Lightweight multi-scale early warning system for influenza A spillovers

Tommaso Alfonsi[1†], Anna Bernasconi[1†*], Matteo Chiara[2], and Stefano Ceri[1]

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy.

[2]Dipartimento di Bioscienze, Università degli Studi di Milano, Milan, Italy.

[*]Corresponding author. Email: anna.bernasconi@polimi.it

[†]These authors contributed equally to this work.

Spillovers of zoonotic Influenza A viruses into farmed animals and humans have the potential to trigger epidemics or even global pandemics. We introduce FluWarning, a highly efficient and elegant computational method based on anomaly detection of codon bias and dinucleotide composition for early identification of divergent viral HA segments. We applied FluWarning to the 2009 influenza pandemic as a test case. FluWarning successfully identified the emergence of pdm09 -the virus that caused the pandemic- with warnings preceding the observed global spread. Applied to H5N1 specimens collected between 2019 and 2025, FluWarning flagged genotypes D1.1 and B3.13, both associated with recent spillovers in dairy cows in the United States. In summary, FluWarning is an effective, lightweight, multi-scale warning system for IAVs, detecting spillovers with few available sequences.

Teaser: Early detection of zoonoses using a few viral sequences is critical for the timely implementation of mitigation strategies.

Early detection of changes in the viral genome is critical for the development of warning systems and the implementation of mitigation strategies. Cross-species transmission events have been shown to be increasingly frequent (*1*) and can escalate to global health emergencies. Evolutionary breakthroughs may lead to sustainable transmission in a new ecological niche and eventually spillover to further species (as previously observed for Ebola (*2*) or coronaviruses (*3*)).

In information science, conflict and disruption are typically observed through anomaly detection methods; anomalies are carriers of critical information, such as extreme weather conditions, fault tracking, and fraud activities. Several surveys of anomaly detection methods have been proposed previously (*4,5*), also for temporal data (*6*). In the context of genomic surveillance, these approaches can be extended to processing information carried by viral genomes such as the composition in %GC, dinucleotide frequencies, and the preference for specific synonymous codons in protein-coding genes. The analysis of patterns of codon preference (exploiting redundant encoding of amino acid residues) was previously proposed to capture patterns of viral adaptation in influenza A viruses (*7, 8*). However, such analyses were retrospective and did not aim to identify the emergence of novel viral strains.

Here we introduce FluWarning, an efficient computational approach that applies anomaly detection to genomic features, flagging viral specimens showing critical changes. We selected `stray` (Search and TRace Anomaly, (*9*)), a powerful unsupervised method for anomaly detection applied to high-dimensional data. Given a set of items, anomalies are defined as items whose distance gap to typical data is significantly larger than the distance between typical items.

We use genome sequences retrieved from the Global Initiative on Sharing All Influenza Data (GISAID, (*10*)). To recognize "anomalous" sequences, the method only requires the collection date metadata and the hemagglutinin (HA) segment sequence, which is the most commonly sequenced one and modulates host specificity (*11*); we collect a few additional metadata (serotype, location, genotype, host type) to help the interpretability of the results. The method is general and adapts to any collection of IAV sequences available at a surveillance center; the method is also multi-scale, as it can detect both macroscopic events, such as the significant prevalence growth of a new clade/genotype within a large population, and microscopic events, such as individual anomalous sequences.

Computational approaches to tracking viral evolution typically focus on single mutations or

groups of mutations and their predicted functional effects on epidemiology, immunology, viral dynamics, and diagnosis/treatment, to identify novel variants of concern (*12*). Here, we leverage codon usage bias and dinucleotide frequency to capture significant changes in genome composition. Each sequence is encoded in a 75-dimensional feature vector, and the `stray` algorithm is applied to detect divergent sequences, which are individually or globally assessed. FluWarning does not aim to predict evolutionary trajectories, as in other studies based on anomaly detection (*13, 14*), but is capable of capturing key epidemiological events at their early stage.

FluWarning is configurable by time window and geographic location of interest (from continents to local surveillance units). The method runs in a streamlined, reproducible pipeline (*15*). The results are organised in a user-friendly data mart (*16*) that allows data navigation and selection of specific countries, weeks, and infected host species.

FluWarning has been designed and validated on the paradigmatic case of the 2009 H1N1 pandemic (pdm2009), as the progression and all marker events are well-known. The method was then employed in the compelling case of the H5N1 influenza virus, which has recently caused an epidemic in dairy cattle in the United States (*17*) and might be close to reaching pandemic potential according to recent studies (*18*). Our work builds upon a recently published, systematic analysis of the complete collection of publicly available complete genome H5N1 sequences collected in North America (*19*) starting from 2019.

Results

Data collection. From GISAID EpiFlu (*10*), we downloaded 3,387 H1N1 specimens collected from mammalian hosts between July 28th, 2008 and January 10th, 2010, and 12,854 H5N1 specimens from any host between January 1st, 2019 and July 25th, 2025. For both datasets, we downloaded only specimens collected in North America. After our data curation pipeline (see *Materials and Methods*) we obtained: 3,034 H1N1 sequences (96.4% from human hosts and 3.6% from swine) and 12,092 H5N1 sequences (79.6% from wild birds, 5.9% from domestic birds, 14.2% from non-human mammals, and 0.3% from human hosts).

Feature Selection. Several studies highlighted how IAVs from different hosts display different genome compositions in terms of di-nucleotides (*20, 21*) and codon usage preferences (*7, 8, 22–25*).

The data processing pipeline computes 75 numeric features for every sequence. The first 59 features correspond to the Relative Synonymous Codon Usage (RSCU, (*26*)) of all the codons, excluding the three stop codons (TAA, TAG, TGA) and the amino acids lacking alternative synonymous codons (Met: ATG; Trp: TGG). Further, 16 features describe dinucleotide frequencies of all the permutations $P(4, 2)$, with repetitions, of the 4 nucleotides, normalized by genome composition.

Anomaly Detection in High Dimensional Data. `Stray` (*9*) focuses on anomaly detection in high-dimensional data. In our application of `stray`, each item is a vector of 75 variables; items are sorted according to a total order by `CollectionDate` and then by `AccessionID`. The algorithm is applied to moving windows consisting of N items. At each new window W_i, the first item of window W_{i-1} is dropped, and the item n_i, denoted as *new item*, is added as last window item; we refer to the collection date of the new item as *characteristic date* of the window. Note that the new item in the window W_i is included in the subsequent $N - 1$ windows, where it progressively scales up to becoming the first item, and then is finally removed.

In `stray`, an important, user-defined parameter is k; an item is deemed an outlier when the distance to its k-th nearest neighbor is above a data-driven threshold, termed "gap", automatically computed (see *Materials and Methods*); k can be interpreted as the maximum size of micro-clusters detected as anomalies. For each window W_i, `stray` produces a vector $\mathcal{V}_i$ of N binary values v_{ij}, where $v_{ij} = 1$ indicates that item j in the window i is anomalous. Finally, a window is associated with a *warning* if the new item is anomalous, i.e., $v_{i,N} = 1$.

Tracking viral strain replacement with `stray`

Fig. 1A shows `stray` applied to 51 consecutive windows, representing a complete substitution of blue items (baseline viral population) with red items (new viral population). We consider $N = 50$ and $k = 10$; `stray` is executed 51 times, starting from 50 blue items (blue), progressively replaced by red items. Note that the blue item in position 1 of window 0 is discarded in window 1, making space for an item in position 50 of window 1. This item is an anomaly, as its distance from all other items is greater than a data-defined gap; all anomalies are rendered with a black triangle.

Anomalies occur only when red items, that progressively form a micro-cluster, are below the user-defined threshold k. Thus, anomalies progressively occur only for windows 1 through 10, all associated with a warning. At window 11, a cluster of size 11 is created, and for each red item, the

10th-neighbor distance drops; red items are no longer anomalies as they belong to a sufficiently large cluster (see Fig. 1B). Then, the substitution of blue items with red items continues, up to window 40. Here, blue items are a minority; they create a micro-cluster of size 10 and become anomalies. However, as warnings are only associated to new items, no warning is thrown.

