Computational and Structural Biotechnology Journal 27 (2025) 2757–2771

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research Article

A codon usage-based approach for the stratification of Influenza A across recent spillovers

Tommaso Alfonsi [a], Matteo Chiara [b], Anna Bernasconi [a,*]

[a] Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy
[b] Department of Biosciences, Università degli Studi di Milano, Milan, Italy

ARTICLE INFO

Dataset link: gisaid.org/EPI_SET_250319su, gisaid.org/EPI_SET_250319ms, gisaid.org/EPI_SET_250616yr, gisaid.org/EPI_SET_250616ms, gisaid.org/EPI_SET_250616nx, https://doi.org/10.5281/zenodo.14561947

Keywords:
Influenza A virus
Codon usage
Genomic surveillance
Viral evolution
Epidemiological events

ABSTRACT

Influenza A virus (IAV) is a highly adaptable pathogen that poses a significant threat to human health. Genomic surveillance of IAVs is complex due to their broad host range, zoonotic potential, and rapid evolution. Strategies based on codon preference analysis have been successfully employed for the discrimination of IAVs with different host specificity in the past. Hence, monitoring changes in codon usage offers a promising strategy for tracking IAVs' host range and identifying significant epidemiological events.

In this study, we developed a computational workflow for the stratification of IAVs based on codon usage profiles by analysing recent IAV-associated epidemiological emergencies: 1) the 2009 H1N1 pandemic in North America, 2) the H7N9 epidemic in China (2013–2017), and 3) the long-term circulation of H5N1 in domestic birds and its subsequent spillover to dairy cows. We explore the application of codon usage metrics for capturing patterns of viral diversification and expand previous related findings in the field. Our results uncovered important differences in genomic features, which are not always reflected in the clade-based nomenclature. Interestingly, a reduced set of amino acids and associated codons was sufficient to summarize salient patterns of IAV genomes across the 3 paradigmatic cases herein considered, suggesting shared evolutionary signatures across IAV serotypes.

Codon usage-based stratification effectively highlighted key epidemiological events and enabled detailed comparisons of genomic features across IAV serotypes. The approach developed in this work provides a scalable framework for IAV genomic surveillance, offering insights into viral evolution and shared patterns of codon usage preferences. Its general applicability makes it suitable for extending to other Influenza A serotypes, particularly those for which available genomic data are limited or a reference nomenclature is not established.

1. Introduction

Influenza A virus (IAV), the etiological agent of avian influenza, is a highly versatile pathogen that can infect a broad range of hosts and pose a significant risk to human health. The two distinct modes of IAV evolution are fundamental to its success. Inter-segment genome reassortment is used to conquer new host niches and secure a broader host range by spillover and antigenic shift. Adaptation to a specific host, instead, is achieved through the gradual accumulation and fixation of mutations in the genome, particularly in the genes encoding surface proteins, leading to minor changes in antigenicity (antigenic drift). Once an endemic circulation is established in a new host, antigenic drift becomes the main evolutionary driver [1]. The tropism for a diversified host range challenges IAV with the need to adapt to different "host environments" constantly.

Adjusting the viral codon usage to mimic that of the host is considered a key mechanism of viral adaptation. It is postulated that when the codon preferences of the virus align with those of its host, viral proteins are translated more efficiently, leading to efficient viral replication [2]. Thus, monitoring shifts in codon usage may represent a promising strategy for the development of computational methods for the genomic surveillance of IAVs and the identification of recent spillover and evolutionary changes, even when data are incomplete and genome sequences are partial. Although influenza A exhibits only moderate biases in codon usage [3–6] strategies based on the analyses of codon usage patterns have been successfully used to recapitulate host range and host specificity in IAVs in the past [7]. IAVs from different hosts displayed a different composition of the genome in terms of mono-, di-, tri-, and tetra-nucleotide [8] or only dinucleotide [9] and slight but systematic changes in codon usage preferences [7,10–13]. Hence, while spillovers

* Corresponding author.
E-mail addresses: tommaso.alfonsi@polimi.it (T. Alfonsi), matteo.chiara@unimi.it (M. Chiara), anna.bernasconi@polimi.it (A. Bernasconi).

https://doi.org/10.1016/j.csbj.2025.06.030
Received 7 April 2025; Received in revised form 16 June 2025; Accepted 16 June 2025
Available online 25 June 2025

are typically discussed in the context of reassortment, it is possible -at least in principle- to identify changes in the host range by the analysis of even a single segment or a few segments of the genome. This work is placed in the context of a larger effort on developing cost-effective genomic surveillance methodologies that are increasingly efficient and lightweight. As the HA segment of the genome is sequenced for nearly all IAV isolates and is sufficiently diagnostic to identify relevant changes, strategies based on the analysis of the HA segment might represent an optimum for empowering IAV genome surveillance.

To evaluate the potential of using codon usage patterns in the HA segment for implementing computational methods in the genomic surveillance of influenza A here we developed a novel approach to capture and stratify the salient features of codon usage and summarize the results in a human-interpretable form. Compared to previous studies such as [13] which employed sophistcated ML models to learn and predict IAV host specified by learning the characteristic features of 6 distinct genome segments and for many different serotypes, the aims of our research are focused in the development of a lightweight and effective method (i.e., only focusing on the most commonly sequenced segment HA), with the potential of detecting ongoing/future spillovers by computing codon preference metrics [11,12].

Three paradigmatic use cases associated with recent epidemiological emergencies were considered to demonstrate the applicability of the proposed approach: (a) the "(H1N1) 2009 pandemic" in North America [14]; (b) the H7N9 "Asian" epidemic of avian origin between 2013 and 2017 in China [15]; and (c) the H5N1 influenza virus spreading in domestic bird species since the year 2000 [16] and its subsequent spillover to dairy cows [17].

Publicly available genomic sequences were retrieved from the EpiFlu database of GISAID, the Global Initiative on Sharing All Influenza Data [18] and genome sequences were stratified into clusters based on the Relative Synonymous Codon Usage (RSCU) of the HA segment. The clusters delineated by our approach captured key epidemiological events and emerging properties of viral strains circulating throughout distinct epidemic/pandemic seasons and viral genome sequences in a more granular and informative way compared to the standing clade-based nomenclature. More importantly, our results outlined key differences in codon preferences and patterns of codon usage across distinct IAV serotypes, which could be used to inform genomic surveillance systems, enabling them to identify and track changes in viral genomes. In conclusion, our results provide solid proof of principle of the applicability of codon usage-based metrics for the genomic surveillance of emerging IAV strains, and recovered consistent results across a diverse set of IAV use cases, including the recent H5N1 spillover in dairy cows [17].

2. Results

We developed a data-driven method to partition influenza A genome sequences based on Relative Synonymous Codon Usage (RSCU), without *a priori* knowledge of epidemiological data and phylogenetic inference. Our choices were driven by the need for a lightweight, easy-to-use, easy-to-interpret method. The conceptual workflow of the method is illustrated in Fig. 1. Briefly: (i) Sequence data and metadata are retrieved and prepared; (ii) RSCU values are computed for all the sequences included in the dataset; (iii) Principal component analysis (PCA) is used to identify and rank the most important components that summarize patterns of codon usage; (iv) Agglomerative clustering and k-means in the space of Principal Components (PCs) are used to determine the most suitable number of clusters – the optimal number is established by evaluating different clustering solutions through the Silhouette and Calinski-Harabasz scores; (v) Once clusters are established, a linear classifier is trained to assign cluster labels based on the RSCUs and determine the relative importance (weight) of every codon for the correct assignment of a specific cluster label. Codon weights are scaled by the standard deviation of their RSCU, to reward codons with the highest inter-cluster variability

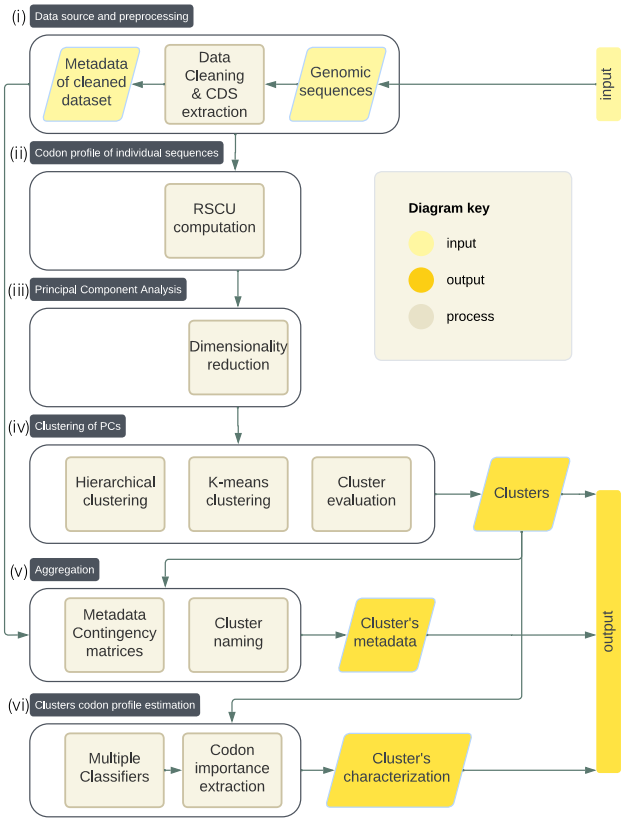

Fig. 1. Schematic overview of methods. We (i) preprocess genomic data and metadata; (ii) compute RSCU values for all the sequences; (iii) run PCA analysis; (iv) cluster along the Principal Components and select the most appropriate number of clusters; (v) assign interpretable labels to clusters; and (vi) characterize clusters by their most important codons.

and calculate an "impact score"; (vi) Codons with the highest score are considered the most impactful codons and key features of the most important codons are inspected to infer common and distinctive patterns of codon usage across clusters. The code used for our analyses is publicly available on Zenodo [19]. In the following sections, we describe the application of our method to three selected use cases:

- H1N1 collected in North America between October 2006 and September 2013 (covering 7 influenza seasons centered around the 2009-2010 "(H1N1) 2009 pandemic");
- H7N9 collected between October 2010 and September 2020 (10 influenza seasons centered around the 2013-2017 "Asian" flu epidemic in China);
- H5N1 collected from: (1) domestic birds from 2000 until May 2024; (2) wild birds from 2000 until May 2024; and (3) complete genome sequences collected in North America from September 2023 to March 2025—see Supplementary File for cases (2) and (3).

Each use case is discussed according to the following rationale: first, we show the results of the PCA of RSCU values (Figs. 2, 5, and 8); then, we discuss the choice of an appropriate number of clusters (Figs. 3, 6, and 9). Results are summarized in two distinct tables: the first maps clusters numerosity onto relevant metadata, e.g., clade, host type, location, flu season, pathogenicity (Tables 1, 3, and 5); the second proposes concise cluster names based on preponderant metadata characteristics (Tables 2, 4, and 6). Finally, we show the temporal evolution of clusters using their first two principal components (Figs. 4, 7, and 10).

