VISTA: Variant Insights, Surveillance and Tracking Analytics, a Visualization Platform for Public Health

Karan Jogi*
Discovery Partners Institute

Stephanie Greenwald†
Argonne National Laboratory

Marco Colombo‡
Discovery Partners Institute

Sarah Owens§
Argonne National Laboratory

Rigel Gjomemo¶
Discovery Partners Institute

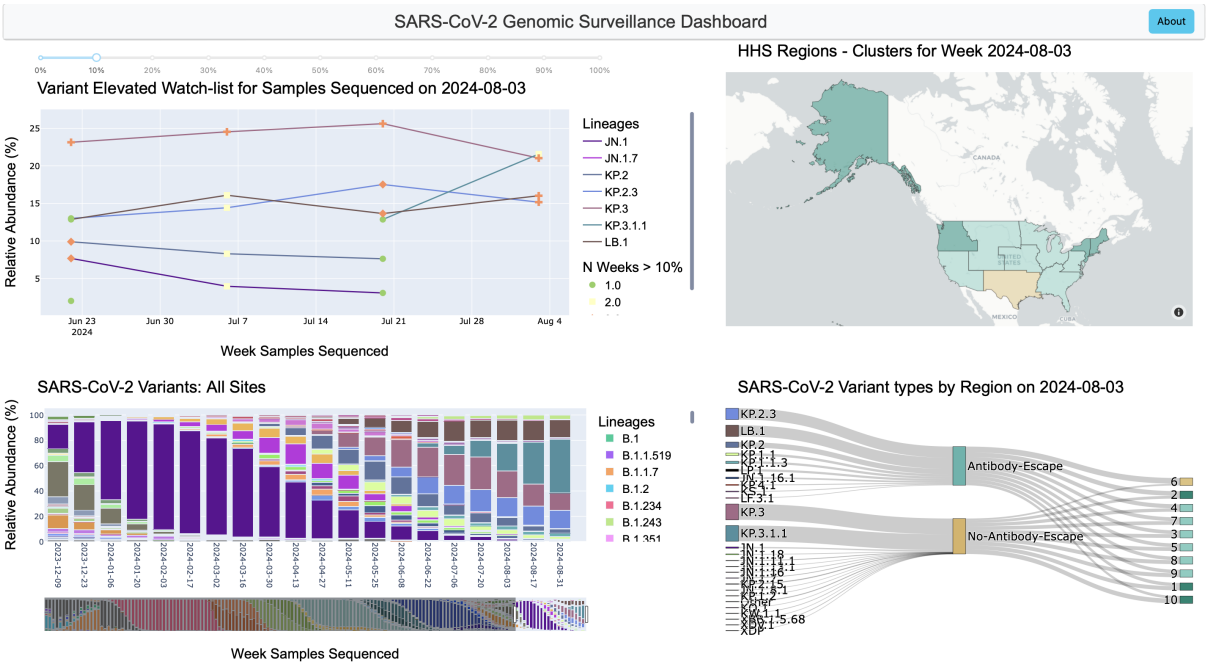


Figure 1: VISTA interface for SARS-CoV-2 Variant Tracking. The figure shows a public CDC dataset subdivided into 10 HHS regions. VISTA can be used with additional datasets (e.g., single-state regions). (A) The elevated watch list displays significant variants of concern (B) The regional similarity panel shows HHS regions clustered based on variant profile. (C) The short and long-term regional trend shows a temporal view of the variant profile (D) The mutation-based variant split shows the regional distribution of different types of variants. (link: https://usa-sarscov2-genomic-dash.azurewebsites.net)

Abstract

The SARS-CoV-2 pandemic gave rise to multiple data analytics strategies and visualizations related to the spread of the virus and community health. These visualizations aim to track SARS-CoV-2 variant diffusion among the population and help public health officials determine what interventions and policies could counter the spread of SARS-CoV-2 variants of interest. Existing visualizations for variant diffusion are typically static and have a rigid workflow. In this paper, we present a new dashboard for the visualization of SARS-CoV-2 variants diffusion. The dashboard, named VISTA, combines multiple datasets in an easy-to-use and intuitive interface that allows users to visually generate numerous analytic tasks, including correlation among regions, comparisons between regional and overall trends, and characterizations of variants with features of interest to public health officials.

