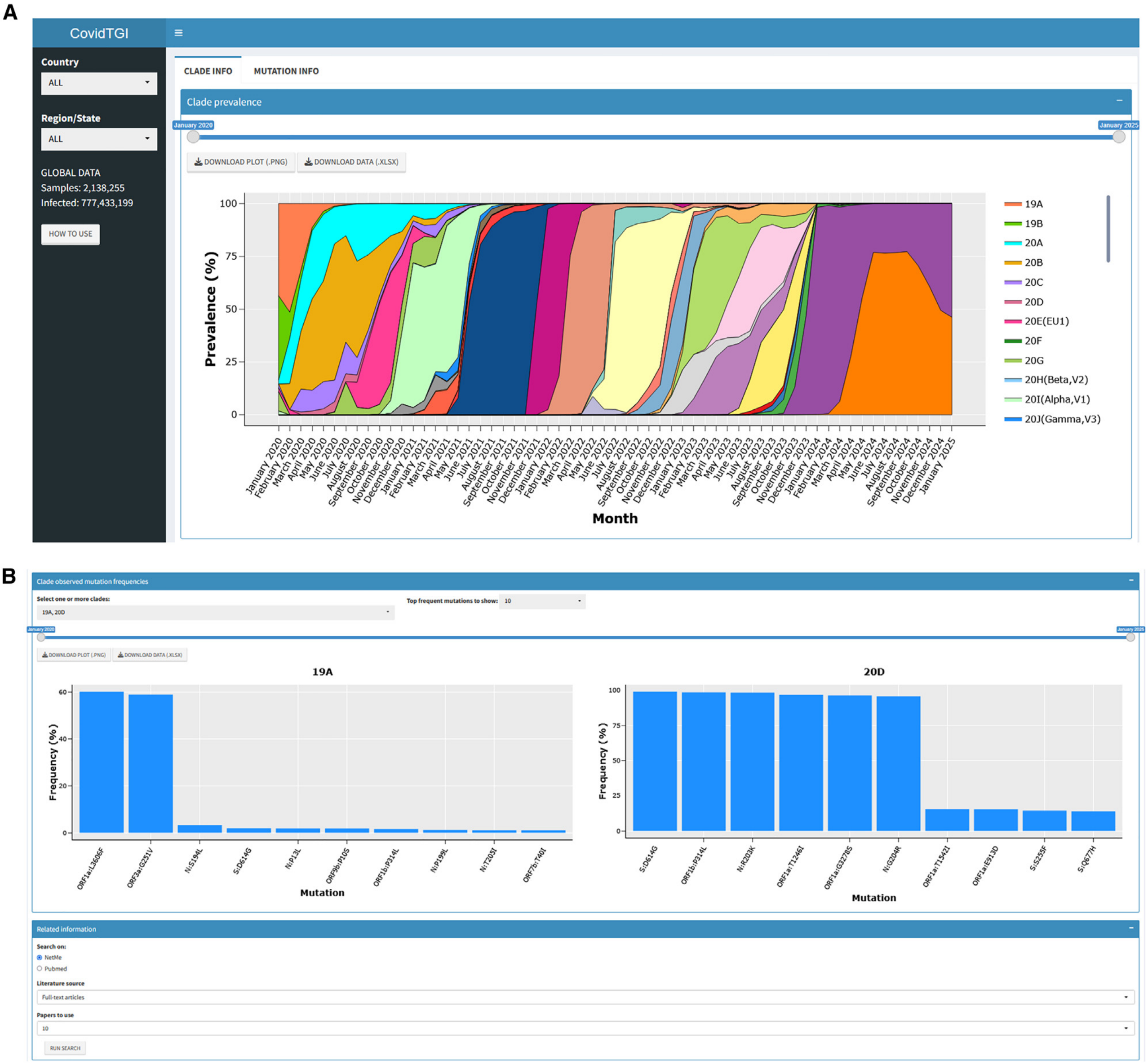
Observed mutations

Utilizing CovidTGI, significant observations emerge. In Figure 2, a striking co-occurrence is evident across all SARS-CoV-2 variants between nucleocapsid mutations R203K and G204R, exhibiting a notably high correlation rate. Directing our focus to Omicron, Figure 4 reveals a consistent pattern: the co-occurrence of N deletion S33 and R32 aligns concordantly with ORF9b deletion E27, N28, and A29. Specifically, N:R32- mirrors the temporal mutational pattern of ORF9b:E27- and ORF9b:N28-, while N:S33- parallels ORF9b:A29-. Exploring anti-correlated mutations, Figure 5 unveils an inverse relationship between N:G204R and M:I82T. An intriguing behavior unfolds with N:D343G and ORF3a:L106F, uniquely mutated solely in Omicron 21K (Figure 6).

DISCUSSION

The emergence of SARS-CoV-2 has resulted in a global pandemic, with the virus mutating and evolving quickly. Monitoring at a global scale, the virus evolution is a vital task in order to understand how it spreads and what are its genetic changes.

The prevalence of specific VOCs emerges as a result of ongoing and intricate genetic changes including strong purifying selection, recombination events, and positive selection.[14] Consequently, CovidTGI allows focusing on the less prevalent



Country: ALL

Region/State: ALL

GLOBAL DATA
Samples: 2,138,255
Infected: 777,433,199

Select one or more clades: 19A, 20D

Top frequent mutations to show: 10

Search on:
● NetMe
○ Pubmed

Literature source: Full-text articles

Papers to use: 10

Figure 1. CovidTGI app interface and "Clade Info" section with example clades in the global population

(A) The sidebar on left lets the user select the country and region/US state in which to analyze data. The "Clade Prevalence" box illustrates the monthly frequency rates of each clade in the chosen geographical area.

