Computational and Structural Biotechnology Journal 27 (2025) 1081–1088

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

VirusWarn: A mutation-based early warning system to prioritize concerning SARS-CoV-2 and influenza virus variants from sequencing data

Christina Kirschbaum [a,*], Kunaphas Kongkitimanon [a,b], Stefan Frank [a], Martin Hölzer [a], Sofia Paraskevopoulou [a], Hugues Richard [a]

[a] Genome Competence Center (MF1), Robert Koch Institute, Nordufer 20, Berlin, 13353, Berlin, Germany
[b] Data Analytics & Computational Statistics, Hasso Plattner Institute, University of Potsdam, Prof.-Dr.-Helmert-Straße 2 - 3, Potsdam, 14482, Brandenburg, Germany

ARTICLE INFO

Keywords:
Warning system
SARS-CoV-2
Influenza virus
Genomic surveillance
Variant prioritization

ABSTRACT

The rapid evolution of respiratory viruses is characterized by the emergence of variants with concerning phenotypes that are efficient in antibody escape or show high transmissibility. This necessitates timely identification of such variants by surveillance networks to assist public health interventions. Here, we introduce *VirusWarn*, a comprehensive system designed for detecting, prioritizing, and warning of emerging virus variants from large genomic datasets. VirusWarn uses both manually-curated rules and machine-learning (ML) classifiers to generate and rank pathogen sequences based on mutations of concern and regions of interest. Validation results for SARS-CoV-2 showed that VirusWarn successfully identifies variants of concern in both assessments, with manual- and ML-derived criteria from positive selection analyses. Although initially developed for SARS-CoV-2, VirusWarn was adapted to Influenza viruses and their dynamics, and provides a robust performance, integrating a scheme that accounts for fixed mutations from past seasons. HTML reports provide detailed results with searchable tables and visualizations, including mutation plots and heatmaps. Because VirusWarn is written in Nextflow, it can be easily adapted to other pathogens, demonstrating its flexibility and scalability for genomic surveillance efforts.

1. Introduction

The importance of genomic surveillance has proven to be instrumental in aiding global efforts to monitor respiratory viruses of public health interest [1]. For example, the COVID-19 pandemic, caused by SARS-CoV-2 that emerged in late 2019, imposed a significant burden on healthcare systems worldwide. The rapid transmission of SARS-CoV-2 along with the initial lack of vaccines and targeted antiviral treatment [2,3] highlighted the need for robust genomic surveillance to track virus evolution and detect emerging variants [4,5]. Similarly, Influenza viruses present a continuous challenge due to their seasonal epidemics and zoonotic potential [6–8]. The ongoing evolution of these viruses highlights the crucial role of efficient and accurate genomic surveillance in pandemic preparedness and response [9]. Global efforts have therefore expanded the use of whole virus genome sequencing, which in turn underscores the need for automated tools to evaluate potentially concerning samples based on sequence data for the timely detection of emerging virus variants [1].

Existing variant warning systems can be classified into “exploratory” and “predictive”. Exploratory tools offer information on the genomic and epidemiological properties of a variant or a group of variants, visualizing mutation patterns and their prevalence in the population. Several such tools were developed during the COVID-19 pandemic and include among others, e.g., outbreak.info [10], nextstrain.org [11], ERVISS [12], cov-spectrum.org [13], CovRadar [14], and sarscoverage.org [15]. These resources have been instrumental in assisting evidence-based public health decisions. However, while these tools provide useful epidemiological insights, they require active user engagement to detect trends in the data and provide confident indications only after a variant is established in a fraction of the population. The HiRiskPred tool [16] extends exploratory analysis by integrating SARS-CoV-2 lineage information with the genetic network of samples in order to predict high risk variants through graph analysis.

Predictive tools inform on the potential risk of virus variants, either through virulence predictions of a given variant or more comprehensive evaluations of their epidemiological impact. One such tool is the

* Corresponding author.
E-mail address: KirschbaumC@rki.de (C. Kirschbaum).

https://doi.org/10.1016/j.csbj.2025.03.010
Received 15 November 2024; Received in revised form 5 March 2025; Accepted 8 March 2025

Available online 12 March 2025

Influenza Risk Assessment Tool (IRAT) [17], which evaluates Influenza virus types and assesses their potential risk for zoonotic emergence and their general impact based on ten risk properties. As it relies on expert knowledge and serves as a reporting framework for a given Influenza variant, it does not support on-demand analyses or the assessment of Influenza virus sequencing data overall. Another predictive tool is VASIL [18], which analyzes SARS-CoV-2 sequences and predicts their potential impact based on the population background immunity. Other tools (VirPreNet [19] and ViPal [20]) use rule-based learning approaches [21] such as weighted convolutional neural networks and combinations of feature vectors with prior knowledge to predict virus-host interactions. VarEPS [22] and the recently-developed VarEPS-Influ [23] are designed to predict viral evolution and assess the potential risks of emerging variants of SARS-CoV-2 and Influenza variants. However, they are only available via a web interface which limits streamlined data processing capabilities that a command-line interface would offer.

Considering all the above tools' strengths and limitations, we understand that there is a need for a command-line tool that can i) analyze a large collection of sequences, ii) report an alert value for each sequence and summarize the results, iii) adapt easily to function for different viruses, and importantly iv) account for any privacy and data protection concerns.

Here, we present *VirusWarn*, an efficient and streamlined Nextflow [24] pipeline that rapidly processes thousands of virus sequences, raises alerts based on genomic properties, and provides user-friendly reports. VirusWarn incorporates a sequence-based ranking system that considers amino acid changes and their overlap with antigenic sites, epitope regions, and sites under positive selection. Samples are ranked into different alert levels according to the predicted impact of the virus variant. We demonstrate its performance and we show in a proof-of-concept experiment how to set up a machine learning-based approach that identifies sites under positive selection. Currently, VirusWarn supports sequence data analyses for variants of SARS-CoV-2, Influenza A (subtypes H1N1 and H3N2), and Influenza B (Victoria strain).

VirusWarn can be easily run in a Docker [25] / Singularity [26] container on any machine and we provide simple tutorials which can be found on the code repository. And finally, as VirusWarn is written in Nextflow, it ensures easy adaptations, reproducibility, and streamlined processing of the genomic data. The code repository and user documentation are open source and accessible at the following link: https://rki-mf1.github.io/viruswarn-doc/.