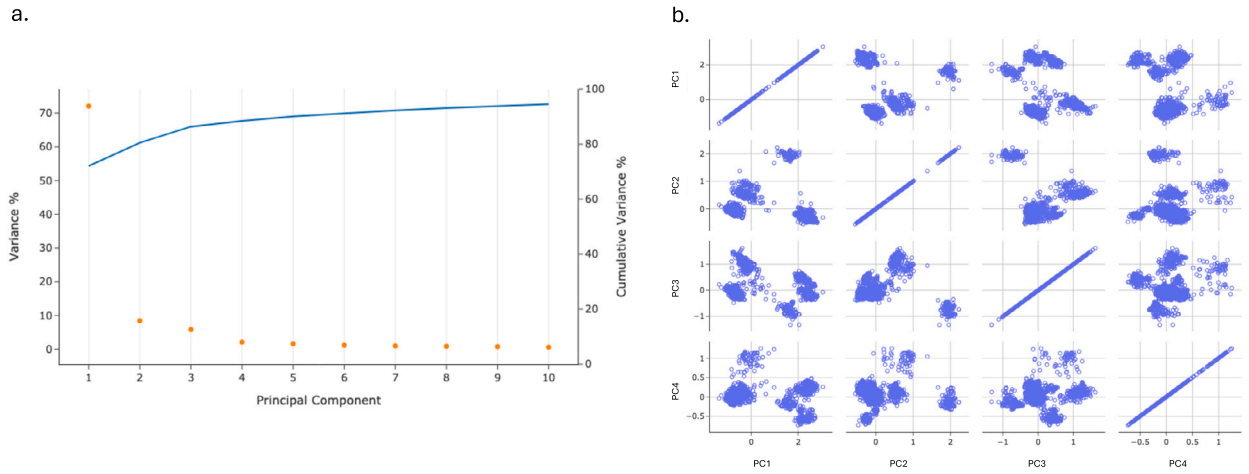

Fig. 2. Principal Components Analysis (H1N1). a) Elbow plot of the PCs. In blue is the cumulative explained variance, and in red is the variance explained by every single component. b) Scatter matrix of the dataset projected on the first 4 PCs.

2.1. The HA segment of H1N1 during the (H1N1) 2009 pandemic

The 2009 H1N1 pandemic was the third recent flu pandemic caused by the H1N1 virus, following the 1918–1920 Spanish flu pandemic and the 1977 Russian flu [14]. The evolutionary origin of the pandemic virus (pdm09) can be traced back to the reassortment in swine in Mexico [20] of at least 3 distinct IAVs circulating in North America, Asia, and Europe. Segments PB2, PB1, PA, NP, and NS are derived from a "triple reassortant" H3N2 swine virus that originated in North American swine during the mid-1990s. The HA (H1) segment originated from the "classical swine" H1N1 lineage that has been circulating in North American swine since the 1918 H1N1 pandemic. The NA (N1) and MP segments were related to those of an avian-like Eurasian swine lineage (EAsw) that emerged in European pigs in the late 1970s [21].

We considered sequences collected between October 2006 and September 2012 in North America and available through the GISAID database [18], to cover the pre-pandemic phase, the initial burst, and the post-pandemic period. The dataset includes 5782 samples, isolated from swine (638), humans (4933), and wild birds (211); 619 sequences were not associated with a defined characterization in the reference nomenclature, while the remaining sequences were assigned to the 6B.1A.6 (838) and the 6B.1 (4325) clades. Exploratory analyses of viral clades and associated hosts recapitulated patterns of viral circulation described before and during the 2009 H1N1 pandemic. Before 2009, the majority of human infections were caused by clade 6B.1A.6, while -in swine- clade 6B.1 was predominant. From 2009 onward, clade 6B.1 replaced 6B.1A.6 and became dominant also in humans. Note that, although the reference nomenclature would place 6B.1 as the ancestor of 6B.1A.6, viral circulation patterns are not compatible with this hypothesis (6B.1A.6 predates 6B.1 according to available data) and a recent study by Ding et al. [22] suggested inconsistencies between the phylogeny of the HA segment of H1N1 IAVs and the clade-based nomenclature.

2.1.1. H1N1: delineation of the optimal number of partitions

As illustrated by Fig. 2a), the first 4 most important PCs in the PCA analysis capture 88% of the variability in RSCU scores profiles in this dataset. Fig. 2b) shows a 2D scatter matrix of the first 4 PCs. Visual inspection of the scatterplots suggests the delineation of 3 (PC2-PC3) to 5 (PC1-PC3 and PC1-PC4) distinct groups.

To delineate the ideal number of clusters, we applied the workflow illustrated in Fig. 3. Panel a) shows a dendrogram of the dataset computed through hierarchical agglomerative clustering based on the first 4 more informative PCs. Following our previous observation, 3 to 5 groups could be defined by cutting the dendrogram at different heights: 40 (3 clusters); 22 (4 clusters); and 15 (5 clusters, see yellow line).

A 10-fold k-means clustering with a k number of clusters ranging between 2 and 9 was also performed; solutions were ranked based on the average score of the Calinski-Harabasz method [23] (Fig. 3b) and the Silhouette method [24] (Fig. 3c). While the Silhouette score displayed limited differences between solutions with 3 to 5 clusters, the Calinski-Harabasz metric peaked at 5. Based on this observation, we set the ideal number of clusters to 5. The data points corresponding to the 5 clusters are shown in the transformed space of the PC1 and PC3 in Fig. 3d) using different colors.

2.1.2. Characteristic features of H1N1 clusters

Metadata features of the sequences assigned to the 5 clusters were used to reconstruct the salient properties of each partition. The following metadata were considered: 1) the viral *clade* defined by the reference nomenclature; 2) the *host* from which the specimen was isolated; and 3) the *flu* epidemic *season* (defined as an interval of time spanning 12 months from October to September of the following year). The results are reported in Table 1, showing the breakdown of the total number of sequences in every cluster by each class of metadata. Any metadata annotation labeling at least 65% sequences in a cluster was considered "prevalent" and used to derive a label for that cluster (unless prevalent in all clusters, see last column "Extracted feature").

Labels were used to derive mnemonic names; specifically, we identify clusters with the pattern `(<cluster id>) (<MAJOR HOST, minor host(s)>) (<flu-season-interval>)`. The analysis of the contingency matrix of metadata led to the names listed in Table 2. When arranged chronologically, the five identified clusters correspond to:

- a group of clade 6B.1 H1N1 viral isolates circulating mostly in swine before, during, and after the 2009 pandemic (**cluster-3**);
- H1N1 IAVs associated with birds (**cluster-2**);
- clade 6B.1A.6 viruses with sustained circulation in humans until the 2006-2007 flu season (**cluster-4**);
- clade 6B.1A.6 IAVs infecting human – with sporadic transmission to swine (**cluster-1**) – before the 2009 pandemic; and
- IAV A(H1N1)pdm09 that caused the outbreak of "swine flu" in humans starting from March 2009 (**cluster-0**) – these isolates are assigned to clade 6B.1 according to the reference nomenclature.

These patterns are illustrated in Fig. 4, where we plot the clusters and their numerosity throughout distinct flu epidemic seasons (from 2006 to 2012), and highlight their differences in terms of codon profile on the transformed space in PC1 and PC3.

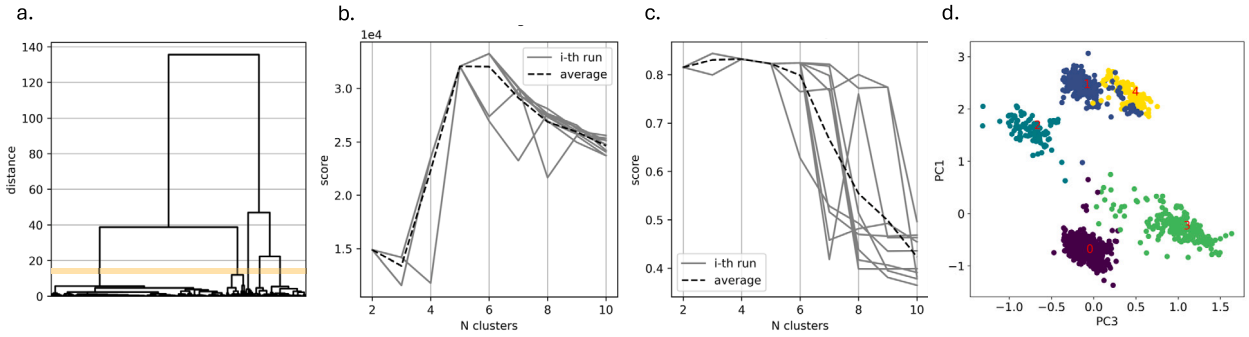

Fig. 3. Clustering (H1N1). a) Dendrogram of the agglomerative clustering algorithm on the H1N1 dataset; b) Calinski-Harabasz scores of the 10-fold k-means clustering algorithm; c) Silhouette scores of the 10-fold k-means clustering algorithm; d) Clusters of sequences highlighted using different colors and plotted on the PC1 and PC3.

Table 1
Partitioning of the H1N1 sequences based on the metadata properties in the 5 clusters. Cluster IDs are reported in the leftmost column. The inner cells report the number of sequences for each metadata value and cluster. The rightmost column contains, if possible, the name of a cluster-describing feature based on the observed occurrences.

Cluster \ Clade	6B.1	6B.1A.6	NA	**Extracted feature**
(cluster-0)	3975	0	18	6B.1
(cluster-1)	2	541	261	6B.1A.6
(cluster-2)	0	0	207	-
(cluster-3)	348	0	56	6B.1
(cluster-4)	0	297	77	6B.1A.6

Cluster \ Host Type	human	swine	wild bird	**Extracted feature**
(cluster-0)	3750	239	4	HUMAN,swine
(cluster-1)	794	10	0	HUMAN,swine
(cluster-2)	0	0	207	wild bird
(cluster-3)	19	385	0	human,SWINE
(cluster-4)	370	4	0	HUMAN

Cluster \ Flu Season	2006-2007	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012	**Extracted feature**
(cluster-0)	1	1	1888	1209	432	462	≥08-09
(cluster-1)	34	258	511	1	0	0	≤08-09
(cluster-2)	27	56	32	43	18	31	-
(cluster-3)	20	29	34	37	97	187	-
(cluster-4)	368	1	2	2	1	0	=06-07

Table 2
Mnemonic cluster names of H1N1, formed according to the "Extracted features" of Table 1.

(cluster-0) 6B.1/HUMAN,swine/≥08-09	**(cluster-1)** 6B.1A.6/HUMAN,swine/≤08-09
(cluster-2) wild bird	**(cluster-3)** 6B.1/human,SWINE **(cluster-4)** 6B.1A.6/HUMAN/=06-07

2.1.3. H1N1: key findings

The (H1N1) 2009 pandemic virus is labeled with clade 6B.1 in our dataset. As reported in Table 1 and Fig. 4, an increasing number of infections caused by 6B.1 was observed from 2006 to 2012 in swine. Subsequently, clade 6B.1 has been observed in human hosts starting from the 2008-2009 epidemic flu season, which corresponds with the start of the 2009 H1N1 pandemic. As illustrated in Fig. 4, by the epidemic season 2009-2010, clade 6B.1 completely replaced 6B.1A.6 and became the most prevalent H1N1 clade infecting humans.

Interestingly, our clustering of H1N1 codon usage profiles identifies two distinct clusters within clade 6B.1, i.e., cluster-0 (6B.1/HUMAN,swine/≥08-09) and cluster-3 (6B.1/human,SWINE). Cluster-0 is associated (almost) exclusively with specimens collected from human hosts starting from the 2008-2009 flu season onward and ideally corresponds with A(H1N1)pdm09. Isolates assigned to cluster-3 were collected starting from the 2006-2007 flu season and are prevalently isolated from swine. This separation is not reflected by the reference nomenclature and might indicate the co-circulation of distinct "classical swine" H1N1 subclades that were not captured by the reference nomenclature. Our analysis also revealed two distinct partitions – cluster-4 (6B.1A.6/HUMAN/=06-07) and cluster-1 (6B.1A.6/HUMAN,swine/≤08-09)– within the 6B.1A.6 clade. Both clusters are prevalently associated with the human host and circulated during consecutive epidemic seasons: 2006-2007 (cluster-4) and 2007-2008 (cluster-1). Cluster-2 (wild bird) is composed exclusively of viral specimens isolated from wild birds and shows large differences in codon preference with respect to all the other clusters.