Sensitivity and specificity analysis. We prepared an *ad hoc* dataset with 1) N-sized baseline group (called $G1$), drawn from an initial set of 130 ordered items (called $X1$); and 2) N-sized penetrating group (called $G2$), drawn from a different set of 100 ordered items (called $X2$). The two sets $X1$ and $X2$ represent viral sequences collected in the same epidemic year, but from different viral clades, retrieved from GISAID; individual IDs are reported in our Zenodo repository (*15*); each sequence is summarized by 75 extracted features representing its HA viral genome. We build the noise experiments by randomly replacing respectively 10%, 20%, or 30% of the items of $G2$ with elements of $X1$ (termed *noise-items*) not used in $G1$.

Table 1 reports results, with $N = (50, 100)$ and $k = (10, 15)$; a complete version of the experiments with 15 combinations of N and k is in Table S1 - but, as discussed in our H1N1 pandemic analysis, smaller N, k pairs are not recommended for epidemic/pandemic analyses. Each table line represents a specific combination of N, k, and noise level (percentage over $G2$ and absolute number of *noise-items*). For each combination, we run FluWarning 50 times and average the results, collected in the last four columns, which represent: 1) the number of warnings – identical to k in all cases; 2) the window number in which the last warning was detected (denoted as Last Warning Ordinal, LWO); 3) sensitivity, defined as k/LWO (i.e., the ideal last warning ordinal divided by the actual last warning ordinal – capturing the ratio of true positives over all the warnings); and 4) specificity, defined as $(N + 1 - LWO)/(N + 1 - k)$ (i.e., the ideal ordinal from which non-warning items are detected, divided by the actual non-warning ordinal – capturing the ratio of true negatives over all the warnings).

As expected, LWO increases with noise, while sensitivity and specificity decrease from the perfect value of 1 (no noise); nonetheless both metrics remain acceptable even with 30% noise, showing that the method is robust. Note that, in any warning system, a few false positives are not a concern, as all sequences that triggered a warning must be further explored.

Data mart-based monitoring. FluWarning is distributed as software supporting *data processing* and *data exploration*. In data processing, the software receives a stream of sequences (e.g, down-

loaded from GISAID by an authorized user or provided from a surveillance laboratory). After sequence alignment, feature selection, and anomaly detection, warning-related sequences and their metadata are collected within a (SQLite) database. For data exploration, the database feeds a *data mart*, explorable through a web-based application (Streamlit), packaged in an easy-to-install container (Docker), see (*15*). The data mart supports the inspection of aggregate counts of warnings across selected hosts, locations, and periods, as well as the extraction of single sequences metadata; the software is developed to facilitate installation and use.

The H1N1 pandemic (2008-2009)

We used the 2009 H1N1 pandemic to identify the most appropriate settings for FluWarning for the detection of epidemiologically important events. The pdm09 virus, which caused the pandemic, originated in late 2008 in Mexico, but was first isolated and sequenced on April 14th, 2009, by the Centers for Disease Control and Prevention (*27*). Following a rapid increase in flu cases associated with pdm09, the World Health Organization (WHO) declared a "public health emergency of international concern" on April 25th, 2009. By June, pdm09 completely replaced other seasonal H1N1 flu strains, and the outbreak was declared a pandemic (*28*).

We analyzed 3,034 H1N1 HA sequences collected from specimens infecting human hosts between August 2008 and January 2010. Data were aggregated in non-overlapping intervals of two weeks. Each interval has a *Starting Week Number* W_i (with incremental numbers, per year), a corresponding *Start Date* D_i, and includes all the sequences with a collection date included in the interval $[D_i, D_{i+1})$. To establish the most suitable settings for identifying the emergence of a novel virus in FluWarning, we visually inspected the total number of warnings emitted by the system at every interval W_i, according to different combinations of N and k.

As shown in Fig. S1, with $N = (50, 100)$ and $k = (10, 15)$, a sharp peak of warnings was observed, specifically at intervals ranging from 2009-11 to 2009-20. These intervals span from March 12th, 2009 to May 13th, 2009 and align exactly with the initial emergence and spread of pdm09. Based on these observations, only values of $N = (50, 100)$ and $k = (10, 15)$ were considered for epidemic/pandemic scale analyses.

We reasoned that windows W_i associated with a statistically significant increase (corrected p-value ≤ 0.05) in the number of detected anomalies likely reflect key epidemiological events such

as -for example- the emergence and spread of a new viral variant. From now on, we will refer to these windows as *Super Warning*s. *Super Warning*s were identified using Fisher's exact test (see *Materials and Methods*) by comparing the proportion of warnings emitted at a given window W_i with the overall proportion of warnings across the entire dataset. P-values were corrected with the Benjamini-Hochberg procedure to control the false discovery rate. An empirical assessment (see Fig. S2) indicated that *Super Warning*s can be reliably detected when an interval contains at least 20 sequences, of which six or more ($\geq$ 30%) are flagged with a warning. To minimize biases introduced by uneven sampling of viral isolates (see large discrepancies in sequencing coverage in Fig. 2), we computed *Super Warning*s only for windows containing at least 20 sequences.

Results are shown in Fig. 2A–C. For all four configurations of `stray`, *Super Warning*s (see dark blue cells) were emitted from 2009-13 to 2009-21. The first pdm09 sequence in the dataset was collected on March 12th, 2009 (week 2009-11) and was identified as a warning by `stray` in all four configurations. A *Super Warning* is instead emitted at 2009-13; by this interval of time, the number of pdm09 sequences in the GISAID database increased to 14, collected in Mexico (n=7) and California (n=7), consistent with the epidemiological history of the H1N1 pandemic of 2009.

Interestingly, starting from week 2009-15 – when pdm09 established itself as the most prevalent H1N1 clade in the dataset – the majority of the warnings emitted by the system flagged seasonal flu viral isolates. By week 2009-25, pdm09 accounted for (almost) the totality of H1N1 samples, and the number of warnings identified by `stray` at any subsequent intervals dropped to levels comparable to those observed before the emergence of pdm09 (see Fig. S3 and Data S1).

Data mart-based monitoring. To showcase the power of our data mart, consider Fig. 2D, obtained by setting $N = 50$ and $k = 10$, which illustrates the geographic locations associated with warnings. In bi-week 2009-13/14 (March 23rd - April 5th, 2009), warnings occurred in California (4 warnings out of 7 isolates), Mexico (2/7), and Illinois (1/1); in bi-week 2009-15/16 (April 6th-19th, 2009), warnings occurred in Hawaii (3/3), Texas (1/11), and Mexico (1/22). By week 2009-17/18 (April 20th - May 3rd, 2009), a substantial number of warnings occurred in California (25/108), and US-collected warnings spread to Minnesota, Kansas, New York, and Texas, marking the start of the pdm09 pandemic. Collectively, these results indicate that FluWarning correctly recapitulates the main events of the 2009 H1N1 pandemic and flags the emergence of new viral clades/strains almost in real time. These considerations prompted us to apply the same analytical workflow to H5N1.

The H5N1 2.3.4.4b clade (2020-2025)

Highly pathogenic (HPAI) strains of H5N1 IAVs are considered a constant pandemic threat since their recurrent outbreaks in mammalian species, including sporadic cases of human infection with reported high mortality rates (*29*). Originally described and isolated in 2013-2014 in Asia, the HPAI H5N1 2.3.4.4b clade recently crossed continental barriers and became increasingly frequent in Europe, North America, and South America (*30*). Besides infecting wild and domestic birds, 2.3.4.4b was associated with several distinct spillovers to mammals throughout 2020 to 2025 and recently caused a large epidemic of avian flu in dairy cattle in the United States, raising a significant concern for human health (*17*). Two distinct genotypes of H5N1 2.3.4.4b have caused independent spillovers of IAV to dairy cows in North America. Genotype B3.13 was responsible for a large-scale outbreak in dairy cattle in 2024; the first isolate of B3.13 was collected from a dairy cow in Texas on March 13th, 2024 (*19*). More recently, in January 2025, genotype D1.1 was detected in dairy cattle in Nevada (*31*); however, this outbreak was relatively limited in scale compared with the one caused by B3.13.