2. Materials and methods

We present here VirusWarn, a command-line tool that processes SARS-CoV-2 and Influenza genomic data to raise subsequent alerts. Genome sequences of SARS-CoV-2 are ranked based on their mutation profile of the spike gene (S). Similarly, we provide a ranking scheme for the Influenza HA segment for three subtypes: A H1N1, A H3N2, and B Victoria. For each query sequence, we infer its mutation profile after comparison to a reference sequence and such inferred profiles are aggregated into a multidimensional profile. The alert raised afterwards is based on the individual sequence profile, following specific classification criteria. The results are summarized in an HTML report for further investigation. Below, we describe in detail each component and scheme of VirusWarn (**Fig. 1a**) and our validation and evaluation approaches.

2.1. Preprocessing of query genome sequences

Whole-genome sequences or sequences of the spike gene for SARS-CoV-2 and the HA segment for Influenza can be given as input files. We identify and annotate amino acid variations relative to reference sequences, focusing only on the spike protein for SARS-CoV-2 and the HA segment (excluding the signal peptide) for Influenza (**Fig. 2a and b**). The reference sequences are listed in **Table S1**. VirusWarn also accepts pre-computed mutation profiles that can be provided as a summary table or derived from *covSonar* [27].

2.2. Annotation scheme of the mutation profile

Each query sequence is described by a nine-dimensional mutation profile as shown in **Fig. 2c**. This profile summarizes the number of substitutions (Subst.), deletions (Del.), or insertions (Ins.) assigned to each of the following three categories:

- Mutations of Concern (MOCs): substitutions, insertions, and deletions that can alter the potential of the virus regarding its virulence and transmissibility, and referenced as experimental results in the literature.
- Regions of Interest (ROIs): important regions of the gene, such as antigenic sites, receptor-binding sites, and glycosylation sites.
- Private Mutations (PMs): substitutions, insertions, and deletions that are neither identified as MOCs nor as ROIs.

The SARS-CoV-2 ROIs are summarized in **Table S2** and the MOCs in **Table S3**. For lineage classification, we use the pangolin lineage assignments [31] and the categories of Variant of Concern (VOC), Variant of Interest (VOI), and Variant under Monitoring (VUM) by the WHO Technical Advisory Group on SARS-CoV-2 Virus Evolution (TAG-VE) [32].

For Influenza variants, the MOC and ROI tables can be found in **Tables S4 to S9**. Clade assignments are given by Nextclade [28], along with the accompanying mutation table.

Because Influenza is a seasonal virus, important mutations, such as MOCs or mutations found in ROIs, might become fixed in the population and thus irrelevant for monitoring. In order to avoid alerts for such mutations, we record those mutations with at least 75% prevalence in each Influenza season (July of one year to June of the following year) and score them separately.

2.3. Ranking and scoring schemes: VirusWarn-Manual and -Auto

The *VirusWarn-Manual* scoring scheme is based on a system of sequential rules to classify concerning variants into three levels: high impact, medium impact, or low impact (see **Fig. 2d**). The rules are manually designed using the following working hypothesis: if a variant accumulates many mutations, especially in crucial genomic regions, such as the spike or the HA, then it is more likely to be concerning. We also consider the sequence's seasonality, i.e., whether it belongs to a variant circulating in a previous season. The rules are detailed in **Table S10** and **Table S11**, respectively. To ease visual interpretations, we output color-coded alerts:

- Red: Variants with many MOCs and ROIs that do not belong to an already-known concerning variant group are considered as high-impact variants.
- Pink: Variants with mutations in the MOCs and ROIs categories, likely to have concerning effects, e.g., emerging variants from an already known VOC, VOI, VUM for SARS-CoV-2, and variants from the previous season for Influenza.
- Orange: Variants with a few MOCs and ROIs/PMs, considered as medium impact variants that could have some concerning properties.
- Yellow: Variants with no MOCs, but with a high number of ROIs or PMs, considered as low-impact variants.
- Grey: Remaining variants that do not trigger any alert.

For each rule, we set cut-off values on the number of MOCs, ROIs, and PMs to ensure that known high impact variants are singled out from the red category of alerts. We used a predetermined set of sequences to identify the thresholds that distinguished the rules according to the