2.2. The HA segment of H7N9 during the 2013-2017 outbreak

Starting from March 2013, recurrent transmission to humans of a novel zoonotic avian influenza A(H7N9) virus was reported by Chinese authorities. As of January 2018, 1566 cases were documented, including 569 deaths. Since the majority of the cases were isolated and no sustained person-to-person transmission was observed, sporadic zoonotic transmission to humans from poultry was considered the most likely explanation for the outbreak [25]. Recurrent spikes of infections were

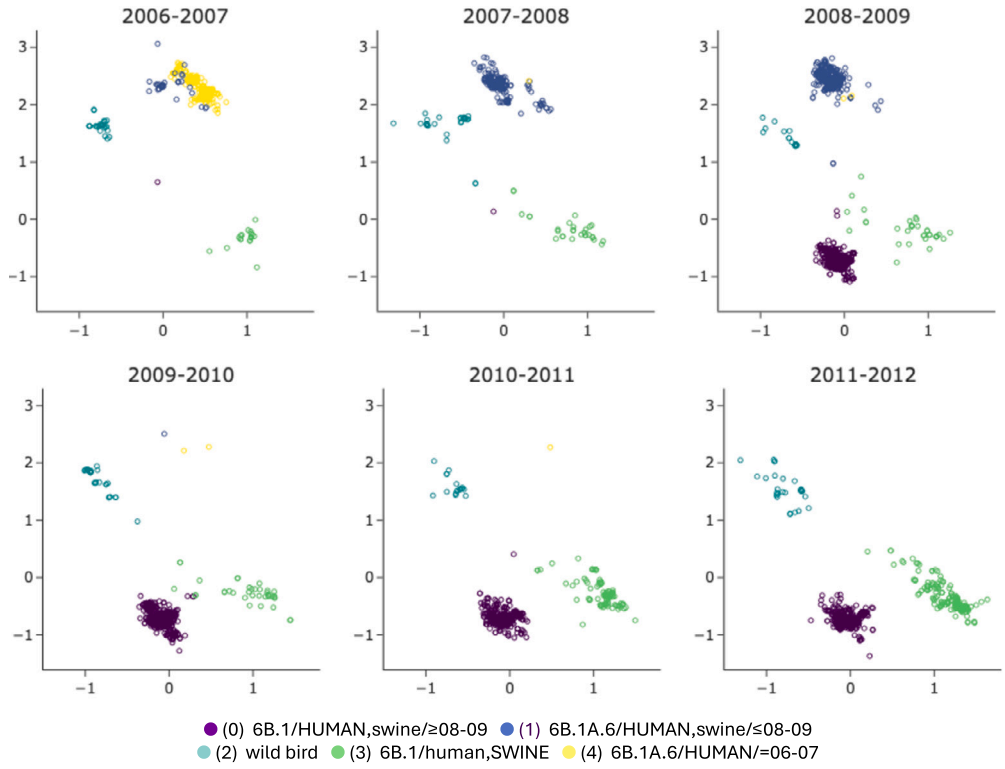

Fig. 4. Codon profiles of clusters (H1N1). The codon profile of the H1N1 sequences is projected on the PC1 (y-axis) and PC3 (x-axis). The scatter plot is organized by flu season. Each sequence is colored according to the assigned cluster.

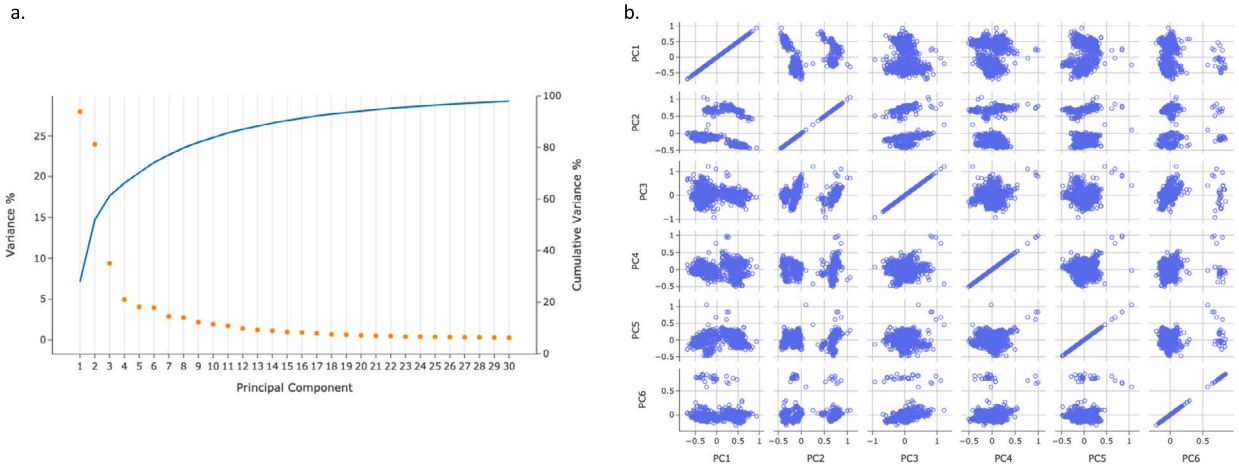

Fig. 5. Principal Components Analysis (H7N9). a) Elbow plot of the PCs. In blue, the cumulative explained variance, in red the variance explained by every single component. b) Scatter matrix of the dataset projected on the first 6 PCs.

observed until 2018 when the emergency was mitigated thanks to a mass vaccination campaign [26]. A Highly Pathogenic (HP) [27] H7N9 strain emerged in the 5th and last wave of the epidemic, whereas all the strains associated with the 1st to 4th waves were classified as lowly pathogenic (LP). Note that, unfortunately, only a very limited number of sequences were collected and made publicly available outside of the temporal span of the 2013-2017 epidemic for H7N9.

We retrieved a total of 1879 sequences of the HA segment of H7N9, obtained from October 2010 to 2019, both from human and avian samples in Asia. Due to the lack of clade-based nomenclature for H7N9, these sequences were not stratified into clades. A total of 1728 sequences were annotated as LP, 150 as HP, and 4 were not labeled. Sequences were sampled from 12 different host species, but the large majority was isolated either from chickens (41%) or humans (55%).

2.2.1. *H7N9: delineation of the optimal number of partitions*

RSCU scores were computed and the first 6 principal components of the PCA were identified as the most informative (Fig. 5a). The selected PCs are plotted in the scatter matrix of Fig. 5b), showing three separable clusters on PC1-PC2.

Consistent with this observation, the hierarchical clustering based on RSCUs, as illustrated in Fig. 6a), shows three stable clusters. Conversely, the Calinski-Harabasz (Fig. 6b) and Silhouette (Fig. 6c) scores do not provide an obvious optimal solution. For our analysis, we selected 3 clusters, as indicated by the global maximum of the Calinski-Harabasz score (for all the runs) and by the average Silhouette score. The scatter plot of the dataset is illustrated in Fig. 6d), with colors denoting three different clusters.

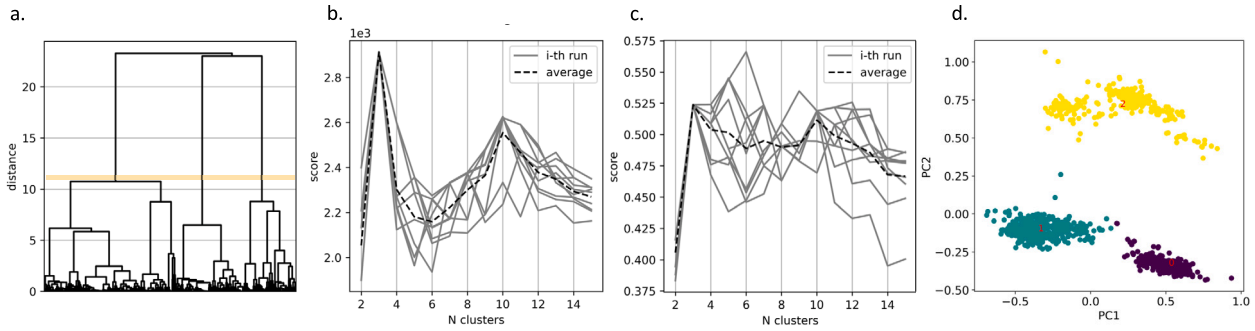

Fig. 6. Clustering (H7N9). a) Dendrogram of the agglomerative clustering algorithm on the H7N9 dataset; b) Calinski-Harabasz scores of the 10-fold k-means clustering algorithm; c) Silhouette scores of the 10-fold k-means clustering algorithm; d) Clusters of sequences highlighted using different colors and plotted on the PC1 and PC2.

Table 3
Partitioning of the H7N9 sequences based on the metadata properties in the 3 clusters. Cluster IDs are reported in the leftmost column. The inner cells report the number of sequences for each metadata value and cluster. The rightmost column contains, if possible, the name of a cluster-describing feature based on the observed occurrences.

Cluster \ Flu Season	2010-2011	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	**Extracted feature**
(cluster-0)	0	0	0	0	39	465	1	0	16-17
(cluster-1)	1	228	539	177	10	22	0	0	12-15
(cluster-2)	1	6	90	24	13	202	20	44	13-19

Cluster \ Host Type	domestic bird	human	wild bird	**Extracted feature**
(cluster-0)	102	385	18	-
(cluster-1)	493	451	33	-
(cluster-2)	181	194	25	-

Cluster \ Pathogenicity	HP	LP	None	**Extracted feature**
(cluster-0)	0	504	1	LP
(cluster-1)	0	976	1	LP
(cluster-2)	150	248	2	HP/LP

Table 4
Mnemonic cluster names of H7N9 according to the "Extracted features" of Table 3.

(cluster-0) 16-17/LP	**(cluster-1)** 12-15/LP	**(cluster-2)** 13-19/HP,LP

2.2.2. *Characteristic features of H7N9 clusters*

Clusters' metadata (*pathogenicity*, *host type*, and *flu season*) is reported in Table 3; clades were not used due to the lack of a reference nomenclature. Cluster-0 comprises LP sequences collected between 2015-2017; cluster-1, LP sequences from 2012-2015, and a limited number of sequences from 2015-2017. Finally, cluster-2 includes both HP and LP sequences, sparsely sampled between September 2013 and October 2019, with the highest number of sequences associated with the 2016-2017 flu season. Consistent with the hypothesis of sporadic zoonotic transmission, we observe that human and poultry hosts are equally represented across all the clusters in this dataset. The metadata partitions of the 3 clusters suggest the names in Table 4.

The temporal span covered by sequences assigned to each cluster is shown in Fig. 7, where cluster-1 is the first detected (flu seasons 12-15), followed cluster-0 (15-16 and 16-17). Cluster-2 spans from range from 2013 to 2019.

2.2.3. *H7N9: key findings*

Although we notice that all of the HP isolates are consistently assigned to cluster-2 (see mnemonic names in Table 4), the 3 clusters identified by our method do not suggest a neat association with a specific host or epidemiological feature (i.e., pathogenicity) in this case. This observation might be in part explained by 1) the coarse-grained/limited sampling; 2) the lack of a reference nomenclature that could aid in the classification of the H7N9 isolates; and -most importantly- 3) the sporadic pattern of transmission from poultry to humans of different viral strains during the 2013-2017 epidemic. The emergence of highly pathogenic IAVs is often associated with the evolution of a furin cleavage site at the interface between the HA1 and HA2 subunits [28]. Interestingly, we notice that 88% (132/150) of the sequences labeled as HP in this dataset carry a non-synonymous A to G single nucleotide substitution at position 1012 in the HA and an insertion of 12 bp (AAACGGACTGCA) at position 1014. The resulting amino acid sequence (PKRKRTARG, see Supplementary File, Fig. S1, panel a)) [29] harbors a polybasic furin cleavage motif. Due to the length of the HA protein (570 amino acid residues), single changes in amino acids and/or small insertions are not likely to shift codon usage significantly. This observation might explain why our solution based on 3 clusters did not partition all of the HP isolates into a single group. This result represents a potential limitation of our automated approach. However, following a careful examination of alternative clustering solutions (see the Silhouette score, panel c), we observed that 10 clusters might also be an adequate cluster choice for this dataset, as illustrated in Fig. 6. By manually setting the number of clusters to 10, the method effectively grouped all the HP sequences into a single cluster spreading from 2017 to 2019 – we invite the interested reader to inspect the results of our analysis repeated using this partition on the H7N9 dataset, provided in the Supplementary File, Table S1.

