Supplemental information

1. **Supplemental information presents detailed information of supplemental results, supplemental Methods, supporting materials and programming scripts for supplemental information.**
2. **Supplemental results include the description about Supplemental tables, the legends of supplemental Figures, Supplemental tables and supporting data.**
3. **Supplemental Methods described in details about Feature Engineering and sequence resampling.**
4. **Supplemental Figures were uploaded as independent files. Supplemental Figure 19 with high resolution was provided at https://github.com/Jamalijama/Predict_IAV_Host.**
5. **Supporting data for figures, supplemental figures and programming scripts were also provided respectively.**

Supplemental results

1. Supplemental Tables

Supplemental table 1. List of the full-length IAV coding sequences within the length range.

Sequence samples with the labels of Host, Subtype and Segment were listed, post the dropout of 8,634 sequences, due to the length range or the repeated sequence IDs. The length range was set as mean ± 3 * std (2280 ± 9, 2274 ± 9, 2151 ± 9, 1695 ± 27, 1497 ± 9 and 1380 ± 33, respectively for PB2, PB1, PA, HA, NP and NA).

Supplemental Table 2-7, cv_score and its rolling mean for ML models for 6 segments.

The cv_score and its rolling mean (moving average) 3 (MA3) were listed for the model of GBRT, MLP, RFC and SVC respectively. For PB1 with MLP model, the second downcross of cv_score with its MA3 (at the 11th feature number) was designated as the threshold. In another word, the best feature number indicated by the MLP model for PB1 was 10. For all models for the six segments, the cv_score and its MA3 value were listed respectively.

2. Legends for supplemental Figures

Supplemental Figure 1. Numbers of the full-length IAV coding sequences from different countries/areas, hosts, subtypes, segments and years.

List of all the full-length influenza A virus (IAV) coding sequences since December 31st, 2018. Samples from different countries/areas (A), hosts (B), subtypes (C), segments (D) and years (E) were counted and presented as histograms. Values were sorted on a descending turn, and the y-axis was set with logarithmic tick for figure subpart A and E, with linear tick for others.

Supplemental Figure 2. Distribution of the IAV sequences, post a random resampling, in the labels of countries/areas, hosts, subtypes, segments and years.

A random resampling was performed to keep an approximate sample ratio of 1:1 for the country of the USA and China. Samples from different countries/areas (A), hosts (B), subtypes (C), segments (D) and years (E) were counted and presented as histograms. Values were sorted with a descending turn, and the y-axis was set with logarithmic tick for figure subpart A and E, with linear tick for others.

Supplemental Figure 3. Heatmap and hierarchical clustering of randomly-sampled human and avian IAV sequences basing on the Euclidean distance of the 60 (d)nts.

59-61 sequence samples were randomly (random state = 1) selected from each segment sequence set (3.59‰ to 5.01‰ of total sequences), and then were clustered with heatmap and hierarchical clustering for PB2 (A) and the other 5 segments (B-F), based on the Euclidean distance of the 48 di-nucleotides and the 12 mono-nucleotides respectively; Sequence identity and (d)nts were clustered respectively. Standardized scaling was performed for data with the function of (x-x.mean)/x.std. Color in the heatmap presented the value for each (d)nt in x-axis, as showing by the color bar in the left-top. The hierarchical relationships for the sampled sequences and for (d)nts were respectively indicated in the left and top side in each image. The red-blue column in the left of heatmap was utilized to show the human (red) and avian (blue) group.

Supplemental Figure 4. Phylogenetic analysis of randomly-sampled IAV sequences with maximum likelihood method.

59-61 sequence samples were randomly (random state = 1) selected from each segment sequence set (3.59‰ to 5.01‰ of total sequences), and then were utilized for phylogenetic analysis with MEGA software (MEGA 7.0.26), for PB2 (A), PB1 (B), PA (C), HA (D), NP (E) and NA (F). The sequence ID was indicated as segment, subtype and the strain name from left to right, respectively; the slash "/" in strain name was automatically replaced with a blank by MEGA software.

Supplemental Figure 5. PCA analysis of the 60 (d)nts between human and avian IAV sequences

The 48 dnts and the 12 nts for PB2 (A), PB1 (B), PA (C), HA (D), NP (E) or NA (F) were converted into two principal components and then were plotted with pairplot (seaborn package, python) (left-down and right-up in each figure subpart). The distribution of principal component 1 (PCA_1) and 2 (PCA_2) of avian (blue) and human (orange) sequences was indicated by kernel density estimation (KDE) (left-up and right-down in each figure subpart), and the separability between avian and human sequences was shown respectively for the six segments (A-F) , with the pairplot and KDE.

Supplemental Figure 6. Sampling times for each of the 60 (d)nts for the PCA / SVC optimizer for each segment.

Characterization of human adaption-associated nucleotide composition of IAVs from the 60 (d)nts was performed with combined PCA and SVC. Independent performing times for each (d)nt for the six segments (A-F) in the 3540 iterations of PCA/SVC.

Supplemental Figure 7. Sorting of the 60 (d)nts by the PCA / SVC optimizer for each segment.

3540 iterations of PCA/SVC were performed with randomly-selected four of the 60 (d)nts reduced into one component classify avian and human IAV sequences. The importance of each (d)nt was sorted according to their area under curve (AUC) score of PCA/SVC (A-F).

Supplemental Figure 8. Difference in the PCA/SVC-optimized (d)nts between avian and human IAV segment sequences.

The relative levels of the 9-13 optimized (d)nts for avian (A) and human (H) sequences were plotted with boxplot, for PB2 (A), PB1 (B), PA (C), HA (D), NP (E) and NA (F). The top whisker, the top boarder, the middle line, the bottom boarder and the bottom whisker were respectively presented the maximum value, 75%-, 50%- and 25%- quantile values and the minimum value, and in which outliers were indicated as diamonds.

Supplemental Figure 9. PCA analysis of the optimized (d)nts for PA and HA between human and avian IAV sequences

The optimized 11 and 13 (d)nts for PA (A) and HA (B), respectively, were converted into two principal components and then were plotted with pairplot (seaborn package, python) (left-down and right-up in each figure subpart). The distribution of principal component 1 (PCA_1) and 2 (PCA_2) of avian (blue) and human (orange) sequences was indicated by kernel density estimation (KDE) (left-up and right-down in each figure subpart), and the separability between avian and human sequences was shown respectively for PA (A) and HA (B), with the pairplot and KDE.

Supplemental Figure 10. PCA analysis of the optimized (d)nts for NP and NA between human and avian IAV sequences

The optimized 10 and 9 (d)nts for NP (A) and NA (B), respectively, were converted into two principal components and then were plotted with pairplot (seaborn package, python) (left-down and right-up in each figure subpart). The distribution of principal component 1 (PCA_1) and 2 (PCA_2) of avian (blue) and human (orange) sequences was indicated by kernel density estimation (KDE) (left-up and right-down in each figure subpart), and the separability between avian and human sequences was shown respectively for NP (A) and NA (B), with the pairplot and KDE.

Supplemental Figure 11-15. Heatmap and hierarchical clustering of human and avian IAV sequences basing on the Euclidean distance of the optimized (d)nts.

59-61 sequence samples were randomly (random state = 1) selected from PB1, PA, HA, NP and NA (respectively for **Supplemental** Figure 11-15) and then were clustered with heatmap and hierarchical clustering, based on the Euclidean distance of the optimized 12, 11, 13, 10 and 9 (d)nts, respectively for PB1, PA, HA, NP and NP; Sequence identity and (d)nts were clustered respectively. Standardized scaling was performed for data with the function of (x-x.mean)/x.std. Color in the heatmap presented the value for each (d)nt in x-axis, as showing by the color bar in the left-top. The hierarchical relationships for the sampled sequences and for (d)nts were respectively indicated in the left and top side in each image. The red-blue column in the left of heatmap was utilized to show the human (red) and avian (blue) group.

Supplemental Figure 16-18. The prediction of human adaption classes (True/False) and the human adaption probability by the GBRT, MLP or RFC model, with optimized (d)nts for the six segments.

The human adaption classes (True/False) and the human adaption probability of avian and human sequences were predicted by SVC with the optimized (best) 9, 12, 11, 13, 10 and 9 (d)nts respectively for PB2, PB1, PA, HA, NP and NP, with same optimized-(d)nt number of tail (worst) (d)nts as control, respectively. The confusion matrix of human adaption class prediction, the Receiver Operating Characteristic (ROC) and Area Under ROC Curve (AUC) for the GBRT (**Supplemental** Figure 16), MLP (**Supplemental** Figure 17) or RFC (**Supplemental** Figure 18), model with the worst or with the best (d)nts were indicated respectively for PB2 (A), PB1 (B), PA (C), HA (D), NP (E) and NA (F).

Supplemental Figure 19. Heatmap and hierarchical clustering of randomly-sampled IAV sequences before with pd09H1N1 sequences basing on the Euclidean distance of the 60 (d)nts.

1000 IAV sequences were randomly-resampled (random state = 1) from the IAV sequences before 2009 for each segment, and then were clustered with the pd09H1N1 sequences by the heatmap and hierarchical clustering methods for PB2 (A), PB1 (B), PA (C), HA (D), NP (E) and NA (F), respectively, based on the Euclidean distance of the optimized 9, 12, 11, 13, 10, 9 (d)nts respectively. H1N1 IAV virus strains isolated on April, 2009 in USA were taken as example sequences. The labels of sequence ID, host, subtype, year, country/area and human-adaption probability were isolated from the sequence name and were indicted as a mixed sequence ID in the Heatmap and hierarchical clustering. Host for all sequences was also indicated as blue (avian or human), green (swine), red (avian or human) and white (pd09H1N1) respectively.

3. **Supplemental tables**

Supplemental table 1. Numbers of the sequence with the label of segment, host and subtype

Host	**Subtypes**	**Sequence number for each segment**						**Total**	**Total**
		PB2	**PB1**	**PA**	**HA**	**NP**	**NA**		
Avian	H1N1	387	380	354	275	328	363	2,087	68,739
	H3N2	223	225	230	198	199	207	1,282	
	Others	12,236	11,993	11,814	11,685	10,478	7,164	65,370	
Human	H1N1	7,274	7,043	7,379	13,426	5,949	7,960	49,031	113,820
	H3N2	9,400	9,343	8,975	13,956	7,948	11,826	61,448	
	Others	505	470	502	776	488	600	3,341	
Swine	H1N1	1,746	1,820	1,687	3,579	1,832	3,409	14,073	34,990
	H3N2	1,293	1,303	1,212	2,336	1,274	2,203	9,621	
	Others	1,340	1,298	1,269	3,198	1,321	2,870	11,296	
Total_Avian	H1N1	12,846	12,598	12,398	12,158	11,005	7,734	68,739	217,549
Total_Human	H3N2	17,179	16,856	16,856	28,158	14,385	20,386	113,820	
Total_Others	Others	4,379	4,421	4,168	9,113	4,427	8,482	34,990	
Total	/	34,404	33,875	33,422	49,429	29,817	36,602	217,549	/

Supplemental Table 2 Cross_validation score and its moving average level for PB2 genomic sequences by Gradient Boosted Regression Trees (GBRT), Multiple Layer Perception Classifier (MLP), Random Forest Classifier (RFC) and support vector classifier (SVC).