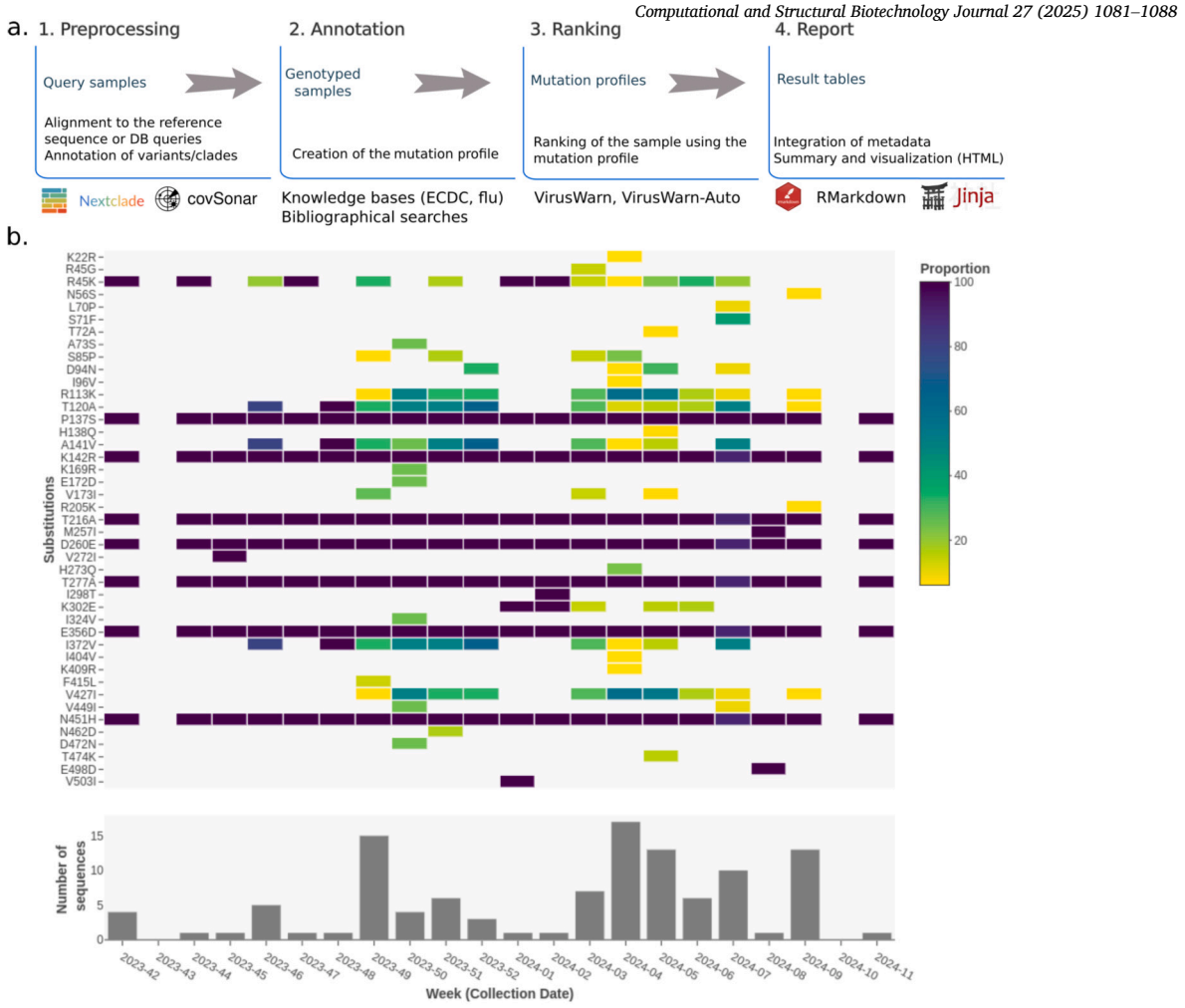

Fig. 1. (**a**) VirusWarn steps: assembled genome sequences and metadata (sampling dates, geographical locations) are used as query samples, which are pre-processed to infer mutation profiles for the monitored proteins. This is achieved either by aligning the sequences to a reference genome (e.g., in covSonar [27]) or via direct database queries (e.g., in Nextclade [28]). The mutation profiles are then ranked according to their concern level and results for query samples are summarized in interactive HTML reports using Jinja2 [29] or RMarkdown [30]. (**b**) Frequency heatmaps show mutation frequencies over time (calendar weeks on the x axis). A histogram showing the number of sequences per calendar week is shown at the bottom.

color-coded alerts ($n = 7{,}475$ SARS-CoV-2 sequences from France collected in the first three months of 2021 and $n = 2{,}615$ Influenza H1N1 HA sequences from Europe collected from April to December 2009). A strict mode, which ignores medium impact variants, can also be activated.

Collecting the MOC annotations requires extensive literature search and expert knowledge, which can be time-consuming. We thus prototyped *VirusWarn-Auto* that ranks concerning variants in an automated manner for SARS-CoV-2. Two aspects were automated: the construction of ranking rules using machine learning algorithms and knowledge extraction based on positive selection. We hypothesized that evolutionary episodes of SARS-CoV-2, i.e., sites that are under positive selection, could replace a manually-curated list of MOCs. We identified such sites within the SARS-CoV-2 spike gene based on the ratio of non-synonymous to synonymous substitution rates (dN/dS), as described previously [33]. We utilized a pipeline originally developed by Pond et al. [34], that comprises the necessary steps of data cleaning, gene selection, alignment, and tree inference to perform a per-site positive selection detection. Positive or negative selection was detected using FEL [34], while MEME [35] was used to identify sites of episodic selection. We then filtered for sites detected as "significantly positively selected" based on either p-value$_{FEL} < 0.05$ or p-value$_{MEME} < 0.05$.

The sites under positive selection were computed using the sequences sampled between 2021-01-01 and 2021-03-31 via the German Electronic Sequence Hub [36]. The compiled list with n = 1273 positively selected sites is available on the GitHub repository.[1]

To automate the variant ranking rules, the scoring scheme was extended to integrate a supervised learning strategy. We considered a binary classification problem, separating "alert" from "no alert", and prepared training and testing sets assuming that a sequence would raise an alert when assigned in the VOC/VOI/VUM categories. Furthermore, we considered an additional classification task to account for the specific case where a VOC emerges as a descendant of another VOC. Considering the low dimensionality of the mutation profiles, we restricted the evaluation to three supervised learning methods: Random Forest, Support Vector Machine (SVM), and Logistic Regression. Those classifiers have the double advantage of being relatively straightforward to train and easier to interpret by nonexperts. All machine learning models were trained using the Python scikit-learn library [37], using default settings. We also performed a grid search for SVM and Logistic Regression to optimize the model parameters. We used a five-fold cross-validation and the macro-averaged score as performance metrics to compare the different parameters. We used the macro-averaged score for performance

[1] https://github.com/rki-mf1/viruswarn-sc2/blob/master/data/sergei_extracted.2021-05-17.csv.