Index Terms: Genomic data, visualization, dashboards, public health.

1 Introduction

Since appearing in 2020, SARS-CoV-2 continues to be a significant public health issue, necessitating multiple community-level tracking methods. Among these, variant tracking from sequencing data is vital for understanding how the virus evolves, adapts and spreads. The results from sequence data can directly impact the effectiveness of public health strategies, vaccines, and treatments. The mutations of each SARS-CoV-2 variant can effect their transmissibility, severity of host infection, and resistance to host immunity. These mutations can potentially lead to new outbreaks or diminish the efficacy of existing vaccines. By closely monitoring these changes through genomic sequencing of the viral RNA extracted from the wastewater, public health officials can identify and track the emergence of variants of concern. Visual analytics is an important method that transforms complex data into actionable insights by helping intuitively identify patterns and trends. SARS-CoV-2 dashboards can

*e-mail: kjogi2@uillinois.edu
†e-mail: smoormann@anl.gov
‡e-mail: mcolom4@uic.edu
§e-mail: sarah.owens@anl.gov
¶e-mail: rgjome1@uic.edu

be used to visualize viral genomic data, support public analytics, and are a convenient way to disseminate results to the general public.

These dashboards have been an important tool utilized by public health officials. The Johns Hopkins University SARS-CoV-2 dashboard provides a global overview of cases, deaths, and recoveries, making it easier to track the pandemic's progression across different regions [2]. Similarly, the CDC's COVID Data Tracker offers detailed insights into trends within the United States, including county-level data, testing, vaccination rates, and variants [3].

The two most common visualizations of genomic data used by public health departments are the phylogenetic tree and the stacked bar chart shown in Figures 2a and 2b[1]. Depicted on the left, the phylogenetic tree provides information on the evolutionary relationship between SARS-CoV-2 lineages, indicating the last common ancestors and their descendants. The stacked bar chart on the right provides information on the relative amount of each lineage at multiple points in time. Although useful, these visualizations often provide limited information about additional concerns for public health, such as immune resistance, comparison of genomic data among multiple locations of interest, and aggregation of relative abundances for regions of interest.

To provide more insight, the visualization of variant genomic data must be further specialized and integrated with additional datasets of public interest, such as population data and geographic information. In addition, such visualizations can enable public health officials to conduct a variety of analytical visualizations. To the best of our knowledge, there are currently few visualizations attempting to integrate genomic data about SARS-CoV-2 into larger workflows and datasets. Some visualization dashboards, such as Nexstrain and GISAID, provide additional filtering capabilities via criteria input forms to produce variations on the phylogenetic tree and stacked bar chart [15, 1]. After reviewing existing public health dashboards [21], we believe that there exists consistent challenges with genomic data visualizations.

In this paper, we describe a visual analytics dashboard for SARS-CoV-2 variant tracking that is designed to support several topics of interest for public health officials in an intuitive format. A view of our dashboard is shown in Figure 1. In particular, the dashboard is composed of four interactive areas, each dedicated to a different view of the data. The dashboard incorporates an interactive workflow with dynamic cross-filtering that supports users by correlating and comparing regional trends in real-time. Users can interact directly with the dashboard elements by zooming in and out of regions to visualize details at different levels of granularity.

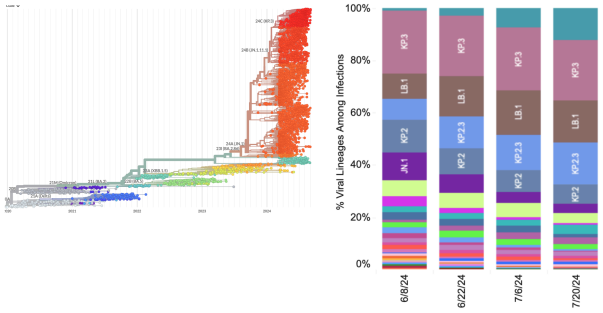
This paper is organized as follows. Section 2 gives an overview of related work. Section 3 describes our dashboard in detail, while section 4 provides some description of future work.