2.3. *The HA segment of highly pathogenic H5N1 IAVs*

Due to the high lethality, virulence, widespread occurrence, and diverse host range, highly pathogenic (HP) H5N1 strains are considered a constant pandemic threat by the WHO. The common ancestor of modern

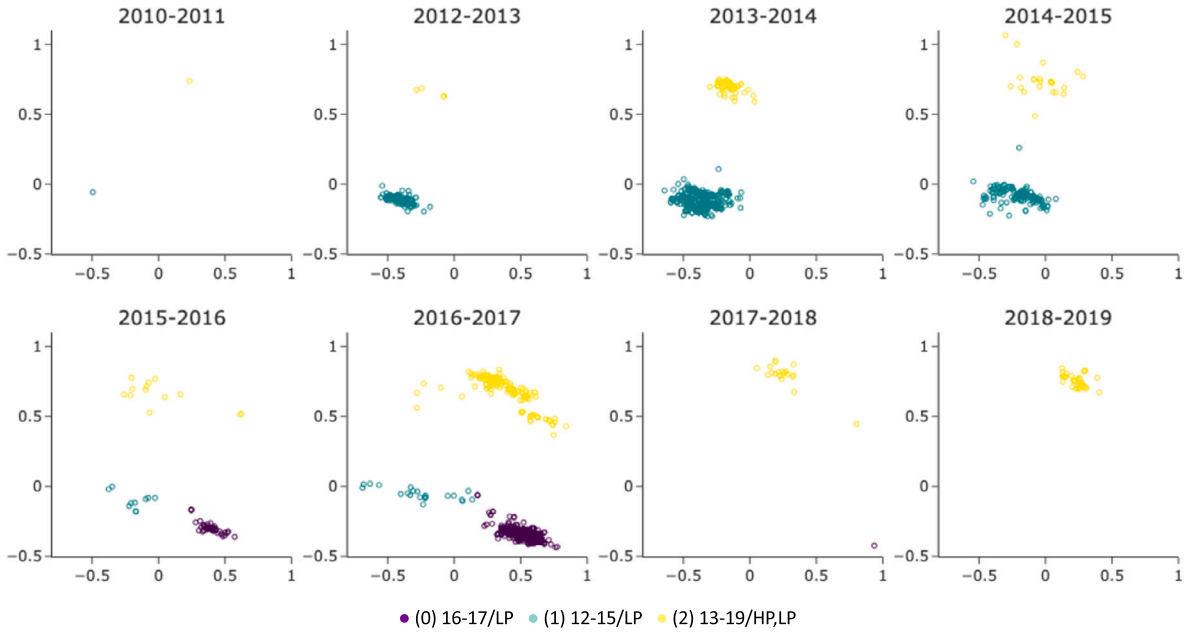

Fig. 7. Codon profiles of clusters (H7N9). The codon profile of the H7N9 sequences is projected on the PC1 (x-axis) and PC2 (y-axis). The scatter plot is organized by flu season. Each sequence is colored according to the assigned cluster.

HP A(H5N1) IAV was first identified in domestic waterfowl in southern China in 1996. Over time, this virus differentiated into multiple clades and subclades. Avian A/H5N1 is now widespread in wild bird populations worldwide, with numerous outbreaks both among domestic and wild birds and mammals. Sporadic cases of human transmission have also been reported, often with high mortality rates (>50%) [30–32,16]. In late 2024 clade 2.3.4.4b -a new strain of HP H5N1- caused an epidemic of avian flu in dairy cows in the USA with multiple spillovers to other farmed animals, raising significant concern by health authorities A single reassortant genotype of 2.3.4.4b, labeled B3.13, was linked with the outbreak [33,17].

To verify whether our approach could flag the emergence of clade 2.3.4.4b in domestic birds (1) and wild birds (2), and its subsequent spillover to dairy cows (3), we analyzed three distinct datasets. For (1), we considered the complete collection of HA segment sequences of H5N1 isolated from 2000 to May 2024 from domestic birds. The dataset (*d1*) is composed of 3475 sequences and includes all the descendant clades (0 to 9) of the Highly Pathogenic Avian Influenza virus labeled Goose Guangdong (GsGd) and associated subclades, as well as eight Eurasian non-Goose Guandong Influenza viruses (EA_nonGsGd) [34,35]. For (2), we performed an equivalent analysis on H5N1 viruses isolated from wild birds in a comparable time period (2000 to May 2024). The dataset (*d2*) includes 12021 sequences of the HA segment. Results are discussed in the Supplementary File, Figs. S2-S4 and Tables S2-S4. Finally, for (3) we analysed complete genomes collected in North America between Sept 2023 and May 2025. This dataset (*d3*) includes 5243 complete sequences (with HA). The choice to analyse complete genome sequence was driven by three key considerations: i) increased data availability: complete genome sequences were available for 94% of the isolates); ii) general applicability: we aimed to explore whether our approach could be applied to other genome segments, not just HA; and iii) genotypes definition: original genotypes were defined by considering complete genome sequences in [17]. Results are shown in the Supplementary File, Figs. S5-S8 and Tables S5-S7. Parametrization and the definition of the optimal number of clusters will be discussed only for (*d1*) in the main text.

2.3.1. H5N1 (domestic birds): delineation of the optimal number of partitions

According to our analytical workflow, the first three PCs of the PCA of the RSCU scores captured the highest proportion of the variability in the data (see Fig. 8a) and were considered for subsequent analyses. These 3 PCs were considered for clustering (see Fig. 8b).

Visual inspection of the dendrogram in Fig. 9a) suggested 3 to 7 clusters. A 10-fold iteration of the k-means clustering algorithm with 2 to 15 clusters was used to derive the optimal clustering solution based on the Calinski-Harabasz and Silhouette scores shown in Figs. 9b) and 9c). Both metrics concurred in the identification of 3 well-defined clusters, which are also illustrated on the PC2 (x-axis) and PC1 (y-axis) of Fig. 9d).

2.3.2. Characteristic features of H5N1 (domestic birds) clusters

Table 5 reports the breakdown of the total number of sequences in every cluster by *flu* epidemic *season*, assigned *clade*, *host type*, *continent*, and *pathogenicity* – high (HP) or low (LP). The analysis of the contingency matrix of cluster metadata derived the mnemonic names reported in Table 6 for every cluster. Cluster-0 included isolates from 2008 onward and the majority of the sequences were either from clades 2.3.2.1a (166) or 2.3.2.1c (622). Cluster-1 featured sequences collected starting from September 2020 and corresponded precisely with clade 2.3.4.4b. Sequences assigned to cluster-2 were all collected before 2016 and were labeled under all the clades (0 to 9) in the H5N1 nomenclature, with the exclusion of clade 2.3.2.1 (and its descendants) and clade 2.3.4.4b. The temporal span of the clusters is shown in Fig. 10.

2.3.3. H5N1: key findings

The nomenclature of H5N1 comprises 10 distinct clades (designated 0–9), along with numerous secondary subclades. Certain subclades of clade 2 (i.e., 2.2 (2005–2006); 2.3.2.1c (2009–2010 and 2014–2015); 2.3.4.4a (2014–2015); and 2.3.4.4b (2016–2017)) were previously linked with widespread intercontinental circulation and waves of avian flu.

From 2020, highly pathogenic avian influenza A(H5N1) viruses of clade 2.3.4.4b emerged and rapidly spread across Africa, Asia, Europe, and the Americas. IAVs assigned to this clade caused recurrent outbreaks in domestic birds and in several mammalian species, including sea lions and cats. More recently, an epidemic caused by a recombinant genotype of 2.3.4.4b, B3.13, was registered in dairy cattle in the USA. In dataset

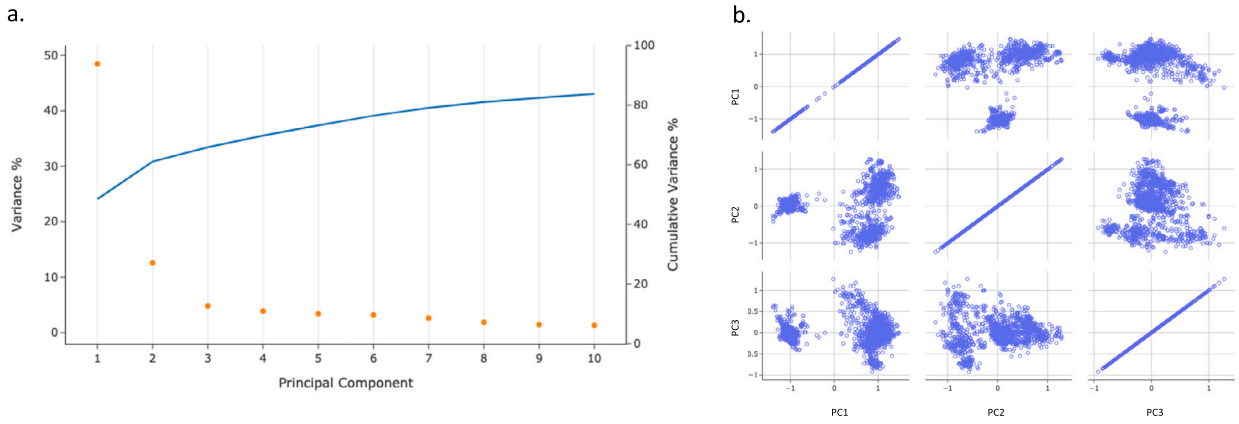

Fig. 8. Principal Components Analysis (H5N1). a) Elbow plot of the PC. In blue, the cumulative explained variance, in red the variance explained by every single component. b) Scatter matrix of the dataset projected on the first 3 PCs.

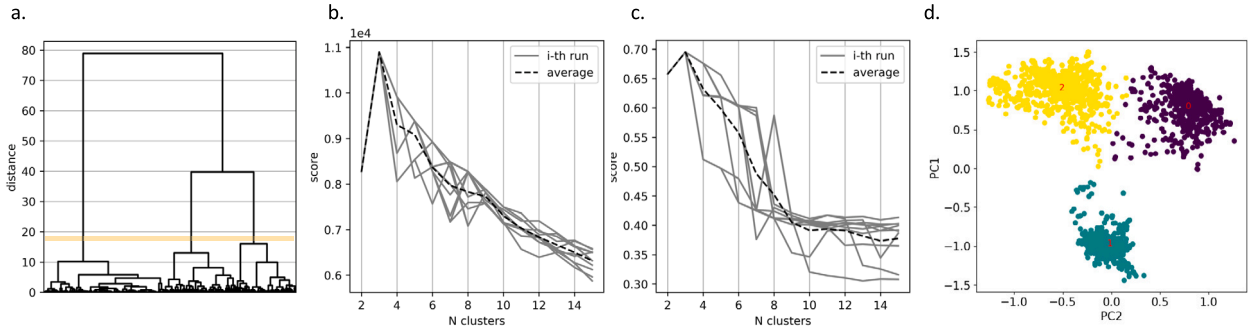

Fig. 9. Clustering (H5N1). a) Dendrogram of the agglomerative clustering algorithm on the H5N1 dataset; b) Calinski-Harabasz scores of the 10-fold k-means clustering algorithm; c) Silhouette scores of the 10-fold k-means clustering algorithm; d) Clusters of sequences highlighted using different colors and plotted on the PC1 and PC2.