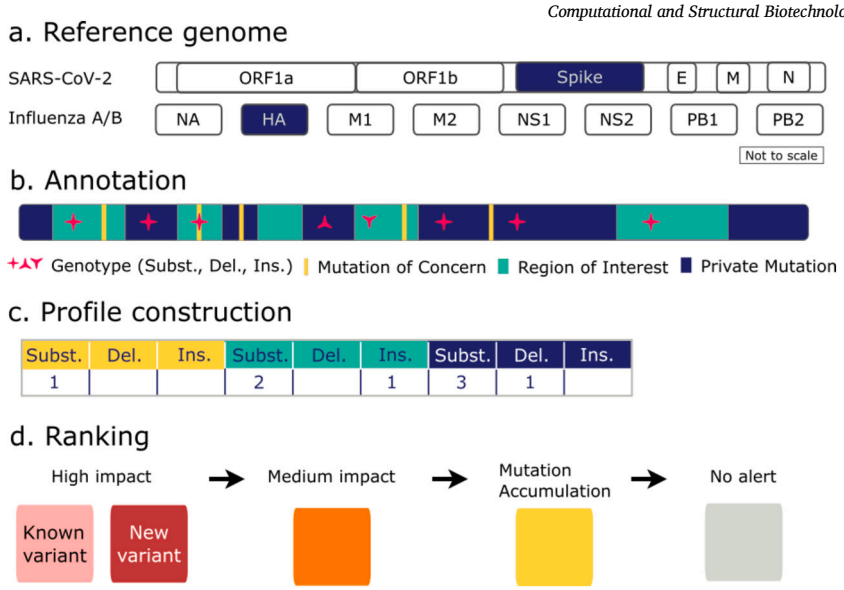

Fig. 2. Steps during annotation and ranking by VirusWarn. (**a**) Selected proteins to monitor (blue) for SARS-CoV-2 and Influenza. (**b**) Example annotations and genomic variation for a monitored protein. Annotations used for scoring: Mutation of Concern (MOC, yellow), Region of Interest (ROI, green), and Private Mutation (PM, blue). Genomic variation types (pink): Substitution, Deletion, or Insertion. (**c**) Example mutation profile: counts of variation types in the different annotation groups (MOC, ROI, PM). (**d**) Ranking to alert color levels: high impact (pink/red), medium impact (orange), low impact (yellow), and no alert (grey).

evaluation as it considers the imbalance between positive and negative classes.

The machine learning models included a set of fourteen features comprising the nine-dimensional mutational profile, as well as statistics integrating information about antibody escape scores from deep mutational scanning data [38]. The calculated antibody escape score shows the mutations' potency to escape antibody recognition, and we computed an average mutation escape score for each site. We also evaluated the classifiers by combining them with a knowledge extraction step where the MOCs are substituted by the sites under positive selection, reducing the number of features from fourteen to eleven.

2.4. Datasets used for validation

2.4.1. Validation on SARS-CoV-2

We validated VirusWarn-Manual using two blind test case-scenarios on randomly sampled SARS-CoV-2 data from Germany in 2021 [36]. For case 1, we used Delta sequences from calendar weeks 13 to 15; for case 2, we used Omicron sequences from calendar weeks 48 to 51 (the time point of Omicron's emergence). To ensure data consistency, we excluded genomes with sampling and submission dates differing by more than two months, resulting in $n = 30$ Delta and $n = 3,446$ Omicron samples. Lineage-defining mutations (specific to Delta, Omicron, and WHO-classified variants) were removed from the annotation files to ensure that the validation is variant-agnostic. Additionally, we validated VirusWarn-Manual's ranking for a given timeframe in the context of a country-wide surveillance network (IMSSC2 Laboratory Network [39]). Aiming to recover alerts for Omicron and recombinant variants, we analyzed all $n = 766$ sequences submitted to the IMSSC2 Laboratory Network between December 1st, 2022 and February 28th, 2023.

To evaluate VirusWarn-Auto performance, we prepared a dataset for a machine learning model using genomic surveillance data from Germany for the first three quarters of 2021 [36]. The data was imported in *covSonar* [27] to extract spike mutation profiles. For model training and validation, we used data from the first quarter of 2021 (January 1st to March 31st, $n = 53,752$ genomes). For testing, we randomly sampled sequences from the second and the third quarter of 2021 (April 1st to September 30th, $n = 5,902$). The problem of detecting concerning variants was simplified into a classification task with two classes, class A: concerning variant, and class B: non-concerning variant. Concerning variants (class A) were defined as samples assigned to either of the VOC/VOI/VUM categories and the remaining samples were assigned to class B. Identical- or low genetic diversity samples were removed to ensure sequence diversity in the training set and avoid redundancy, and a balanced number of class A and class B samples was used for training, validation, and testing.

2.4.2. Validation on Influenza viruses

The datasets used to evaluate VirusWarn on Influenza viruses were obtained from GISAID EpiFlu™ [40] for Influenza A (H1N1 and H3N2) and Influenza B (Victoria), respectively. Because Influenza seasons differ depending on the hemisphere, we restricted our dataset to regions primarily located in the northern hemisphere: Europe, North America, and Asia. For testing and validation of Influenza A H1N1, we downloaded a dataset of $n = 38,514$ sequences from Europe, $n = 49,785$ sequences from North America, and $n = 25,529$ sequences from Asia. To identify mutations that became fixed in the population, we used sequences collected between April 2009 and June 2015, as well as from July 2019 to June 2024, from this combined dataset. We calculated the frequency of every mutation per season to determine whether it had become fixed in the population. For Influenza A H3N2, we analyzed $n = 61,535$ sequences from the northern hemisphere from July 2021 to June 2024. We used the same time frame for Influenza B Victoria and analyzed $n = 19,792$ sequences. Accession numbers for the GISAID EpiFlu™ dataset can be retrieved by applying the filters described in **Table S12**.

2.5. Reporting and visualization of the results

VirusWarn generates CSV tables and HTML reports summarizing the results for further expert analysis. The SARS-CoV-2 reports feature additionally a table highlighting top sequence clusters for the red and pink alert levels. For Influenza, reports include visualizations such as lollipop plots that show the number of mutations at each position of the HA segment and heatmaps showing the mutation frequencies over time (**Fig. 1b**). Example reports and screenshots for SARS-CoV-2[2] and Influenza[3] are available in our code repositories.

[2] https://github.com/rki-mf1/viruswarn-sc2/tree/master/example.
[3] https://github.com/rki-mf1/VirusWarn-Flu/tree/main/example.