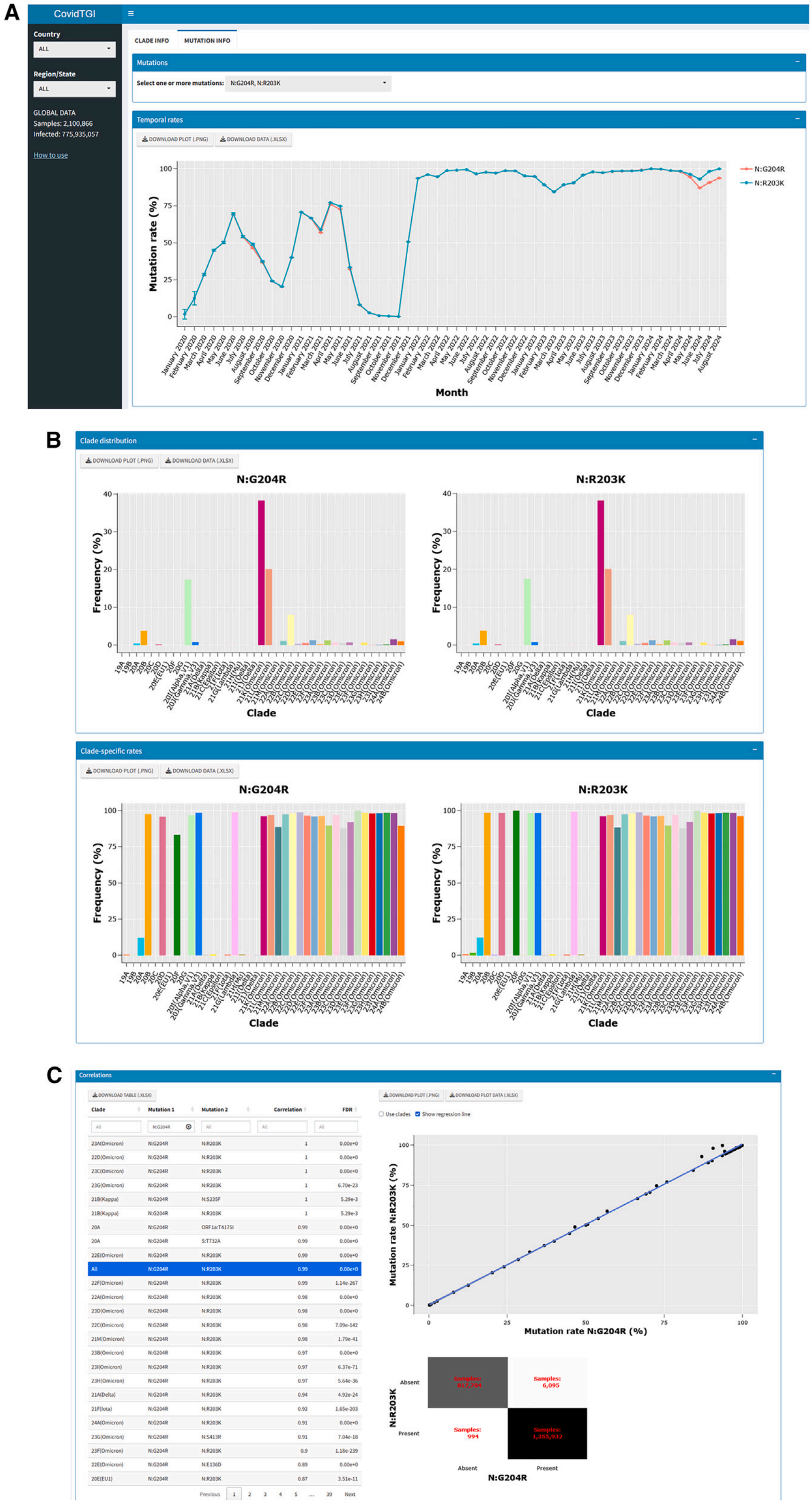
(B) "Clade Observed Mutation Frequencies" box shows the most frequent mutations for user-selected clades in the geographical area of interest.

clades or variants that exist between dominant VOCs, as they may serve as carriers of advantageous mutations.

Another crucial reason to monitor SARS-CoV-2 evolution is to identify mutations that may alter its antigenic properties, affecting the ability to evade the immune system and leading to the reduction of vaccine effectiveness.[15–17] While scientific tools and methods are continually evolving, the bottleneck is often driven by policy-related issues surrounding data sharing. These limitations must be addressed to fully leverage the power of bioinformatics in near-real-time variant surveillance. Therefore, readiness and accuracy remain critical to keep the pace of SARS-CoV-2 evolution.

Our case study highlights that, although Omicron appears phylogenetically closer to Alpha than to Delta, it displays a unique mutational pattern sharing genetic features with both these variants as well as harboring new mutations. Interestingly, we observed that some mutations shared by Alpha and Omicron almost disappeared during the Delta period. In particular, as it can be seen on our app, ORF1a.T3255I and S.T478K substitutions occurred for the first time during Delta dominance and were maintained over time with an increasing prevalence due to their involvement in viral replication and transmission.

Mutations such as S.Y144-, S.H69-, S.V70-, S.N501Y, S.P681H, N.G204R, ORF1a.F3677-, ORF1a.G3676-, and N.R203K, which



Country: ALL

Region/State: ALL

GLOBAL DATA
Samples: 2,100,866
Infected: 775,935,057

How to use

Select one or more mutations: N:G204R, N:R203K

□ Use clades ☑ Show regression line

Clade	Mutation 1	Mutation 2	Correlation	FDR
All	N:G204R	All	All	All
23A(Omicron)	N:G204R	N:R203K	1	0.00e+0
22D(Omicron)	N:G204R	N:R203K	1	0.00e+0
23C(Omicron)	N:G204R	N:R203K	1	0.00e+0
23G(Omicron)	N:G204R	N:R203K	1	6.70e-23
21B(Kappa)	N:G204R	N:S235F	1	5.29e-3
21B(Kappa)	N:G204R	N:R203K	1	5.29e-3
20A	N:G204R	ORF1a:T4175I	0.99	0.00e+0
20A	N:G204R	S:T732A	0.99	0.00e+0
22E(Omicron)	N:G204R	N:R203K	0.99	0.00e+0
All	N:G204R	N:R203K	0.99	0.00e+0
22F(Omicron)	N:G204R	N:R203K	0.99	1.14e-267
22A(Omicron)	N:G204R	N:R203K	0.98	0.00e+0
23D(Omicron)	N:G204R	N:R203K	0.98	0.00e+0
22C(Omicron)	N:G204R	N:R203K	0.98	7.09e-142
21M(Omicron)	N:G204R	N:R203K	0.98	1.79e-41
23B(Omicron)	N:G204R	N:R203K	0.97	0.00e+0
23I(Omicron)	N:G204R	N:R203K	0.97	6.37e-71
23H(Omicron)	N:G204R	N:R203K	0.97	5.64e-36
21A(Delta)	N:G204R	N:R203K	0.94	4.92e-24
21F(Iota)	N:G204R	N:R203K	0.92	1.65e-203
24A(Omicron)	N:G204R	N:R203K	0.91	0.00e+0
23G(Omicron)	N:G204R	N:S413R	0.91	7.04e-18
23F(Omicron)	N:G204R	N:R203K	0.9	1.18e-239
22E(Omicron)	N:G204R	N:E136D	0.89	0.00e+0
20E(EU1)	N:G204R	N:R203K	0.87	3.51e-11

Previous 1 2 3 4 5 … 39 Next

N:R203K \ N:G204R	Absent	Present
Absent	Samples: 817,704	Samples: 6,095
Present	Samples: 994	Samples: 1,355,932

Figure 2. “Mutation Info” section of CovidTGI app with example mutations in the Italian population

(A) “Temporal rates” box depicts monthly frequency rates of the mutations selected by the user in the “Mutations” box.

(B) “Clade Distribution” and “Clade specific rates” boxes show, respectively, the clade distribution and the clade-specific rates of user-selected mutations.

