A codon usage-based approach for the stratification of Influenza A

Tommaso Alfonsi[1], Matteo Chiara[2], Anna Bernasconi[1]*

[1]Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy.
[2]Department of Biosciences, Università degli Studi di Milano, Milan, Italy.

*Corresponding author(s). E-mail(s): anna.bernasconi@polimi.it;
Contributing authors: tommaso.alfonsi@polimi.it; matteo.chiara@unimi.it;

Supplementary File

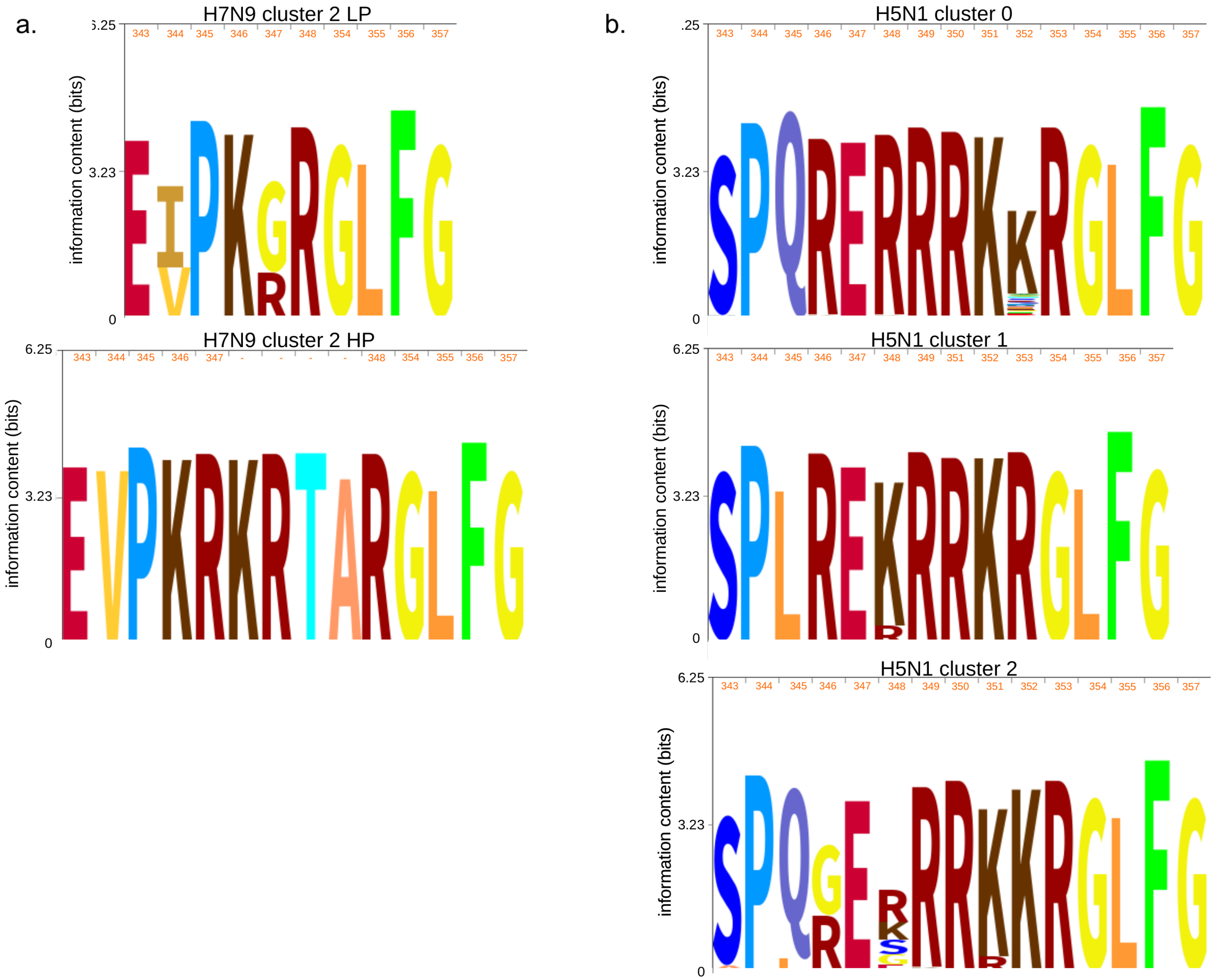

Fig. S1: **Sequence logo of hemagglutinin (HA) H1/H2 subunits cleavage site H7N9 and H5N1 domestic bird case studies.** Panel a) H7N9 cluster2: the sequence logo for LP and HP sequences are displayed. A sequence motif compatible with a furin cleavage site (KRKRTAR) is observed in HP sequences. Panel b) H5N1: comparison of furin cleavage sites of cluster0, cluster1, and cluster2. For both panels a) and b), numbers in orange are used to indicate relative positions in the HA protein alignment.