Table 5
Partitioning of the H5N1 clusters on the metadata values. Cluster IDs are reported in the leftmost column. The inner cells report the number of sequences for each metadata value and cluster. In the *clade* subtable, we sorted clades according to the alpha-numeric order and grouped the ones with low representativeness into ordered macro-categories. The first contains clades from 0 to 2.3.2; then, a series of specific clades are represented (2.3.2.1 and its subclades labelled a, b, and c); another macro-category includes the subclades rooted in 2.3.4, until 2.3.4.4; the clade 2.3.4.4b is represented in a separate column; all alphabetically subsequent clades (from 2.3.4.4c to 9) are merged in one column; finally, we have EA_nonGsGd. The rightmost column contains, if possible, the name of a cluster-describing feature based on the observed occurrences (the symbol * captures the clade 2.3.2.1 and its subclades a,b,c).

Cluster \ Flu Season	2001-2008	2008-2013	2013-2014	2014-2017	2017-2020	2020-2024	**Extracted feature**
(cluster-0)	7	218	65	396	83	51	≥ 08
(cluster-1)	0	0	13	2	0	1635	≥ 20
(cluster-2)	434	420	31	116	2	2	≤ 16

Cluster \ Clade	0 to 2.3.2	2.3.2.1	2.3.2.1a	2.3.2.1b	2.3.2.1c	2.3.4 to 2.3.4.4	2.3.4.4b	2.3.4.4c to 9	EA_nonGsGd	**Extracted feature**
(cluster-0)	4	14	166	4	622	1	1	3	5	2.3.2.1*
(cluster-1)	0	0	0	0	0	13	1637	0	0	2.3.4.4b
(cluster-2)	867	0	0	0	0	74	1	60	3	≠2.3.2.1*,≠2.3.4.4b

Cluster \ Host Type	anas plat.	anser anser	cairina moschata	chicken	domestic goose	guineafowl	**Extracted feature**
(cluster-0)	5	0	53	758	0	4	-
(cluster-1)	20	23	3	1571	15	18	-
(cluster-2)	0	2	8	995	0	0	-

Cluster \ Continent	Africa	Asia	Europe	North America	South America	**Extracted feature**
(cluster-0)	218	592	9	1	0	Afr,As
(cluster-1)	181	298	883	198	90	global
(cluster-2)	439	556	10	0	0	Afr,As

Cluster \ Pathogenicity	HP	LP	**Extracted feature**
(cluster-0)	814	6	-
(cluster-1)	1650	0	-
(cluster-2)	995	10	-

Table 6
Mnemonic cluster names of H5N1 according to the "Extracted features" of Table 5.

(cluster-0) 2.3.2.1*/≥ 08/Afr,As	**(cluster-1)** 2.3.4.4b/≥ 20/global	**(cluster-2)** ≠2.3.2.1*,≠2.3.4.4b/≤ 16/Afr,As

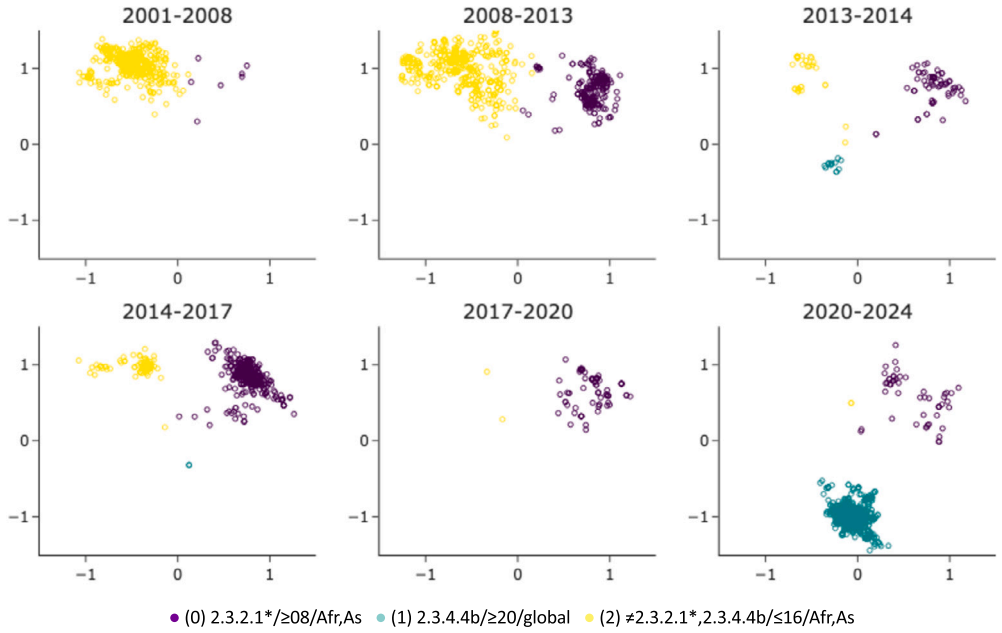

Fig. 10. Codon profiles of clusters (H5N1). The codon profile of the H5N1 sequences is projected on the PC2 (x-axis) and PC1 (y-axis). The scatter plot is organized by flu season. Each sequence is colored according to the assigned cluster.

(*d1*) (domestic birds), clustering of viral isolates based on codon usage profiles clearly separated clade 2.3.4.4b (cluster-1, labelled 2.3.4.4b/≥ 20/global) and clade 2.3.2.1 and descendant clades (cluster-0, labelled 2.3.2.1*/≥ 08/Afr,As) w.r.t. all the other clades in H5N1 (cluster-2). Sequences from clade 2.3.4.4 did not form a single cluster, and clades 2.3.4-2.3.4.4 and 2.3.4.4c were assigned to cluster-2 rather than cluster-1 with 2.3.4.4b.

Cluster-2 and cluster-1 circulated during distinct and non-overlapping intervals of time. Our results indicate differences in codon usage between clades 2.3.4.4 and 2.3.4.4b, consistent with the distinct phylogeographic origin and diversification of 2.3.4.4b compared to other 2.3.4.4x clades [36]. As illustrated in the Supplementary File, Fig. S1, panel b), we also notice that sequences assigned to cluster-0, cluster-1, and cluster-2 show slight differences at the furin cleavage site of the HA protein: RERRRKKR for cluster-0; REKRRKR for cluster-1 and G/RER/KRRKKR for cluster-2.

For wild birds (dataset (*d2*)) our approach delineated five distinct clusters. Each of these five clusters could be easily aligned with those defined in domestic birds (dataset (*d1*)), based on the similarity of their respective mnemonic names (Supplementary File, Table S4). Interestingly, sequences collected after 2020 and assigned to clade 2.3.4.4b were partitioned into two distinct clusters in dataset (*d2*): cluster-0 2.3.4.4b/≠anas/≥ 20/Am and cluster-1 2.3.4.4b/≥ 20, compared with the single cluster formed in the domestic birds dataset (*d1*). One of these (*d2*)-specific clusters was included for the majority of sequences collected from the American continent, hinting at the circulation of one or more geographically distinct lineages of 2.3.4.4b in wild birds in America starting from 2020.

The analysis of complete genome sequences, collected in North America from September 2023 to March 2025, clearly delineated 5 clusters (Supplementary File, Fig. S6, panel d)). Among these, cluster-3 incorporated all sequences from genotype B3.13 (see Supplementary File, Table S5), including every specimen isolated from the avian flu outbreak in dairy cows. This pattern is clearly reflected in cluster-3's mnemonic name: B3.13/dairy cow/23-25. Although the remaining clusters were less epidemiologically relevant than cluster-3, we observed that all genotypes defined by [17] corresponded almost exactly with only one of our clusters. Since genotypes were derived based on the complete genome phylogeny of H5N1 in North America, our analysis of RSCU preferences broadly recapitulates these phylogenetic groups, thereby providing indirect validation of our approach. Importantly, when the same analyses were repeated using only the HA segment (Supplementary File, Table S7 and Fig. S8), a single cluster (cluster-3, mnemonic name B3.13/dairy cow/23-25) still contained all isolates from the dairy cow epidemic. Although there were minor inconsistencies, the mnemonic names of these resulting clusters easily matched those derived from the complete genome analysis. In our opinion, these results indicate that even single genome segments (like HA) are sufficiently diagnostic to capture broad changes in IAV genomics features. However, when available, analyses based on complete genome sequences likely provide more robust results.

2.4. Common trends of codon usage across IAV serotypes

Clusters delineated by our approach broadly correspond to patterns of codon usage profiles and capture the most salient differences in codon preference across different IAVs. Interestingly, our results indicate how changes in codon usage across the 3 distinct pandemic/epidemic spillovers and IAV serotypes herein considered are explained only by a limited number of amino acids (and associated codons), hence pointing to common trends of codon usage. In particular, arginine displayed a high discriminative power for the delineation of viral groups, and 5 out of the 6 synonymous codons for this amino acid were ranked as highly informative in one or more of cases considered in our analyses: H1N1 (AGA, AGG); H7N9 (CGG); and H5N1 (AGA, CGT, CGA). Conversely, codons CCG (proline, H1N1), ACG (threonine, H7N9), and GGC (glycine, H5N1) displayed a high discriminative power only in a specific use case/dataset. We computed and visually inspected RSCU values for all alternative codons of each amino acid across all clusters defined by our method in H1N1, H5N1, and H7N9 (Supplementary File, Fig. S9). Interestingly, we observed that the codons independently flagged by our approach were consistently associated with amino acids displaying the most pronounced codon usage biases and the largest differences in RSCU across all IAV serotypes included in our analysis. This finding supports the effectiveness of our method in detecting key differences in codon usage metrics. For the sake of completeness, impact scores are reported in Supplementary File, Table S8.

Fig. 11 reports RSCU values for all the highly informative amino acid and associated codons across H5N1, H1N1, and H7N9. The codon AGA is by far the preferred codon for arginine in all the IAV serotypes herein considered. This pattern is more evident in H1N1 cluster-0 (6B.1/HU-

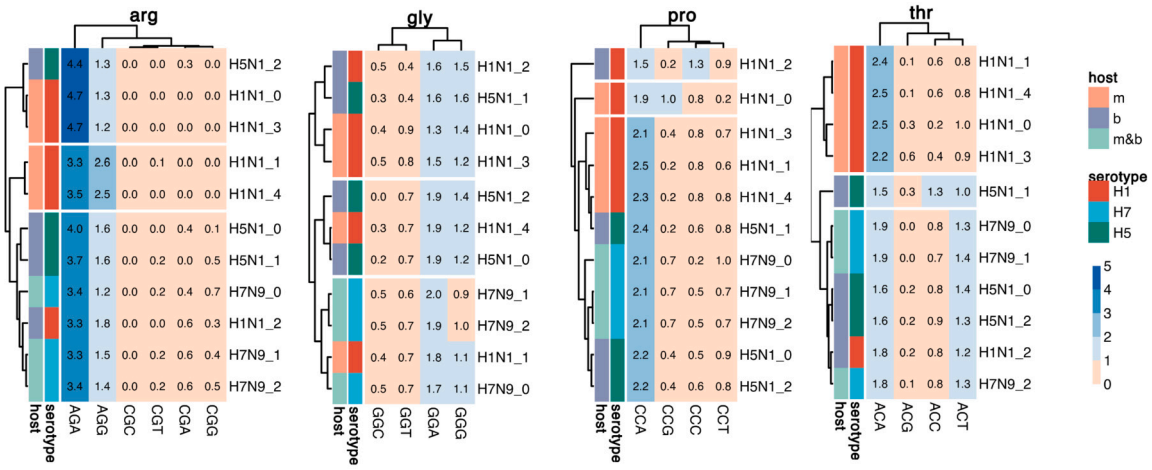

Fig. 11. Average RSCU values for all the synonymous codons of the four most informative amino acids. Each heatmap shows, on a color scale from pink (low) to dark blue (high) blue, the RSCU values for all the synonymous codons of selected amino acids. Each row represents the values of one of the possible codons encoding the amino acid, in one of the clusters defined by our approach in three use cases (H1N1, H7N9, and H5N1). Bars on the left are used to indicate the serotype and host (m = mammals, b = birds, m&b = mammal and birds) according to the color code in the legend, for every cluster.