(C) “Correlations” box contains a table on the left reporting all significant (false discovery rate [FDR] ≤ 0.05) pairwise correlations between user-selected mutations with their Pearson correlation coefficient. Selecting a pair of significant mutations, a scatterplot of their clade distribution rates and the contingency table of their frequencies in the selected geographical area are shown on the right.

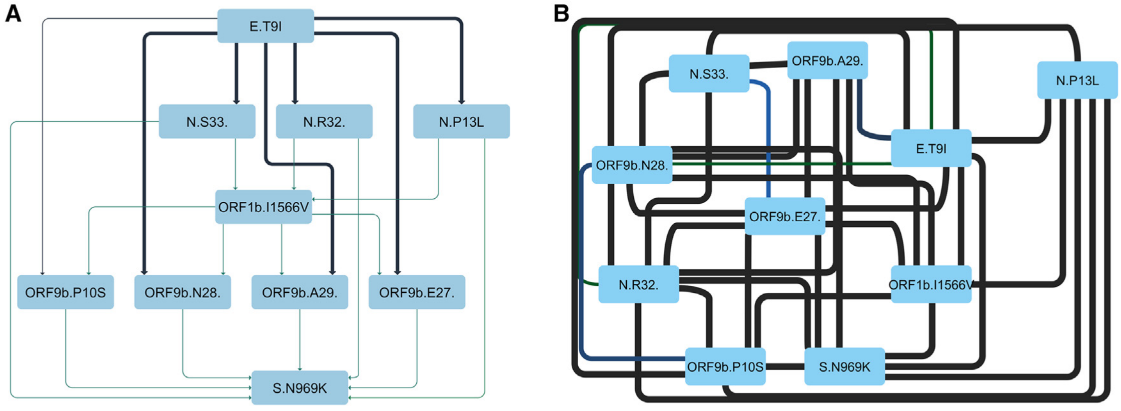


Figure 3. Bayesian network of Omicron clade

Bayesian network of (A) Sicilian mutations from May 2021 to June 2022 (see also Figure S1) and (B) rest of the World with a link strength of at least 0.8. Edges colors range from green for 0.8, blue for 0.9, and black for 1 as the link strength increases. Moreover, while the strength grows, the thickness of the edges increases.

declined during the Delta variant dominance, have re-emerged with Omicron, underscoring their biological significance. S.H69 and S.V70 can reduce monoclonal antibody binding to the spike protein, potentially compromising antibody-based treatments and vaccine efficacy.[18] The S.P681H mutation, altering the furin cleavage site, likely increases infectivity and transmission, especially with S.D614G, and contributes to resistance against type I interferon by evading interferon-induced transmembrane protein restriction.[19]

Mutations in ORF1a may potentially have an impact on viral replication and stability.[18,20]

For instance, the couple N.R203K/N.G204R (see Figure 4) has been found highly mutated in Alpha and Omicron. Its role is to promote virus replication efficiency acting in coordination with S.N501Y.[21] During Delta, the prevalent mutation was instead N.R203M.

This finding led us to suppose that the high observed mutation frequency, the remarkable number of mutational hotspots, and

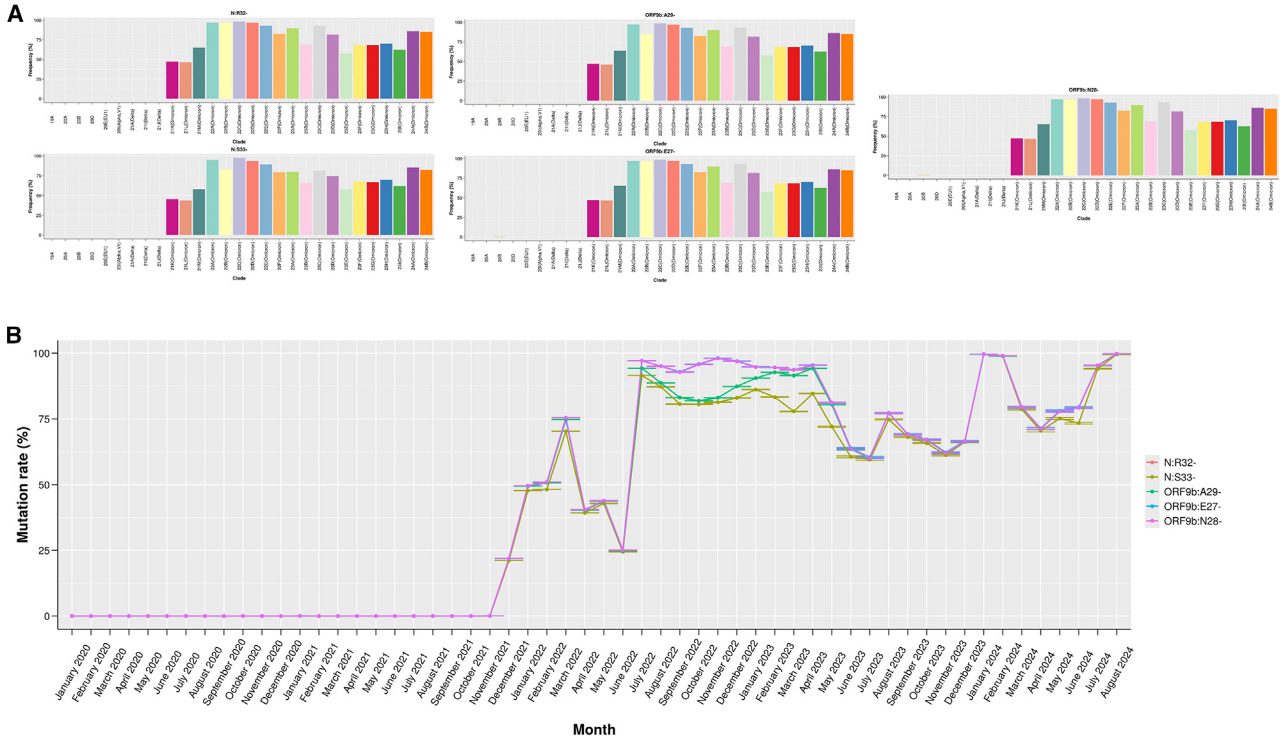


Figure 4. Example of bar plot and temporal observed mutation frequency of N.S33, N.R23, ORF9b.E27, ORF9b.N28, and ORF9b.A29

(A) Bar plot of clade-specific rate and (B) temporal observed mutation frequency of N deletion S33 and R32 and ORF9b deletion E27, N28, and A29.