2 Related Work

The two most common visualizations for variant tracking of genomic data are the phylogenetic tree and the stacked bar chart, depicted in Figures 2a and 2b.

A phylogenetic tree describes the evolutionary relationship between biological entities in a graphical format [19, 20]. In the case of SARS-CoV-2, phylogenetic trees depict the evolutionary relationship between the viral variants to the original Wuhan-Hu-1 genome (MN908947.3) and to each other. Relationships between viral variants can be inferred by the branching order and amount of evolutionary change inferred by the branch lengths between nodes. Phylogenetic trees can trace their origins to medieval family trees [17], with many prominent naturalists and biologists, such as Lamarck, Darwin, Woese, Kandlet, Wheelis, Hillis, and Hedges (among many others), modifying and improving the model

[1]Images' source: https://covid.cdc.gov/covid-data-tracker/



(a) Phylogenetic Tree (b) Stacked Bar Chart

Figure 2: Popular visualizations of genomic data.

of trees we use today. Phylogenetic trees are a popular visualization of genomic data we use today. These trees the evolutionary relationship of different species to each other and are often made using the genetic information generated through generation sequencing technology. Phylogenetic trees are great tools that show evolutionary changes over time. and showing evolutionary relationships. They cannot, however, visualize the health and community impacts of individual variants of a virus over time. Moreover, phylogenetic trees cannot diagram convergent evolution, where two variants on different branches of the tree independently have the same mutation.

A stacked bar chart is a common statistical tool that many SARS-CoV-2 dashboards have employed to visualize their data. Unlike the standard bar chart it diagrams numeric values across two categorical variables, instead of one. The first categorical variable is found on the X-axis. The second categorical variable is found within the single bar. Each bar is divided into sub-bars stacked end to end, with each sub-bar corresponding to a level of the second categorical variable. Here the first categorical value is time on the X-axis, and the second categorical variable is the relative abundance of each of the SARS-CoV-2 in the sample at the given time point. This visualization method allows for a lot of complex data to be shown in an intuitive format and allows us to see trends over time. It does not, however, communicate in an intuitive way variable health outcomes that each variant confers.

Several visualization dashboards have been created for tracking SARS-CoV-2 genetic variants for public health purposes. Among these, CDC's COVID Data Tracker, specifically its variant summary dashboard, is perhaps one of the most prominent [3]. This dashboard provides a detailed overview of the prevalence and distribution of different SARS-CoV-2 variants across the United States. It includes visualizations of variant proportions over time and by region, allowing users to monitor the relative abundance and phylogenetic trees for multiple regions of the USA.

The Global Initiative on Sharing Avian Influenza Data (GISAID) offers a dashboard that tracks the global evolution of SARS-CoV-2 variants [1]. GISAID provides global and regional insights into variant distribution, mutation patterns, and phylogenetic trees. Users' interaction with GISAID consists of forms for specifying query parameters (such as locations of interest, and time intervals) and data visualization in output in the form of graphs.

Nextstrain is an online visualization dashboard for genomic data that provides different visualizations of phylogenetic information filtering by clades, geographic region, and date range [15]. The users interact with the system by entering different filtering criteria in a form and receiving the results in output in the form of graphs.

A survey and comparative analysis of the dashboards, found at COVIDPoops19 (as of September 2024) [4, 21], shows that most of the dashboards only depict viral-load concentration data and very few have genomic variants. The visualizations used by known dashboards include: tabular representation of the data, concentration scatter plot or density plot split with variants visually encoded by

color, stacked bar charts, and pie charts of relative abundances. Although these visualizations are effective for representing a single view of the data, the overall analysis workflow needs to include interactively and dynamic analysis for a user. Some of the visual encoding used makes it difficult to cross-reference the visualizations with other USA-wide dashboards like the CDC's data tracker [3]. Those dashboards with filtering capabilities only have standalone filters which restrict a user's analyzing ability to a single view.