We considered a total of 12,092 HA segment sequences and associated metadata of H5N1 collected between September 15th, 2020, and May 29th, 2025, in North America. As a side note, we extracted GISAID data up to July 25th, 2025; 1,097 isolates submitted in the last three months of 2025 had incomplete geographic and collection date metadata, and could not be evaluated. A total of 1,068 distinct warnings were emitted by FluWarning in at least one of the $N = 50, 100$ and $k = 10, 15$ configurations (see Fig. S4 and Data S1); sequences assigned to the B3.13 genotype accounted for 62.83% of the total number of warnings and 90.05% of the warnings emitted by FluWarning between 2024-11 and 2024-37. These windows were associated with *Super Warning*s according to all four selected configurations of `stray`, see Fig. 3C (complete heatmap in Fig. S5). As observed from the breakdown of the circulation of 2.3.4.4b genotypes in North America, displayed in Fig. 3A-B, this interval aligns almost exactly with the emergence and spread of B3.13. Several other genotypes, including for example B2.1, B3.2, B3.6, display a widespread and sustained circulation in North America (estimated prevalence ≥ 0.2 for more than 4 weeks) in the time-frame considered in our analyses, but none of these exhibited a statistically significant association with *Super Warning*s.

Unlike B3.13, which accounted for the majority of warnings, only 23 D1.1 sequences were collectively associated with a warning, representing approximately 0.2% of the total (Data S1). However, 20 of these sequences were collected between the end of 2024 and the first half of 2025 (weeks 2024-50 to 2025-25); in this dataset, only four D1.1 sequences originate from dairy cows, of which the first one – collected in Nevada on January 21st, 2025 – was flagged as a warning, aligning with the observation reported in (*31*). All D1.1 warnings were detected using a window size of $N = 50$, and no warnings were detected for D1.1 with $N = 100$ (Fig. S4).

As shown in Fig. 3, the configuration with $N = 50$ and $k = 10$ emitted *Super Warning*s at the end of 2024/beginning of 2025 (2024-52 to 2025-01), coinciding with the reported outbreak of D1.1 in dairy cattle in the United States. D1.1 sequences accounted for a substantial proportion (∼35%) of warnings recorded across this interval of time, consistent with a *Super Warning* for this genotype. Configurations with $N = 100$ also produced a *Super Warning* in 2025; however, this occurred later and was associated solely with B3.13.

These observations suggest that, due to the large number of sequences with incomplete metadata, the patchy distribution and relatively limited number of available sequences starting from February 2025, configurations of `stray` with ($N = 50$) are better suited for analyzing 2025 data and can effectively flag the D1.1 spillover to dairy cattle.

Data mart-based analysis for individual B3.13 sequences. FluWarning can be applied to surveillance on smaller scales (US States in our dataset, but -in principle- regional surveillance centers); with more granular analysis, `stray` can be exploited with lower k parameters. When focusing only on B3.13 sequences, a progressive movement of warnings (using N=50 and k=1) is observed in the same bi-weekly intervals – see Fig. 3D, showing subsequent US maps and the corresponding B3.13 warning counts (automatically generated by our data mart). Each panel shows geo-located warnings (red circle) included in collected sequences (black circle). Numbers are explicitly indicated in the tables below; grey cards specify warnings thrown by sequences found on hosts other than dairy cows.

Starting March 7th, 2024, Texas collected one warning sequence from chicken, then on March 10th, 2024, three from dairy cows; the following intervals see an increase in warnings in Texas (1 wild bird and 1 chicken in 2024-12; 1 human in 2024-13 ID: "*EPI_ISL_19027114_A/Texas/37/2024*"). At weeks 2024-11/12, we observe warnings in New Mexico (March 15/16th), Ohio (March 15th),

and subsequently Kansas (March 20th), New Mexico (feline on March 20th), and Michigan (March 21st). In the interval 2024-13/14, warnings were also recorded in Idaho (March 27th), North Carolina (April 4th), and South Dakota (April 5th). In the interval 2024-15/16, warnings appear in Colorado (April 10th). The last two maps depict a much later period, showing the emergence of warnings in California (2024-35/36), precisely on August 26th, 2024 (ID: "*EPI_ISL_19387789_A/dairy_cow/USA/24_024712-003/2024*").

Warnings emitted for 5 US States are shown in Fig. S6. For every row, we report bi-weekly counts, according to all possible assignments of parameters N and k. Note the peculiar situation of California (Fig. S6B): warnings are emitted for a long period, from August 2024 to February 2025, and then very few sequences from complete isolates were collected (7 during bi-week starting March 24th, 2025 and 1 during bi-week starting May 19th, 2025), in spite of the reported critical situation of dairy cattle (*19*).

Multi-segment analysis. With the aim of assessing whether non-HA segments provide distinct or complementary signals in the identification of divergent/novel viral variants, we applied our early warning system to all eight genomic segments of H5N1 clade 2.3.4.4b. Only viral isolates for which all eight segments of the genome were included in this analysis, for a total of 12,092 viruses. For segments (PB1, PA, MP, and NS) with overlapping coding sequences, we considered only the longest CDS. For all the segments, we applied `stray` with $N = 50$ and $k = 10$, which had previously yielded accurate and timely results for HA.

We compared the overall number of warnings (hundreds per segment) and the number of *Super Warning*s derived by the analysis of each segment (Fig. 4A). HA and NA generated a number of *Super Warning*s similar to the other segments, but with markedly fewer warnings. Indeed, while the other segments generally produced approximately one *Super Warning* every ∼100 warnings, the HA and NA ratios were nearly threefold lower, indicating a more favorable signal-to-noise profile for these two structural segments.

We then inspected the distribution of clade 2.3.4.4b genotypes associated with warnings (Fig. 4B). HA and NA flagged almost exclusively genotype B3.13, whereas the other segments displayed a broader warning profile, encompassing additional genotypes, including B1.1, B1.2, B3.2, and A1, none of which have epidemiological relevance according to current knowledge. Genotypes D1.1 and B3.6 (i.e., the direct ancestor of B3.13) accounted for a relatively small

proportion of warnings across all segments, with the exception of the M1 segment, where their representation was slightly elevated.

We computed *Super Warning*s over the two-year interval from week 2023-19 to week 2025-21, spanning the recent spillover of 2.3.4.4b into dairy cattle for all eight segments, and compared their temporal distribution (Fig. 4C). A high degree of concordance was observed among segments, with the majority of *Super Warning*s clustered between weeks 2024-11 and 2024-39 and corresponding to the emergence of B3.13. Genotypes associated with every *Super Warning*, derived by the analysis of each independent segment, were recorded and inspected.

According to Fig. S7, HA primarily flagged B3.13 but also detected B3.6 and D1.1. The latter's *Super Warning* occurred in January 2025, coinciding with its independent spillover into dairy cattle, and aligned well with epidemiological observations. NA, NP, and NS1 exhibited temporal and genotype-specific *Super Warning* patterns broadly consistent with those detected for HA. Nonetheless, these segments additionally identified genotypes not currently regarded as epidemiologically relevant, including A3 and B3.7 (NP), C2.1 (NS1), and D1.3 (NA). Furthermore, although all three segments generated *Super Warning*s for D1.1, in contrast to HA, these signals were distributed across an extended temporal interval and did not temporally coincide with the reported spillover of this lineage into dairy cattle.

Segment M1 (Fig. S8) flagged additional non-epidemiologically relevant genotypes (A.3 and C2.1) and detected B3.13 with a delay of 2 weeks compared to HA, whereas it identified B3.6 and D1.1 earlier than HA. Among other segments (Fig. S8), PB2 exhibited a limited number of *Super Warning*s and did not identify B3.13 in a timely manner. PB1 and PA exhibited a more patchy temporal profile, and their *Super Warning*s were linked not only to epidemiologically relevant genotypes but also to A2, C2.1, and other minor genotypes.

Taken together, these findings support the HA segment as the best single-segment choice for a lightweight warning system. Its advantages include a reduced total number of warnings, robust identification of all currently epidemiologically relevant genotypes in clade 2.3.4.4b, and ease of interpretation, given the large record of potentially epidemiologically relevant mutations already characterised for the HA gene. NA and NP could be viable alternatives, with overall similar performance metrics, but with reduced interpretability compared to HA. In contrast, non-structural segments (PA, PB2, PB1, and NS1) and the M1 segment are less suitable due to their higher warning

counts and the broader spectrum of genotypes - many of which are not epidemiologically relevant - associated with their *Super Warning*s. This wider genotype range may reflect higher reassortment rates for these internal proteins, consistent with previous reports of segment-specific reassortment dynamics in H5N* avian influenza viruses (*32, 33*).