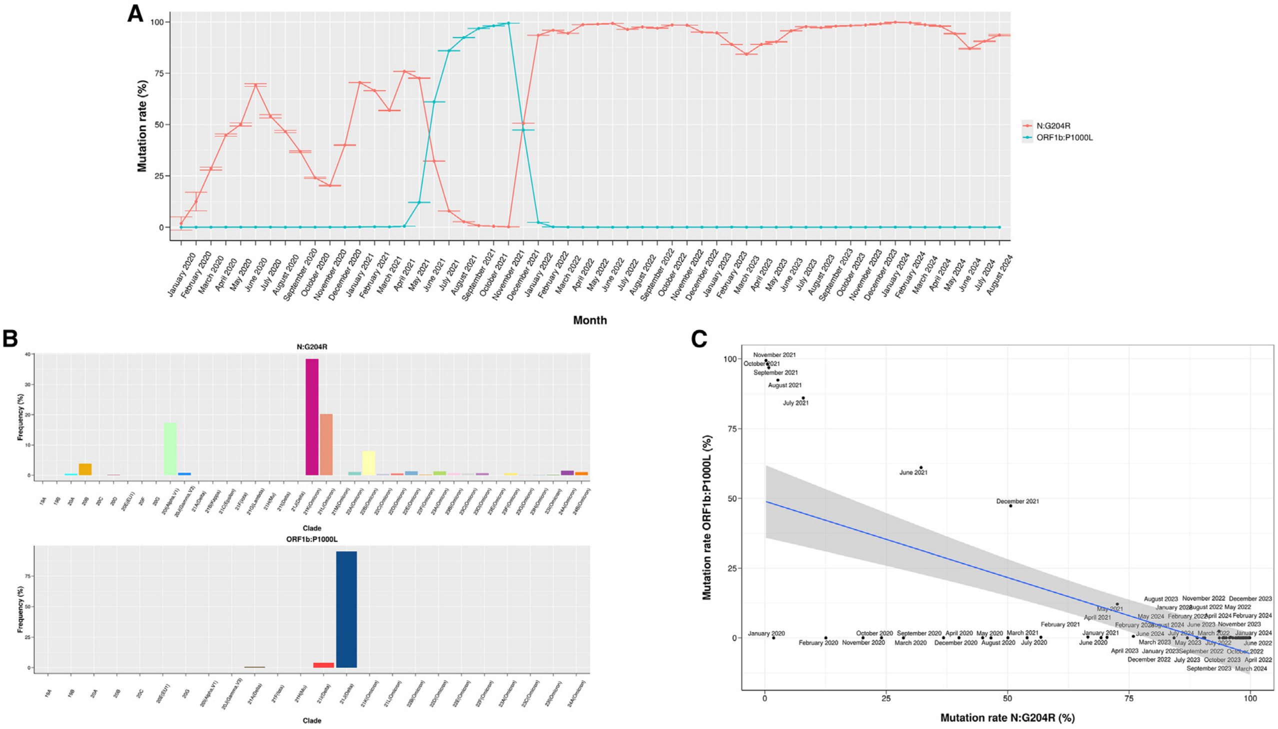


Figure 5. Observation of mutations N.G204R and ORF1b.P1000L

(A) Temporal observed mutation frequency, (B) clade distribution bar plot, and (C) correlation plot of N:G204R and ORF1b:P1000L. These mutations result anti-correlated during almost all period with an R = −0.73.

natural selection are responsible for the "genetic instability time interval" characterized by the emergence of new SNPs, while a specific variant is prevalent. The phase of genetic instability usually corresponds to social/epidemiological factors such as the relaxation of restrictive measures, increased domestic and international movement, and increased social interactions.[22]

This evolutionary path is largely driven by natural selection, which favors viruses that are able to adapt to their hosts and spread, not necessarily causing severe illness or death. This observation is consistent with the evolutionary history of some respiratory viruses such as influenza virus as well as other coronaviruses.[23]

By making use of a Bayesian network analysis, we investigated the possible causal relations between pairs of amino acid mutations. In particular, we selected the pairs with higher observed mutation frequency and higher strength to create a network. In Figure 3, we report the network created on top of Omicron mutations with a strength of at least 0.8. These associations have been confirmed using our app.

We have observed (see Figure S1) a significant mutation (0.54 strength), namely S:S477N, which plays a crucial role in the binding of the ACE2 receptor, thereby increasing the virus's contagiousness.[24] This mutation also aids in evading various forms of immune protection, such as vaccines, monoclonal antibodies, and polyclonal antibodies. Our analysis reveals a connection between this mutation and other genetic changes, including N deletions (R32 and S33), mutation P13L, ORF9b deletions (E27, N28, and A29) (see Figure 5), mutation P10S, and ORF1a mutation P3395H, with a strength between 0.52 and 0.8.

Notably, the N mutations primarily occur in the N arm zone, which is located before the N-terminal domain (NTD). The N arm plays a crucial role in regulating RNA binding, and as a result, these mutations are likely to impact the binding process.

On the other hand, it is worth noting that ORF9b plays a significant role in inhibiting type I interferon responses by interacting with TOM70. Consequently, the mutations occurring in the N terminus are likely to contribute to the stabilization of protein stability.[25]

Additionally, the E27 deletion of ORF9b is associated with spike mutations present in the receptor-binding domain (RTB), NTD, and S1/S2 regions. Particularly, the N440K mutation (see Figure S1), which is also linked to N:P13L, plays a significant role in enhancing infectivity and aiding the virus in evading vaccines, monoclonal antibodies, and polyclonal antibodies. These mutations are further associated with S.Q498R, S.H655Y, and S.N679K, which are already known to augment virus infectivity.[26]

Interestingly, the mutations N.D343G and ORF3a.L106F (Figure 6) have exclusively been detected together in the Omicron 21K variant. It is well established that these two mutations act as viral transmission suppressors, which may explain the relatively lower temporal trajectory observed for this specific variant.

Our research demonstrates that CovidTGI app is an effective tool for analyzing variant of interest and tracking their mutation