	GBRT		MLP		RFC		SVC	
(d)nt_num	cv_score	MA3_cv_score	cv_score	MA3_cv_score	cv_score	MA3_cv_score	cv_score	MA3_cv_score
0	0.872	0.872	0.839	0.839	0.872	0.872	0.571	0.571
1	0.964	0.918	0.941	0.89	0.963	0.918	0.891	0.731
2	0.979	0.938	0.961	0.914	0.98	0.939	0.865	0.776
3	0.992	0.978	0.983	0.962	0.993	0.979	0.913	0.89
4	0.993	0.988	0.977	0.974	0.993	0.989	0.934	0.904
5	0.994	0.993	0.979	0.98	0.995	0.994	0.971	0.939
6	0.995	0.994	0.978	0.978	0.995	0.994	0.966	0.957
7	0.995	0.994	0.984	0.98	0.995	0.995	0.966	0.968
8	0.995	0.995	0.967	0.977	0.995	0.995	0.967	0.966
9	0.995	0.995	0.986	0.979	0.996	0.995	0.969	0.967
10	0.995	0.995	0.988	0.98	0.996	0.995	0.972	0.969
11	0.995	0.995	0.989	0.988	0.995	0.995	0.976	0.972
12	0.995	0.995	0.991	0.989	0.996	0.995	0.972	0.973
13	0.995	0.995	0.987	0.989	0.996	0.995	0.974	0.974
14	0.995	0.995	0.99	0.989	0.995	0.995	0.974	0.973
15	0.995	0.995	0.987	0.988	0.995	0.995	0.972	0.973
16	0.995	0.995	0.99	0.989	0.995	0.995	0.989	0.978
17	0.995	0.995	0.99	0.989	0.995	0.995	0.989	0.983
18	0.995	0.995	0.988	0.99	0.995	0.995	0.989	0.989
19	0.994	0.994	0.988	0.989	0.996	0.995	0.989	0.989
20	0.995	0.994	0.989	0.989	0.996	0.996	0.99	0.989
21	0.995	0.994	0.989	0.989	0.996	0.996	0.99	0.99
22	0.995	0.995	0.99	0.99	0.996	0.996	0.99	0.99
23	0.995	0.995	0.99	0.99	0.995	0.996	0.99	0.99
24	0.995	0.995	0.99	0.99	0.996	0.996	0.99	0.99
25	0.995	0.995	0.992	0.99	0.995	0.996	0.99	0.99
26	0.995	0.995	0.989	0.99	0.995	0.996	0.99	0.99
27	0.995	0.995	0.989	0.99	0.996	0.996	0.99	0.99
28	0.995	0.995	0.989	0.989	0.995	0.995	0.99	0.99
29	0.995	0.995	0.988	0.989	0.996	0.996	0.99	0.99
30	0.995	0.995	0.99	0.989	0.996	0.996	0.989	0.99
31	0.995	0.995	0.99	0.99	0.996	0.996	0.99	0.99
32	0.995	0.995	0.989	0.99	0.996	0.996	0.99	0.99
33	0.994	0.995	0.99	0.99	0.996	0.996	0.991	0.99
34	0.995	0.995	0.99	0.989	0.995	0.996	0.99	0.99
35	0.995	0.995	0.99	0.99	0.996	0.996	0.99	0.99
36	0.995	0.995	0.993	0.991	0.996	0.996	0.991	0.991

37	0.995	0.995	0.992	0.992	0.996	0.996	0.991	0.991
38	0.995	0.995	0.99	0.992	0.996	0.996	0.991	0.991
39	0.995	0.995	0.993	0.992	0.996	0.996	0.991	0.991
40	0.996	0.995	0.991	0.991	0.996	0.996	0.991	0.991
41	0.996	0.996	0.993	0.992	0.995	0.996	0.991	0.991
42	0.994	0.995	0.987	0.99	0.996	0.996	0.991	0.991
43	0.996	0.995	0.991	0.99	0.996	0.996	0.991	0.991
44	0.995	0.995	0.99	0.989	0.996	0.996	0.991	0.991
45	0.995	0.995	0.99	0.99	0.996	0.996	0.991	0.991
46	0.996	0.995	0.994	0.991	0.996	0.996	0.991	0.991
47	0.995	0.995	0.991	0.991	0.996	0.996	0.991	0.991
48	0.995	0.995	0.992	0.992	0.996	0.996	0.991	0.991
49	0.996	0.995	0.994	0.992	0.995	0.996	0.991	0.991
50	0.995	0.995	0.992	0.993	0.996	0.996	0.991	0.991
51	0.995	0.995	0.991	0.992	0.996	0.996	0.991	0.991
52	0.995	0.995	0.993	0.992	0.996	0.996	0.991	0.991
53	0.994	0.995	0.989	0.991	0.996	0.996	0.992	0.991
54	0.995	0.995	0.992	0.991	0.996	0.996	0.993	0.992
55	0.992	0.994	0.992	0.991	0.996	0.996	0.993	0.992
56	0.992	0.993	0.989	0.991	0.995	0.996	0.993	0.993
57	0.992	0.992	0.994	0.992	0.996	0.996	0.993	0.993
58	0.994	0.993	0.991	0.991	0.996	0.996	0.993	0.993
59	0.992	0.993	0.992	0.992	0.996	0.996	0.993	0.993

Supplemental Table 3 Cross_validation score and its moving average level for PB1 genomic sequences by GBRT, MLP, RFC and SVC.

	GBRT		MLP		RFC		SVC	
(d)nt_num	cv_score	MA3_cv_score	cv_score	MA3_cv_score	cv_score	MA3_cv_score	cv_score	MA3_cv_score
0	0.912	0.912	0.88	0.88	0.92	0.92	0.878	0.878
1	0.97	0.941	0.963	0.922	0.971	0.946	0.931	0.904
2	0.981	0.954	0.973	0.939	0.983	0.958	0.934	0.914
3	0.984	0.978	0.972	0.969	0.985	0.98	0.935	0.933
4	0.984	0.983	0.976	0.974	0.988	0.985	0.939	0.936
5	0.985	0.984	0.976	0.975	0.989	0.987	0.939	0.938
6	0.985	0.985	0.971	0.974	0.989	0.988	0.962	0.947
7	0.987	0.985	0.973	0.973	0.99	0.989	0.965	0.955
8	0.989	0.987	0.98	0.974	0.992	0.99	0.975	0.967
9	0.994	0.99	0.987	0.98	0.993	0.992	0.984	0.975
10	0.994	0.992	0.98	0.982	0.995	0.993	0.982	0.98
11	0.995	0.994	0.989	0.985	0.994	0.994	0.987	0.984
12	0.994	0.995	0.988	0.986	0.994	0.995	0.987	0.985
13	0.995	0.995	0.99	0.989	0.994	0.994	0.988	0.987
14	0.995	0.995	0.99	0.989	0.995	0.994	0.988	0.988

15	0.995	0.995	0.99	0.99	0.995	0.994	0.988	0.988
16	0.995	0.995	0.991	0.99	0.995	0.995	0.988	0.988
17	0.995	0.995	0.99	0.99	0.995	0.995	0.988	0.988
18	0.996	0.995	0.99	0.99	0.995	0.995	0.987	0.988
19	0.996	0.995	0.99	0.99	0.995	0.995	0.988	0.988
20	0.995	0.995	0.989	0.99	0.995	0.995	0.988	0.988
21	0.995	0.995	0.99	0.99	0.994	0.995	0.989	0.988
22	0.995	0.995	0.992	0.99	0.995	0.995	0.989	0.989
23	0.996	0.995	0.992	0.991	0.994	0.995	0.988	0.989
24	0.995	0.995	0.994	0.993	0.995	0.995	0.988	0.988
25	0.996	0.996	0.996	0.994	0.995	0.995	0.988	0.988
26	0.995	0.996	0.993	0.994	0.995	0.995	0.988	0.988
27	0.996	0.996	0.995	0.995	0.994	0.995	0.988	0.988
28	0.996	0.996	0.995	0.994	0.994	0.995	0.99	0.989
29	0.996	0.996	0.993	0.995	0.996	0.995	0.99	0.989
30	0.996	0.996	0.994	0.994	0.995	0.995	0.99	0.99
31	0.996	0.996	0.995	0.994	0.996	0.995	0.99	0.99
32	0.996	0.996	0.995	0.995	0.995	0.995	0.99	0.99
33	0.996	0.996	0.996	0.995	0.996	0.995	0.99	0.99
34	0.996	0.996	0.992	0.994	0.995	0.995	0.99	0.99
35	0.996	0.996	0.996	0.995	0.995	0.995	0.99	0.99
36	0.996	0.996	0.993	0.994	0.995	0.995	0.99	0.99
37	0.996	0.996	0.994	0.994	0.996	0.995	0.99	0.99
38	0.996	0.996	0.996	0.994	0.996	0.995	0.99	0.99
39	0.996	0.996	0.99	0.993	0.995	0.995	0.99	0.99
40	0.996	0.996	0.993	0.993	0.995	0.995	0.99	0.99
41	0.996	0.996	0.994	0.992	0.995	0.995	0.99	0.99
42	0.996	0.996	0.994	0.994	0.995	0.995	0.991	0.99
43	0.996	0.996	0.995	0.995	0.996	0.995	0.991	0.99
44	0.997	0.996	0.993	0.994	0.996	0.996	0.99	0.991
45	0.997	0.996	0.995	0.995	0.996	0.996	0.99	0.991
46	0.996	0.997	0.995	0.994	0.996	0.996	0.991	0.991
47	0.996	0.996	0.996	0.995	0.995	0.996	0.991	0.991
48	0.995	0.996	0.993	0.994	0.996	0.996	0.991	0.991
49	0.995	0.996	0.996	0.995	0.996	0.996	0.991	0.991
50	0.995	0.995	0.993	0.994	0.995	0.996	0.991	0.991
51	0.996	0.996	0.994	0.994	0.996	0.996	0.991	0.991
52	0.996	0.996	0.995	0.994	0.996	0.996	0.991	0.991
53	0.996	0.996	0.992	0.994	0.995	0.996	0.991	0.991
54	0.996	0.996	0.995	0.994	0.996	0.996	0.992	0.991
55	0.996	0.996	0.993	0.993	0.996	0.996	0.992	0.992
56	0.995	0.996	0.996	0.994	0.996	0.996	0.992	0.992
57	0.996	0.996	0.992	0.994	0.996	0.996	0.992	0.992
58	0.995	0.995	0.993	0.994	0.996	0.996	0.992	0.992

59	0.996	0.996	0.996	0.994	0.996	0.996	0.992	0.992

* PB1, mlp,cv_score,rolling, amended value = 10

Supplemental Table 4 Cross_validation score and its moving average level for PA genomic sequences by GBRT, MLP, RFC and SVC.

	GBRT		MLP		RFC		SVC	
(d)nt_num	cv_score	MA3_cv_score	cv_score	MA3_cv_score	cv_score	MA3_cv_score	cv_score	MA3_cv_score
0	0.972	0.972	0.972	0.972	0.972	0.972	0.573	0.573
1	0.974	0.973	0.978	0.975	0.975	0.973	0.573	0.573
2	0.989	0.979	0.957	0.969	0.987	0.978	0.787	0.644
3	0.99	0.984	0.937	0.957	0.989	0.984	0.854	0.738
4	0.989	0.989	0.963	0.952	0.989	0.989	0.855	0.832
5	0.987	0.989	0.94	0.947	0.99	0.989	0.843	0.851
6	0.987	0.988	0.939	0.947	0.99	0.99	0.865	0.854
7	0.988	0.987	0.96	0.946	0.991	0.99	0.933	0.88
8	0.987	0.988	0.963	0.954	0.99	0.99	0.934	0.91
9	0.99	0.989	0.966	0.963	0.992	0.991	0.929	0.932
10	0.988	0.989	0.966	0.965	0.991	0.991	0.95	0.938
11	0.988	0.989	0.963	0.965	0.991	0.991	0.942	0.94
12	0.989	0.989	0.974	0.968	0.992	0.991	0.941	0.945
13	0.991	0.989	0.983	0.973	0.992	0.992	0.96	0.948
14	0.993	0.991	0.984	0.98	0.993	0.993	0.96	0.954
15	0.992	0.992	0.978	0.981	0.992	0.993	0.96	0.96
16	0.993	0.992	0.978	0.98	0.993	0.993	0.96	0.96
17	0.993	0.993	0.981	0.979	0.993	0.993	0.967	0.962
18	0.992	0.993	0.988	0.982	0.993	0.993	0.971	0.966
19	0.993	0.993	0.985	0.984	0.993	0.993	0.972	0.97
20	0.993	0.993	0.987	0.987	0.993	0.993	0.978	0.974
21	0.994	0.993	0.983	0.985	0.994	0.993	0.978	0.976
22	0.994	0.994	0.986	0.985	0.993	0.993	0.98	0.979
23	0.994	0.994	0.988	0.986	0.994	0.994	0.981	0.98
24	0.993	0.994	0.985	0.987	0.993	0.994	0.98	0.98
25	0.992	0.993	0.988	0.987	0.994	0.994	0.978	0.98
26	0.994	0.993	0.978	0.984	0.994	0.994	0.976	0.978
27	0.993	0.993	0.984	0.984	0.994	0.994	0.977	0.977
28	0.992	0.993	0.986	0.983	0.994	0.994	0.977	0.977
29	0.993	0.993	0.983	0.984	0.994	0.994	0.979	0.977
30	0.992	0.992	0.995	0.988	0.994	0.994	0.987	0.981
31	0.992	0.992	0.991	0.99	0.994	0.994	0.987	0.984
32	0.992	0.992	0.993	0.993	0.994	0.994	0.987	0.987
33	0.993	0.992	0.994	0.993	0.994	0.994	0.989	0.988
34	0.992	0.992	0.994	0.994	0.993	0.994	0.989	0.988
35	0.992	0.993	0.988	0.992	0.994	0.994	0.99	0.989