Cluster and mutation analysis. The HaploCoV (*34*) software was used to partition 2.3.4.4b genotypes, as defined by Nguyen et al. (*19*), into clusters of similar HA sequences. Distinct clusters were delineated when at least 100 sequences differed by five or more non-shared nucleotide variants. A total of 10 distinct clusters were formed (Fig. 5A). Interestingly, 3 of these 10 clusters (HA.N1, HA.N2, and HA.N10) were composed almost exclusively of sequences assigned to the B3.13 genotype. The B3.6 genotype -ancestor of B3.13 according to (*19*)- corresponded with a single HaploCoV cluster labelled HA.N3. All HA segments of genotype D1.1 were assigned to a single HaploCoV cluster (HA.N5).

As observed in Fig. S9, the B3.6 (HA.N3) and D1.1 (HA.N5) genotypes were widespread across many states; HA.N1 was associated with specimens collected from dairy cattle in Texas, Colorado, Idaho, Michigan, and Minnesota; instead, the HA.N10 and HA.N2 clusters, detected starting from week 2024-35, were observed only in California.

We designate as defining nucleotide variants those observed in at least 50% of the sequences of each HaploCoV cluster; these nucleotide variants were inspected (see Fig. 5B). Consistent with phylogenetic analyses, HA.N3 (B3.6) and HA.N1, HA.N10, HA.N2 (B3.13) shared a similar pattern of defining nucleotide variants and formed a well-defined group. HA.N1, HA.N10, and HA.N2 displayed five additional characteristic nucleotide variants compared to HA.N3; none of them resulted in an amino acid substitution. However, HA.N2 and HA.N10 displayed two additional non-synonymous substitutions,i.e., S336N and D104G, compared to HA.N1. S336N was previously reported as a virulence marker, with an increased prevalence in viruses isolated from dairy cows (*19*). D104G was recently associated with a marked increase in immune escape (*35*). Genotype D1.1 (HA.N5) exhibited a mutational profile closely related to that of genotype A3 (HA.N11), consistent with the known evolutionary history of these lineages (*36*). However, D1.1 displayed two additional characteristic amino acid changes: T52A in the H1 segment and N146D in the H2 segment. To date, neither of these mutations has been associated with any known functional or epidemiological relevance.

Taken together, these analyses provide preliminary evidence of potential host adaptation of the B3.13 lineage in dairy cows in California and highlight specific HA mutations in the D1.1 lineage that warrant further investigation.

Discussion

Our FluWarning method issues warnings whenever a new viral sequence is significantly different from a baseline of N previously collected background sequences; differences are measured in a dimensional space of 75 variables, describing dinucleotide frequencies and codon preferences. The method does not differentiate among warnings; these may reflect spillovers, new clades, or new lineages. Investigation responsibility is passed to users who interpret the warning. The method does not make use of phylogenetic analysis but leverages compositional properties of genomes. This makes it able to scale up to big datasets, being computationally faster and generally very accurate–a similar approach has been proven successful in previous research (*37–39*) that allowed us to detect major evolutionary changes (e.g., variants and recombinations) in SARS-CoV-2 genomes.

The method is multi-scale, as it considers different levels of alerting: *simple warnings* are issued on single sequences (that may be anomalous in groups of k elements); *Super Warning*s are based on bi-weekly batches of collected sequences and require a statistically significant size and p-value. Simple warnings can be observed on different location scales, host populations, and genotypes, while *Super Warning*s are available at the whole-dataset level (but could be adapted to different scales). We illustrated the robustness of the method by computing sensitivity and specificity with increasing amounts of noise, starting from an ideal case of replacement of a baseline population by a new viral population.

While multiple genome segments can yield timely and informative signals, the HA segment alone provides an efficient, interpretable, and high-specificity basis for a lightweight genomic surveillance tool, such as FluWarning. Expanding surveillance to multiple segments may still offer incremental benefits, particularly for capturing early reassortment events or detecting rare genotype emergences, but would entail a higher noise level and analytic complexity.

FluWarning is conceived for supporting surveillance systems that sample domestic and wild animals globally; it is most likely to succeed when applied to pathogens with an existing surveillance

network (now in place only for seasonal influenza). Even under the best surveillance systems (like the WHO's seasonal influenza GISRS system), sequences routinely get backfilled for earlier dates after the sample collection. Therefore, should public health officials run FluWarning over public data, the input would be relatively incomplete and noisy.

Thanks to its lightweight code, user-defined parametrization, and easy-to-inspect data mart analytics, FluWarning has the potential of being used by many laboratories or regional-scale genomic surveillance institutions. The accuracy of our tool shows that broader and deeper support for global surveillance could pay off, enabling significant small-scale discoveries.

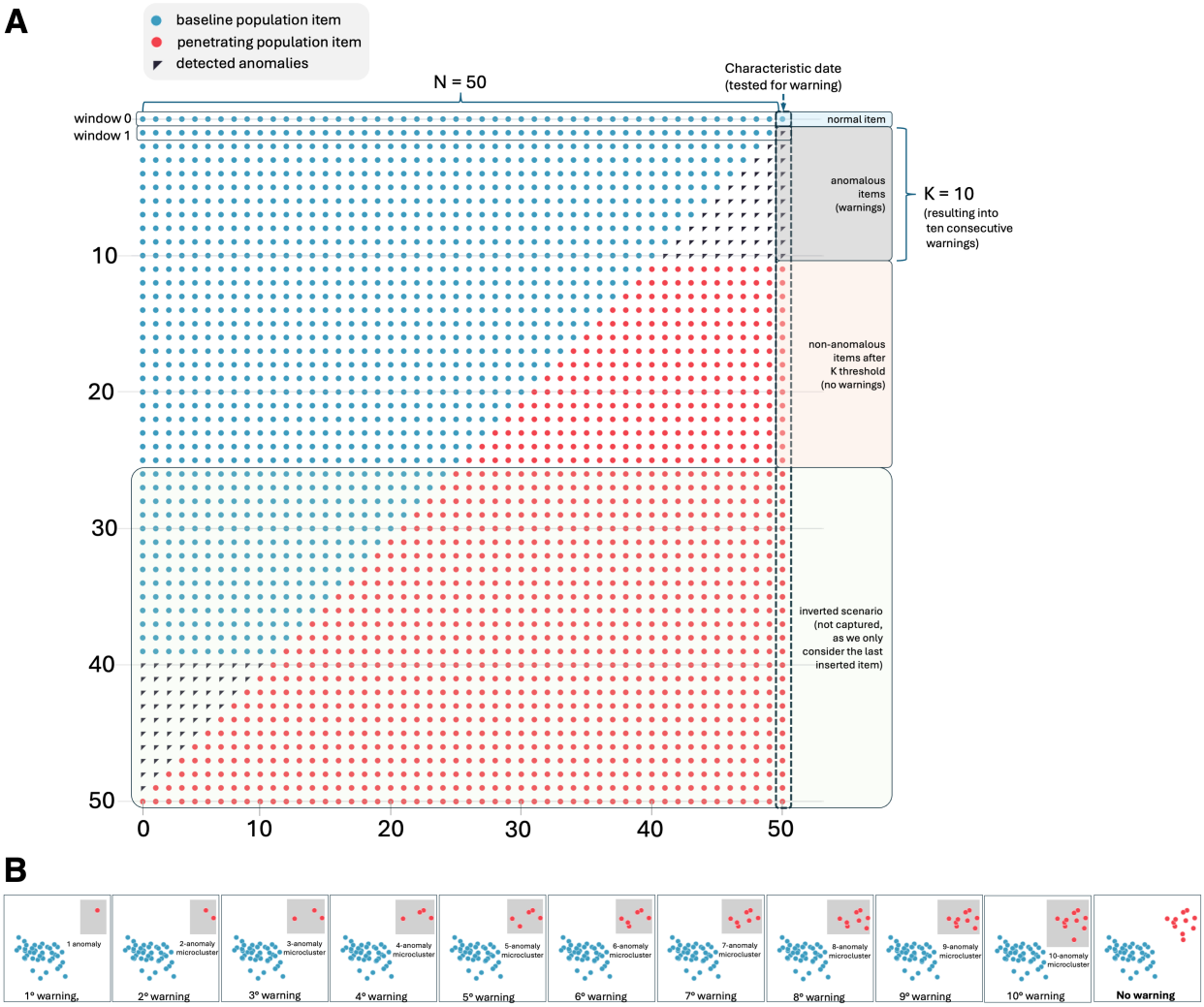


Figure 1: **`Stray` method explained.** (**A**) Complete substitution of blue items (a baseline viral population) with red items (a new viral population). With N = 50 and k = 10, `stray` is executed 51 times, starting from a window that contains 50 normal items (blue). Anomalies occur only when red items, that progressively form a micro-cluster, are below the user-defined threshold k - then, in windows 1 through 10, all are associated with a warning. At window 11, a cluster of size 11 is created, red items cannot be considered anomalous because they have created a sufficiently large cluster. Then, the population of red items grows up to window 40, where blue items are a minority; they create a micro-cluster of size 10, and all blue items become anomalies. However, as we are only interested in new items (in position 50), no warning is thrown. (**B**) shows an "ideal" case where blue and red sequences are well-separated in the 75-dimensional space. Here, we capture distances on two highly-representative dimensions (e.g., the two principal components after dimensional reduction). Red micro-clusters are recognized as anomalous until their size reaches k = 10.