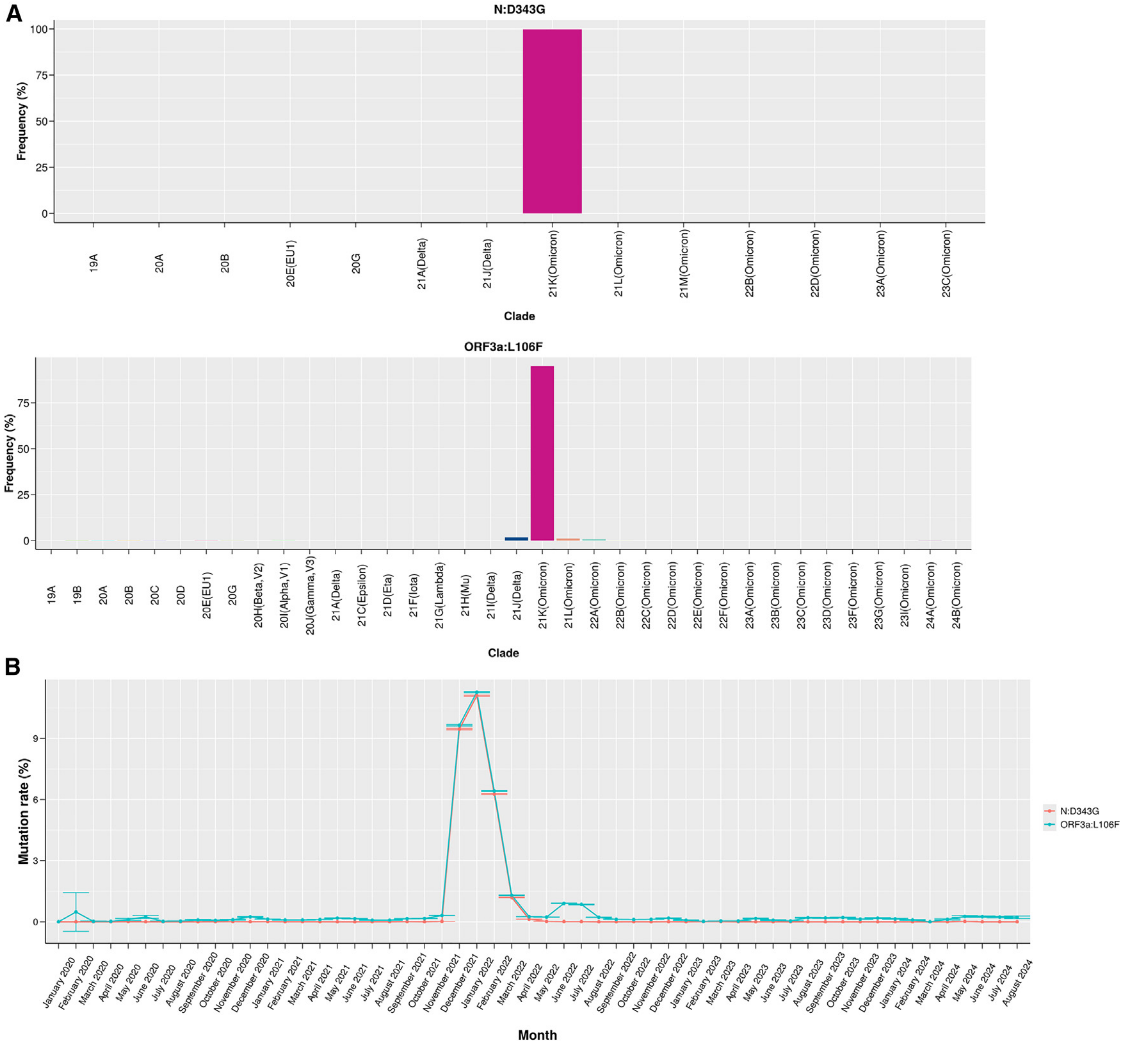


Figure 6. Bar plot and lineplot describing mutations N.D343G and ORF3a.L106F during the month and in several clades
(A) Clade distribution bar plot and (B) temporal observed mutation frequency of N:D343G and ORF3a:L106F.

patterns over time globally. Our aim is to incorporate all available worldwide samples, and we are currently working on adding GISAID[27] data to fill gaps in regional information.

Monitoring SARS-CoV-2 evolution is critical for understanding the spread of the virus, its antigenic changes, and treatment implications. The mere succession of the well-known VOCs usually does not reflect the real evolution of SARS-CoV-2, but it is only an epiphenomenon that hides a more tangled scenario characterized by a constant genetic movement. Therefore, regardless the decrease in worldwide COVID-19 mortality, it is crucial to keep tracing the emergence, the spread, and the potential implications of SARS-CoV-2 mutations with special regard to the local appearance of mutations involved in person-to-person transmission and/or immune evasion, even if they are harbored by subordinate clades/variants.

It is also important to highlight that the virus can follow multiple, simultaneous evolutionary pathways. This means that while the virus may evolve to become less virulent over time, it could also evolve in other ways that make it more transmissible or more resistant to treatments. Overall, the evolution of SARS-CoV-2 is still an active area of research, and this study underscores the importance of developing new tools and strategies to deal with emerging infectious diseases as well as the need of a faster and more accurate genomic epidemiology.

In conclusion, CovidTGI outstands as a robust platform for users to explore and analyze mutation data comprehensively, empowering researchers and healthcare professionals with valuable insights into the dynamics of mutation spread and its implications for disease management and control strategies.

Limitations of the study

While CovidTGI provides valuable insights into the temporal genetic instability of SARS-CoV-2, certain limitations should be considered. First, the study relies on publicly available sequencing data from the NCBI database and a limited set of in-house Sicilian samples, which may not fully represent all geographical and temporal variations. Additionally, while the Bayesian network analysis highlights key mutation co-occurrences, the causal relationships between these mutations remain speculative and require further experimental validation. The accuracy of the tool is also influenced by the completeness and quality of sequencing data, which can vary across regions and over time. In addition, the platform currently lacks integration with other major genomic databases, such as GISAID, which could enhance the comprehensiveness of the analysis. Finally, despite its advanced analytical capabilities, the study underscores the challenge of real-time variant surveillance, emphasizing the need for continuous updates and improvements in global data-sharing policies.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Alfredo Pulvirenti (alfredo.pulvirenti@unict.it).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data used have been downloaded from the NCBI virus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) in the SARS-CoV-2 Data Hub. Sicilian samples data can be found on Zenodo at: https://doi.org/10.5281/zenodo.14905657 (May and June 2021), https://doi.org/10.5281/zenodo.14907009 (June 2021), https://doi.org/10.5281/zenodo.14909193 (June and July 2021), https://doi.org/10.5281/zenodo.14909537 (July and November 2021), https://doi.org/10.5281/zenodo.14916165 (November and December 2021), https://doi.org/10.5281/zenodo.14917141 (December 2021 and January 2022), https://doi.org/10.5281/zenodo.14917375 (January 2022), https://doi.org/10.5281/zenodo.14923771 (January and February 2022),https://doi.org/10.5281/zenodo.14925875 (February and March 2022), https://doi.org/10.5281/zenodo.14926333 (March, April and May 2022), https://doi.org/10.5281/zenodo.14930257 (May 2022), and https://doi.org/10.5281/zenodo.14871965 (June 2022).
- All the original code has been deposited at GitHub and is publicly available as of the date of publication at the following link https://github.com/knowmics-lab/covid-app. CovidTGI is available at https://covidapp.dioncogen.eu/.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This research was partially supported by EU funding with the MUR PNRR Extended Partnership Initiative of Emerging Infection Disease Project no. PE 00000007 INF-ACT.

AUTHOR CONTRIBUTIONS

S.S., A.P., and G.S. conceived and coordinated the research. G.F.P. developed the SARS-CoV-2 analysis pipeline and performed bioinformatics analysis. G.M. and S.A. performed the statistical analysis. G.M. developed the CovidTGI app. C.B. and N.M. conducted the laboratory experiments. D.B. and P.G.B. conducted the NGS sequencing and biological data analysis. N.M., S.S., and A.P. conceived the experiments. G.F.P., G.M., and C.B. wrote draft manuscript. G.F.P., G.M., C.B., A.P., N.M., G.S., and S.S. reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - ○ Identification of mutations
  - ○ Analysis of sicilian samples
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ SHINY APP

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.112315.