Among public health dashboards, the Wisconsin Department of Health Services (WI DHS) has developed a state-specific dashboard to track SARS-CoV-2 variants [6]. This dashboard focuses on wastewater surveillance data, offering a unique perspective on viral trends at a community level and county level. It includes different modes of interacting with users, including toggling variants of interest on and off, filtering the data included in the visualizations by city, and time intervals.

From our review of these existing dashboards, we believe that there is room to further enhance these visualizations that would allow us to support additional analytics for multiple types of users such as public health officials, research scientists, and the general public. We present our tool, VISTA, which is a work-in-progress prototype being developed by the Discovery Partners Institute, University of Illinois System and Argonne National Laboratory to incorporate different enhancements to variant visualizations [10, 9].

3 Our Solution

In this section, we present details about our dashboard, VISTA is built to support public health tasks including correlating and comparing variant information among different regions and visualizing multiple features related to the variants. An overview of the dashboard is shown in Figure 1.

3.1 Users and tasks

The variant tracking dashboard is designed for a range of users in the public health domain, including health department stakeholders, scientists, and public health experts, who require real-time insights into the spread of SARS-CoV-2 variants and their prevalence in the community. These users rely on data-driven decision-making to facilitate surveillance programs, which enable public health departments to make more informed decisions and study viral spread patterns across different communities. By providing both detailed and aggregate views of the data, the dashboard supports the following tasks:

1. **Highlight variants of concern V_j**: The system identifies key variants to monitor based on their emerging prevalence, allowing users to track potential threats in real-time.
2. **Comparing a region R_i to the entire state S**: This task allows users to contrast the variant distribution in a specific region with the country-wide trend to detect local deviations.
3. **Comparing the abundance of specific types of variants V_j in a region R_i**: Users can explore which type of variants are more prevalent in a particular region, helping prioritize surveillance and resource allocation.
4. **Identify regions R_i and R_k that are similar in variant distribution V_j** : Users can detect similarities between regions based on their variant profiles, helping to understand geographical spread or common factors influencing variant dynamics.

This functionality provides a comprehensive platform for monitoring and responding to variant trends, aiding in public health decision-making and research efforts.

3.2 Data Ingestion

For this system, we utilize several data sources. First, we use the publicly available CDC dataset on SARS-CoV-2 variants in wastewater, which provides weekly aggregated data for 10 Health and Human Services (HHS) Regions from 2021 to present [8]. This dataset includes sequencing data aggregated from multiple sites from states in the HHS Regions and contains the weekly percentage share of variants in an HHS region and an estimate of the low and high ranges of the shares of the variants [11]. In addition, we use data on pre-defined key variants of interest, where we can collapse several low-prevalence lineages into their corresponding parent pangolin lineages [22, 7]. This allows us to focus on significant variants while preserving overall abundance patterns.

To normalize the variant proportions, we also created a synthetic dataset representing SARS-CoV-2 viral concentration in wastewater for each HHS Region. This type of data is usually available to public health departments that track SARS-CoV-2 in wastewater. This dataset was generated using publicly available regional wastewater concentration data, allowing for normalizing the variant abundances by region and week. This approach ensures a more accurate representation of variant spread relative to the overall viral load in each region.

To determine the effects of the mutations each variant possesses, we use a combination of two publicly available tools to determine if the mutations comprising each variant have some particular characteristics that may have an impact on public health, outbreak.info [5], COV-Spectrum [18] along with guidance from our colleagues at RIPHL [16]. For example, as of September 4, 2024, LB.1 is predicted to confer antibody escape due to the presence of the following mutations: S:R346 + S:F456. These mutations have been shown in the literature to confer antibody escape per several publications linked with outbreak.info, notably [23, 13].

When we identify a variant as having increased antibody escape, severity, or transmissibility, we compare it directly to other variants circulating simultaneously. For example, in August 2023, we labeled EG.5 as a variant with a known antibody escape mutations, F456L. However, by November 2023, we removed EG.5 from our antibody escape tracker because it no longer showed significantly more antibody escape compared to other circulating variants. This change occurred because nearly all variants had the F456L mutation at that time. We update our mutation tracker weekly, considering the proportion of samples with a particular mutation. If a mutation becomes common in the majority of variants in our population, we remove it from the tracker, as it then represents the baseline level.