Table 1: **Sensitivity and specificity analysis, with $N = (50, 100)$ and $k = (10, 15)$.** Each line represents a specific combination of N (column 1), k (column 2), and noise level as a percentage over $G2$ (column 3). The absolute number of *noise-items* is also reported (column 4). For each combination, we run FluWarning 50 times; we report the number of warnings (column 5); averages of the window number in which the last warning was detected, denoted as Last Warning Ordinal - LWO (column 6); sensitivity (column 7); and specificity (column 8).

N	K	%Noise	#NoiseItems	#Warnings	LWO	Sensitivity	Specificity
50	10	0 %	0	10	10.00	1	1
		10 %	5	10	11.16	0.901	0.972
		20 %	10	10	12.68	0.798	0.935
		30 %	15	10	13.84	0.736	0.906
	15	0 %	0	15	15.00	1	1
		10 %	5	15	16.70	0.900	0.953
		20 %	10	15	18.60	0.812	0.900
		30 %	15	15	21.10	0.719	0.831
100	10	0 %	0	10	10.00	1	1
		10 %	10	10	11.00	0.915	0.989
		20 %	20	10	11.98	0.849	0.978
		30 %	30	10	14.52	0.709	0.950
	15	0 %	0	15	15.00	1	1
		10 %	10	15	16.52	0.913	0.982
		20 %	20	15	18.22	0.829	0.963
		30 %	30	15	21.70	0.706	0.922

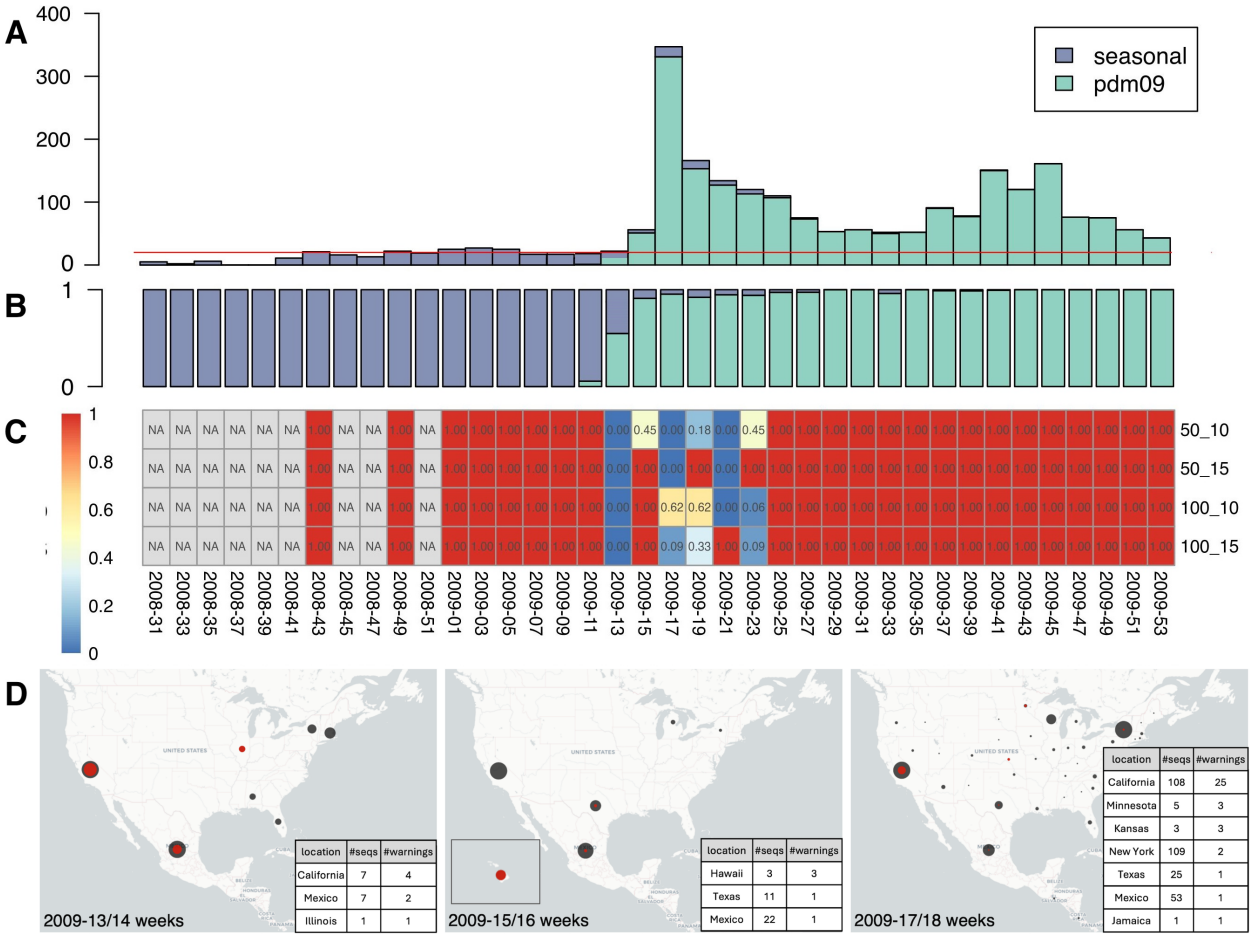
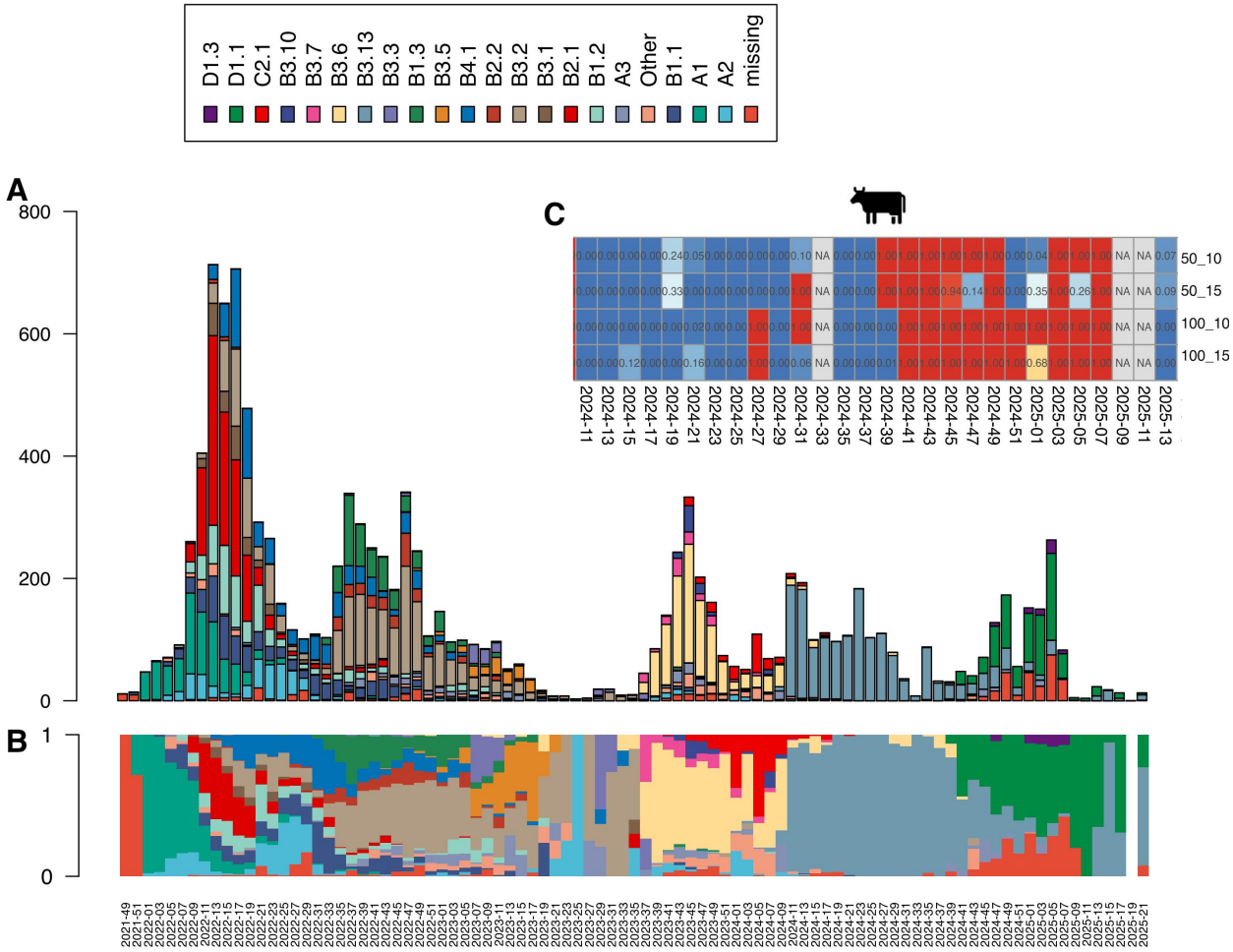
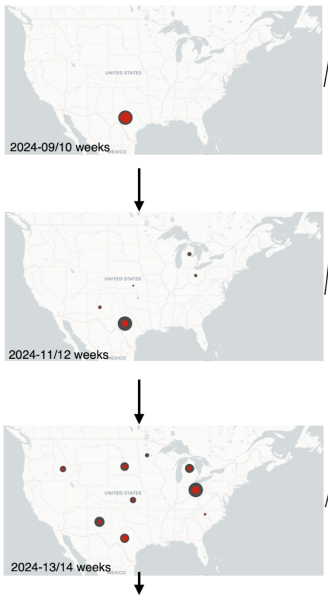