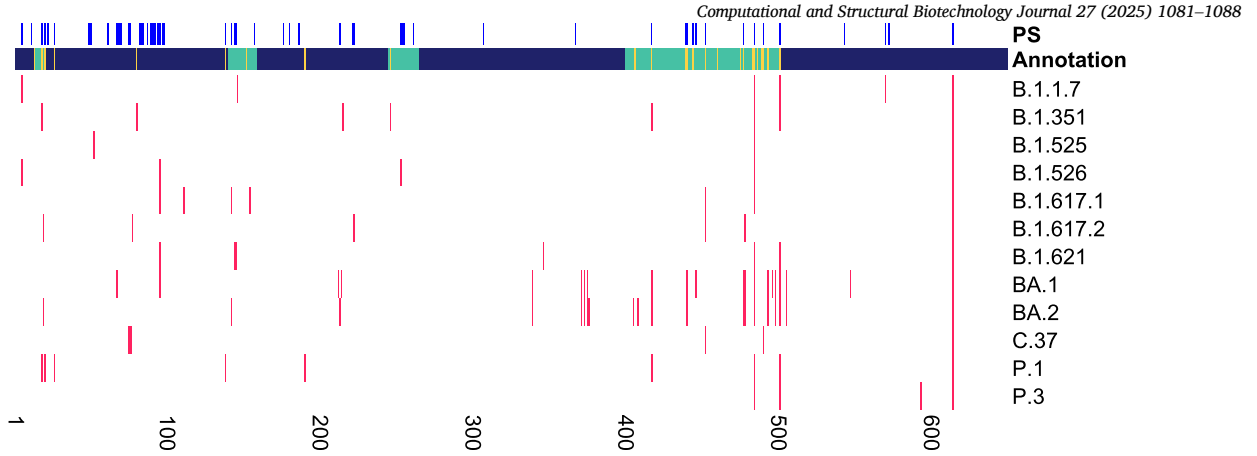

Fig. 3. SARS-CoV-2 partial spike protein (the first 650 amino acids, incl. receptor-binding sites) The MOC, ROI, and PM color annotations appear as in Fig. 2), and are contrasted to sites under positive selection (PS, blue bars on top) and lineage-defining mutations for selected VOCs from the ECDC table [42] (red bars).

3. Results

3.1. VirusWarn-Manual accurately warns about concerning SARS-CoV-2 variants

To evaluate VirusWarn-Manual, we used two test case-scenarios aiming to detect the emergence of two major SARS-CoV-2 VOCs in 2021 in data from Germany: Delta (B.1.617.2, case 1) and Omicron (B.1.1.529, case 2). In both cases, VirusWarn-Manual correctly raised red alerts for the majority of sequences (Delta: 30/30 (100%), Omicron: 3,372/3,446 (97%)). In case 1, we also detected a set of $n = 21$ sequences primarily assigned to the Delta sister lineages B.1.617.1 (Kappa) and B.1.617.3, which were shortly considered as potentially concerning before the establishment of Delta. This shows that in early 2021, VirusWarn-Manual could have been used by experts to alert on the first occurrences of Delta variants before they were officially declared as VOCs. Additionally, the proportion of false positives was relatively low in our evaluation, as virtually all VirusWarn-Manual alerts were relevant, even those for Kappa and B.1.617.3.

3.2. SARS-CoV-2 lineage-level evaluation in Germany's IMSSC2 laboratory network

After the pandemic was deescalated in 2023, SARS-CoV-2 genomic surveillance in Germany was scaled down and since then relies on the medium-scale IMSSC2 Laboratory Network. Despite scaling down the sample and thus data collection, previous studies have demonstrated that the IMSSC2 Laboratory Network can effectively support genomic surveillance efforts at a national level [39]. Here, we investigated whether this data could trigger VirusWarn alerts on potentially concerning variants soon after these emerge. Therefore, we reset the mutations, lineages, and concerning variants to reflect Germany's SARS-CoV-2 genomic variant background as of November 30th, 2022, and applied the VirusWarn-Manual model to rank all SARS-CoV-2 sequences from December 2022 to February 2023. VirusWarn accurately detected concerning variants, identifying 205 red, seven pink, and 554 grey sequences in a total of 766 sequences analyzed from the IMSSC2 Laboratory Network (Supplementary **Table S13**).

In detail, the red level primarily consisted of lineages descending from BA.2 (e.g., BQ lineages), BA.5 (e.g., BN and CH lineages), and XBB variants. Except for one XBB sequence classified as grey, all other $n = 87$ sequences that belong to XBB-descendant lineages (e.g., XBB.1.5 and XBB.1.9) were classified as red. It is worth noting that XBB was classified as VOI by the WHO TAG-VE on December 8th, 2022 while our analysis reflects the status from November 30th, 2022, highlighting that VirusWarn raised a red alert for concerning lineages before they were assigned as such.

Pink alerts were raised for lineages BA.2, BQ.1.1, BQ.1.5, BA.5.11, and BA.5.2.32. However, not all tested genomes from these lineages were classified as pink; some were classified as grey because they did not carry MOCs. For instance, the pink BA.2 sequence carried S:F468P, distinguishing it from the grey-classified sequences. It is also worth noting that until early 2023, descendant lineages of VOCs/VOIs/VUMs inherited the VOC/VOI/VUM status from their parental lineages irrespectively from their epidemic impact, and this was only later revised by the WHO in March 15th, 2023 [41].

3.3. Positive selection sites can accurately rank concerning SARS-CoV-2 variants

Replacing MOCs with inferred sites under positive selection could further automate VirusWarn. Therefore, we explored this potential using the same underlying principle on ranking and scoring, and focused our positive selection analysis only on substitutions, ignoring insertions and deletions. We observed that the inferred sites under positive selection capture most of the MOCs and approximately half of the lineage-defining mutations on the spike protein (see **Fig. 3**). This is a much higher proportion than the average share of lineage-defining mutations (~5%) which highlights that sites under positive selection provide a good proxy for MOCs. Our analysis was based on data from Germany to simulate the situation experienced by public health institutes during an epidemic.

3.4. VirusWarn-Auto evaluation on concerning SARS-CoV-2 variants

We evaluated VirusWarn-Auto using Germany's genomic surveillance data from the first quarter of 2021 for training and testing (January to March 2021). Positive samples were defined as those with a VOC/VOI/VUM status. The performance of Logistic Regression, Random Forest, and Support Vector Machine was validated using data from the second and third quarters of 2021: April to June and July to September (total $n = 5,902$ samples).

For all tests in default mode, each classifier performed satisfactorily well, with almost all concerning sequences being recovered. This came at the cost of precision, with around half of the positive cases marked as false positives (Table 1). We then examined the impact of using manually curated MOCs on the classifiers' performance and observed that the use of MOCs didn't impact the results with the Logistic Regression classifier (Table 1). Finally, we explored the possibility of automatically obtaining and using sites under positive selection as a replacement for MOCs. Interestingly, replacing the MOCs with the sites under positive selection didn't affect the model's performance in our tests (Table 1).