Cluster \ Flu Season	2010-2011	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019
(0)	0	218	205	3	2	1	0	0
(1)	0	2	201	6	0	0	0	0
(2)	0	6	37	8	1	0	0	0
(3)	0	6	90	19	9	5	0	0
(4)	0	0	38	122	6	8	0	0
(5)	1	2	58	38	2	15	0	0
(6)	0	0	0	0	39	184	0	0
(7)	0	0	0	0	3	107	1	0
(8)	0	0	0	0	0	279	1	0
(9)	1	0	0	5	0	90	19	44

Cluster \ Pathogenicity	HPIV	LPIV	None
(0)	0	428	1
(1)	0	209	0
(2)	0	52	0
(3)	0	129	0
(4)	0	174	0
(5)	0	116	0
(6)	0	222	1
(7)	0	111	0
(8)	0	280	0
(9)	150	7	2

Table S1: **High resolution analysis for the H7N9 dataset.** By repeating the analysis on the H7N9 dataset with 10 clusters, we obtain a more detailed overview of the genomic heterogeneity of the sequences. Intriguingly, the 10 clusters reflect the evolution of the H7N9 virus over time, with each cluster spanning a period of 1 or 2 years on average. More importantly, all of the HPIV sequences are recognized and isolated into a single cluster. This table shows the distribution of sequences over time (flu seasons) and pathogenicity. The *host type* metadata has not been reported here, since samples captured from domestic birds, wild birds, and human hosts were generally present in all clusters. Cluster IDs are reported in the leftmost column. The inner cells report the number of sequences for each metadata value and cluster.