Figure 2: **Circulation of H1N1 strains from 2008-33 to 2009-53 and warnings by `stray`.** (**A**) Barplot showing the total number of H1N1 sequences classified as "seasonal-flu" or "pdm09" in the GISAID EpiFlu database at each time interval. The red line indicates the threshold (20 sequences) used to compute *Super Warning*s (**B**) Relative proportions (percentage with respect to the total number of sequences) of seasonal-flu and pdm09 sequences. (**C**) Heatmap (with `stray` settings $N = 50, 100$ and $k = 10, 15$) displaying the FDR corrected p-values for the statistically significant increase in the number of warnings. Time intervals are shown on columns, parameters of `stray` on the rows. NA: *Super Warning*s are not computed due to an insufficient number of sequences. (**D**) Maps drawn for North and Central America, representing warnings during the 2009-13 to 2009-17 weeks period. One can note the diffusion, initially in Mexico and Southern California (as it can be read in location metadata), up to other States of the US.



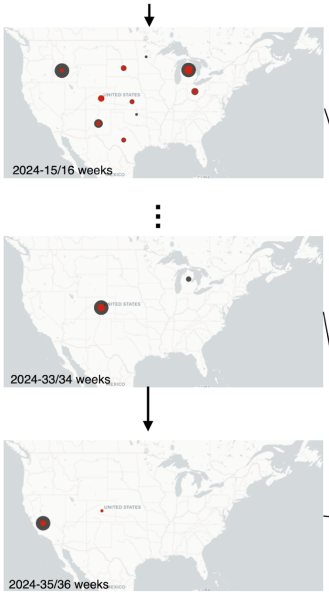
D



location	#seqs	#warnings
USA, Texas	7	4 (1 chicken)

location	#seqs	#warnings
USA, Texas	146	30 (1 chicken, 1 grackle)
USA, New Mexico	10	3 (1 feline)
USA, Ohio	8	2
USA, Michigan	13	2
USA, Kansas	3	1

location	#seqs	#warnings
USA, Ohio	56	16
USA, Texas	23	10 (1 human, 1 chicken, 1 grackle)
USA, South Dakota	20	6
USA, Michigan	22	5
USA, New Mexico	28	4
USA, Idaho	11	3
USA, Kansas	11	2
USA, North Carolina	2	1



location	#seqs	#warnings
USA, Michigan	25	9 (1 chicken, 2 turkey, 2 cats)
USA, Colorado	5	5
USA, Ohio	6	4
USA, South Dakota	4	3 (cats)
USA, New Mexico	9	3 (2 chicken, 1 cat)
USA, Texas	3	2
USA, Kansas	3	2 (1 chicken)
USA, Idaho	25	2

location	#seqs	#warnings
USA, Colorado	7	2

location	#seqs	#warnings
USA, California	81	14
USA, Colorado	4	4

Figure 3: **Circulation of 2.3.4.4b H5N1 strains from 2020 to 2025 and warnings by `stray`.** (**A**) Barplot showing the total number of H5N1 sequences at every time interval, classified according to the genotypes recently associated with H5N1 in North America. (**B**) Relative proportions (percentage with respect to the total number of sequences) of genotypes, shown to facilitate the direct comparison of their prevalence. (**C**) Heatmap displaying the FDR-corrected p-values for the statistically significant increase in the number of warnings (*Super Warning*). Only the interval associated with the diffusion in non-human mammals (dairy cattle) is displayed (complete analyses in Fig. S5). Time intervals are shown in columns. Parameters of `stray` on the rows. NA: *Super Warning*s were not computed due to an insufficient number of sequences. (**D**) Maps drawn for North and Central America, representing warnings during the 2024-09 to 2024-36 weeks period. One can note the diffusion, initially in Texas (week 9-10), then also in New Mexico, Ohio, Kansas and Michigan (week 11-12), then also in Idaho, North Carolina and South Dakota (week 13-14), then also in Colorado (week 15-16); we move then to weeks 33-36 to show the offspring of warnings in California, first presented in weeks 35-36.