3.3 System Overview

To support the tasks mentioned in Section 3.1, the dashboard is split into 4 sections described below.

The Variant Elevated Watch-List visualization (Figure 1.A is designed to track emerging SARS-CoV-2 variants across all regions under monitoring. A variant is flagged when its aggregated abundance across all regions exceeds $x\%$, where x is the threshold chosen by the user. This allows the user to focus on the most prevalent variants. It remains on the watch-list until its aggregated abundance drops below the threshold for four consecutive weeks, ensuring the visualization captures mid-term trends rather than reacting to short-term fluctuations. This approach helps identify potentially threatening variants that may require closer monitoring. The visualization employs a line-chart format, with specialized markers indicating the number of weeks a variant has remained above the threshold. This provides a visual cue to the confidence in the observed trend. The color scheme is aligned with that of the CDC's variant tracker, facilitating easy cross-reference and comparison with national data.

User interactions include a slider for setting the threshold of the elevated watch-list along with zoom in/out, legend-based toggles and hover for variants share and time information.

Regional Similarity Map (Figure 1.B). When making a public health decision: resource allocation, guidelines, or analyzing the overall spread of the variants, the expert would like to compare the different regions and cluster them based on the similarity of vari-

ant share-profile. To model regional similarity, we represent each region R_i at time t in a vector embedding space of variants, where the relative abundance of each variant is $a_1, \ldots, a_n$. For each time t, the viral-load normalized concentration *conc*, is represented in a matrix A_{ij}:

$$A_{ij} = \frac{a_{ij}}{conc_j}$$

Where:
- A_{ij} is the normalized abundance of variant i at site j.
- a_{ij} is the abundance of the variant i at site j for time t.
- $conc_j$ is the concentration at site j for time t.

This normalization helps account for the differences in the prevalence of SARS-CoV-2 between sites, allowing for the discovery of similar sites based on the distance between sites. We use the cosine distance to find the pairwise distance, $D_M(x_j, x_k)$, between the unit vector for the sites x_j and x_k:

$$D_M(x_j, x_k) = 1 - S_c(x_j, x_k)$$

Where:
- $S_c(x_j, x_k)$ is the cosine similarity between sites x_j and x_k

$$S_c(x_j, x_k) = \frac{\mathbf{x_j} \cdot \mathbf{x_k}}{\|\mathbf{x_j}\| \|\mathbf{x_k}\|}$$

Where:
  - $\mathbf{x_j} \cdot \mathbf{x_k}$ is the dot products of vectors x_j and x_k
  - $\|\mathbf{x_j}\| = \sqrt{\sum_{n=1}^{n} A_{jn}^2}$ is the magnitude of vector x_j
  - $\|\mathbf{x_k}\| = \sqrt{\sum_{n=1}^{n} A_{kn}^2}$ is the magnitude of vector x_k

The distance between sites that are more similar in terms of variant distribution is smaller. Then, we cluster similar sites based on a hierarchical clustering algorithm, which aids in identifying regions with comparable variant profiles. We visualize similar regions with similar colors.

User interactions include clicking on Section 1.C to select regions and zoom in/out on the map with hover information for each region.

Short and Long Term Regional Trends. The short and long term regional trends view (see Figure 1.C) provides a detailed visualization of the temporal patterns in variant prevalence between regions. It enables users to examine both immediate (short-term) and sustained (long-term) trends in variant dynamics. The view supports cross-filtering with the Regional Similarity view, allowing users to compare temporal trends across regions that exhibit similar variant distributions. This facilitates a deeper understanding of how variants evolve in different geographical areas. The color scheme is consistent with the Variant Elevated Watch-List and the CDC's tracker, ensuring seamless comparison between regional and overall spatial trends in variant prevalence.