36	0.992	0.992	0.995	0.992	0.994	0.994	0.99	0.989
37	0.992	0.992	0.989	0.991	0.994	0.994	0.991	0.99
38	0.992	0.992	0.994	0.993	0.994	0.994	0.991	0.991
39	0.992	0.992	0.992	0.991	0.994	0.994	0.991	0.991
40	0.992	0.992	0.994	0.993	0.994	0.994	0.991	0.991
41	0.992	0.992	0.994	0.993	0.994	0.994	0.991	0.991
42	0.992	0.992	0.994	0.994	0.994	0.994	0.993	0.992
43	0.991	0.992	0.991	0.993	0.994	0.994	0.993	0.992
44	0.991	0.991	0.995	0.993	0.994	0.994	0.993	0.993
45	0.992	0.992	0.994	0.993	0.994	0.994	0.992	0.993
46	0.991	0.992	0.993	0.994	0.994	0.994	0.993	0.993
47	0.992	0.992	0.994	0.994	0.995	0.995	0.993	0.993
48	0.991	0.991	0.994	0.994	0.994	0.994	0.993	0.993
49	0.99	0.991	0.994	0.994	0.994	0.994	0.993	0.993
50	0.99	0.991	0.993	0.994	0.994	0.994	0.993	0.993
51	0.991	0.991	0.995	0.994	0.994	0.994	0.993	0.993
52	0.992	0.991	0.995	0.994	0.994	0.994	0.993	0.993
53	0.992	0.992	0.993	0.994	0.994	0.994	0.993	0.993
54	0.992	0.992	0.992	0.993	0.994	0.994	0.993	0.993
55	0.991	0.992	0.995	0.993	0.994	0.994	0.993	0.993
56	0.991	0.991	0.995	0.994	0.994	0.994	0.993	0.993
57	0.992	0.991	0.993	0.994	0.994	0.994	0.993	0.993
58	0.992	0.991	0.994	0.994	0.994	0.994	0.994	0.993
59	0.992	0.992	0.996	0.994	0.995	0.994	0.994	0.994

Supplemental Table 5 Cross_validation score and its moving average level for HA genomic sequences by GBRT, MLP, RFC and SVC.

	GBRT		MLP		RFC		SVC	
(d)nt_num	cv_score	MA3_cv_score	cv_score	MA3_cv_score	cv_score	MA3_cv_score	cv_score	MA3_cv_score
0	0.785	0.785	0.647	0.647	0.895	0.895	0.696	0.696
1	0.881	0.833	0.827	0.737	0.935	0.915	0.800	0.748
2	0.960	0.875	0.871	0.781	0.976	0.936	0.814	0.770
3	0.965	0.936	0.888	0.862	0.980	0.964	0.840	0.818
4	0.979	0.968	0.931	0.897	0.987	0.981	0.835	0.829
5	0.987	0.977	0.953	0.924	0.993	0.987	0.898	0.858
6	0.991	0.986	0.962	0.949	0.994	0.991	0.906	0.880
7	0.993	0.990	0.962	0.959	0.995	0.994	0.917	0.907
8	0.994	0.993	0.964	0.963	0.995	0.995	0.916	0.913
9	0.994	0.994	0.979	0.968	0.995	0.995	0.918	0.917
10	0.995	0.994	0.980	0.974	0.995	0.995	0.918	0.917
11	0.995	0.995	0.982	0.980	0.996	0.995	0.943	0.926
12	0.993	0.994	0.985	0.982	0.996	0.995	0.961	0.941
13	0.993	0.994	0.986	0.984	0.997	0.996	0.968	0.957

14	0.993	0.993	0.985	0.985	0.996	0.996	0.968	0.966
15	0.992	0.993	0.989	0.987	0.996	0.996	0.970	0.969
16	0.995	0.993	0.992	0.989	0.996	0.996	0.976	0.971
17	0.995	0.994	0.993	0.992	0.996	0.996	0.978	0.975
18	0.995	0.995	0.989	0.991	0.996	0.996	0.978	0.978
19	0.994	0.995	0.992	0.991	0.997	0.996	0.979	0.978
20	0.995	0.995	0.992	0.991	0.997	0.997	0.981	0.979
21	0.995	0.995	0.990	0.991	0.997	0.997	0.982	0.981
22	0.995	0.995	0.993	0.992	0.997	0.997	0.981	0.982
23	0.995	0.995	0.993	0.992	0.997	0.997	0.982	0.982
24	0.995	0.995	0.994	0.994	0.997	0.997	0.984	0.983
25	0.992	0.994	0.994	0.994	0.997	0.997	0.984	0.983
26	0.996	0.994	0.993	0.994	0.997	0.997	0.984	0.984
27	0.995	0.994	0.993	0.993	0.997	0.997	0.986	0.985
28	0.995	0.995	0.993	0.993	0.997	0.997	0.986	0.985
29	0.995	0.995	0.988	0.991	0.997	0.997	0.986	0.986
30	0.996	0.995	0.993	0.991	0.997	0.997	0.988	0.987
31	0.996	0.996	0.993	0.991	0.997	0.997	0.988	0.987
32	0.996	0.996	0.994	0.993	0.997	0.997	0.989	0.988
33	0.996	0.996	0.994	0.994	0.997	0.997	0.989	0.989
34	0.996	0.996	0.990	0.993	0.996	0.997	0.989	0.989
35	0.996	0.996	0.994	0.993	0.997	0.997	0.990	0.990
36	0.997	0.996	0.993	0.992	0.997	0.997	0.992	0.990
37	0.996	0.996	0.995	0.994	0.997	0.997	0.992	0.991
38	0.997	0.997	0.996	0.995	0.997	0.997	0.993	0.992
39	0.997	0.997	0.996	0.996	0.997	0.997	0.993	0.992
40	0.997	0.997	0.994	0.995	0.997	0.997	0.993	0.993
41	0.997	0.997	0.994	0.995	0.997	0.997	0.993	0.993
42	0.997	0.997	0.995	0.994	0.996	0.997	0.993	0.993
43	0.997	0.997	0.995	0.994	0.997	0.997	0.994	0.993
44	0.997	0.997	0.994	0.995	0.997	0.997	0.993	0.993
45	0.997	0.997	0.995	0.994	0.997	0.997	0.993	0.993
46	0.996	0.997	0.994	0.994	0.997	0.997	0.993	0.993
47	0.996	0.996	0.996	0.995	0.997	0.997	0.993	0.993
48	0.997	0.996	0.995	0.995	0.997	0.997	0.993	0.993
49	0.997	0.997	0.996	0.996	0.997	0.997	0.993	0.993
50	0.996	0.997	0.994	0.995	0.997	0.997	0.993	0.993
51	0.997	0.997	0.996	0.995	0.997	0.997	0.993	0.993
52	0.996	0.996	0.994	0.995	0.997	0.997	0.993	0.993
53	0.996	0.996	0.995	0.995	0.997	0.997	0.993	0.993
54	0.997	0.996	0.995	0.995	0.997	0.997	0.994	0.994
55	0.997	0.997	0.996	0.995	0.997	0.997	0.994	0.994
56	0.997	0.997	0.996	0.996	0.997	0.997	0.994	0.994
57	0.997	0.997	0.995	0.996	0.997	0.997	0.994	0.994

58	0.996	0.996	0.995	0.995	0.997	0.997	0.994	0.994
59	0.997	0.996	0.996	0.995	0.997	0.997	0.994	0.994

Supplemental Table 6 Cross_validation score and its moving average level for NP genomic sequences by GBRT, MLP, RFC and SVC.

	GBRT		MLP		RFC		SVC	
(d)nt_num	cv_score	MA3_cv_score	cv_score	MA3_cv_score	cv_score	MA3_cv_score	cv_score	MA3_cv_score
0	0.737	0.737	0.668	0.668	0.822	0.822	0.657	0.657
1	0.916	0.827	0.829	0.749	0.952	0.887	0.664	0.661
2	0.938	0.864	0.794	0.764	0.966	0.913	0.673	0.665
3	0.974	0.942	0.921	0.848	0.977	0.965	0.616	0.651
4	0.991	0.968	0.967	0.894	0.993	0.979	0.859	0.716
5	0.994	0.986	0.972	0.954	0.995	0.988	0.962	0.812
6	0.995	0.993	0.982	0.974	0.995	0.994	0.982	0.934
7	0.995	0.995	0.993	0.982	0.995	0.995	0.994	0.979
8	0.994	0.994	0.993	0.989	0.995	0.995	0.994	0.99
9	0.993	0.994	0.994	0.993	0.995	0.995	0.995	0.994
10	0.993	0.993	0.994	0.994	0.995	0.995	0.995	0.995
11	0.994	0.993	0.994	0.994	0.995	0.995	0.995	0.995
12	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995
13	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995
14	0.995	0.995	0.994	0.994	0.996	0.995	0.995	0.995
15	0.995	0.995	0.994	0.994	0.996	0.996	0.995	0.995
16	0.994	0.995	0.994	0.994	0.995	0.995	0.995	0.995
17	0.994	0.994	0.993	0.994	0.996	0.995	0.995	0.995
18	0.994	0.994	0.994	0.994	0.996	0.996	0.995	0.995
19	0.994	0.994	0.994	0.994	0.996	0.996	0.996	0.995
20	0.994	0.994	0.994	0.994	0.996	0.996	0.996	0.995
21	0.994	0.994	0.993	0.994	0.996	0.996	0.996	0.996
22	0.994	0.994	0.993	0.993	0.996	0.996	0.996	0.996
23	0.993	0.994	0.992	0.993	0.996	0.996	0.996	0.996
24	0.994	0.994	0.994	0.993	0.996	0.996	0.996	0.996
25	0.994	0.994	0.994	0.994	0.996	0.996	0.995	0.996
26	0.994	0.994	0.993	0.994	0.995	0.996	0.996	0.996
27	0.994	0.994	0.993	0.994	0.996	0.996	0.996	0.996
28	0.994	0.994	0.994	0.994	0.995	0.995	0.996	0.996
29	0.993	0.994	0.995	0.994	0.996	0.996	0.996	0.996
30	0.994	0.994	0.994	0.994	0.995	0.995	0.996	0.996
31	0.993	0.994	0.995	0.994	0.996	0.996	0.996	0.996
32	0.993	0.994	0.994	0.994	0.995	0.995	0.996	0.996
33	0.994	0.994	0.994	0.994	0.995	0.995	0.996	0.996
34	0.994	0.994	0.995	0.994	0.995	0.995	0.995	0.996
35	0.993	0.994	0.994	0.994	0.996	0.995	0.995	0.995