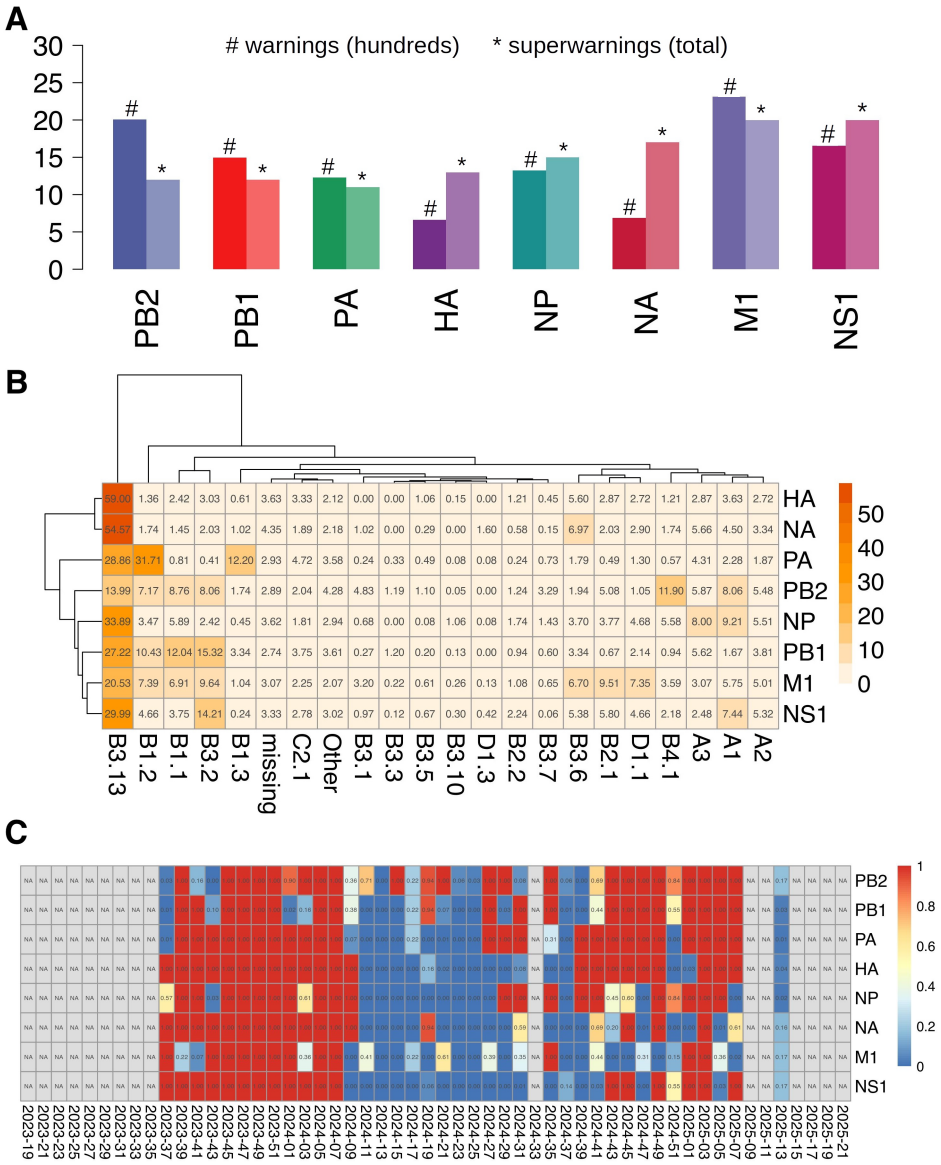


Figure 4: **Comparison of warnings and *Super Warnings* obtained from the analysis of all eight segments of H5N1 2.3.4.4b influenza A genomes.** (**A**) Barplot showing the total number of warnings (in hundreds, #) and the number of *Super Warnings*, displayed side by side for each segment. (**B**) Heatmap illustrating the overall proportion (%) of warnings associated with distinct 2.3.4.4b genotypes for each genome segment. Genotypes are reported on the columns, segments on the rows. A two-way hierarchical clustering groups segments and genotypes with similar patterns of warnings detected by `stray`. (**C**) Heatmap of *Super Warnings* associated with each segment from 2023-19 to 2025-21. Time intervals are shown on the columns and segments on the rows. Corrected p-values for *Super Warnings* are reported in each cell.

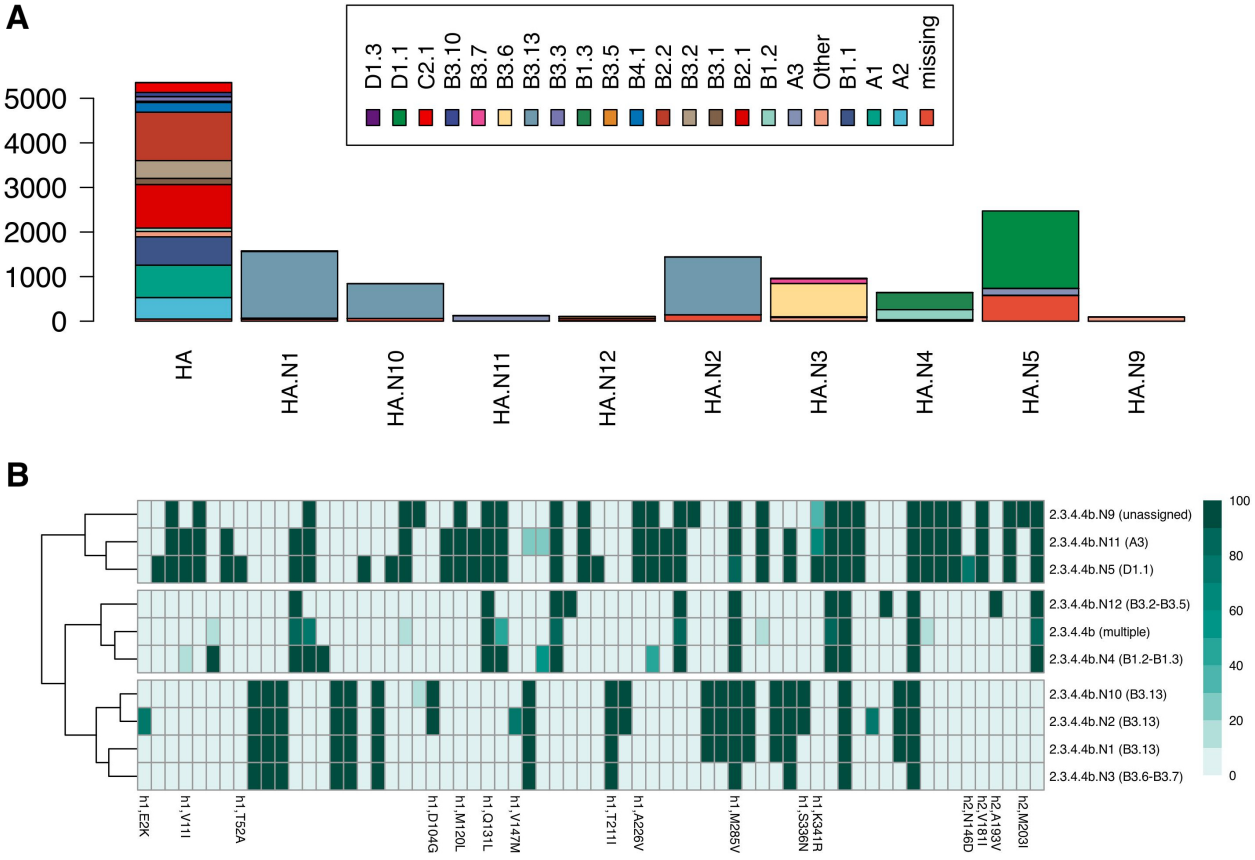


Figure 5: **Number of sequences of each genotype corresponding to a HaploCoV-defined cluster and characteristic mutations of HaploCoV clusters.** (**A**) For every HaploCoV cluster, the barplot displays the total number of sequences from each distinct genotype assigned to that cluster. Clusters HA.N1, HA.N10, and HA.N2 are formed almost exclusively by sequences assigned to the B3.13 genotype. Similarly, a clear correspondence between HA.N3 and B3.6 can be observed. (**B**) The heatmap illustrates the prevalences (% min 0 max 100) of characteristic nucleotide variants in each distinct HaploCoV cluster. HaploCoV clusters are reported on the columns; nucleotide variants on the rows. For nucleotide variants associated with non-synonymous substitutions, the predicted amino acid substitution is reported (complete annotation in Data S1). Consistent with phylogenetic analyses, clusters associated with B3.6 (HA.N3) and B3.13 (HA.N1, HA.N10, and HA.N2) form a neat and well-separated group and are defined by a similar pattern of characteristic nucleotide variants.

References and Notes

1. A. J. Meadows, N. Stephenson, N. K. Madhav, B. Oppenheim, Historical trends demonstrate a pattern of increasingly frequent and severe spillover events of high-consequence zoonotic viruses. *BMJ Global Health* **8** (11), e012026 (2023).

2. S. Mursel, *et al.*, Estimation of Ebola's spillover infection exposure in Sierra Leone based on sociodemographic and economic factors. *Plos one* **17** (9), e0271886 (2022).

3. G. T. Keusch, *et al.*, Pandemic origins and a One Health approach to preparedness and prevention: Solutions based on SARS-CoV-2 and other RNA viruses. *Proceedings of the National Academy of Sciences* **119** (42), e2202871119 (2022).

4. V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey. *ACM computing surveys (CSUR)* **41** (3), 1–58 (2009).

5. C. C. Aggarwal, *Otlier analysis (2nd ed.)* (Cham: Springer) (2017).

6. M. Gupta, J. Gao, C. C. Aggarwal, J. Han, Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and data Engineering* **26** (9), 2250–2267 (2013).

7. E. H. Wong, D. K. Smith, R. Rabadan, M. Peiris, L. L. Poon, Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. *BMC evolutionary biology* **10**, 1–14 (2010).

8. T. Alfonsi, M. Chiara, A. Bernasconi, A codon usage-based approach for the stratification of Influenza A across recent spillovers. *Computational and Structural Biotechnology Journal* (2025).