Received: November 4, 2024
Revised: December 21, 2024
Accepted: March 25, 2025
Published: March 28, 2025

REFERENCES

1. Cucinotta, D., and Vanelli, M. (2020). WHO declares COVID-19 a pandemic. Acta Biomed. *91*, 157–160. https://doi.org/10.23750/abm.v91i1.9397.
2. John, G., Sahajpal, N.S., Mondal, A.K., Ananth, S., Williams, C., Chaubey, A., Rojiani, A.M., and Kolhe, R. (2021). Next-Generation Sequencing (NGS) in COVID-19: A Tool for SARS-CoV-2 Diagnosis, Monitoring New Strains and Phylodynamic Modeling in Molecular Epidemiology. Curr. Issues Mol. Biol. *43*, 845–867. https://doi.org/10.3390/cimb43020061.
3. Chiara, M., D'Erchia, A.M., Gissi, C., Manzari, C., Parisi, A., Resta, N., Zambelli, F., Picardi, E., Pavesi, G., Horner, D.S., and Pesole, G. (2021). Next generation sequencing of SARS- 363 CoV-2 genomes: challenges, applications and opportunities. Briefings Bioinf. *22*, 616–630. https://doi.org/10.1093/bib/bbaa297.
4. Manzanares-Meza, L.D., and Medina-Contreras, O. (2020). SARS-CoV-2 and influenza: a comparative overview and treatment implications. Bol. Med. Hosp. Infant. Mex. *77*, 262–273. https://doi.org/10.24875/bmhim.20000183.
5. Kawasaki, Y., Abe, H., and Yasuda, J. (2023). Comparison of genome replication fidelity between sars-cov-2 and influenza a virus in cell culture. Sci. Rep. *13*, 13105. https://doi.org/10.1038/s41598-023-40463-4.
6. Keusch, G.T., Amuasi, J.H., Anderson, D.E., Daszak, P., Eckerle, I., Field, H., Koopmans, M., Lam, S.K., Das Neves, C.G., Peiris, M., et al. (2022). Pandemic origins and a one health approach to preparedness and prevention: Solutions based on sars-cov-2 and other rna viruses. Proc. Natl. Acad. Sci. USA *119*, e2202871119. https://doi.org/10.1073/pnas.2202871119.
7. Kannan, S.R., Spratt, A.N., Sharma, K., Chand, H.S., Byrareddy, S.N., and Singh, K. (2022). Omicron SARS-CoV-2 variant: Unique features and their impact on pre-existing antibodies. J. Autoimmun. *126*, 102779. https://doi.org/10.1016/j.jaut.2021.102779.
8. Wu, J., Scarabel, F., Majeed, B., Bragazzi, N.L., and Orbinski, J. (2021). The impact of public health interventions on delaying and mitigating against replacement by SARS-CoV-2 variants of concern. SSRN Journal. https://doi.org/10.2139/ssrn.3779007.

9. Chen, A.T., Altschuler, K., Zhan, S.H., Chan, Y.A., and Deverman, B.E. (2021). COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest. Elife *10*, e63409. https://doi.org/10.7554/eLife.63409.

10. Gangavarapu, K., Latif, A.A., Mullen, J.L., Alkuzweny, M., Hufbauer, E., Tsueng, G., Haag, E., Zeller, M., Aceves, C.M., Zaiets, K., et al. (2023). Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. Nat. Methods *20*, 512–522. https://doi.org/10.1038/s41592-023-01769-3.

11. Tsueng, G., Mullen, J.L., Alkuzweny, M., Cano, M., Rush, B., Haag, E., Lin, J., Welzel, D.J., Zhou, X., Qian, Z., et al. (2023). Outbreak.info Research Library: a standardized, searchable platform to discover and explore COVID-19 resources. Nat. Methods *20*, 536–540. https://doi.org/10.1038/s41592-023-01770-w.

12. Cilibrasi, L., Pinoli, P., Bernasconi, A., Canakoglu, A., Chiara, M., and Ceri, S. (2022). ViruClust: direct comparison of SARS-CoV-2 genomes and genetic variants in space and time. Bioinformatics *38*, 1988–1994. https://doi.org/10.1093/bioinformatics/btac030.

13. Pinoli, P., Canakoglu, A., Ceri, S., Chiara, M., Ferrandi, E., Minotti, L., and Bernasconi, A. (2023). VariantHunter: a method and tool for fast detection of emerging SARS-CoV-2 variants. Database *2023*, baad044. https://doi.org/10.1093/database/baad044.

14. Frutos, R., Pliez, O., Gavotte, L., and Devaux, C.A. (2022). There is no "origin" to SARS-CoV-2. Environ. Res. *207*, 112173. https://doi.org/10.1016/j.envres.2021.112173.

15. Malik, J.A., Ahmed, S., Mir, A., Shinde, M., Bender, O., Alshammari, F., Ansari, M., and Anwar, S. (2022). The SARS-CoV-2 mutations versus vaccine effectiveness: New opportunities to new challenges. J. Infect. Public Health *15*, 228–240. https://doi.org/10.1016/j.jiph.2021.12.014.

16. Hafiz, I., Illian, D.N., Meila, O., Utomo, A.R.H., Susilowati, A., Susetya, I.E., Desrita, D., Siregar, G.A., and Basyuni, M. (2022). Effectiveness and Efficacy of Vaccine on Mutated SARS-CoV-2 Virus and Post Vaccination Surveillance: A Narrative Review. Vaccines *10*, 82. https://doi.org/10.3390/vaccines10010082.

17. Noh, J.Y., Jeong, H.W., and Shin, E.C. (2021). SARS-CoV-2 mutations, vaccines, and immunity: implication of variants of concern. Signal Transduct. Targeted Ther. *6*, 203. https://doi.org/10.1038/s41392-021-00623-2.

18. Biswas, S., Mahmud, S., Mita, M.A., Afrose, S., Hasan, M.R., Paul, G.K., Shimu, M.S.S., Uddin, M.S., Zaman, S., Park, M.N., et al. (2022). The Emergence of SARS-CoV-2 Variants With a Lower Antibody Response: A Genomic and Clinical Perspective. Front. Med. *9*, 825245. https://doi.org/10.3389/fmed.2022.825245.

19. Lista, M.J., Winstone, H., Wilson, H.D., Dyer, A., Pickering, S., Galao, R.P., De Lorenzo, G., Cowton, V.M., Furnon, W., Suarez, N., et al. (2022). The P681H Mutation in the Spike Glycoprotein of the Alpha Variant of SARS-CoV-2 Escapes IFITM Restriction and Is Necessary for Type I Interferon Resistance. J. Virol. *96*, e0125022. https://doi.org/10.1128/jvi.01250-22.