36	0.994	0.994	0.995	0.994	0.996	0.995	0.996	0.996
37	0.993	0.994	0.994	0.994	0.996	0.996	0.996	0.996
38	0.993	0.994	0.995	0.995	0.995	0.996	0.996	0.996
39	0.994	0.993	0.994	0.994	0.996	0.996	0.996	0.996
40	0.994	0.994	0.995	0.994	0.995	0.995	0.996	0.996
41	0.994	0.994	0.995	0.994	0.995	0.995	0.996	0.996
42	0.994	0.994	0.994	0.994	0.995	0.995	0.996	0.996
43	0.993	0.993	0.995	0.994	0.996	0.995	0.996	0.996
44	0.993	0.993	0.994	0.994	0.995	0.995	0.996	0.996
45	0.994	0.993	0.994	0.994	0.996	0.996	0.996	0.996
46	0.992	0.993	0.994	0.994	0.996	0.995	0.996	0.996
47	0.994	0.993	0.993	0.994	0.995	0.995	0.996	0.996
48	0.994	0.993	0.994	0.994	0.995	0.995	0.996	0.996
49	0.994	0.994	0.994	0.994	0.995	0.995	0.996	0.996
50	0.994	0.994	0.994	0.994	0.996	0.995	0.996	0.996
51	0.994	0.994	0.995	0.994	0.995	0.995	0.996	0.996
52	0.993	0.994	0.993	0.994	0.994	0.995	0.996	0.996
53	0.995	0.994	0.993	0.994	0.995	0.995	0.996	0.996
54	0.994	0.994	0.995	0.994	0.995	0.995	0.996	0.996
55	0.995	0.995	0.995	0.994	0.996	0.995	0.996	0.996
56	0.994	0.994	0.996	0.995	0.996	0.995	0.996	0.996
57	0.995	0.995	0.994	0.995	0.995	0.995	0.996	0.996
58	0.994	0.994	0.994	0.995	0.995	0.995	0.996	0.996
59	0.995	0.995	0.995	0.994	0.995	0.995	0.996	0.996

Supplemental Table 7 Cross_validation score and its moving average level for NA genomic sequences by GBRT, MLP, RFC and SVC.

	GBRT		MLP		RFC		SVC	
(d)nt_num	cv_score	MA3_cv_score	cv_score	MA3_cv_score	cv_score	MA3_cv_score	cv_score	MA3_cv_score
0	0.93	0.93	0.909	0.909	0.943	0.943	0.905	0.905
1	0.954	0.942	0.939	0.924	0.962	0.953	0.912	0.908
2	0.984	0.956	0.976	0.941	0.986	0.964	0.973	0.93
3	0.987	0.975	0.978	0.964	0.987	0.979	0.973	0.953
4	0.987	0.986	0.978	0.978	0.988	0.987	0.978	0.975
5	0.992	0.989	0.988	0.981	0.993	0.99	0.988	0.98
6	0.993	0.991	0.989	0.985	0.994	0.992	0.989	0.985
7	0.993	0.993	0.989	0.989	0.994	0.994	0.989	0.988
8	0.993	0.993	0.989	0.989	0.994	0.994	0.989	0.989
9	0.993	0.993	0.988	0.989	0.994	0.994	0.988	0.989
10	0.993	0.993	0.987	0.988	0.994	0.994	0.989	0.989
11	0.994	0.993	0.988	0.988	0.995	0.994	0.989	0.989
12	0.994	0.994	0.991	0.989	0.995	0.995	0.991	0.99
13	0.994	0.994	0.99	0.99	0.994	0.995	0.991	0.99

14	0.994	0.994	0.991	0.991	0.995	0.995	0.991	0.991
15	0.995	0.994	0.992	0.991	0.995	0.995	0.994	0.992
16	0.994	0.994	0.993	0.992	0.995	0.995	0.994	0.993
17	0.994	0.994	0.993	0.993	0.995	0.995	0.994	0.994
18	0.994	0.994	0.99	0.992	0.995	0.995	0.994	0.994
19	0.994	0.994	0.991	0.991	0.995	0.995	0.994	0.994
20	0.994	0.994	0.992	0.991	0.995	0.995	0.994	0.994
21	0.994	0.994	0.994	0.992	0.995	0.995	0.994	0.994
22	0.994	0.994	0.992	0.992	0.995	0.995	0.995	0.994
23	0.995	0.994	0.994	0.993	0.996	0.995	0.995	0.995
24	0.994	0.994	0.993	0.993	0.995	0.995	0.995	0.995
25	0.994	0.994	0.994	0.994	0.995	0.996	0.995	0.995
26	0.994	0.994	0.994	0.994	0.996	0.996	0.995	0.995
27	0.993	0.994	0.993	0.994	0.996	0.996	0.995	0.995
28	0.994	0.994	0.994	0.994	0.996	0.996	0.995	0.995
29	0.994	0.994	0.994	0.994	0.995	0.996	0.995	0.995
30	0.994	0.994	0.995	0.994	0.995	0.995	0.995	0.995
31	0.994	0.994	0.992	0.993	0.995	0.995	0.995	0.995
32	0.995	0.994	0.995	0.994	0.996	0.995	0.995	0.995
33	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995
34	0.994	0.994	0.994	0.994	0.996	0.995	0.995	0.995
35	0.994	0.994	0.995	0.995	0.995	0.995	0.995	0.995
36	0.993	0.994	0.994	0.994	0.996	0.996	0.995	0.995
37	0.995	0.994	0.994	0.994	0.996	0.996	0.995	0.995
38	0.994	0.994	0.993	0.994	0.995	0.996	0.995	0.995
39	0.994	0.994	0.992	0.993	0.995	0.996	0.995	0.995
40	0.994	0.994	0.993	0.993	0.995	0.995	0.995	0.995
41	0.993	0.994	0.993	0.992	0.995	0.995	0.995	0.995
42	0.994	0.993	0.994	0.993	0.996	0.995	0.995	0.995
43	0.994	0.994	0.993	0.993	0.996	0.996	0.995	0.995
44	0.995	0.994	0.991	0.993	0.996	0.996	0.995	0.995
45	0.994	0.994	0.994	0.993	0.996	0.996	0.995	0.995
46	0.994	0.994	0.994	0.993	0.995	0.995	0.995	0.995
47	0.995	0.994	0.993	0.994	0.996	0.996	0.995	0.995
48	0.995	0.995	0.994	0.994	0.997	0.996	0.995	0.995
49	0.995	0.995	0.993	0.993	0.996	0.996	0.995	0.995
50	0.995	0.995	0.994	0.993	0.995	0.996	0.995	0.995
51	0.995	0.995	0.991	0.993	0.996	0.995	0.995	0.995
52	0.995	0.995	0.993	0.993	0.995	0.995	0.995	0.995
53	0.995	0.995	0.993	0.992	0.995	0.996	0.995	0.995
54	0.995	0.995	0.994	0.993	0.996	0.996	0.995	0.995
55	0.995	0.995	0.99	0.992	0.996	0.996	0.995	0.995
56	0.994	0.995	0.991	0.992	0.997	0.996	0.995	0.995
57	0.995	0.995	0.991	0.991	0.996	0.996	0.995	0.995

58	0.995	0.995	0.993	0.992	0.996	0.996	0.995	0.995
59	0.995	0.995	0.992	0.992	0.996	0.996	0.995	0.995

Supplemental Methods

Supplemental information about Feature Engineering and sequence resampling

Supplemental information about Feature Engineering

Feature engineering and feature selection were most important for machine learning analysis. Biologically, there is a species barrier for human and avian influenza viruses, and there should be a linear separability of genotype and phenotype between both groups of viruses. Here, we supposed that the genomic composition of mono- or di-nucleotide is associated with the linear separability. In another word, there should be a hyper plane with a margin between avian and human viruses in genomic composition. We supposed that the human/avian-IAV-separability should be consistently linear and make sense biologically. In this context, support vector classifier (SVC) was the best choice. In the case of SVC, data points are viewed as n-dimensional vectors multiply m-number, and it is to separate such points with a hyperplane with maximum-margin. The nonlinear separators, Gradient Boosted Regression Trees (GBRT), Random Forest Classifier (RFC) and Multiple Layer Perception Classifier (MLP), which are based on neural network (MLP) or decision tree (RFC and GBRT), are grown very deep tend to learn highly irregular patterns, at the expense of a small increase in the bias and some loss of interpretability, let alone the biological separability. However, to avoid over-fitting, we adjusted the optimized (d)nt number of SVC, via averaging it with the optimized (d)nt number with MLP, RFC and GBRT classifiers.

SVC was the optional model. Thus, SVC was used as main supervised machine learning model for both feature selection and sample classification. SVC was used firstly for (d)nt sorting, secondly for (d)nt optimization, along with principal component analysis (PCA), thirdly as train final classifier with the optimized (d)nts. The (d)nt optimization was performed using four types of machine learning approaches, SVC, GBRT, RFC and MLP. methods.

As Supplementary Figure 3 shown, avian and human sequences were not well classified separately with the 60 (d)nt features. Moreover, as compositional information, the 60 (d)nt features were theoretically not independent of each other, and there was a feature redundancy for the 60 (d)nts. Thus, PCA is used to reduce the dimensionality of batches of (d)nt features before SVC analysis for the feature selection. If there was a higher dependence/correlation between/among a batch of (d)nt features, the AUC score of SVC would be lower post dimensionality reduction of (d)nt features by PCA. In addition, it is time-saving for the calculation of only one PCA value, rather a feature matrix.

Theoretically, to identify every possible dependence of (d)nt features, every possible combination of (d)nt features, with various feature number (a combination of m features from n features, 2< = m <= 30,since combination (60,m)= combination (60,(60-m))), should be utilized for the PCA/SVC feature selection. However, it is a huge job to exhaust all combinations. Here, we selected 2*combination (60, 2) (3,540) as sampling times for a random sampling of four features from the 60 features for the feature selection with PCA/SVC. As shown in Supplementary Figure 6, more than 200 times were sampled for each of the 60 features in such process. For each time of PCA/SVC analysis, AUC score was taken as the feature importance value for each of the four sampled features. According to the average (n>200) AUC score, the 60 (d)nt features were sorted.

Finally, SVC, MLP, RFC and GBRT with accumulating (d)nt features were performed again for (d)nt number optimization. The feature list was updated for each round of SVC analysis, with top n (n = n +1 for n in range [1,59]) (d)nt features from the sorted feature list. The 60 AUC score value of the 60 iteration of machine learning analysis were utilized for the final (d)nt number optimization.

Supplemental information about sequence resampling

Resampling was performed via pandas.DataFrame.sample (Python) with a float ratio multiplying the segment sequence number, and the final sequence number was an integral number (the Integral function in python is just removing the float, not same as the Rounding function). Thus, 59-61 segment sequence samples were produced for phylogeny and hierarchical clustering analysis, 46, 042 human-adaptive sequences and 46, 488 human-inadaptive avian sequences were produced for feature extraction and model building, with not the same sample number for avian and human sets.