User interactions include clicking on a particular bar to filter by variant and area selected 1.A, 1.B and 1.D with the data for the selected time t. Users can also brush and zoom in/out on a specific portion with hovering to visualize the share of variants.

Mutation-based Variants split (Figure 1.D) When viewing the genetic sequencing of SARS-COV-2 through a public health lens, arguably one of the most important aspects is determining if the variants circulating in the population have one of three significant adverse health outcomes. (1) Variants that have mutations that confer a greater antibody escape potential. (2) Variants that have mutations that confer increased transmissibility within the population. And (3) variants that confer an increased severity of illness in infected individuals.

We depict this distribution of different types of variants as a Sankey plot, which connects concentration-normalized share of the variants to variants types, and variants types to HHS Regions. This visualization provides an intuitive and simplified view of the complex relationship that exists among the three datasets of variants, regions, and features of interest.

User interactions include hovering to visualize the aggregate share of the variant/ variant-type.

3.4 User Feedback and Analysis Workflow

The dashboard was presented to the stakeholders and feedback was collected based on their usage. This data visualization offers users a unique way to break down complex data. The users can take a very large and interconnected dataset and break it into individual parts. This allows a user to narrow their focus to specific questions. Without this tool, users would have to go through a time-consuming process to be able to look at a specific snapshot of the data. This tool allows users to quickly test hypotheses or ideas they have regarding the data and its trends and to react to the outcome in real-time.

The interactivity of the Figure 1.C particularly allows the user to filter out and focus only on individual variants. Having the ability to see the full history of specific variants, rather than viewing them as part of the whole, offers valuable insight into the evolution and trajectory of SAR-COV-2.

An example analytics workflow for public health stakeholders, users begin on the landing page, where they can immediately identify watch-list variants showing regional predominance. In the regional similarity plot, color-coded regions with similar variant dynamics quickly highlight areas requiring comparable intervention, guiding resource allocation decisions. A stacked bar chart further reveals regional variant trends over time, with comparisons to national data, contextualizing regional outbreaks within the broader public health landscape. The mutation-based variant split feature identifies regions with a high prevalence of antibody escape mutations, offering insights into potential vaccine-evasion hotspots and guiding targeted resource shifts.

4 Future Work

We believe VISTA can be further enhanced with additional visual analytics.

Correlation detection. This analysis detects if there is a correlation in variants spread among different regions. This feature would help provide a better understanding of whether *sentinel* sites can be used to monitor the variants spread for a region rather than multiple sites in that region. This would allow public health departments to toggle monitoring on and off for multiple sites according to resource needs and results present at the sentinel.

Multiple algorithms for clustering. We plan to test additional algorithms for clustering and similarity. For instance, the Mahalanobis distance or Jaccard similarity can be used instead of the cosine similarity to take into account additional features of the data [14, 12].

Integration with additional datasets. Another enhancement includes creating visual analytics that integrate additional datasets such as population vulnerability indexes, correlation with clinical data, etc.

Acknowledgments

This publication was supported by the Centers for Disease Control and Prevention of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award totaling $188,346,657 with 100 percent funded by CDC/HHS. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by CDC/HHS, or the U.S. Government.

Acknowledgments: Chicago Department of Public Health and Regional Innovative Public Health Laboratory.

REFERENCES

[1] Gisaid. https://gisaid.org, 2008. Accessed: 2024-09-05. 2

[2] Coronavirus resource center. https://coronavirus.jhu.edu/map.html, 2020. Accessed: 2024-09-05. 2

[3] Coronavirus resource center. https://covid.cdc.gov/covid-data-tracker/datatracker-home, 2020. Accessed: 2024-09-05. 2, 3

[4] Covidpoops19 summary of global sars-cov-2 wastewater monitoring efforts by uc merced researchers. https://www.arcgis.com/apps/dashboards/c778145ea5bb4daeb58d31afee389082, 2020. Accessed: 2024-10-24. 2

[5] Outbreak.info: A project from the hughes , su , wu , and andersen labs at scripps research to unify covid-19 and sars-cov-2 epidemiology and genomic data, published research, and other resources. https://outbreak.info/, 2020. Accessed: 2024-09-09. 3