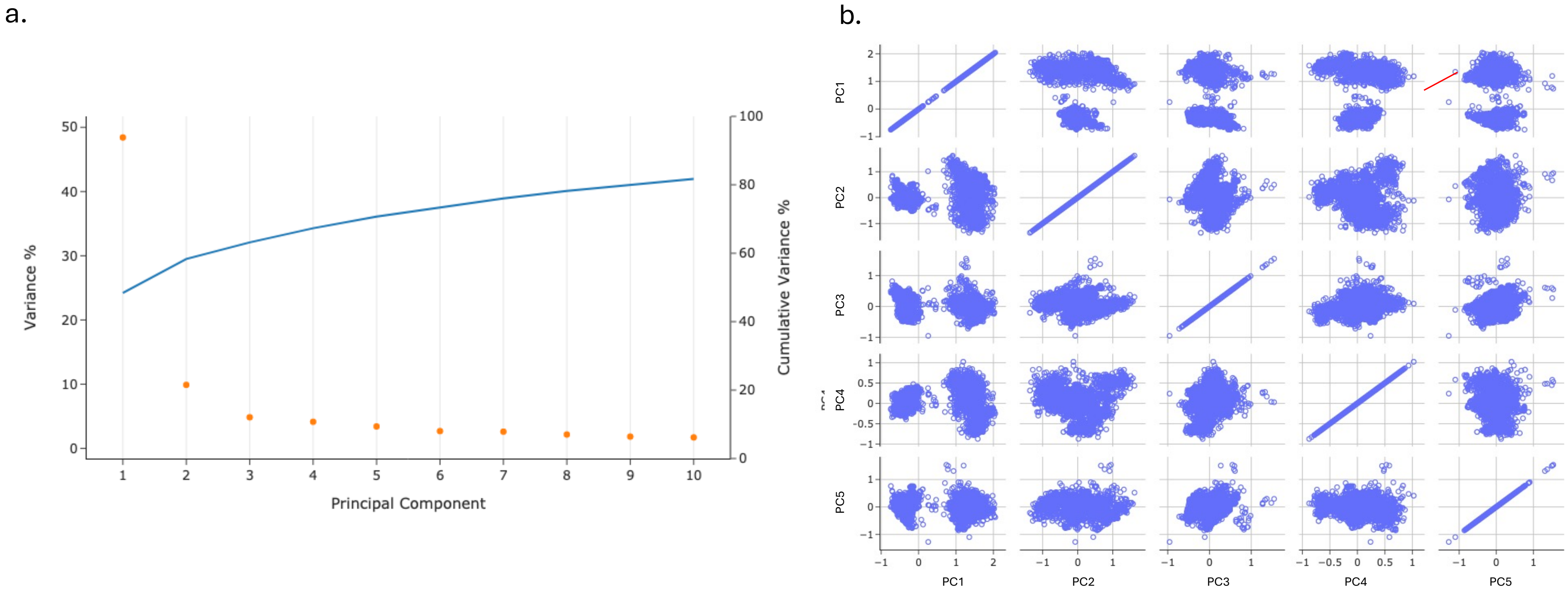

Fig. S2: **Principal Components Analysis (H5N1, wild bird).** We retrieved a dataset of 12021 sequences collected from H5N1-infected hosts of 129 wild bird species, with the oldest sequences dating back to 2000, January 1st. According to panel a), the first five PCs were ranked as the most important in the PCA of RSCU values. Two groups of data points are immediately recognizable in panel b).

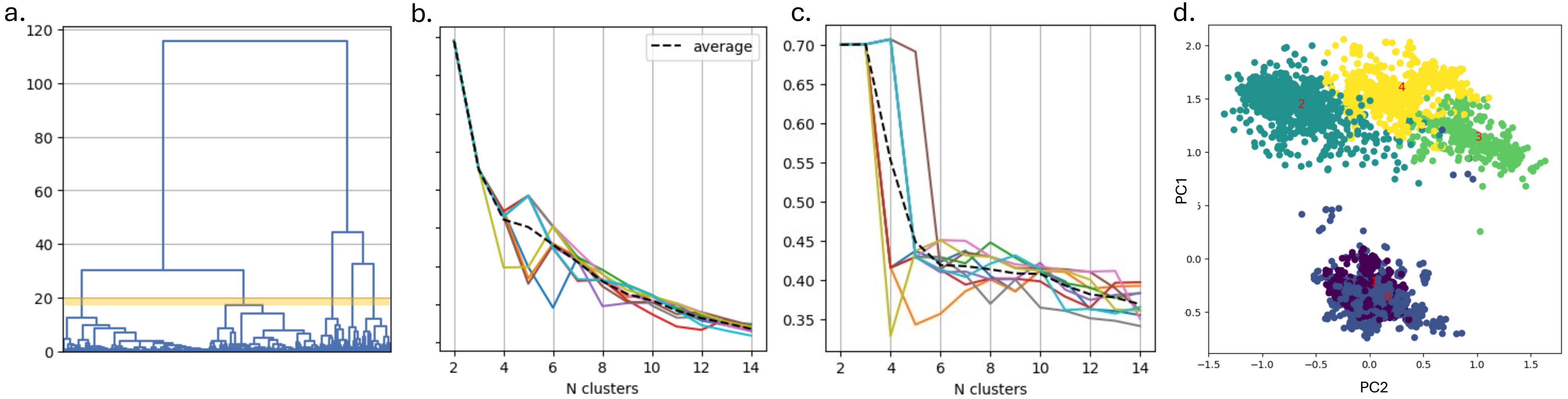

Fig. S3: **Clustering (H5N1, wild birds).** The groups recognized in Fig. S2 correspond to the two main clusters found at a distance ≥ 45 in the dendrogram (see panel a)), also confirmed by the peak scores of the Kalinski-Harabasz and Silhouette metrics for n. 2 clusters (see panels b) and c)). However, using hierarchical agglomerative clustering, we delineate up to 5 clusters at intra-cluster distances between 17 and 30. These emerge as sub-partitions of the two previously mentioned groups. When plotted on PCs 1 and 2 (see panel d)), these clusters appear close to each other, partially overlapping, and sometimes they do not have a spherical shape. Clusters as such are notoriously hard to recognize using k-means clustering. Indeed, the average score of Calinski-Harabaz suggests the existence of only two clusters. The Silhouette score confirms the previous assessment but also suggests three clusters as an equally good solution. Overall, the k-means method struggles to identify the sub-partitions within the two groups of data points. Upon these observations, we conclude that k-means may not be the best clustering method for this specific dataset and proceed with the analysis workflow using the clusters identified by hierarchical clustering. Panel d) illustrates the five clusters identified by hierarchical clustering, plotted on PC1 and PC2. Clusters 0 and 1 appear to overlap largely as they are hierarchically related.