Supporting data

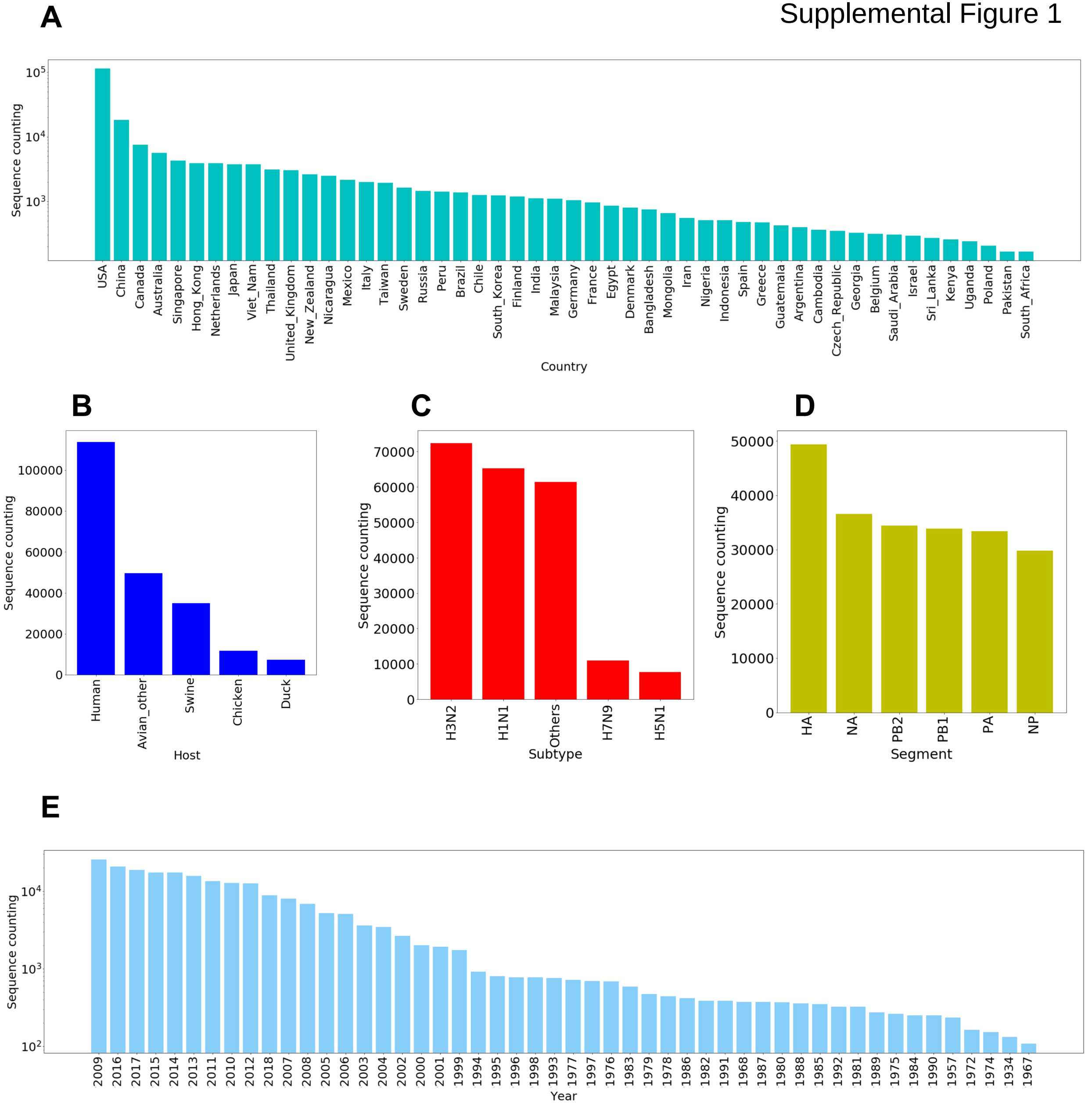
Supporting data includes the sequence ID table, the supporting data for Figures and for supplementary Figures. Supporting data was available online: https://github.com/Jamalijama/Predict_IAV_Host.

Code availability

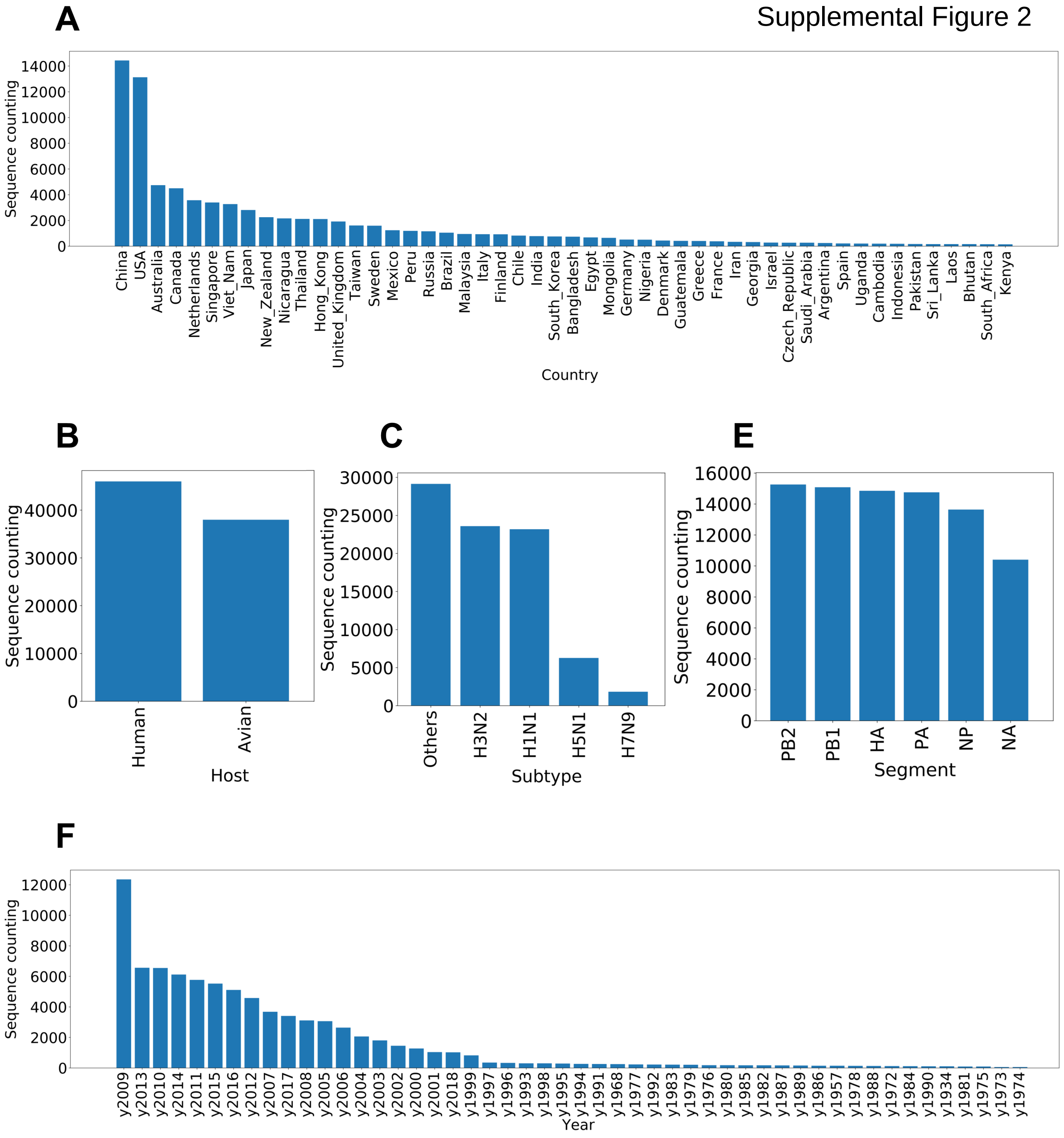
The project code available at following website:

https://github.com/Jamalijama/Predict_IAV_Host.