20. Thakur, S., Sasi, S., Pillai, S.G., Nag, A., Shukla, D., Singhal, R., Phalke, S., and Velu, G.S.K. (2022). SARS-CoV-2 Mutations and Their Impact on Diagnostics, Therapeutics and Vaccines. Front. Med. *9*, 815389. https://doi.org/10.3389/fmed.2022.815389.

21. Wu, H., Xing, N., Meng, K., Fu, B., Xue, W., Dong, P., Tang, W., Xiao, Y., Liu, G., Luo, H., et al. (2021). Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. Cell Host Microbe *29*, 1788–1801.e6. https://doi.org/10.1016/j.chom.2021.11.005.

22. Zhang, M., Wang, S., Hu, T., Fu, X., Wang, X., Hu, Y., Halloran, B., Li, Z., Cui, Y., Liu, H., et al. (2022). Human mobility and COVID-19 transmission: a systematic review and future directions. Spatial Sci. *28*, 501–514. https://doi.org/10.1080/19475683.2022.2041725.

23. King, A. (2021). The coronavirus could end up mild like a common cold. New Sci. *249*, 12–13. https://doi.org/10.1016/S0262-4079(21)00084-1.

24. Mondeali, M., Etemadi, A., Barkhordari, K., Mobini Kesheh, M., Shavandi, S., Bahavar, A., Tabatabaie, F.H., Mahmoudi Gomari, M., and Modarressi, M.H. (2023). The role of S477N mutation in the molecular behavior of SARS-CoV-2 spike protein: An in-silico perspective. J. Cell. Biochem. *124*, 308–319. https://doi.org/10.1002/jcb.30367.

25. Hossain, A., Akter, S., Rashid, A.A., Khair, S., and Alam, A.S.M.R.U. (2022). Unique mutations in SARS-CoV-2 Omicron subvariants' non-spike proteins: Potential impacts on viral pathogenesis and host immune evasion. Microb. Pathog. *170*, 105699. https://doi.org/10.1016/j.micpath.2022.105699.

26. Yang, H.C., Wang, J.H., Yang, C.T., Lin, Y.C., Hsieh, H.N., Chen, P.W., Liao, H.C., Chen, C.H., and Liao, J.C. (2022). Subtyping of major SARS-CoV-2 variants reveals different transmission dynamics based on 10 million genomes. PNAS nexus *1*, pgac181. https://doi.org/10.1093/pnasnexus/pgac181.

27. Khare, S., Gurry, C., Freitas, L., Schultz, M.B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R.T., Yeo, W., et al. (2021). Gisaid's role in pandemic response. China CDC Wkly. *3*, 1049–1051. https://doi.org/10.46234/ccdcw2021.255.

28. Aksamentov, I., Roemer, C., and Hodcroft, E. (2021). Nextclade: clade assignment, mutation calling and quality control for viral genomes. J. Open Source Softw. *6*, 3773. https://doi.org/10.21105/joss.03773.

29. Nextclade. https://clades.nextstrain.org.

30. Toole, A.O.'., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J.T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B., et al. (2021). Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evolution *7*, veab064. https://doi.org/10.1093/ve/veab064.

31. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

32. Trimgalore. https://github.com/FelixKrueger/TrimGalore.

33. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv. 1303.3997. https://doi.org/10.48550/arXiv.1303.3997.

34. Picard. http://broadinstitute.github.io/picard/.

35. Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., and Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. *40*, 11189–11201. https://doi.org/10.1093/nar/gks918.

36. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

37. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at bioRxiv. https://doi.org/10.1101/201178.

38. vafator. GitHub - TRON-Bioinformatics/vafator: Annotate variants in a VCF file with technical annotations from one or more BAMs. https://github.com/TRON-Bioinformatics/vafator.

39. Scutari, M. (2010). Learning bayesian networks with the bnlearn r package. J. Stat. Software *35*, 1–22.

40. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

41. Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2024). shiny: Web Application Framework for R. R package version 1.10. https://github.com/rstudio/shiny, https://shiny.posit.co/.

42. Pubmed. https://pubmed.ncbi.nlm.nih.gov/.

43. Muscolino, A., Di Maria, A., Rapicavoli, R.V., Alaimo, S., Bellomo, L., Billeci, F., Borzì, S., Ferragina, P., Ferro, A., and Pulvirenti, A. (2022). NETME: on-the-fly knowledge network construction from biomedical literature. Appl. Netw. Sci. *7*, 1. https://doi.org/10.1007/s41109-021-00435-x.

44. Di Maria, A., Bellomo, L., Billeci, F., Cardillo, A., Alaimo, S., Ferragina, P., Ferro, A., and Pulvirenti, A. (2024). NetMe 2.0: a web-based platform for extracting and modeling knowledge from biomedical literature as a labeled graph. Bioinformatics *40*, btae194, 04 ISSN 1367-4811. https://doi.org/10.1093/bioinformatics/btae194.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
Agilent RNA 6000 Nano Reagents	Agilent Technologies	Cat. No. 5067-1511
QIAseq DIRECT SARS-CoV-2 Library Kit	QIAGEN	Cat. No. 333891
Agilent High Sensitivity DNA Kit	Agilent Technologies	Cat. No. 5067-4626
MiSeq® v3 reagent Kit	Illumina	Cat. No. 15043895
Deposited data		
NCBI SARS-CoV-2 fasta sequences	NCBI SARS-CoV-2 Data Hub	https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2697049
Sicilian Samples	This paper	https://doi.org/10.5281/zenodo.14905657 https://doi.org/10.5281/zenodo.14907009 https://doi.org/10.5281/zenodo.14909193 https://doi.org/10.5281/zenodo.14909537 https://doi.org/10.5281/zenodo.14916165 https://doi.org/10.5281/zenodo.14917141 https://doi.org/10.5281/zenodo.14917375 https://doi.org/10.5281/zenodo.14923771 https://doi.org/10.5281/zenodo.14925875 https://doi.org/10.5281/zenodo.14926333 https://doi.org/10.5281/zenodo.14930257 https://doi.org/10.5281/zenodo.14871965
Software and algorithms		
Nextclade	Aksamentov et al.[28,29]	https://clades.nextstrain.org/
Pangolin	O'Toole et al.[30]	https://cov-lineages.org/resources/pangolin.html
R Statistical Software	R Foundation	http://www.R-project.org
FastQC	Andrews[31]	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
TrimGalore	Krueger[32]	https://github.com/FelixKrueger/TrimGalore
BWA	Li[33]	https://bio-bwa.sourceforge.net/
Picard	Broad Institute[34]	http://broadinstitute.github.io/picard/
LoFreq	Wilm at al.[35]	https://csb5.github.io/lofreq/
BCFtools	Danecek et al.[36]	https://samtools.github.io/bcftools/
GATK	Poplin et al.[37]	https://gatk.broadinstitute.org/hc/en-us
Vafator	TRON-Bioinformatics[38]	https://github.com/TRON-Bioinformatics/vafator
bnlearn	Scutari[39]	https://www.bnlearn.com/
Cytoscape	Shannon et al.[40]	https://cytoscape.org/
Shiny	Chang et al.[41]	https://shiny.posit.co/
CovidTGI app	This paper	https://covidapp.dioncogen.eu/ https://github.com/knowmics-lab/covid-app
Other		
Data about the monthly number of infected people observed in a specific region or country	github	https://github.com/owid/616.covid-19-data
Pubmed	NCBI	https://pubmed.ncbi.nlm.nih.gov/
NetMe	Muscolino et al.[33] Di Maria et al.[34]	https://netme.click/#/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study did not contain any experimental model and study participant.