[6] Sars-cov-2 wastewater genomic dashboard. https://dataportal.slh.wisc.edu/sc2-ww-dashboard, 2020. Accessed: 2024-09-05. 3

[7] Phylogenetic assignment of named global outbreak lineages. https://cov-lineages.org/index.html, 2021. Accessed: 2024-09-10. 3

[8] Sars-cov-2 variant proportions. https://data.cdc.gov/Laboratory-Surveillance/SARS-CoV-2-Variant-Proportions/jr58-6ysp/about_data, 2021. Accessed: 2024-09-09. 3

[9] Argonnne national laboratory. https://www.anl.gov, 2024. Accessed: 2024-09-05. 3

[10] Discovery partners institute. https://dpi.uillinois.edu, 2024. Accessed: 2024-09-05. 3

[11] Hhs regional offices and regional maps, 2024. Accessed: 2024-09-10. 3

[12] Jaccard similarity. https://www.learndatasci.com/glossary/jaccard-similarity/, 2024. Accessed: 2024-09-10. 4

[13] Lb.1 page. https://outbreak.info/situation-reports?xmin=2024-03-05xmax=2024-09-05pango=LB.1, 2024. Accessed: 2024-09-10. 3

[14] Mahalanobis distance. https://www.statisticshowto.com/mahalanobis-distance/, 2024. Accessed: 2024-09-10. 4

[15] Nextstrain: Real-time tracking of pathogen evolution. https://nextstrain.org, 2024. Accessed: 2024-09-05. 2

[16] 'regional innovative public health laboratory'. https://www.rushu.rush.edu/research-rush-university/research-collaborations/regional-innovative-public-health-laboratory, 2024. Accessed: 2024-11-02. 3

[17] ANDERSSON, J. O. The New Foundations of Evolution: On the Tree of Life. *Systematic Biology 60*, 1 (10 2010), 114–115. 2

[18] CHEN, C., NADEAU, S., YARED, M., VOINOV, P., XIE, N., ROEMER, C., AND STADLER, T. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics 38*, 6 (12 2021), 1735–1737. 3

[19] EDWARDS, R. J. Phylogenetic tree rooting. In *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds. Academic Press, Oxford, 2019, pp. 727–735. 2

[20] HEDGES, S. B., MARIN, J., SULESKI, M., PAYMER, M., AND KUMAR, S. Tree of Life Reveals Clock-Like Speciation and Diversification. *Molecular Biology and Evolution 32*, 4 (03 2015), 835–845. 2

[21] NAUGHTON, C. C., ROMAN, FERNANDO A, J., ALVARADO, A. G. F., TARIQI, A. Q., DEEMING, M. A., KADONSKY, K. F., BIBBY, K., BIVINS, A., MEDEMA, G., AHMED, W., KATSIVELIS, P., ALLAN, V., SINCLAIR, R., AND ROSE, J. B. Show us the data: global COVID-19 wastewater monitoring efforts, equity, and gaps. *FEMS Microbes 4* (01 2023), xtad003. 2

[22] O'TOOLE, , SCHER, E., UNDERWOOD, A., JACKSON, B., HILL, V., MCCRONE, J. T., COLQUHOUN, R., RUIS, C., ABU-DAHAB, K., TAYLOR, B., YEATS, C., DU PLESSIS, L., MALONEY, D., MEDD, N., ATTWOOD, S. W., AANENSEN, D. M., HOLMES, E. C., PYBUS, O. G., AND RAMBAUT, A. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution 7*, 2 (07 2021), veab064. 3

[23] PLANAS, D., STAROPOLI, I., PLANCHAIS, C., YAB, E., JEYARAJAH, B., RAHOU, Y., PROT, M., GUIVEL-BENHASSINE, F., LEMOINE, F., ENOUF, V., SIMON-LORIERE, E., MOUQUET, H., RAMEIX-WELTI, M.-A., AND SCHWARTZ, O. Escape of sars-cov-2 variants kp1.1, lb.1 and kp3.3 from approved monoclonal antibodies. *bioRxiv* (2024). 3