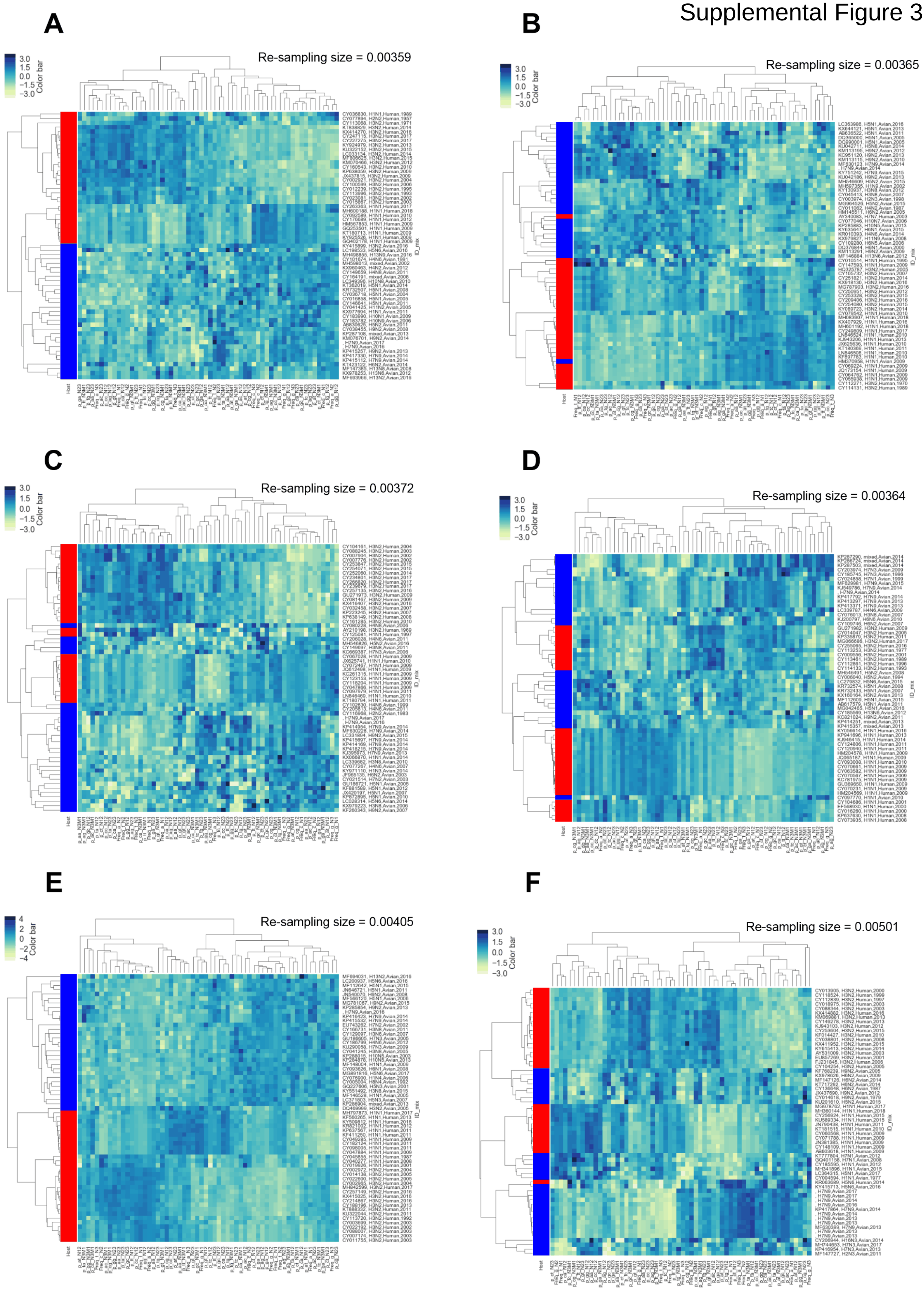
Supplemental Figure 1



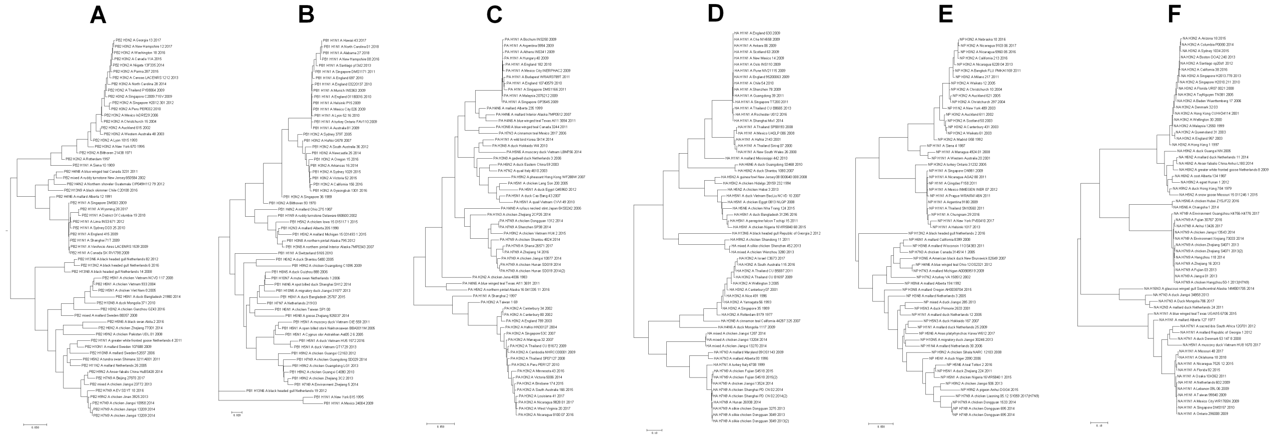
Supplemental Figure 2



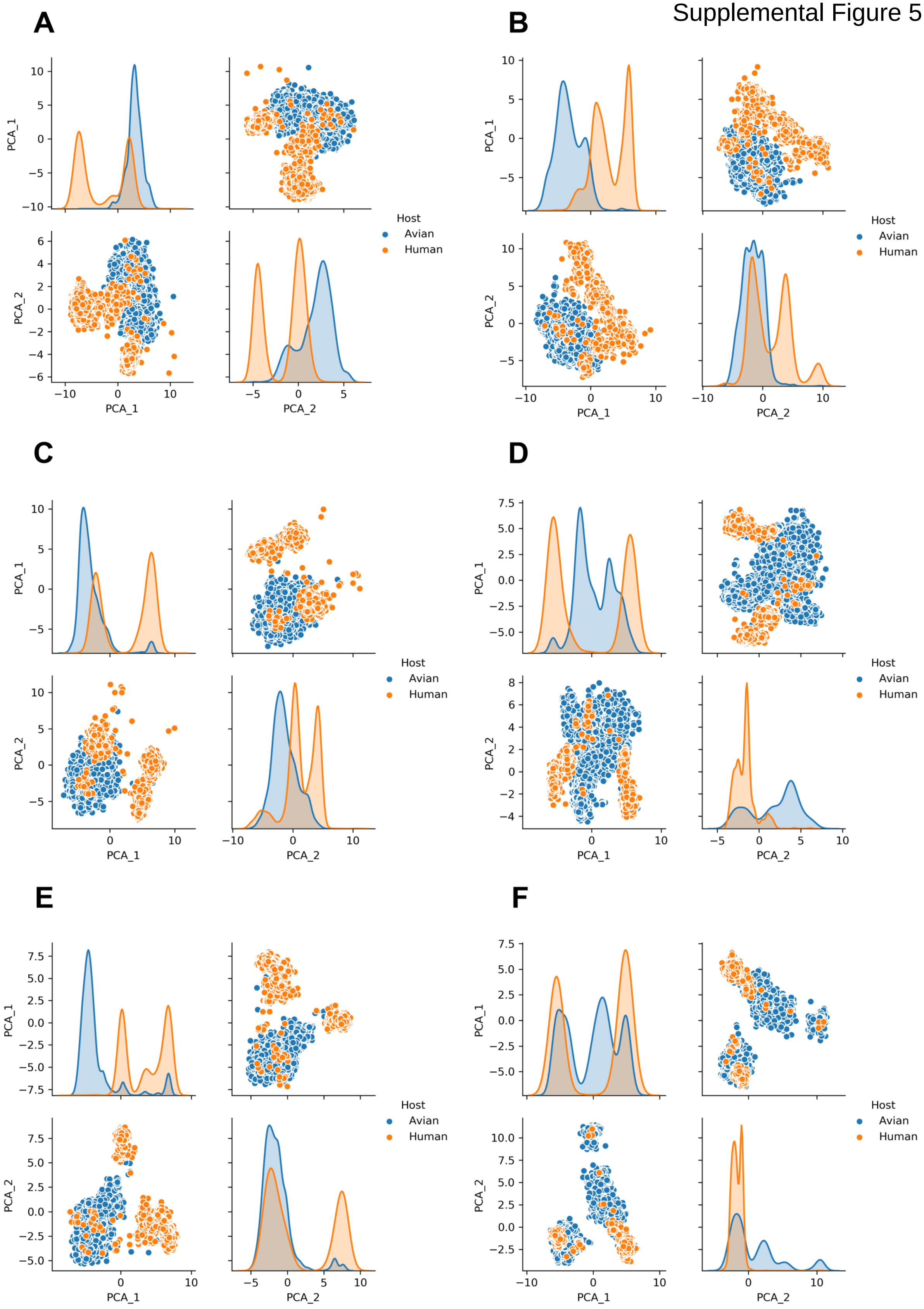
Supplemental Figure 3



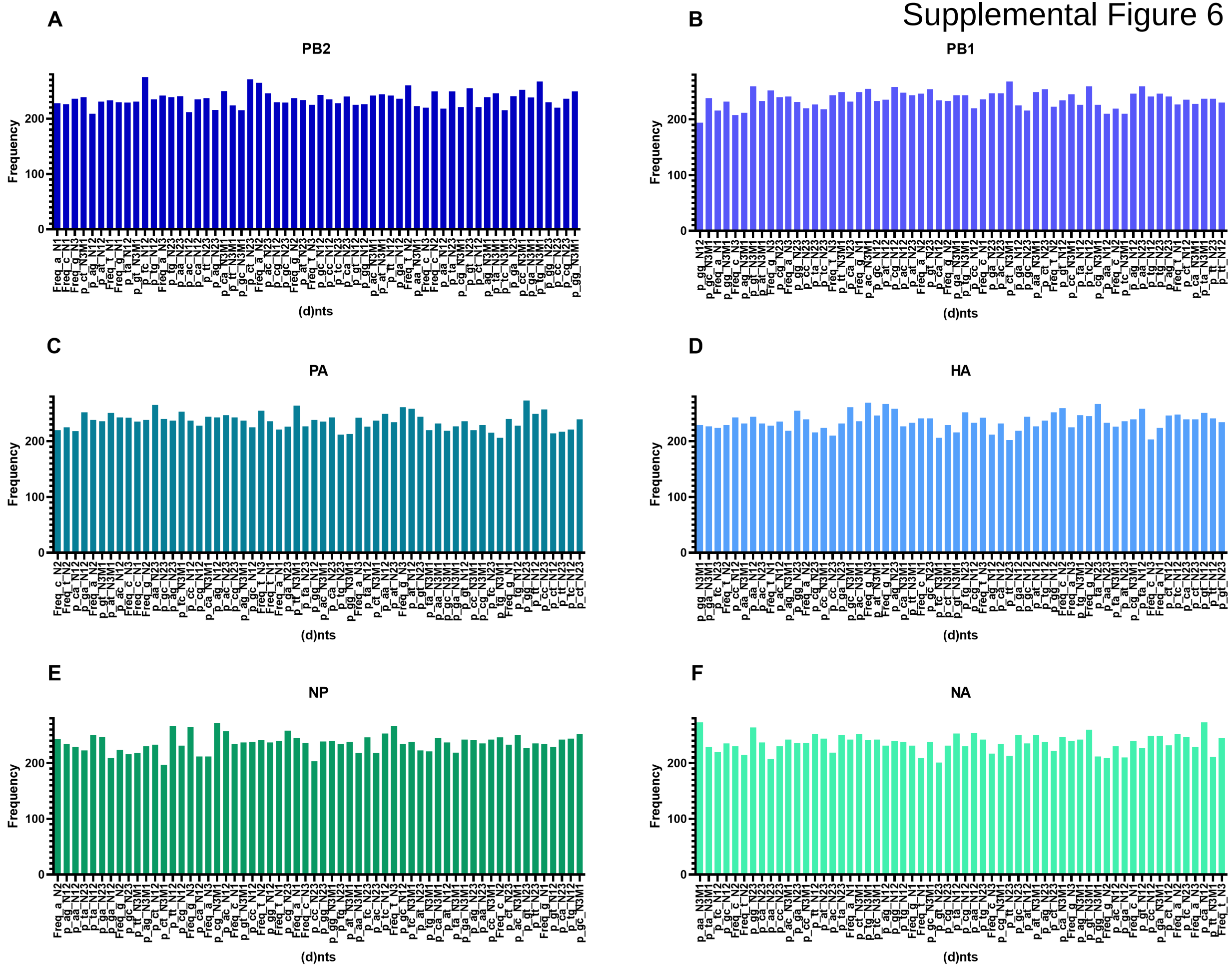
Supplemental Figure 4


A
PB2 H3N2 A Georgia 13 2017
PB2 H3N2 A New Hampshire 12 2017
PB2 H3N2 A Washington 18 2016
PB2 H3N2 A Canada 11A 2015
PB2 H3N2 A Niigata 13F335 2014
PB2 H3N2 A Parma 287 2015
PB2 H3N2 A Canoas LACENRS 1212 2013
PB2 H3N2 A North Carolina 28 2014
PB2 H3N2 A Thailand PY08064 2009
PB2 H3N2 A Singapore C2009.715V 2009
PB2 H3N2 A Singapore H2012.301 2012
PB2 H3N2 A Peru PER032 2010
PB2 H3N2 A Mexico InDRE29 2006
PB2 H3N2 A Christchurch 15 2004
PB2 H3N2 A Auckland 615 2002
PB2 H3N2 A Western Australia 48 2003
PB2 H3N2 A Lyon 1815 1993
PB2 H3N2 A New York 670 1995
PB2 H3N2 A Bilthoven 21438 1971
PB2 H2N2 A Rotterdam 1957
PB2 H1N1 A Siena 10 1989
PB2 H4N8 A blue winged teal Canada 3231 2011
PB2 mixed A ruddy turnstone New Jersey 650584 2002
PB2 H4N2 A Northern shoveler Guatemala CIP049H112 79 2012
PB2 H13N9 A black skimmer Chile C20108 2016
PB2 H4N6 A mallard Alberta 12 1991
PB2 H1N1 A Singapore DMS63 2009
PB2 H1N1 A Wyoming 28 2017
PB2 H1N1 A District Of Columbia 19 2018
PB2 H1N1 A Lima INS3 671 2012
PB2 H1N1 A Sydney DD3 25 2010
PB2 H1N1 A England 415 2009
PB2 H1N1 A Shanghai 71T 2009
PB2 H1N1 A VenAncio Aires LACENRS 1539 2009
PB2 H1N1 A Canada SK RV1798 2009
PB2 H13N6 A black headed gull Netherlands 82 2012
PB2 H13N2 A black headed gull Netherlands 6 2016
PB2 H13N8 A black headed gull Netherlands 14 2008
PB2 H5N1 A chicken Vietnam NCVD 117 2008
PB2 H5N1 A chicken Vietnam 933 2004
PB2 H5N1 A chicken Viet Nam 8 2005
PB2 H5N1 A duck Bangladesh 21980 2014
PB2 H10N8 A duck Mongolia 371 2010
PB2 H3N2 A chicken Ganzhou GZ43 2016
PB2 mixed A mallard Sweden 80057 2008
PB2 H5N6 A black swan Akita 2 2016
PB2 H6N2 A chicken Zhejiang 77001 2014
PB2 H9N2 A chicken Pakistan UDL 01 2008
PB2 H1N1 A greater white fronted goose Netherlands 4 2011
PB2 H10N1 A mallard Sweden 107688 2009
PB2 H10N9 A mallard Sweden 52837 2006
PB2 H5N2 A tundra swan Shimane 3211 A001 2011
PB2 H11N2 A mallard Netherlands 26 2005
PB2 H9N2 A Anser fabalis China HuBS428 2014
PB2 H7N9 A Beijing 27870 2017
PB2 mixed A chicken Jiangxi 23772 2013
PB2 H7N9 A EV SDYT 10 2016
PB2 H9N2 A chicken Jinan 3925 2013
PB2 H7N9 A chicken Jiangxi 10958 2014
PB2 H7N9 A chicken Jiangxi 13209 2014
PB2 H7N9 A chicken Jiangxi 13209 2014
0.050
B
PB1 H1N1 A Hawaii 43 2017
PB1 H1N1 A North Carolina 01 2018
PB1 H1N1 A Alabama 27 2018
PB1 H1N1 A New Hampshire 08 2016
PB1 H1N1 A Santiago p13d2 2013
PB1 H1N1 A Singapore DMS1171 2011
PB1 H1N1 A England 697 2010
PB1 H1N1 A England 03220137 2010
PB1 H1N1 A Munich INS363 2009
PB1 H1N1 A England 01180016 2010
PB1 H1N1 A Helsinki P15 2009
PB1 H1N1 A Mexico City 026 2009
PB1 H1N1 A Lyon 52.16 2010
PB1 H1N1 A turkey Ontario FAV110 2009
PB1 H1N1 A Australia 81 2009
PB1 H3N2 A Sydney 379T 2005
PB1 H3N2 A HaNoi Q578 2007
PB1 H3N2 A South Australia 36 2012
PB1 H3N2 A Newcastle 25 2014
PB1 H3N2 A Oregon 15 2016
PB1 H3N2 A Arkansas 16 2014
PB1 H3N2 A Sydney 1029 2015
PB1 H3N2 A Victoria 52 2015
PB1 H3N2 A California 156 2016
PB1 H3N2 A Gyeongbuk 1301 2016
PB1 H3N2 A Singapore 36 1989
PB1 H3N2 A Bilthoven 93 1970
PB1 H4N2 A mallard Ohio 275 1987
PB1 H1N9 A ruddy turnstone Delaware 650600 2002
PB1 H5N2 A chicken Iowa 15 015117 1 2015
PB1 H2N3 A mallard Alberta 205 1998
PB1 H5N2 A mallard Michigan 15 031493 1 2015
PB1 H3N8 A northern pintail Alaska 795 2012
PB1 H3N8 A northern pintail Interior Alaska 7MP0343 2007
PB1 H1N1 A Switzerland 5165 2010
PB1 H6N2 A duck Shantou 5480 2005
PB1 H9N2 A chicken Guangdong C1096 2009
PB1 H6N5 A duck Guizhou 888 2006
PB1 H10N7 A mute swan Netherlands 1 2006
PB1 H4N6 A spot billed duck Shanghai SH12 2014
PB1 H10N5 A migratory duck Jiangxi 31577 2013
PB1 H6N1 A duck Bangladesh 25767 2015
PB1 H7N7 A Netherlands 219 03
PB1 H6N1 A chicken Taiwan SP1 00
PB1 H5N8 A goose Zhejiang 925037 2014
PB1 H5N1 A muscovy duck Vietnam OIE 559 2011
PB1 H5N1 A open billed stork Nakhonsawan BBA3011M 2005
PB1 H5N1 A Cygnus olor Astrakhan Ast05 2 6 2005
PB1 H5N1 A duck Vietnam HU5 1572 2016
PB1 H5N1 A duck Vietnam QT1729 2013
PB1 H9N2 A chicken Guangxi C2163 2012
PB1 H7N9 A chicken Guangdong SD029 2014
PB1 H9N2 A chicken Guangdong LG1 2013
PB1 H9N2 A chicken Guangxi C4080 2010
PB1 H9N2 A chicken Zhejiang 3C2 2013
PB1 H7N9 A Environment Zhejiang 5 2014
PB1 H13N6 A black headed gull Netherlands 19 2012
PB1 H1N1 A New York 615 1995
PB1 H1N1 A Mexico 24004 2009
0.020
C
PA H1N1 A Bochum INS250 2009
PA H1N1 A Argentina 8994 2009
PA H1N1 A Athens INS341 2009
PA H1N1 A Hungary 40 2009
PA H1N1 A England 182 2010
PA H1N1 A Mexico City INERPHAC2 2009
PA H1N1 A Budapest WRAIR3789T 2011
PA H1N1 A England 10740579 2010
PA H1N1 A Singapore DMS1166 2011
PA H1N1 A Malaysia 2076212 2009
PA H1N1 A Singapore GP3545 2009
PA H4N6 A mallard Alberta 235 1999
PA H4N6 A mallard Interior Alaska 7MP0512 2007
PA H4N6 A blue winged teal Texas AI11 3094 2011
PA H3N8 A blue winged teal Canada 3244 2011
PA H7N3 A cinnamon teal Mexico 2817 2006
PA H1N1 A wild bird Korea SK14 2014
PA H3N8 A duck Hokkaido W4 2010
PA H5N6 A muscovy duck Vietnam LBM756 2014
PA H3N8 A gadwall duck Netherlands 3 2006
PA H6N2 A duck Eastern China 59 2003
PA H7N2 A quail Italy 4610 2003
PA H9N2 A pheasant Hong Kong WF288W 2007
PA H5N1 A chicken Lang Son 200 2005
PA H5N1 A duck Egypt Q4596D 2012
PA H5N1 A duck Cao Bang 43 2007
PA H5N1 A quail Vietnam CVVI 49 2010
PA H4N8 A rufous necked stint Japan 6KS0242 2006
PA H1N3 A chicken Zhejiang 2CP25 2014
PA H7N9 A chicken Dongguan 1312 2014
PA H7N9 A Shenzhen SP38 2014
PA H9N2 A chicken Vietnam HU4 2 2015
PA H7N9 A chicken Shantou 4824 2014
PA H7N9 A Shanxi 26571 2017
PA H7N9 A Zhejiang 14 2016
PA H7N9 A chicken Jiangxi 10877 2014
PA H7N9 A chicken Hunan SD019 2014
PA H7N9 A chicken Hunan SD019 2014(2)
PA H2N2 A chicken Jena 4836 1983
PA H4N6 A blue winged teal Texas AI11 3691 2011
PA H5N2 A northern pintail Alaska 16 041335 11 2016
PA H1N1 A Shanghai 2 1997
PA H3N2 A Taiwan 1 69
PA H3N2 A Canterbury 34 2002
PA H3N2 A Canterbury 80 2002
PA H3N2 A England 789 2003
PA H3N2 A HaNoi HN30127 2004
PA H3N2 A Singapore 53C 2007
PA H3N2 A Managua 32 2007
PA H3N2 A Thailand CU B1672 2009
PA H3N2 A Cambodia NHRCC00001 2009
PA H3N2 A Thailand SP07127 2008
PA H3N2 A Peru PER127 2010
PA H3N2 A Minnesota 43 2016
PA H3N2 A Victoria 5006 2014
PA H3N2 A Brisbane 174 2015
PA H3N2 A South Australia 166 2015
PA H3N2 A Louisiana 41 2017