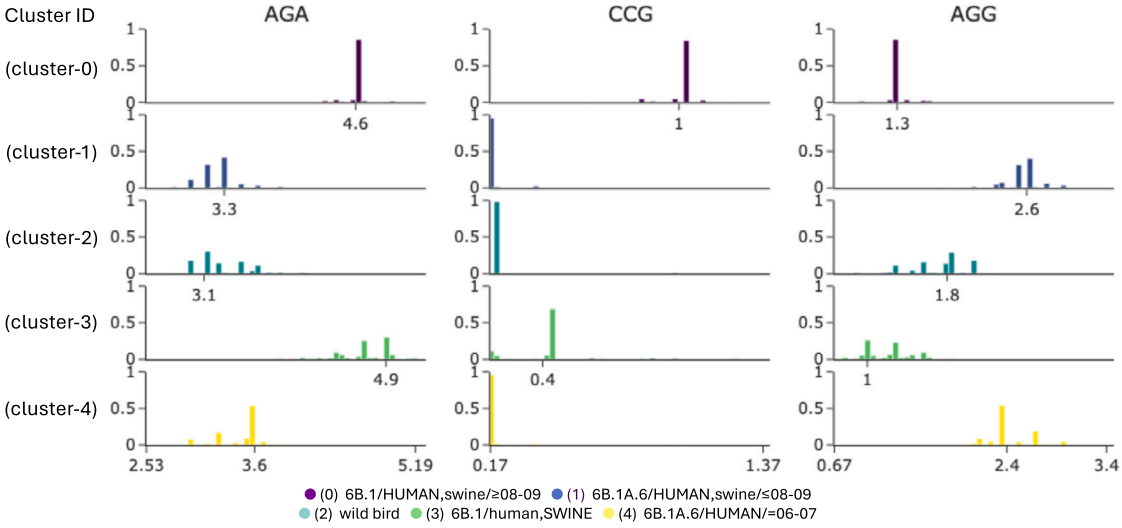

Fig. 12. Codons separating clusters (H1N1). Distribution of RSCUs (exp. value = 1) for codons AGA, CCG, and AGG (i.e., codons for which RSCU distribution mostly differs between the 5 clusters). See boxplot in the Supplementary File, Fig. S10.

MAN,swine/≥08-09) and cluster-3 (6B.1/human,SWINE) and in H5N1 cluster-2. Interestingly, the CGC codon is systematically avoided (RSCU close to 0) by all the serotypes, and CG* codons for arginine are generally avoided in IAVs isolated from mammals (see H1N1 clusters 0, 1, 3, and 4) but are more tolerated in viruses isolated from birds (see H1N1 cluster-2, H5N1 and H7N9). A more detailed discussion of codon usage patterns observed in each serotype is reported in the following sections.

2.4.1. Codon usage in H1N1 clusters

RSCU scores associated with 3 codons – AGA (arginine), CCG (proline), and AGG (arginine) – were sufficient to discriminate the 5 clusters defined by our analytical workflow, with each cluster showing clearly detectable differences in patterns of codon usage as illustrated by Fig. 12: CCG is scarcely used by all the clusters with the exception of cluster-0; AGG is associated with an increased RSCU in clusters 1 and 4 w.r.t. the other clusters, while AGA is highly preferred in cluster-0 and cluster-3.

Notwithstanding the relative differences, AGA is consistently (RSCU range 3.2-4.7) the preferred codon for arginine in H1N1 and the differences in RSCU values (~3 in clusters 1, 2, and 4; and ~5 in clusters 0 and 3) reflect a relative increase (clusters 1, 2, and 4) or decrease (clusters 0 and 3) in the usage of the alternative synonymous codon AGG. Interestingly, the CCG codon for proline is avoided by clusters 1-4, but is not under-represented in cluster-0.

2.4.2. Codon usage in H7N9 clusters

The ACG (threonine) and CGG (arginine) codons were ranked as the most important and could ideally discriminate the 3 clusters identified by our method in H7N9. The corresponding distribution of RSCU values is illustrated in Fig. 13. Cluster-2 displays a very low preference for ACG (RSCU close to 0), while cluster-0 and cluster-1 show an RSCU close to 1, which is in line with the expected value. Relative differences in the preference for CGG discriminate between cluster-0 (RSCU close to 0.8) and cluster-1 (RSCU close to 0.4).

2.4.3. Codon usage in H5N1 (domestic birds) clusters

The RSCUs for the codons CGT, AGA, CGA (arginine), and GGC (glycine) are sufficient to discriminate between the 3 H5N1 clusters Fig. 14. The codon CGT is largely avoided in cluster-1, where it shows an RSCU close to 0, but is not under-represented in cluster-0 and cluster-2, where the RSCU is close to 1. Cluster-0 and cluster-2 instead show dif-

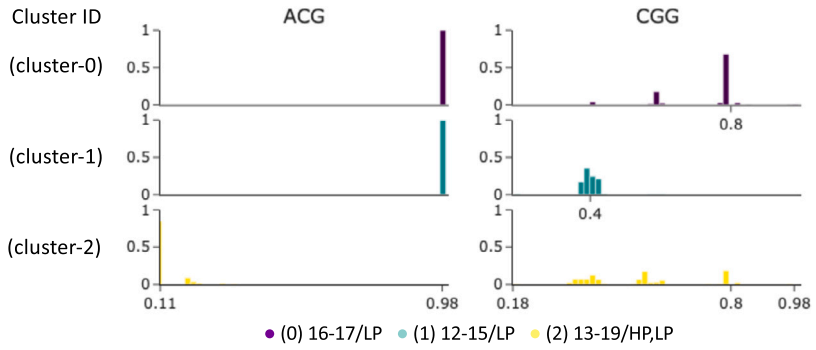

Fig. 13. Codons separating clusters (H7N9). Distribution of RSCUs (exp. value = 1) for codons ACG and CGG (i.e., those for which RSCU distribution mostly differs among the 3 clusters). See boxplot in the Supplementary File, Fig. S11.

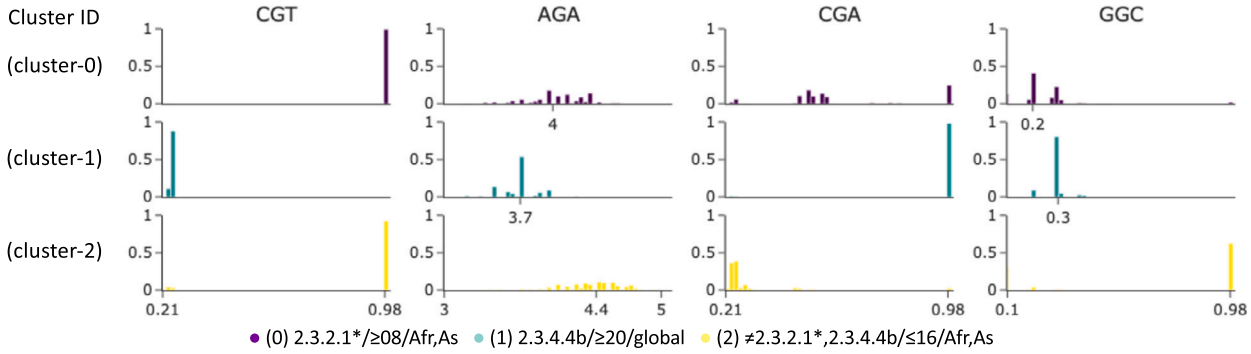

Fig. 14. Codons separating clusters (H5N1). Distribution of RSCUs (exp. value = 1) for four codons CGT, AGA, CGA, and GGC, for which RSCU distribution mostly differs among the 3 clusters. See the boxplot in the Supplementary File, Fig. S12. Corresponding charts for the H5N1 wild birds case are reported in the Supplementary File, Figs. S13-S14.

ferences in the usage of GGC (avoided in cluster-0), CGA (avoided in cluster-2), and AGA (highest RSCU in cluster-2).

3. Discussion

We conceptualized and developed a computational workflow for the stratification of IAVs based on analyzing RSCU and codon preferences. The method was applied to the HA segment of three different influenza A serotypes: H1N1, H7N9, and H5N1 all associated with recent spillover and critical epidemiological emergencies (the 2009 flu epidemic, the 2013-2018 avian flu epidemic in China, and the ongoing H5N1 flu epidemic respectively for H1N1, H7N9 and H5N1). Our approach systematically captured relevant differences in codon preferences in IAVs, identified relevant groups of viruses associated with key epidemiological events, and in some cases had a higher resolution compared to the standing clade-based nomenclature.

In the case of H1N1, our analyses clearly discriminated between the human pandemic virus (2009pdm) and related viruses in the 6B.1 clade circulating in swine before the 2009 pandemic. These results highlighted some key features of the 2009pdm virus and a characteristic preference for the CCG proline codon that is not observed in any other H1N1 IAV included in our dataset. In the absence of additional data, appreciate the evolutionary and epidemiological significance of these data. We observe, however, that these findings provide a proof of principle for the potential application of metrics based on codon preference for the identification and tracking of novel emerging IAVs. Importantly, even in the absence of a corresponding clade in the reference nomenclature, all the H1N1 IAVs isolated from birds were correctly placed in the same cluster.

Notwithstanding the availability of a limited number of sequences sampled unevenly over time, our method could stratify 3 distinct viral groups also in H7N9. Two of these groups included exclusively LP isolated in the early phases of the 2013-2017 flu epidemic and during different epidemic seasons, while HP isolates and viruses from the last wave of the epidemic were all placed in a single cluster. Manual inspection of these sequences also indicated the presence of a furin cleavage site in H7N9 HP. Interestingly, a broad classification into 3 main lineages (A, B, and C) was originally proposed for H7N9 IAVs from the 2013-2017 epidemic [37]. More recently, a study by Lu et al. [38] suggested that a subset of isolates from lineage C acquired high pathogenicity. We observe that our results are substantially in line with these findings. Although in this case codon usage profiles could not set apart viruses isolated from humans and viruses isolated from birds, we believe that this result does not represent a limitation of our method, and can be explained by the zoonotic nature of the 2013-2017 H7N9 epidemic and the lack of sustained human-to-human transmission of H7N9. Indeed, as observed by Millman et al., “A(H7N9) person-to-person spread has been limited to 2 or possibly 3 generations of transmission” [39].

H5N1 IAVs are considered a global pandemic threat for their ability to recurrently infect mammals, as also demonstrated by the recent spillover of H5N1 clade 2.3.4.4b in dairy cows in the United States [40]. We conducted several complementary analyses to validate and assess our approach in identifying and correctly labeling H5N1 viruses associated with important epidemiological events. First, a broad-spectrum analysis of sequences collected from 2000 to May 2024 was performed to stratify viral clades circulating in domestic birds (3475 sequences) and wild birds (12021 sequences). Both datasets formed highly consistent partitions, clearly indicating the worldwide emergence and spread of the novel 2.3.4.4b clade starting in 2020. Notably, a cluster of 2.3.4.4b specific to the American Continent was detected exclusively in the wild bird sequences. These findings align well with known epidemiological events, suggesting that codon usage-based metrics are effective in capturing major shifts in influenza A virus (IAV) evolution [41]. Subsequently, we focused on complete genome sequences of 2.3.4.4b IAVs isolated in North America from September 2023 to March 2025. The objective of this targeted analysis was to verify if our analytical workflow could specifically flag the emergence of B3.13, a reassortant 2.3.4.4b

genotype known to have caused a large outbreak of avian influenza in dairy cows.

Our approach revealed a clear correspondence between the B3.13 genotype and a single distinct cluster, indicating that the results of our workflow are highly concordant with established phylogenetic analyses. This finding further supports the utility of our method in identifying novel emerging viral clades. Importantly, these findings were substantially corroborated by independent analyses performed solely on the hemagglutinin (HA) segment, demonstrating the utility of single-segment analyses for inferring key whole-genome evolutionary shifts.