Table 1
Results of sequence classification from the VirusWarn-Auto SARS-CoV-2 test set (5,902 samples uniformly drawn from April 1st to September 30th 2021). The Precision (Prec.), Recall (Rec.) and the F1 Score (F1) are listed for Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF).

	with MOC			only Bloom scores			with PS		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
LR	0.5	0.95	0.66	0.51	0.94	0.66	0.51	0.98	0.67
SVM	0.5	0.92	0.65	0.42	0.05	0.08	0.51	0.98	0.67
RF	0.51	0.5	0.44	0.44	0.03	0.05	0.51	0.98	0.67

3.5. Masking fixed seasonal mutations refines alerts for Influenza viruses

Variants which are already known to be concerning because they carry crucial mutations (MOCs), may evolve to carry additional MOCs. However, MOCs with high prevalence in the virus population during one season may become fixed in the next season's circulating viruses. We have therefore developed a scheme that masks MOCs with $> 75\%$ prevalence, assuming that those become fixed in the virus population of the following season. We validated this approach by retrospectively assessing VirusWarn alerts on Influenza H1N1 variants circulating in Europe from July 2010 to June 2011 and from July 2011 to June 2012. Before the application of the $> 75\%$ prevalence scheme, we find four red, 543 orange, and 110 grey sequences in season 2010-2011, and four red, 91 orange, and 31 grey sequences in season 2011-2012. This is already an overwhelming number of alerts that dilutes the information on truly concerning variants.

With the 75% prevalence scheme in place, orange alerts disappear and the number of red alerts drops to one sequence for season 2010-2011. Pink alerts now ($n = 546$) hold the rest of previously-red and the orange sequences, demonstrating the importance and role of the pink level. The same accounts for season 2011-2012: only two are still classified as red, 93 appear now as pink instead of orange, and 31 as grey. Example MOCs that are masked with this scheme are: HA:S203T, which in season 2009-2010 occurred in over 80% of all sequences in the northern hemisphere with its prevalence rising to 95% in the next season, and HA:E374K with 77% prevalence in season 2010-2011. These results provide overall a more concise and biologically meaningful report.

Therefore, we generated tables of fixed H1N1 mutations based on the mutations with $> 75\%$ prevalence for the time periods 2009-2015[4] and 2019-2024.[5] Fixed mutations were calculated for the time period 2021-2024 for H3N2[6] and Influenza B (Victoria).[7]

4. Discussion

We introduce here *VirusWarn*, a robust warning system for rapid assessment of concerning SARS-CoV-2 and Influenza variants in large genomic datasets. Our validation results for SARS-CoV-2 demonstrate the accuracy of the corresponding raised alerts which successfully identified the Delta/Omicron VOCs as well as lineage-level concerning variants. The performance of VirusWarn is consistent even when expert knowledge of key mutations (MOCs) is substituted with positively selected sites on the spike gene that are automatically calculated. Also, our defined MOCs and ROIs for Influenza viruses lie within genomic regions of antigenic drift, which are already known to be under positive selection [43,44]. Thus, irrespective of whether the -Auto or the -Manual modes are used, VirusWarn can be considered a practical tool for generating alert reports, effectively complementing genomic surveillance efforts.

VirusWarn has proven its performance, using a manually-curated set of rules, but also with an automated ranking scheme via a machine learning classifier. Our SARS-CoV-2 analysis suggests that increasing the model's complexity by using machine-learning methods does not necessarily improve the detection of concerning variants. In addition, incorporating information from deep mutational scanning experiments did not improve the classifier's performance, most likely because our evaluation datasets stem from 2021 when the majority of MOCs had already emerged in different variants of SARS-CoV-2. This aligns with recent trends in SARS-CoV-2 evolution, where many variants converge on a specific set of mutations in the spike protein [45]. The high accuracy of VirusWarn on SARS-CoV-2, combined with the simple but powerful scoring system, highlights VirusWarn's extension possibilities to metagenomic datasets such as wastewater or bio-terror samples that often have shallow sequencing depth of the target pathogen.

VirusWarn's accurate warning on the emergence of the VOCs Delta and Omicron opens the discussion on how thresholds for detecting concerning variants should be adjusted over time. As virus evolution progresses, it may be necessary to implement more stringent classification thresholds or even consider using a different reference genome to infer MOCs. Similarly to how vaccines are periodically updated according to the circulating variants, it is essential to update VirusWarn's knowledge base, too. Our preliminary experiments with VirusWarn-Auto show that it is possible to do so efficiently using positive selection results; a direction for future research.

Because we initially programmed VirusWarn for SARS-CoV-2 and later adapted it to Influenza, we identified two likely challenges that may be encountered when adapting VirusWarn to other viruses. First, while SARS-CoV-2 has a universal reference sequence (Wuhan-Hu-1), the reference sequences for Influenza viruses are frequently changing due to the seasonality and reassortment of the virus. Second, the amount of genomic data for Influenza is considerably lower compared to SARS-CoV-2, which adds a burden in identifying suitable datasets for validation. The time period of the 2009 Influenza A pandemic was the best validation dataset we could use, as it was the most recent Influenza pandemic [46,47]. However, data quality, size, and curation are not comparable to those available for the COVID-19 pandemic. Additionally, it is worth noting that published tools, such as VarEPS, do not account for sample redundancy between training and testing sets, which can induce information leakage and lead to inflated performance measures [48]. VirusWarn evaluations explicitly addressed this problem by temporarily separating the training and testing sets, and showing good performances.

VirusWarn is designed to be easily extended to other viruses and for example, Influenza A H5N1 could be integrated to support genomic surveillance efforts on zoonotic Influenza viruses [49]. Moreover, extending VirusWarn to process Influenza virus segments beyond the hemagglutinin (HA) may be necessary as e.g., polymerase proteins contain key viral adaptation and evolution markers [50]. Genetic profiles from the PB2 segment could complement information derived from the HA segment, as they have been shown to contribute to the prediction of pathogenicity [21,19]. As VirusWarn uses Nextclade for Influenza clade and mutation annotations, the extension to other viruses supported by Nextclade (such as Mpox or the Respiratory Synctial Virus, RSV) can be done, provided that tables for MOCs and ROIs are generated and the scoring scheme is adjusted accordingly. However, challenges encountered in using Influenza virus data, such as frequent reference changes and limited availability of genomic data, are likely to be encountered in Mpox, RSV, and other viruses.