PA H3N2 A Nicaragua 9828 01 2017
PA H3N2 A West Virginia 20 2017
PA H3N2 A Nicaragua 8100 07 2016
0.050
D
HA H1N1 A England 630 2009
HA H1N1 A Che NIV658 2009
HA H1N1 A Ankara 06 2009
HA H1N1 A Scotland 63 2009
HA H1N1 A New Mexico 14 2009
HA H1N1 A Oslo INS110 2009
HA H1N1 A Pune NIV21115 2009
HA H1N1 A England 95200063 2009
HA H1N1 A Chile 54 2010
HA H1N1 A Shenzhen 78 2009
HA H1N1 A Guangdong 39 2011
HA H1N1 A Singapore TT200 2011
HA H1N1 A Thailand CU B8665 2013
HA H1N1 A Rochester U012 2016
HA H1N1 A Shanghai Mix1 2014
HA H1N1 A Thailand SP08193 2008
HA H1N1 A Mexico UASLP 006 2008
HA H1N1 A HaNoi 2143 2001
HA H1N1 A Thailand Siriraj 07 2000
HA H1N1 A New South Wales 26 2000
HA H1N1 A mallard Mississippi 442 2010
HA H6N6 A duck Guangdong S3468 2010
HA H6N2 A duck Shantou 1080 2007
HA H5N2 A guinea fowl New Jersey 08 000640 008 2008
HA H5N2 A chicken Hidalgo 28159 232 1994
HA H5N2 A chicken Hebei 3 2013
HA H5N1 A duck Vietnam BacLiu NCVD 10 2007
HA H5N1 A chicken Egypt 0813 NLQP 2008
HA H5N6 A chicken Nha Trang 124 2015
HA H5N1 A duck Bangladesh 31295 2016
HA H5N1 A peregrine falcon Tochigi 15 2011
HA H5N1 A chicken Nigeria 16VIR5840 68 2015
HA H13N6 A black headed gull Republic of Georgia 2 2012
HA H9N2 A chicken Shandong 11 2011
HA mixed A silkie chicken Shenzhen 452 2013
HA mixed A chicken Wenzhou 3280 2013
HA H3N2 A Israel C3673 2017
HA H3N2 A South Australia 115 2016
HA H3N2 A Thailand CU B5697 2011
HA H3N2 A Thailand CU B1697 2009
HA H3N2 A Wellington 3 2005
HA H3N2 A Canterbury 07 2001
HA H3N2 A Nice 491 1996
HA H3N2 A Yamagata 56 1993
HA H3N2 A Singapore 35 1989
HA H3N2 A Rotterdam 8179 1977
HA H3N8 A cinnamon teal California 44287 325 2007
HA H4N6 A duck Mongolia 1117 2009
HA mixed A chicken Jiangxi 1287 2014
HA mixed A chicken Jiangxi 13204 2014
HA mixed A chicken Jiangxi 13270 2014
HA H7N3 A mallard Maryland 09OS1143 2009
HA H7N3 A mallard Alberta 93 1996
HA H7N1 A turkey Italy 4708 1999
HA H7N9 A chicken Fujian S4518 2015
HA H7N9 A chicken Fujian S4518 2015(2)
HA H7N9 A chicken Jiangxi 13524 2014
HA H7N9 A chicken Shanghai PD CN 02 2014
HA H7N9 A chicken Shanghai PD CN 02 2014(2)
HA H7N9 A Hunan 26938 2014
HA H7N9 A silkie chicken Dongguan 3275 2013
HA H7N9 A silkie chicken Dongguan 3049 2013
HA H7N9 A silkie chicken Dongguan 3049 2013(2)
0.10
E
NP H3N2 A Nebraska 10 2016
NP H3N2 A Nicaragua 9103 06 2017
NP H3N2 A Nicaragua 5960 05 2016
NP H3N2 A California 213 2016
NP H3N2 A Nicaragua 6228 04 2013
NP H3N2 A Bangkok FLU PMKA1169 2011
NP H3N2 A Milano 217 2011
NP H3N2 A Waikato 12 2005
NP H3N2 A Christchurch 10 2004
NP H3N2 A Auckland 621 2005
NP H3N2 A Christchurch 297 2004
NP H1N2 A New York 489 2003
NP H3N2 A Auckland 611 2002
NP H3N2 A Scotland 50 2003
NP H3N2 A Canterbury 431 2003
NP H3N2 A Waikato 61 2003
NP H3N2 A Madrid G58 1992
NP H1N1 A Siena 4 1987
NP H1N1 A Managua 4924.01 2008
NP H1N1 A Western Australia 20 2001
NP H3N2 A turkey Ontario 31232 2005
NP H1N1 A Singapore ON981 2009
NP H1N1 A Nicaragua AGA2 80 2011
NP H1N1 A Qingdao F158 2011
NP H1N1 A Mexico INMEGEN INER 07 2012
NP H1N1 A Prague WRAIR4146N 2011
NP H1N1 A Argentina 9180 2009
NP H1N1 A Thailand SN10560 2011
NP H1N1 A Chungnam 29 2016
NP H1N1 A New York PV00410 2017
NP H1N1 A Helsinki 1017 2013
NP H13N2 A black headed gull Netherlands 2 2016
NP H6N1 A mallard California 8399 2008
NP H3N8 A mallard Wisconsin 11OS4383 2011
NP H7N3 A chicken Canada 314514 1 2005
NP H3N6 A American black duck New Brunswick 02649 2007
NP H4N6 A blue winged teal Ohio 12OS2321 2012
NP H7N3 A mallard Michigan A00869519 2009
NP H7N2 A turkey VA 158512 2002
NP H8N4 A mallard Alberta 194 1992
NP H3N8 A mallard Oregon AH0038704 2015
NP H3N8 A mallard Netherlands 3 2005
NP mixed A duck Jiangxi 285 2013
NP H5N3 A duck Primorie 2633 2001
NP H1N1 A mallard duck Netherlands 12 2005
NP H5N3 A duck Hokkaido 167 2007
NP H1N1 A mallard duck Netherlands 25 2009
NP H5N6 A Anas platyrhynchos Korea W612 2017
NP H10N5 A migratory duck Jiangxi 30246 2013
NP H1N4 A mallard Netherlands 30 2006
NP H9N2 A chicken Sihala NARC 12103 2008
NP H5N1 A duck Niger 2090 2006
NP H5N6 A teal Tottori 2 2016
NP H5N1 A duck Zhejiang 224 2011
NP H5N1 A chicken Nigeria 16VIR5840 1 2015
NP H9N2 A chicken Jiangxi 506 2013
NP H9N2 A pigeon Anhui DGG4 2015
NP H7N9 A chicken Liaoning 05.12 SY069 2017(H7N9)
NP H7N9 A chicken Dongguan 1533 2014
NP H7N9 A chicken Dongguan 695 2014
NP H7N9 A chicken Dongguan 695 2014
0.050
F
NA H3N2 A Arizona 18 2015
NA H3N2 A Columbia P0080 2014
NA H3N2 A Sydney 1034 2015
NA H3N2 A Boston DOA2 240 2013
NA H3N2 A Santiago op20d1 2012
NA H3N2 A California 38 2016
NA H3N2 A Singapore H2013.778 2013
NA H3N2 A Singapore H2010.211 2010
NA H3N2 A Florida UR07 0021 2008
NA H3N2 A TayNguyen TN381 2005
NA H3N2 A Baden Wuerttemberg 17 2006
NA H3N2 A Denmark 32 03
NA H3N2 A Hong Kong CUHK34114 2001
NA H3N2 A Wellington 30 2000
NA H3N2 A Malaysia 12560 1999
NA H3N2 A Queensland 31 2003
NA H3N2 A England 967 2003
NA H3N2 A Hong Kong 1 1997
NA H9N2 A duck Guangxi NN 2005
NA H6N2 A mallard duck Netherlands 11 2014
NA H6N2 A Anser fabalis China Anhui L180 2014
NA H6N2 A greater white fronted goose Netherlands 8 2009
NA H6N2 A coot Alberta 134 1987
NA H9N2 A egret Hunan 1 2012
NA H9N2 A duck Hong Kong 784 1979
NA H5N2 A snow goose Missouri 15 011246 1 2015
NA H5N6 A chicken Hubei ZYSJF22 2016
NA H5N6 A Changsha 1 2014
NA H7N9 A Environment Guangzhou K4756-K4776 2017
NA H7N9 A Fujian 30767 2016
NA H7N9 A Anhui 13426 2017
NA H7N9 A chicken Jiangxi 13543 2014
NA H7N9 A Environment Xinjiang 73033 2014
NA H7N9 A chicken Zhejiang S4071 2013
NA H7N9 A chicken Zhejiang S4071 2013(2)
NA H7N9 A Hangzhou 118 2014
NA H7N9 A Zhejiang 16 2013
NA H7N9 A Fujian 03 2013
NA H7N9 A Jiangxi 01 2013
NA H7N9 A chicken Hangzhou 50-1 2013(H7N9)
NA H16N3 A glaucous winged gull Southcentral Alaska 14MB01705 2014
NA H7N3 A duck Jiangxi 34858 2013
NA H7N3 A Duck Mongolia 786 2017
NA H2N3 A mallard duck Netherlands 24 2011
NA H1N1 A blue winged teal Texas UGAI15 6706 2015
NA H1N1 A mallard Alberta 127 1977
NA H7N1 A sacred ibis South Africa 120701 2012
NA H1N1 A mallard Republic of Georgia 1 2012
NA H7N1 A duck Denmark 53 147 8 2008
NA H5N1 A muscovy duck Vietnam HU8 1670 2017
NA H1N1 A Missouri 48 2017
NA H1N1 A Oklahoma 18 2018
NA H1N1 A Nicaragua 7525 12 2015
NA H1N1 A Florida 92 2015
NA H1N1 A Osaka 10K362 2011
NA H1N1 A Netherlands 602 2009
NA H1N1 A Lebanon 09L 06 2009
NA H1N1 A Taiwan 99840 2009
NA H1N1 A Mexico City WR1765N 2009
NA H1N1 A Singapore DMS157 2010
NA H1N1 A Ontario 296008 2009
0.10