Cluster \ Flu Season	2000-2008	2008-2010	2010-2017	2017-2020	2020-2024	**Extracted feature**
(0)	0	0	0	0	5360	≥ 20
(1)	0	5	11	2	4052	≥ 20
(2)	585	189	252	3	4	≤ 17
(3)	0	0	221	121	168	≥ 10
(4)	7	102	816	21	102	≥ 08

Cluster \ Clade	0-2.3.2	2.3.2.1	2.3.2.1a	2.3.2.1b	2.3.2.1c	2.3.3-2.3.4.4	2.3.4.4b	2.3.4.4c-9	nonGsGd	**Extracted feature**
(0)	0	0	0	0	0	0	5360	0	0	2.3.4.4b
(1)	0	0	0	0	0	10	4055	0	5	2.3.4.4b
(2)	660	0	2	5	0	278	0	40	47	oth.clades
(3)	0	0	506	4	0	0	0	0	0	2.3.2.1a
(4)	4	20	177	50	796	0	1	0	0	2.3.2.1*

Cluster \ Host Type	anas pla.	anas sp.	branta le.	cygnus ol.	eagle	goose	gull	swan	turkey	**Extracted feature**
(0)	0	0	9	3	112	311	35	17	254	<>anas
(1)	48	3	99	121	25	121	124	114	302	
(2)	55	7	0	9	0	39	3	20	32	
(3)	22	152	0	0	0	4	0	0	6	
(4)	45	40	0	2	1	16	5	7	5	

Cluster \ Continent	Africa	Antar.	Asia	Europe	North Am.	Oceania	South Am.	**Extracted feature**
(0)	4	18	4	35	5141	0	158	Am
(1)	176	0	636	3032	225	0	1	
(2)	175	0	748	71	38	1	0	Afr,As
(3)	0	0	510	0	0	0	0	Afr,As
(4)	39	0	995	14	0	0	0	Afr,As

Cluster \ Pathogen.	HP	LP	**Extracted feature**
(0)	5360	0	
(1)	4064	6	
(2)	982	51	
(3)	510	0	
(4)	1046	2	

Table S2: **Partitioning of the H5N1 wild bird clusters on the metadata properties in the 5 clusters.** Here we show the contingency matrix generated by the comparison of the cluster assignments with metadata-based partitions. Cluster IDs are reported in the leftmost column. The inner cells report the number of sequences for each metadata value and cluster. All the meaningful metadata attributes available in the dataset have been included in the comparison: the *flu season*, *clade*, *host type*, *continent*, and *pathogenicity* (high (HP) or low (LP)). The rightmost column contains, when possible, the name of a cluster-describing feature based on the observed numbers. The "<>" symbol means "anything but". The *Host Type* subtable shows only the most represented species, although the dataset contains 129 species.

Cluster names	**(0)** (2.3.4.4b) (≥ 20) (Am)	**(1)** (2.3.4.4b) (≥ 20)	**(2)** (<>2.3.2.1*,<>2.3.4.4b) (≤ 17) (Afr,As)	**(3)** (2.3.2.1a) (≥ 10) (Afr,As)	**(4)** (2.3.2.1*) (≥ 08) (Afr,As)

Table S3: **Mnemonic cluster names.** These are formed according to the characteristics observed in the column "Extracted feature" of Table S2.