CRediT authorship contribution statement

Christina Kirschbaum: Writing – original draft, Writing – review & editing, Visualization, Validation, Software, Methodology,

[4] https://github.com/rki-mf1/VirusWarn-Flu/blob/main/data/A(H1N1)pdm09/fixed_2009-2015_California.csv.
[5] https://github.com/rki-mf1/VirusWarn-Flu/blob/main/data/A(H1N1)pdm09/fixed_2019-2024_Wisconsin.csv.
[6] https://github.com/rki-mf1/VirusWarn-Flu/blob/main/data/A(H3N2)/fixed_2021-2024_Darwin.csv.
[7] https://github.com/rki-mf1/VirusWarn-Flu/blob/main/data/B(Victoria)/fixed_2021-2024_Brisbane.csv.

Conceptualization. **Kunaphas Kongkitimanon:** Validation, Software, Methodology. **Stefan Frank:** Software. **Martin Hölzer:** Writing – original draft, Validation, Supervision, Resources, Data curation. **Sofia Paraskevopoulou:** Writing – review & editing, Supervision, Conceptualization, Validation, Data curation. **Hugues Richard:** Writing – original draft, Writing – review & editing, Visualization, Validation, Data curation, Supervision, Software, Methodology, Conceptualization.

Funding

K.K. and S.P. were supported by European Union's EU4Health program (ECDC/HERA/2021/008 ECD.12222). H.R. was supported by a visiting grant from the Japan Society for the Promotion of Science [BRIDGE BR220203]. This work was supported by co-funding from the European Union's EU4Health program under project no. 101113012 (IMS-HERA2) and has also received financial support from the German Federal Ministry of Health (project: IMS-RKI) on the basis of a resolution of the German Bundestag.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully acknowledge all data contributors, i.e., the authors and their originating laboratories for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative. We want to thank the SARS-CoV-2 AG EVO group at the RKI (Martin Hölzer, Sébastien Clavignac-Spencer, Max von Kleist, Thorsten Wolff) for proposing this problem, Stefan Kröger for comments on the reports, as well as Matthew Huska and Stephan Fuchs for fruitful discussions and help with debugging and continuous software support. We also thank Marianne Wedde for providing informative material on Influenza viruses, and Thorsten Wolff for general discussions and insightful comments on our work. We thank all German Electronic Sequence Data Hub (DESH) contributors and the IMSSC2 Laboratory Network. Finally, we are grateful to the Sequencing Laboratory of the Genome Competence Center, RKI, for outstanding sequencing support.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2025.03.010.

References

[1] Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 2018;19(1):9–20. https://doi.org/10.1038/nrg.2017.88.

[2] Wouters OJ, Shadlen KC, Salcher-Konrad M, Pollard AJ, Larson HJ, Teerawattananon Y, et al. Challenges in ensuring global access to COVID-19 vaccines: production, affordability, allocation, and deployment. Lancet 2021;397(10278):1023–34. https://doi.org/10.1016/S0140-6736(21)00306-8.

[3] Global dashboard for vaccine equity |. Data Futures Exchange. https://data.undp.org/insights/vaccine-equity. [Accessed 14 November 2024].

[4] World Health Organization. Global genomic surveillance strategy for pathogens with pandemic and epidemic potential, 2022–2032. World Health Assembly, Geneva. 2022.

[5] Subissi L, von Gottberg A, Thukral L, Worp N, Oude Munnink BB, Rathore S, et al. An early warning system for emerging SARS-CoV-2 variants. Nat Med 2022;28(6):1110–5. https://doi.org/10.1038/s41591-022-01836-w.

[6] McHardy AC, Adams B. The role of genomics in tracking the evolution of influenza A virus. PLoS Pathog 2009;5(10):e1000566. https://doi.org/10.1371/journal.ppat.1000566.

[7] Nypaver C, Dehlinger C, Carter C. Influenza and Influenza Vaccine: A Review. J Midwifery Women's Health 2021;66(1):45–53. https://doi.org/10.1111/jmwh.13203.

[8] Kang M, Li H-p, Tang J, Wang X-y, Wang L-f, Baele G, et al. Changing epidemiological patterns in human avian influenza virus infections jul 2024. The Lancet Microbe:100918. https://doi.org/10.1016/S2666-5247(24)00158-7.

[9] World Health Organization. Considerations for developing a national genomic surveillance strategy or action plan for pathogens with pandemic and epidemic potential. World Health Organization; 2023.

[10] Gangavarapu K, Latif AA, Mullen JL, Alkuzweny M, Hufbauer E, Tsueng G, et al. Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. Nat Methods 2023;20(4):512–22. https://doi.org/10.1038/s41592-023-01769-3.

[11] Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34(23):4121–3. https://doi.org/10.1093/bioinformatics/bty407.

[12] European Centre for Disease Prevention and Control. Erviss: European respiratory virus surveillance summary. https://erviss.org/. [Accessed 25 February 2025].

[13] Chen C, Nadeau S, Yared M, Voinov P, Xie N, Roemer C, et al. CoV-spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Bioinformatics 2022;38(6):1735–7. https://doi.org/10.1093/bioinformatics/btab856.

[14] Wittig A, Miranda F, Hölzer M, Altenburg T, Bartoszewicz JM, Beyvers S, et al. CovRadar: continuously tracking and filtering SARS-CoV-2 mutations for genomic surveillance. Bioinformatics 2022;38(17):4223–5. https://doi.org/10.1093/bioinformatics/btac411.

[15] Norwood K, Deng Z-L, Reimering S, Robertson G, Foroughmand-Araabi M-H, Goliaei S, et al. In silico genomic surveillance by coVerage predicts and characterizes SARS-CoV-2 variants of interest. bioRxiv 2024. https://doi.org/10.1101/2024.03.07.583829.

[16] Li L, Li C, Li N, Zou D, Zhao W, Luo H, et al. Machine learning early detection of sars-cov-2 high-risk variants. Adv Sci 2024;11(45):2405058. https://doi.org/10.1002/advs.202405058.

[17] Burke SA, Trock SC. Use of influenza risk assessment tool for prepandemic preparedness. Emerg Infect Dis 2018;24(3):471–7. https://doi.org/10.3201/eid2403.171852.

[18] Raharinirina NA, Gubela N, Börnigen D, Smith MR, Oh D-Y, Budt M, et al. SARS-CoV-2 evolution on a dynamic immune landscape. Nature 2025:1–9. https://doi.org/10.1038/s41586-024-08477-8.