METHOD DETAILS

Identification of mutations

We collected from NCBI almost all the covid samples from January 2020 to January 2025 obtaining 2,138,255 samples. The dataset is monthly updated with new NCBI samples. Using Nextclade (v3.7)[28,29] and Pangolin (v4.0.6),[30] we analyzed the FASTA file derived from NCBI to obtain clade, lineage, and amino acid substitution information. Patients' mutations have been then stored into matrices represented in R (v.4.2.1), including the country and the region (or US state) where the samples were collected.

Analysis of sicilian samples

We collected 625 samples from Oriental Sicily as a case study (further information about samples extraction and sequencing can be found in supplementary document S1). A custom pipeline was developed for the identification of clade-specific mutations through FASTQ analysis. Initially, all samples were assessed using FastQC,[31] and only those passing the quality filter were selected for further processing. The selected FASTQ files were then trimmed with TrimGalore[32] and aligned using BWA-MEM,[33] with the "Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1" (NC_045512.2) genome from NCBI as the reference.

Subsequently, the aligned SAM files were reordered and sorted using Picard.[34] Variant calling was performed using three different tools: LoFreq,[35] BCFtools,[36] and GATK HaplotypeCaller,[36,37] each configured with a minimum base quality and mapping quality of 20. Vafator[38] was used to annotate allele frequency (AF) for each mutation. Variants with an AF below 0.5 were filtered using BCFtools filter, and consensus sequences were generated with BCFtools consensus.

Clade, lineage, and amino acid substitution information were obtained using Nextclade (v3.7)[28,29] and Pangolin (v4.0.6).[30] Finally, a Bayesian network was constructed to analyze the relationships among Sicilian mutations, employing the R package bnlearn[39] with the hill-climbing (HC) algorithm and performing three bootstrap replicates. To identify the most prevalent mutation pairs for each Variant of Concern (VOC), the Bayesian network was analyzed using Cytoscape.[40]

QUANTIFICATION AND STATISTICAL ANALYSIS

For each country and each region, we calculated the monthly mutational rates of all mutations observed in samples of that geographical area. Due to the incompleteness of sequencing data and the low prevalence of Sars-Cov-2 in some countries/regions, we also calculated (where data were available) error margins for the monthly mutational rates related to the population size of infected in a particular area. Specifically, we first collected from https://github.com/owid/covid-19-data data about the monthly numbers of infected people observed in a specific region or country. For each month M, we performed a correction of sample size for finite population, where a sample is represented by the number of patients sequenced for Sars-Cov-2 in month M. Next, we calculated error margins using the Cochran formula, based on a 99% confidence interval.

To further investigate the prevalence of a mutation in certain clades, we also calculated the clade distribution and the clade-specific rates of all mutations for each geographical area. Given a mutation *m*, the clade distribution is the frequency distribution of *m* across all Sars-Cov-2 clades, while the clade-specific rate is the relative number of samples in a clade *C* having mutation *m*.

We also performed a correlation analysis in order to find all significant global and clade-specific pairwise correlations between observed mutation frequencies of all mutations observed in the same geographical area. In order to filter only mutation pairs with significant correlations, we performed a 3-step analysis. First we ran a Fisher exact test, starting from a 2x2 contingency table of the frequencies of the two mutations in a specific geographical area. P-values returned by the Fisher test were then corrected for multiple testing using Benjamini-Hockberg method. Finally, for mutation pairs with False Discovery Rate (FDR) ≤ 0.05 we calculated the Pearson correlation.

All the rates and pairwise correlations were computed considering only mutations with population frequency $\geq 0.1\%$ in the considered geographical area. The statistical analysis was also performed by aggregating samples of all regions of a given country and samples of all countries worldwide.

SHINY APP

The results of the statistical analyses are easily accessible through CovidTGI, an user-friendly Shiny[41] R web-app designed for comparison purposes.

The app is divided into two sections, named "Clade Info" and "Mutation Info", letting the user browse the data starting from specific clades or specific mutations, respectively (Figures 1 and 2). By default, all the statistics reported by the app refer to the world population, but by using the app's sidebar, users can optionally restrict the analysis to specific regions, countries, or US states.

The "Clade Info" section (Figure 1) contains three informative boxes: "Clade prevalence", "Clade observed mutation frequencies" and "Related information". "Clade prevalence" box (Figure 1A) contains an area plot with a time slider illustrating which clades are circulating in a certain period of time in the user-selected geographical area. The "Clade observed mutation frequencies" box

(Figure 1B) contains bar plots illustrating the most frequent mutations observed in one or more user-selected clades in a specific time interval that can be adjusted using a slider. The "Related Information" box (Figure 1B) features convenient functionalities for users to further explore the selected clades. By querying the Pubmed database[42] it is possible to retrieve additional scientific literature on the selected clades, expanding the knowledge base. Alternatively, users can query the NetMe tool,[43,44] to construct a comprehensive knowledge network. By analyzing the most relevant PubMed full-texts or abstracts about a query term, NetMe generates a network encompassing genes, diseases, drugs, and other biological entities related to the user's query, facilitating a deeper understanding of clade-related topics.

The "Mutation Info" section (Figure 2) let the user retrieve information about one or more specific mutations selected by the user through the "Mutations" box. The "Temporal rates" box (Figure 2A) depicts monthly observed mutation frequency of the chosen mutations, allowing for a clear understanding of how these mutations fluctuate over time. Where available, error margins for the monthly observed mutation frequency are included as error bars in the plot. The next two boxes "Clade distribution" and "Clade-specific rates" (Figure 2B) contain bar plots illustrating the clade distribution and the clade-specific rates of the mutations in each clade where they are observed. The bar plots help the user enhance its understanding of mutation prevalence across different genetic backgrounds. The "Correlations" box (Figure 2C) offers a comprehensive table showcasing significant pairwise correlations (FDR $\leq$ 0.05) between observed mutation frequencies. These correlations can be explored on both a global and clade-specific level, providing valuable insights into potential interactions between mutations. Clicking on a specific row within the correlation table grants users access to the underlying contingency table, aiding in the interpretation of Fisher exact p-values. Additionally, users can visualize the correlation through a scatter plot, either monthly or clade-by-clade (based on clade distribution rates), enabling a deeper analysis of observed mutation frequency trends. At the end of the "Mutation Info" section, the "Related information" box let the user retrieve additional information about the selected mutations by querying Pubmed database or NetMe tool.