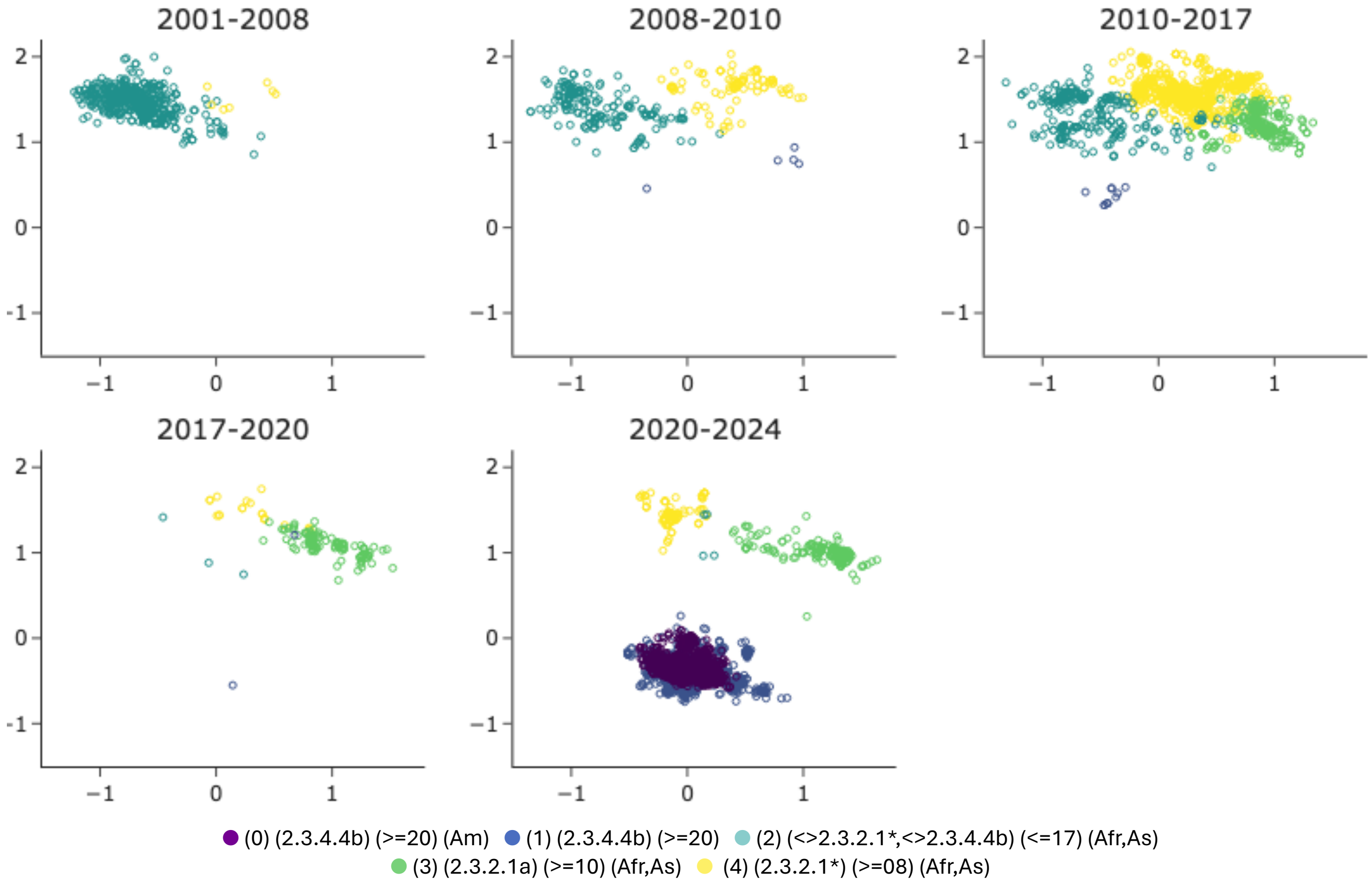

Fig. S4: **Codon profiles of clusters (H5N1, wild birds).** The codon profile of the H5N1 sequences is projected on PC2 (x-axis) and PC1 (y-axis). The scatter plot is organized by flu seasons (see Table S2). Each sequence is colored according to the assigned cluster.