[19] Yin R, Luo Z, Zhuang P, Lin Z, Kwoh CK. VirPreNet: a weighted ensemble convolutional neural network for the virulence prediction of influenza A virus using all eight segments. Bioinformatics 2021;37(6):737–43. https://doi.org/10.1093/bioinformatics/btaa901.

[20] Yin R, Luo Z, Zhuang P, Zeng M, Li M, Lin Z, et al. ViPal: a framework for virulence prediction of influenza viruses with prior viral knowledge using genomic sequences. J Biomed Inform 2023;142:104388. https://doi.org/10.1016/j.jbi.2023.104388.

[21] Ivan FX, Kwoh CK. Rule-based meta-analysis reveals the major role of PB2 in influencing influenza A virus virulence in mice. BMC Genomics 2019;20(9):973. https://doi.org/10.1186/s12864-019-6295-8.

[22] Sun Q, Shu C, Shi W, Luo Y, Fan G, Nie J, et al. VarEPS: an evaluation and prewarning system of known and virtual variations of SARS-CoV-2 genomes. Nucleic Acids Res 2022;50(D1):D888–97. https://doi.org/10.1093/nar/gkab921.

[23] Shu C, Sun Q, Fan G, Peng K, Yu Z, Luo Y, et al. VarEPS-influ:an risk evaluation system of occurred and virtual variations of influenza virus genomes. Nucleic Acids Res 2024;52(D1):D798–807. https://doi.org/10.1093/nar/gkad912.

[24] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol 2017;35(4):316–9. https://doi.org/10.1038/nbt.3820.

[25] Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux J 2014;2014(239):2.

[26] Kurtzer GM, cclerget, Bauer M, Kaneshiro I, Trudgian D, Godlove D. hpcng/singularity: singularity 3.7.3. https://doi.org/10.5281/zenodo.4667718.

[27] CovSonar. https://github.com/rki-mf1/covsonar. [Accessed 7 June 2024].

[28] Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw 2021;6(67):3773. https://doi.org/10.21105/joss.03773.

[29] Jinja – Jinja Documentation (3.1.x). https://jinja.palletsprojects.com/en/3.1.x/. [Accessed 14 August 2024].

[30] Xie Y. R markdown: the definitive guide. https://doi.org/10.1201/9781138359444, 2018.

[31] O'Toole A, Pybus OG, Abram ME, Kelly EJ, Rambaut A. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. BMC Genomics 2022;23(1):121. https://doi.org/10.1186/s12864-022-08358-2.

[32] Technical Advisory Group on SARS-CoV-2 Virus Evolution. https://www.who.int/groups/technical-advisory-group-on-sars-cov-2-virus-evolution. [Accessed 20 August 2024].

[33] Maher MC, Bartha I, Weaver S, di Iulio J, Ferri E, Soriaga L, et al. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. Sci Transl Med 2022;14(633):eabk3445. https://doi.org/10.1126/scitranslmed.abk3445.

[34] Kosakovsky Pond SL, Frost SDW. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 2005;22(5):1208–22. https://doi.org/10.1093/molbev/msi105.

[35] Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. Detecting individual sites subject to episodic diversifying selection. PLoS Genet 2012;8(7):e1002764. https://doi.org/10.1371/journal.pgen.1002764.

[36] Robert Koch-Institut. SARS-CoV-2 Sequenzdaten aus Deutschland. https://doi.org/10.5281/zenodo.11487868, 2024.
[37] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.
[38] Greaney AJ, Starr TN, Bloom JD. An antibody-escape calculator for mutations to the SARS-CoV-2 receptor-binding domain. bioRxiv: The Preprint Server for Biology 2021. https://doi.org/10.1101/2021.12.04.471236.
[39] Oh DY, Hölzer M, Paraskevopoulou S, Trofimova M, Hartkopf F, Budt M, et al. Advancing precision vaccinology by molecular and genomic surveillance of severe acute respiratory syndrome coronavirus 2 in Germany, 2021. Clin Infect Dis 2022;75(Suppl 1):S110–20. https://doi.org/10.1093/cid/ciac399.
[40] Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data – from vision to reality. Euro Surveill 2017;22(13):30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.
[41] Statement on the update of WHO's working definitions and tracking system for SARS-CoV-2 variants of concern and variants of interest. https://www.who.int/news/item/16-03-2023-statement-on-the-update-of-who-s-working-definitions-and-tracking-system-for-sars-cov-2-variants-of-concern-and-variants-of-interest.
[42] SARS-CoV-2 variants of concern. https://www.ecdc.europa.eu/en/covid-19/variants-concern. [Accessed 12 October 2024].
[43] Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GCM, Vervaet G, et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science 2013;342(6161):976–9. https://doi.org/10.1126/science.1244730.
[44] Huddleston J, Barnes JR, Rowe T, Xu X, Kondor R, Wentworth DE, et al. Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. eLife 2020;9:e60067. https://doi.org/10.7554/eLife.60067.
[45] Focosi D, Quiroga R, McConnell SA, Johnson MC, Casadevall A. Convergent evolution in SARS-CoV-2 spike creates a variant soup that causes new COVID-19 waves. bioRxiv. https://doi.org/10.1101/2022.12.05.518843, dec 2022.
[46] Saunders-Hastings PR, Krewski D. Reviewing the history of pandemic influenza: understanding patterns of emergence and transmission. Pathogens 2016;5(4):66. https://doi.org/10.3390/pathogens5040066.
[47] Piret J, Boivin G. Pandemics throughout history. Frontiers in Microbiology 2021;11. https://doi.org/10.3389/fmicb.2020.631736.
[48] Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the pitfalls of applying machine learning in genomics. Nat Rev Genet 2022;23(3):169–81. https://doi.org/10.1038/s41576-021-00434-9.
[49] The Lancet. H5N1: international failures and uncomfortable truths. Lancet 2024;403(10443):2455. https://doi.org/10.1016/S0140-6736(24)01184-X.
[50] Mänz B, Schwemmle M, Brunotte L. Adaptation of avian influenza A virus polymerase in mammals to overcome the host species barrier. J Virol 2013;87(13):7200–9. https://doi.org/10.1128/jvi.00980-13.