9. P. D. Talagala, R. J. Hyndman, K. Smith-Miles, Anomaly detection in high-dimensional data. *Journal of Computational and Graphical Statistics* **30** (2), 360–374 (2021).

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

11. A. Gambaryan, *et al.*, Evolution of the receptor binding phenotype of influenza A (H5) viruses. *Virology* **344** (2), 432–438 (2006).

12. Y. Turakhia, *et al.*, Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nature* **609** (7929), 994–997 (2022).

13. A. Forna, *et al.*, Sequence-based detection of emerging antigenically novel influenza A viruses. *Proceedings B* **291** (2028), 20240790 (2024).

14. A. Garjani, *et al.*, Forecasting influenza hemagglutinin mutations through the lens of anomaly detection. *Scientific Reports* **13** (1), 14944 (2023).

15. T. Alfonsi, A. Bernasconi, M. Chiara, S. Ceri, Supporting data and code for "Lightweight multi-scale early warning system for influenza A spillovers", `https://doi.org/10.5281/zenodo.15498406`, last accessed: Aug 24th, 2025.

16. A. Bonifati, F. Cattaneo, S. Ceri, A. Fuggetta, S. Paraboschi, Designing Data Marts for Data Warehouses. *ACM Transactions on Software Engineering and Methodology* **10** (4), 452–483 (2001).

17. L. C. Caserta, *et al.*, Spillover of highly pathogenic avian influenza H5N1 virus to dairy cattle. *Nature* pp. 1–8 (2024).

18. D. M. Morens, J. Park, J. K. Taubenberger, Many potential pathways to future pandemic influenza. *Science Translational Medicine* **15** (718), eadj2379 (2023).

19. T.-Q. Nguyen, *et al.*, Emergence and interstate spread of highly pathogenic avian influenza A(H5N1) in dairy cattle in the United States. *Science* **388** (6745), eadq0900 (2025).

20. Y. Iwasaki, T. Abe, Y. Wada, K. Wada, T. Ikemura, Novel bioinformatics strategies for prediction of directional sequence changes in influenza virus genomes and for surveillance of potentially hazardous strains. *BMC infectious diseases* **13**, 1–9 (2013).

21. S. A. Babayan, R. J. Orton, D. G. Streicker, Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science* **362** (6414), 577–580 (2018).

22. H. Deka, S. Chakraborty, Compositional constraint is the key force in shaping codon usage bias in hemagglutinin Gene in H1N1 subtype of influenza a Virus. *International journal of genomics* **2014** (1), 349139 (2014).

23. H. Gu, R. L. Fan, D. Wang, L. L. Poon, Dinucleotide evolutionary dynamics in influenza A virus. *Virus evolution* **5** (2), vez038 (2019).

24. J. Sun, *et al.*, Analysis of the codon usage pattern of HA and NA genes of H7N9 influenza A virus. *International Journal of Molecular Sciences* **21** (19), 7129 (2020).

25. J. Li, *et al.*, Machine learning methods for predicting human-adaptive influenza a viruses based on viral nucleotide compositions. *Molecular biology and evolution* **37** (4), 1224–1236 (2020).

26. P. M. Sharp, W.-H. Li, An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of molecular evolution* **24**, 28–38 (1986).

27. I. Mena, *et al.*, Origins of the 2009 H1N1 influenza pandemic in swine in Mexico. *elife* **5**, e16777 (2016).

28. M. P. Girard, J. S. Tam, O. M. Assossou, M. P. Kieny, The 2009 A (H1N1) influenza virus pandemic: A review. *Vaccine* **28** (31), 4895–4902 (2010).

29. T. P. Peacock, *et al.*, The global H5N1 influenza panzootic in mammals. *Nature* **637** (8045), 304–313 (2025).

30. A. Fusaro, *et al.*, High pathogenic avian influenza A (H5) viruses of clade 2.3.4.4b in Europe—Why trends of virus evolution are more difficult to predict. *Virus evolution* **10** (1), veae027 (2024).

31. Animal and Plant Health Inspection Service - U.S. Department of Agriculture, APHIS Confirms D1.1 Genotype in Dairy Cattle in Nevada, `https://www.aphis.usda.gov/news/program-update/aphis-confirms-d11-genotype-dairy-cattle-nevada-0`, last accessed: Aug 24th, 2025.

32. L. Lu, S. J. Lycett, A. J. Leigh Brown, Reassortment patterns of avian influenza virus internal segments among different subtypes. *BMC evolutionary biology* **14** (1), 16 (2014).

33. D. He, *et al.*, Genome-wide reassortment analysis of influenza a H7N9 viruses circulating in China during 2013–2019. *Viruses* **14** (6), 1256 (2022).

34. M. Chiara, D. S. Horner, E. Ferrandi, C. Gissi, G. Pesole, HaploCoV: unsupervised classification and rapid detection of novel emerging variants of SARS-CoV-2. *Communications Biology* **6**, 443 (2023).

35. K. Miyakawa, *et al.*, Emergence of Antigenic Variants in Bovine H5N1 Influenza Viruses. *Journal of Medical Virology* **97** (5), e70394 (2025).

36. A. Mostafa, A. Nogales, L. Martinez-Sobrido, Highly pathogenic avian influenza H5N1 in the United States: recent incursions and spillover to cattle. *npj Viruses* **3** (1), 54 (2025).

37. A. Bernasconi, L. Mari, R. Casagrandi, S. Ceri, Data-driven analysis of amino acid change dynamics timely reveals SARS-CoV-2 variant emergence. *Scientific Reports* **11**, 21068 (2021).

38. P. Pinoli, *et al.*, VariantHunter: a method and tool for fast detection of emerging SARS-CoV-2 variants. *Database* **2023**, baad044 (2023).

39. T. Alfonsi, A. Bernasconi, M. Chiara, S. Ceri, Data-driven recombination detection in viral genomes. *Nature Communications* **15** (1), 3313 (2024).

40. L. Wilkinson, Visualizing big data outliers through distributed aggregation. *IEEE transactions on visualization and computer graphics* **24** (1), 256–266 (2017).

41. I. Weissman, Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association* **73** (364), 812–815 (1978).

42. K. T. Schwarz, *Wind dispersion of carbon dioxide leaking from underground sequestration, and outlier detection in eddy covariance data using extreme value theory* (University of California, Berkeley) (2008).

43. SKTime, STRAY, `https://www.sktime.net/en/v0.31.0/api_reference/auto_generated/sktime.annotation.stray.STRAY.html`, last accessed: Aug 24th, 2025.

Acknowledgments

We gratefully acknowledge all data contributors, i.e. the Authors and their Originating Laboratories responsible for obtaining the specimens, and their Submitting Laboratories that generated the genetic sequence and metadata and shared via the GISAID Initiative the data on which part of this research is based. The authors are grateful to Ilaria Capua (Johns Hopkins University), the group of Alice Fusaro and Isabella Monne (Istituto Zooprofilattico Sperimentale delle Venezie), and Manuela Sironi (IRCCS Eugenio Medea), for the fruitful discussions inspiring this research.

Funding: This work was supported by Ministero dell'Università e della Ricerca (PRIN PNRR 2022 "SENSIBLE" project, n. P2022CNN2J), funded by the European Union, Next Generation EU, within PNRR M4.C2.1.1. Politecnico di Milano, CUP D53D23017400001; Università degli Studi di Milano, CUP G53D23006690001. AB Principal Investigator, MC co-Principal Investigator.

Author contributions: TA and AB are co-primary authors; TA performed data collection and processing; TA, AB, and SC adapted `stray` for the purposes of this paper and performed the specificity and sensitivity analysis; TA and AB designed the data mart. All authors discussed the bi-weekly setting and the presentation of H1N1 and H5N1 cases. MC performed the bioinformatic analyses of H1N1 and H5N1, in particular leading to the identification of relevant mutations for H5N1 and devised the statistical test for the identification of *Super Warning*s. All authors contributed to the writing; AB is the contact author; SC coordinated the research.

Competing interests: All other authors declare they have no competing interests.

Data and materials availability: Original sequences and metadata are publicly accessible through the GISAID platform (H1N1: `https://doi.org/10.55876/gis8.250523as` and H5N1: `https://doi.org/10.55876/gis8.250824cw`). The list of sample accessions and FluWarning code are provided on our Zenodo repository (*15*).

Supplementary materials

Materials and Methods

Figs. S1 to S9

Table S1

Data S1