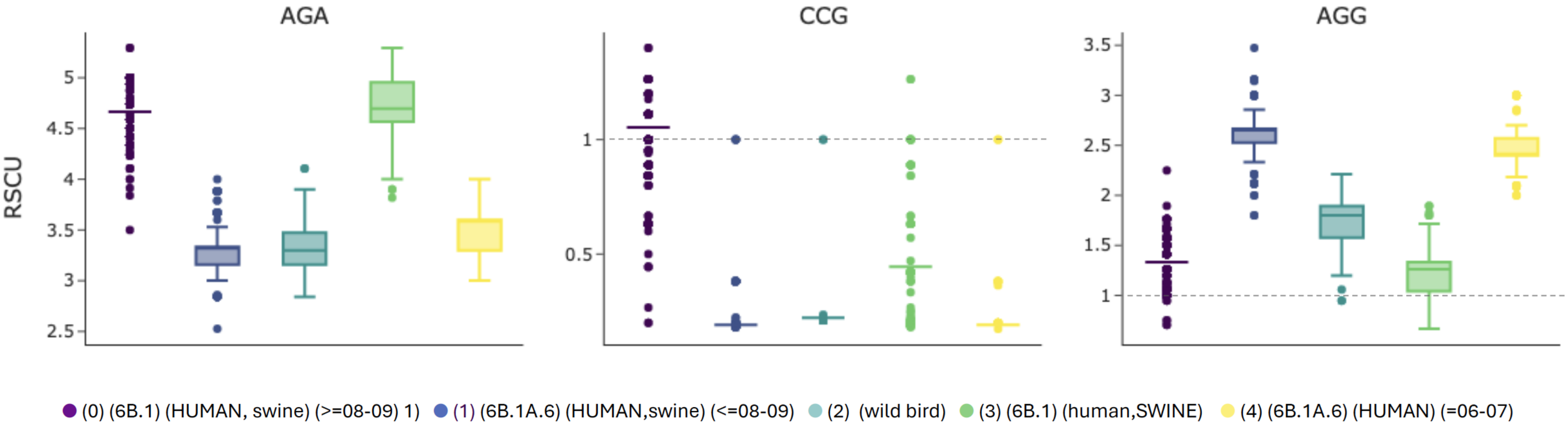

Fig. S5: **Codons separating clusters, boxplot (H1N1).** Distribution of RSCUs for codons AGA, CCG, and AGG (i.e., codons for which RSCU distribution mostly differs between the 5 clusters). The dashed line corresponds to the neutral value of RSCU, i.e., when the observed frequency agrees with the hypothesis that all the synonymous codons for the same amino acids are used equally.

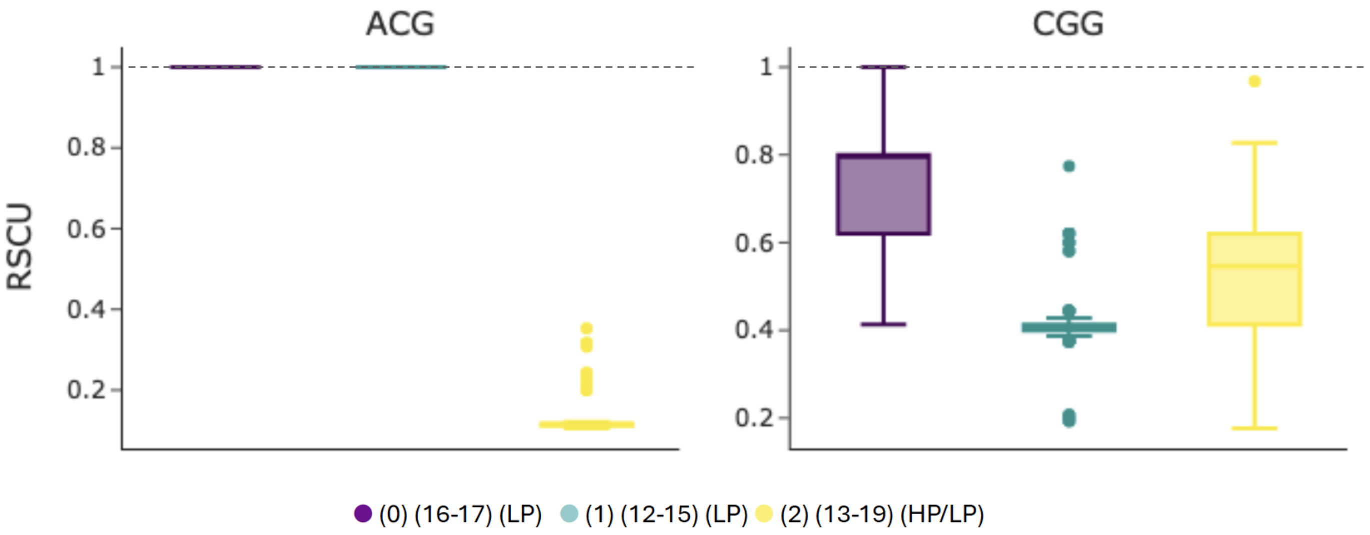

Fig. S6: **Codons separating clusters, boxplot (H7N9).** Distribution of RSCUs for codons ACG and CGG. These codons are the ones for which RSCU distribution mostly differs among the 3 clusters. The dashed line corresponds to the neutral value of RSCU, i.e., when the observed frequency agrees with the hypothesis that all the synonymous codons for the same amino acids are used equally