Another interesting observation is that -at least for the IAVs considered in this study- similar preferences in arginine synonymous codon usage were observed across all serotypes. Whereas the 4 CG* codons for arginine are tendentially avoided and show an RCSU largely lower than 1, the synonymous AGA and AGG codons display RSCU values in excess of 2 or higher, indicating a systematic preference for these codons. This pattern was previously reported by independent studies in the field, including, for example, Wong et al. [7] and Gu et al. [11], and is more pronounced in IAVs isolated from mammals.

In the light of these results, it is tempting to speculate that the codon preference for arginine might represent a strong selective pressure for Influenza A viruses. Similarly to other viruses that can infect mammals, the observed under-representation of CG* arginine codons could reflect an evolutionary pressure exerted by the host zinc finger antiviral protein (ZAP), which is known to restrict viral replication by binding to the CpG-rich regions of viral RNA [42,43].

Notwithstanding these consistent results, we acknowledge that our proposed workflow and approach present some limitations. Here, we only focus on the hemagglutinin segment. This is partly due to the reduced availability of other genome segments. Indeed, if we considered only complete isolates (with all segments available), the proportion of data employed in this study would drop by 42.3% for H1N1, 10.1% for H7N9, and 31.7% for H5N1 – before applying data quality filtering on the non-HA segments (as our analytical framework would require). While this choice can be considered a limitation, at the same time, it is an asset of our method, making it applicable to larger datasets and posing the basis for lightweight methods that can be used for early warning systems [44]. Here, we do not claim that codon usage in HA is sufficient to completely infer host adaptation, but that it is sufficiently diagnostic to identify it -as also demonstrated by our results on H5N1 dataset (*d3*)- where a good level of agreement was observed between the results obtained from the analysis of HA in isolation and the complete genome sequence. Detailed analyses of each individual segments, and reconciliation of the main findings, remain for future work. Further, we recognize that the RSCU metric can be insensitive to small changes in sequence composition. This was evident in the case of H7N9, where we had to manually identify an alternative number of clusters to discern the virus' pathogenicity.

4. Conclusion

All in all, the approach discussed in this study holds significant promise for the development of novel strategies for the monitoring and tracking of major changes in IAV genomes and the development of orthogonal tools for genomic surveillance of influenza viruses that could complement phylogenetic analyses and the clade-based nomenclature. Our findings indicate that, while sensitive to genetic variation, phylogenetic analysis and the current lineage-based nomenclature of IAVs may be insensitive to (possibly) sparse but systematic changes in codon usage potentially associated with host adaptation. In this study, by interpolating our results with metadata describing contextual viral sequences information (e.g., at the time of collection, *host type*, *date*, *location* and at the time of classification, *clade*, *pathogenicity*), we derive patterns of diversification of genomic features that are not captured by the standing clade-based classification. Groups defined by our approach provide a more fine-grained classification compared to the originally assigned lineages in the paradigmatic cases included in our analyses (enriched by the cases of H5N1 infecting wild birds and the recent B3.13 spillover – see Supplementary File) and facilitate a high-level comparison of genomic features across a diverse range of influenza A serotypes, all linked with reported spillovers. Interestingly, a reduced number of codons and amino acids was sufficient to summarize salient features and recurrent patterns of codon usage throughout all the use cases, suggesting common selective pressures.

In conclusion, our analyses highlighted the effectiveness of codon usage-based metrics for the stratification of influenza A viruses and for capturing impactful/important epidemiological events associated with recent spillovers. The analytical workflow developed in this study represents a blueprint for future analyses aimed at the precise characterization of groups of viral sequences based on codon usage metrics. As a natural extension of this work, we already started investigating how to embed the proposed diagnostic method within systematic scanners of continuously produced data to inform early warning systems [45]. Due to its general applicability, this approach could also be extended to other Influenza A serotypes for which genomic data are scarce and/or a reference nomenclature is not established.

5. Methods

Data source and preprocessing. We download viral influenza sequences from GISAID [18] (both data and metadata) using the filters available on the web portal to select the data of interest. Data downloaded from GISAID were processed according to this analytical workflow: removal of sequences with duplicate `Isolate_ID` and `Virus_Name`, unknown host or clade (when the information is essential for selecting the input data of the experiment); removal of sequences with length $> \pm 3\%$ dissimilar from the median value; annotation of the hemagglutinin coding sequence (CDS); and removal of sequences with incomplete/truncated CDS (i.e., when the length is not a multiple of 3). Specifically, we produced five datasets with the following characteristics: 5782 H1N1 HA segments collected from Oct. 2006 to Sep. 2012 in North America (see first use case); 1882 H7N9 HA segments collected from Oct. 2010 to Dec. 2019 (see second use case); 3475 H5N1 HA segments collected from 2000 to May 2024 from domestic bird hosts (see third use case, dataset (*d1*)); 12021 H5N1 HA segments collected from 2000 to May 2024 from wild bird hosts (see dataset (*d2*), Supplementary File, Figs. S2-S4 and Tables S2-S4); 5243 H5N1 complete genomes collected from Sept. 2023 to Mar. 2025 in North America (see dataset (*d3*), Supplementary File, Figs. S5-S8 and Tables S5-S7). For the analysis of complete genomes in H5N1, we concatenated the coding sequences (CDS) of all the 8 segments of each isolate. While segments PB2, HA, NP, and NA contain only one CDS, the other 4 (PB1, PA, MP, and NS) contain two overlapping CDS; for those four segments, we considered only the longest CDS among the two to avoid double-counting the same codons while computing the RSCUs. On average, the extracted sequences were 12,676 nb long for each isolate.

Computation of RSCU. To measure the codon usage profile of individual viral sequences, we used Relative Synonymous Codon Usage (RSCU) [46]. This metric – different from absolute counts or frequency ratios of codon usage – allows for theoretical comparison between viral species. The RSCU is an algebraic transformation of the codon frequency in a sequence and reflects the under/overexpression of a synonymous codon under the assumption of even usage. The equation $RSCU(j) = (X_{ij})/(\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij})$ indicates the RSCU of the j^{th} codon, where X_{ij} is the frequency of incidence of the j^{th} codon encoding the i^{th} amino acid and n_i is the sum of synonymous codons that encode for the i^{th} amino acid. For every sequence, we computed RSCU values corresponding to the 59 non-synonymous codons, i.e., the 64 permutations with repetitions of 4 nucleotides by 3 after removing 3 stop codons (TAA, TAG, TGA) and 2 non-synonymous codons (ATG, TGG). RSCU

values were subsequently analyzed through multiple algorithms and analytical approaches to identify potential clusters of sequences based on similar codon profiles.

Principal component analysis. We use the PCA as a dimensionality reduction technique before using any clustering algorithm. This additional step helps to reduce the number of input features from the 59 RSCU values to less than 10 typically. PCA aims to identify a new set of uncorrelated variables, the *principal components* (PCs), capturing the most significant source of variability in the input data. Each PC is a linear transformation of the feature vectors, with the first PC accounting for the maximum variance, followed by subsequent PCs explaining progressively less variation. To correctly parametrize the PCA, we perform a Scree Test [47]. The test consists of plotting the explained variance ratio of every PC in descending order (called the scree plot or "elbow plot") to identify the "elbow", i.e., the point at which the values seem to level off. The PCs at the left of the elbow are deemed the most significant ones. The input data can be plotted in the transformed space using any two PCs as axes.

Hierarchical agglomerative clustering. For a dataset of size n, the hierarchical agglomerative clustering algorithm produces a variable number of clusters from n to 1. This approach has the advantage of not requiring the number of clusters to be defined *a priori*. The agglomerative clustering algorithm initially assumes that each data point is a cluster itself. Then, two data points (or single-item clusters) are progressively merged into a cluster at each iteration. At every algorithm step, only the two data points leading to the minimum possible increase of the total within-cluster variance are clustered (see Ward's method [48]). The algorithm terminates when only a single cluster is left. The intra-cluster distance (or cluster's cohesion) is computed using the Euclidean distance of the input PCs and shown on a dendrogram at each iteration. Therefore, the intra-cluster distance is 0 at the beginning of the algorithm, and maximum when at the end, i.e., when all the data points are part of the same cluster. The number of clusters depends on the set maximum inter-cluster distance threshold. The threshold evaluation is subjective, but generally, long branches indicate different clusters, while short branches correspond to clusters that can be easily considered as one. Hierarchical clustering is used here to identify a range for an appropriate number of well-separated clusters.

K-means clustering. The k-means clustering algorithm is used to create the final clusters and label the data. The k-means clustering algorithm assumes that the number of clusters is known a priori and creates a corresponding number of randomly generated centroids, i.e., cluster representatives. At every iteration, the distance between the centroids and all the other points is computed, the closest point to a centroid is merged with the respective cluster, and the centroids are updated until all the points are assigned. We run k-means once for every number of clusters in the range of plausible cluster numbers identified through hierarchical clustering.

Cluster evaluation. Using the hierarchical clustering algorithm, we compute a dendrogram that suggests the range of potentially correct numbers of clusters. We search for the final number of clusters in this range by running k-means once for every value in the aforementioned range and evaluating the goodness of the clusters. Among simple evaluation metrics for validating the clusters, we consider the within-cluster sum of squares (WCSS) and the between-cluster sum of squares (BCSS). The former is the sum of squared Euclidean distances between the data points and their respective cluster centroids; the latter is the sum of squared Euclidean distances between each cluster centroid and the overall data centroid, weighted by the cluster's size. In our case, the goodness of the clusters is evaluated using two metrics:

- the Calinski-Harabasz score [23] derives from the ratio of the BCSS and the WCSS, normalized by their respective degrees of freedom. Ideally, a good clustering is made of well-separated (high BCSS) and compact clusters (low WCSS), therefore, corresponding to a high Calinski-Harabasz score.
- the Silhouette score [24] still compares intra-cluster similarity and inter-cluster distance, but using slightly different formulae. For a single data point, it measures how well it fits into the assigned cluster compared to other clusters. For the entire dataset, it is the average of all scores. This value ranges from -1 (i.e., low intra-cluster cohesion and low inter-cluster distance) to 1 (i.e., high intra-cluster cohesion and high inter-cluster distance). As a rule of thumb, a clustering is considered "strong" with a Silhouette score > 0.7; "reasonable" if > 0.5; weak otherwise.

In essence, these measures correspond to a tradeoff between the similarity of the data points within each cluster and the dissimilarity of the data points belonging to different clusters. They both provide a far better evaluation metric than using the simple WCSS and BCSS on their own because, for example, high intra-cluster similarity is achievable also with partially overlapping clusters. Secondly, these are both internal evaluation metrics – meaning that no external knowledge is required to assess the quality of the clustering – which is a desirable property as, oftentimes, no ground truth labels are available for comparison.

Finally, we repeat the assessment procedure 10 times for each number of clusters (10-fold evaluation) to average the scores and mitigate the dependency from the random initialization of the cluster centroids in k-means.

Identification of the most important codons. Salient features of every cluster need to be characterized in order to derive the most relevant changes in codon usage between the different groups. This could be done by simply averaging the RSCUs of the sequences in each cluster. However, we reasoned that comparing 59 features across several clusters might not provide an ideal/easy-to-interpret representation of the most important differences in codon usage. To determine only the most important RSCUs for the delineation of clusters we use a multinomial linear classifier. A multinomial classifier combines the output prediction of multiple binary classification models using the softmax function to return the predicted group. Each binary model learns a vector of weights that, multiplied by the RSCUs of a sequence i, returns the likelihoods of i to be included in one of the clusters. The model is typically built by considering only a subset of the available data, typically 80%, and its fitness is measured on the remaining 20% of the data. Then, the vectors of weights are extracted from each binary model. Since the classifier is multinomial, we build an aggregated vector of "impact-scores" as the sum of the absolute weights of each model, weighted by the standard deviation of the input. A Scree Test on the impact-scores vector returns the subset of the most important codons for the classification task. To compare the typical RSCUs of such codons in different clusters, we compute a matrix of histograms whose columns correspond to the important codons, and rows correspond to the clusters. The plot allows us to effectively characterize a cluster of sequences using only the RSCUs of a few codons.