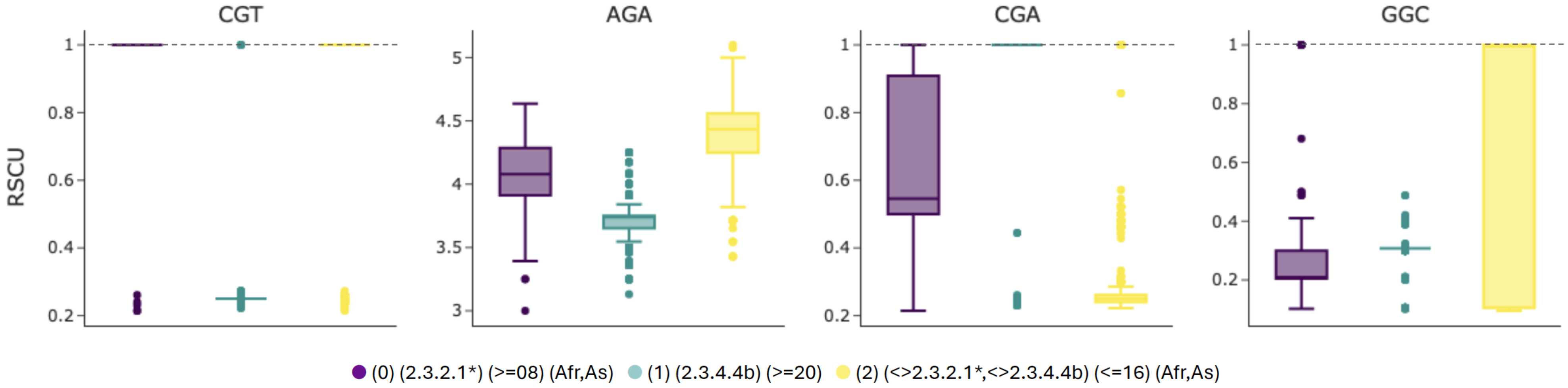

Fig. S7: **Codons separating clusters, boxplot (H5N1, domestic birds)** Distribution of RSCUs for codons CGT, AGA, CGA, and GGC, for which RSCU distribution mostly differs among the 3 clusters. The dashed line corresponds to the neutral value of RSCU, i.e., when the observed frequency agrees with the hypothesis that all the synonymous codons for the same amino acids are used equally