Software implementation. The data source and processing stage was implemented using Python 3.10.12 and the software libraries Pandas 2.2.1 and Numpy 1.26. The computation of RSCUs was implemented purely in Python, without external libraries. The Scikit-Learn 1.3.0 library was used for the algorithms of PCA, hierarchical clustering, and k-means clustering, for evaluating cluster centroids, euclidean distance, and Silhouette score, and for the classification task. All plots have been generated using Plotly 5.17.0, except for the dendrogram (with Matplolib 3.8.0).

Protein sequence alignment and sequence logo. We computed a global multiple sequence alignment of the conceptual translation of the HA CDS for all the 11139 sequences included in our analyses (H1N1, H7N9, H5N1 domestic birds) was computed with muscle [49], using default parameters. Alignments were visualized in seaview [50] and sequence logos were computed with WebLogo [51].

CRediT authorship contribution statement

Tommaso Alfonsi: Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization. **Matteo Chiara:** Writing – review & editing, Methodology, Investigation, Funding acquisition, Conceptualization. **Anna Bernasconi:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Funding

The 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. Principal Investigator A.B., co-Principal Investigator M.C.

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 responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

The authors are grateful to Ilaria Capua, Stefano Ceri, and colleagues at the Istituto Zooprofilattico Sperimentale delle Venezie (specifically, Alice Fusaro and Isabella Monne) for the useful advice and motivation during this research.

Appendix A. Supplementary material

In the Supplementary File, we provide figures and tables for additional analyses to complement the results in the main manuscript.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2025.06.030.

Data availability

The findings of this study are based on metadata associated with three sets of sequences available on GISAID up to 2024/05/10 H1N1 (gisaid.org/EPI_SET_250319su), H7N9 (gisaid.org/EPI_SET_250319ms), and H5N1 (gisaid.org/EPI_SET_250616yr, gisaid.org/EPI_SET_250616ms, and gisaid.org/EPI_SET_250616nx); the notebooks used for computations are made available in the Zenodo repository https://doi.org/10.5281/zenodo.14561947 [19].

References

[1] Taubenberger JK, Kash JC. Influenza virus evolution, host adaptation, and pandemic formation. Cell Host Microbe 2010;7:440–51.
[2] Bahir I, Fromer M, Prat Y, Linial M. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Mol Syst Biol 2009;5:311.
[3] Jenkins GM, Holmes EC. The extent of codon usage bias in human rna viruses and its evolutionary origin. Virus Res 2003;92:1–7.
[4] Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. Patterns of evolution and host gene mimicry in influenza and other rna viruses. PLoS Pathog 2008;4:e1000079.
[5] Li Z-p, Ying D-q, Li P, Li F, Bo X-c, Wang S-q. Analysis of synonymous codon usage bias in 09h1n1. Virol Sin 2010;25:329–40.
[6] Zhou T, Gu W, Ma J, Sun X, Lu Z. Analysis of synonymous codon usage in h5n1 virus and other influenza A viruses. Biosystems 2005;81:77–86.
[7] Wong EH, Smith DK, Rabadan R, Peiris M, Poon LL. Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus. BMC Evol Biol 2010;10:1–14.
[8] Iwasaki Y, Abe T, Wada Y, Wada K, Ikemura T. Novel bioinformatics strategies for prediction of directional sequence changes in influenza virus genomes and for surveillance of potentially hazardous strains. BMC Infect Dis 2013;13:1–9.
[9] Babayan SA, Orton RJ, Streicker DG. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in rna virus genomes. Science 2018;362:577–80.
[10] Deka H, Chakraborty S. Compositional constraint is the key force in shaping codon usage bias in hemagglutinin gene in h1n1 subtype of influenza A virus. Int J Genomics 2014;2014:349139.
[11] Gu H, Fan RL, Wang D, Poon LL. Dinucleotide evolutionary dynamics in influenza A virus. Virus Evol 2019;5:vez038.
[12] Sun J, Zhao W, Wang R, Zhang W, Li G, Lu M, et al. Analysis of the codon usage pattern of ha and na genes of H7N9 influenza A virus. Int J Mol Sci 2020;21:7129.
[13] Li J, Zhang S, Li B, Hu Y, Kang X-P, Wu X-Y, et al. Machine learning methods for predicting human-adaptive influenza A viruses based on viral nucleotide compositions. Mol Biol Evol 2020;37:1224–36.
[14] Girard MP, Tam JS, Assossou OM, Kieny MP. The 2009 a (H1N1) influenza virus pandemic: a review. Vaccine 2010;28:4895–902.
[15] Wang Q, Xu K, Xie W, Yang L, Chen H, Shi N, et al. Seroprevalence of H7N9 infection among humans: a systematic review and meta-analysis. Influenza Other Respir Viruses 2020;14:587–95.
[16] Charostad J, Rezaei Zadeh Rukerd M, Mahmoudvand S, Bashash D, Hashemi SMA, Nakhaie M, et al. A comprehensive review of highly pathogenic avian influenza (HPAI) H5N1: an imminent threat at doorstep. Trav Med Infect Dis 2023;55:102638.
[17] Nguyen T-Q, Hutter CR, Markin A, Thomas M, Lantz K, Killian ML, et al. Emergence and interstate spread of highly pathogenic avian influenza A (H5N1) in dairy cattle in the United States. Science 2025;388:eadq0900.
[18] Shu Y, McCauley J. Gisaid: global initiative on sharing all influenza data–from vision to reality. Euro Surveill 2017;22:30494.
[19] Alfonsi T, Chiara M, Bernasconi A. Supplementary material for "A codon bias data approach for the stratification of Influenza A virus clades". https://doi.org/10.5281/zenodo.14561947, 2024, last accessed: June 1st, 2025.
[20] Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, et al. Antigenic and genetic characteristics of swine-origin 2009 a (h1n1) influenza viruses circulating in humans. Science 2009;325:197–201.
[21] Mena I, Nelson MI, Quezada-Monroy F, Dutta J, Cortes-Fernández R, Lara-Puente JH, et al. Origins of the 2009 H1N1 influenza pandemic in swine in Mexico. eLife 2016;5.
[22] Ding X, Liu J, Jiang T, Wu A. Transmission restriction and genomic evolution coshape the genetic diversity patterns of influenza A virus. Virol Sin 2024.
[23] Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat, Theory Methods 1974;3:1–27.
[24] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.
[25] Goneau L, Mehta K, Wong J, L'Huillier A, Gubbay J. Zoonotic influenza and human health—part 1: virology and epidemiology of zoonotic influenzas. Curr Infect Dis Rep 2018;20:1–12.
[26] Shi J, Deng G, Ma S, Zeng X, Yin X, Li M, et al. Rapid evolution of h7n9 highly pathogenic viruses that emerged in China in 2017. Cell Host Microbe 2018;24:558–68.
[27] Qi W, Jia W, Liu D, Li J, Bi Y, Xie S, et al. Emergence and adaptation of a novel highly pathogenic h7n9 influenza virus in birds and humans from a 2013 human-infecting low-pathogenic ancestor. J Virol 2018;92:10–1128.
[28] Food and Agricultural Organization of the United States and World Organization for Animal Health. Influenza A cleavage sites, OFFLU - network of expertise on animal influenza. https://www.offlu.org/wp-content/uploads/2022/01/Influenza-A-Cleavage-Sites-Final-04-01-2022.pdf, 2022. last accessed: June 1st, 2025.
[29] Stieneke-Gröber A, Vey M, Angliker H, Shaw E, Thomas G, Roberts C, et al. Influenza virus hemagglutinin with multibasic cleavage site is activated by furin, a subtilisin-like endoprotease. EMBO J 1992;11:2407–14.
[30] World Health Organization. Human infection with avian influenza A(H5) viruses. Avian influenza weekly update number 970; 2024.
[31] Pawestri HA, Eggink D, Isfandari S, Thanh TT, Rogier van Doorn H, Setiawaty V, et al. Viral factors associated with the high mortality related to human infections with clade 2.1 influenza A/H5N1 virus in Indonesia. Clin Infect Dis 2020;70:1139–46.
[32] Gambotto A, Barratt-Boyes SM, de Jong MD, Neumann G, Kawaoka Y. Human infection with highly pathogenic h5n1 influenza virus. Lancet 2008;371:1464–75.
[33] Faverjon C, Fanelli A, Cameron A. Expansion of the early warning system for avian influenza in the eu to evaluate the risk of spillover from wild birds to poultry. EFSA Support Publ 2024;21:9114E.

[34] Lee D-H, Bertran K, Kwon J-H, Swayne DE. Evolution, global spread, and pathogenicity of highly pathogenic avian influenza h5nx clade 2.3.4.4. J Vet Sci 2017;18:269–80.
[35] Shepard SS, Davis CT, Bahl J, Rivailler P, York IA, Donis RO. Label: fast and accurate lineage assignment with assessment of h5n1 and h9n2 influenza A hemagglutinins. PLoS ONE 2014;9:e86921.
[36] Xie R, Edwards KM, Wille M, Wei X, Wong S-S, Zanin M, et al. The episodic resurgence of highly pathogenic avian influenza h5 virus. Nature 2023;622:810–7.
[37] Lam TT-Y, Zhou B, Wang J, Chai Y, Shen Y, Chen X, et al. Dissemination, divergence and establishment of H7N9 influenza viruses in China. Nature 2015;522:102–5.
[38] Lu J, Raghwani J, Pryce R, Bowden TA, Thézé J, Huang S, et al. Molecular evolution, diversity, and adaptation of influenza A(H7N9) viruses in China. Emerg Infect Dis 2018;24:1795.
[39] Millman AJ, Havers F, Iuliano AD, Davis CT, Sar B, Sovann L, et al. Detecting spread of avian influenza A(H7N9) virus beyond China. Emerg Infect Dis 2015;21:741.
[40] Department for Environment Food & Rural Affairs, Animal & Plant Health Agency. Influenza A (H5N1) infection in mammals: suspect case definition and diagnostic testing criteria. Department for Environment; 2022. https://www.gov.uk/government/publications/listed-diseases-in-animals-case-definitions-testing-and-reporting/influenza-a-h5n1-infection-in-mammals-suspect-case-definition-and-diagnostic-testing-criteria. last accessed: June 1st, 2025.
[41] Lewis NS, Banyard AC, Whittard E, Karibayev T, Al Kafagi T, Chvala I, et al. Emergence and spread of novel h5n8, h5n5 and h5n1 clade 2.3.4.4 highly pathogenic avian influenza in 2020. Emerg Microbes Infect 2021;10:148–51.
[42] Bohn JA, Meagher JL, Takata MA, Gonçalves-Carneiro D, Yeoh ZC, Ohi MD, et al. Functional anatomy of zinc finger antiviral protein complexes. Nat Commun 2024;15:10834.
[43] Luo X, Wang X, Gao Y, Zhu J, Liu S, Gao G, et al. Molecular mechanism of rna recognition by zinc-finger antiviral protein. Cell Rep 2020;30:46–52.
[44] Bernasconi A, Mari L, Casagrandi R, Ceri S. Data-driven analysis of amino acid change dynamics timely reveals SARS-CoV-2 variant emergence. Sci Rep 2021;11(1):21068. https://doi.org/10.1038/s41598-021-00496-z.
[45] Alfonsi T, Bernasconi A, Chiara M, Ceri S. Multi-scale early warning system for influenza A spillovers. bioRxiv 2025:2025-05. https://doi.org/10.1101/2025.05.24.655955.
[46] Sharp PM, Li W-H. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 1986;24:28–38.
[47] Cattell RB. The scree test for the number of factors. Multivar Behav Res 1966;1:245–76.
[48] Ward Jr JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963;58:236–44.
[49] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–7.
[50] Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 2010;27:221–4.
[51] Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14:1188–90.