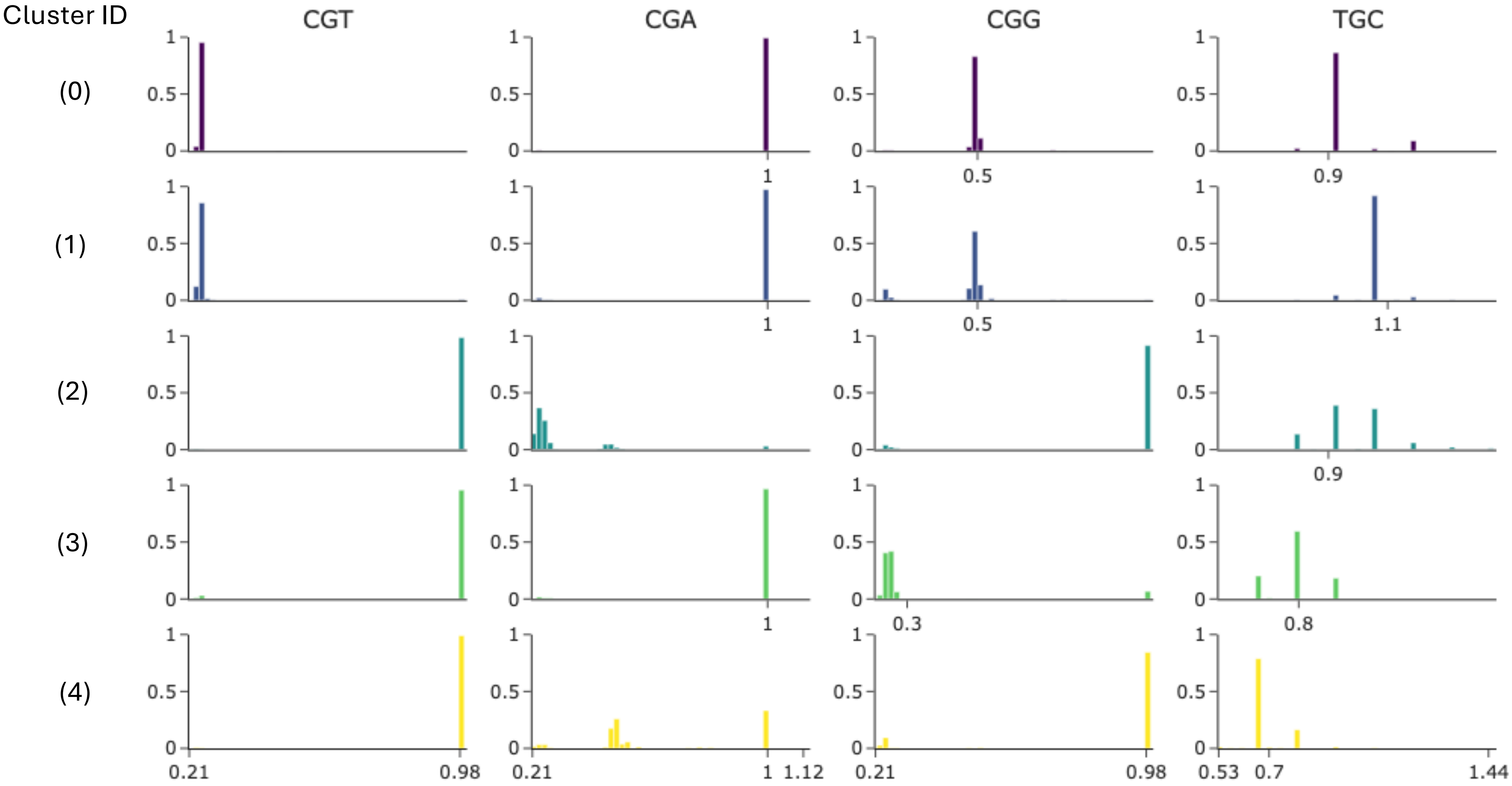

Fig. S8: **Codons separating clusters (H5N1, wild birds).** A multinomial logistic regression trained to recognize the five clusters based on the RSCU usage was used to rank the most important codons for the task of distinguishing the clusters. Here we report the four codons with the highest weight assigned by the classifier. In particular, we notice that the codon usage of CGT is largely different between the group of clusters 0,1 and 2,3,4. Clusters 2 and 3 can be distinguished for the usage of codon CGA; clusters 3 and 4 for that of CGG; and finally, clusters 0 and 1 for that of TGC.

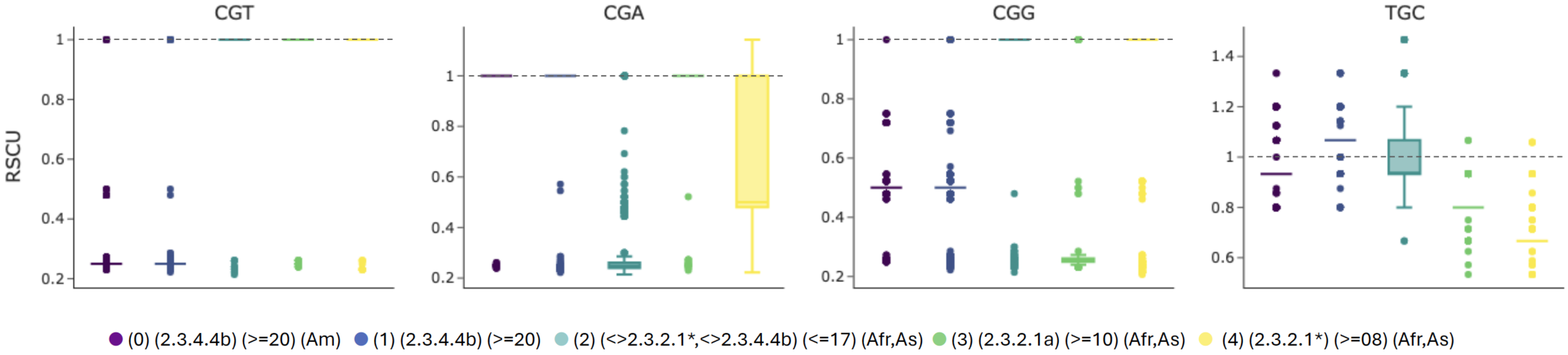

Fig. S9: **Codons separating clusters, boxplot (H5N1, wild birds)** The Figure shows the distribution of RSCUs for four codons CGT, CGA, CGG, and TGC, for which RSCU distribution mostly differs among the 5 clusters. The dashed line corresponds to the neutral value of RSCU, i.e., when the observed frequency agrees with the hypothesis that all the synonymous codons for the same amino acids are used equally.