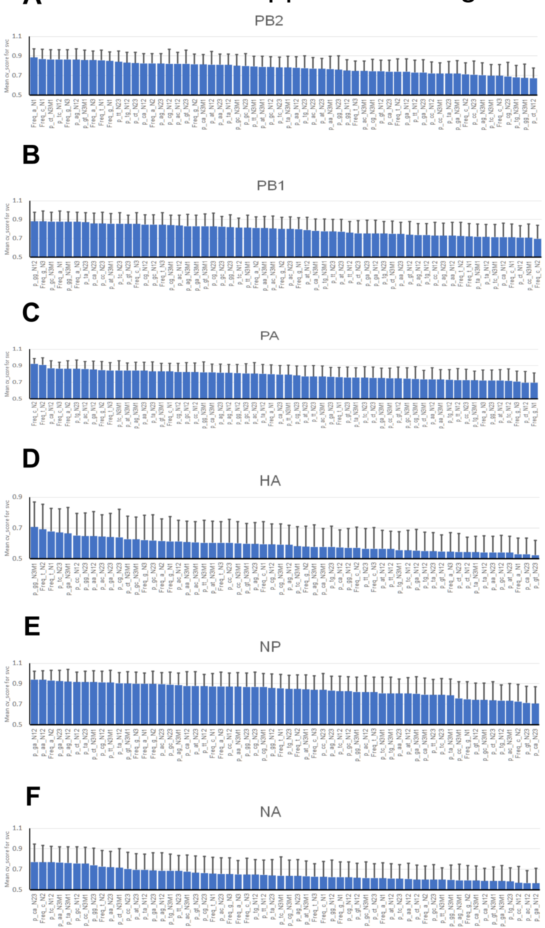
Supplemental Figure 5



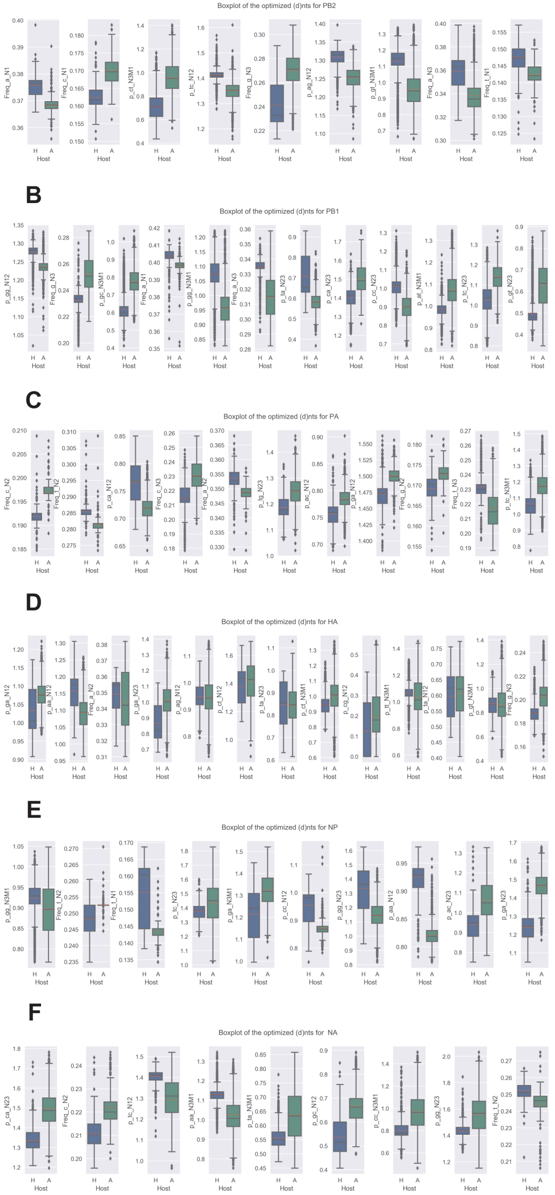
Supplemental Figure 6



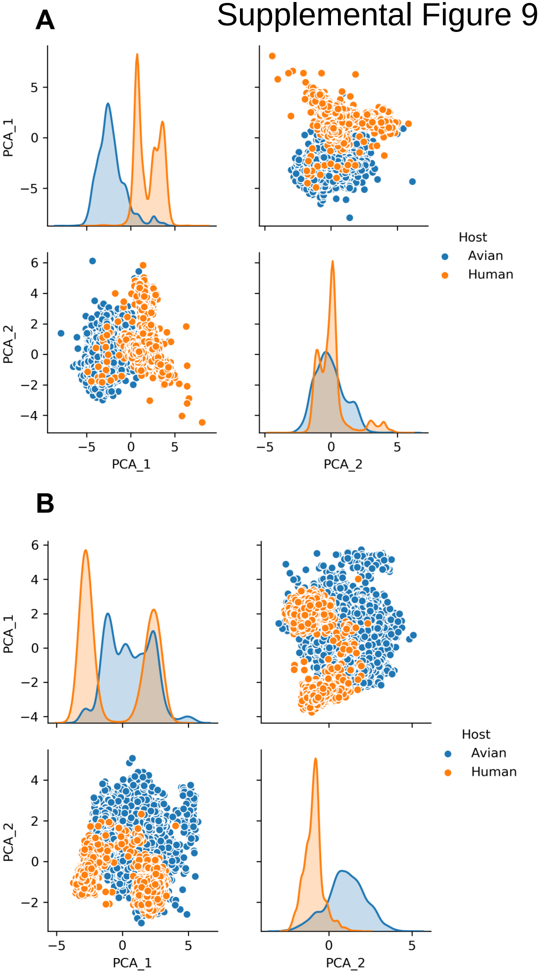
Supplemental Figure 7



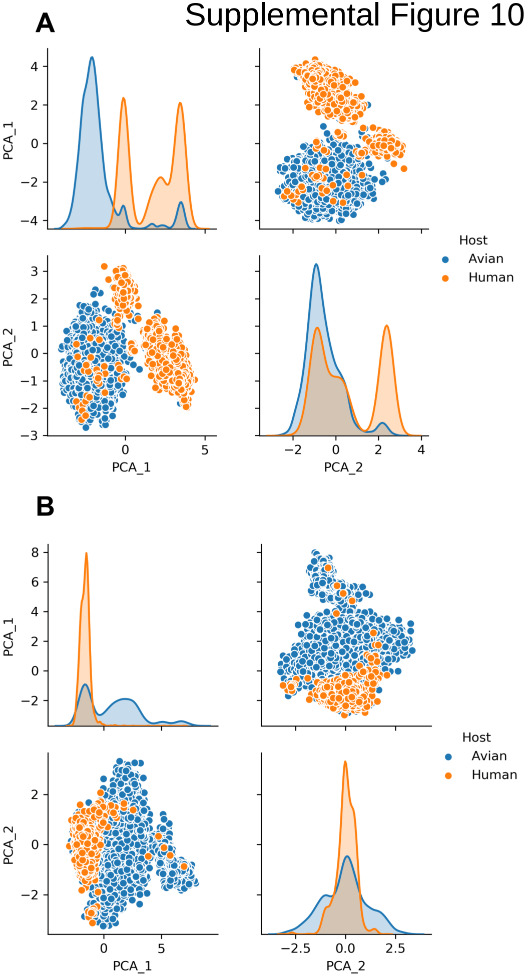
Supplemental Figure 8



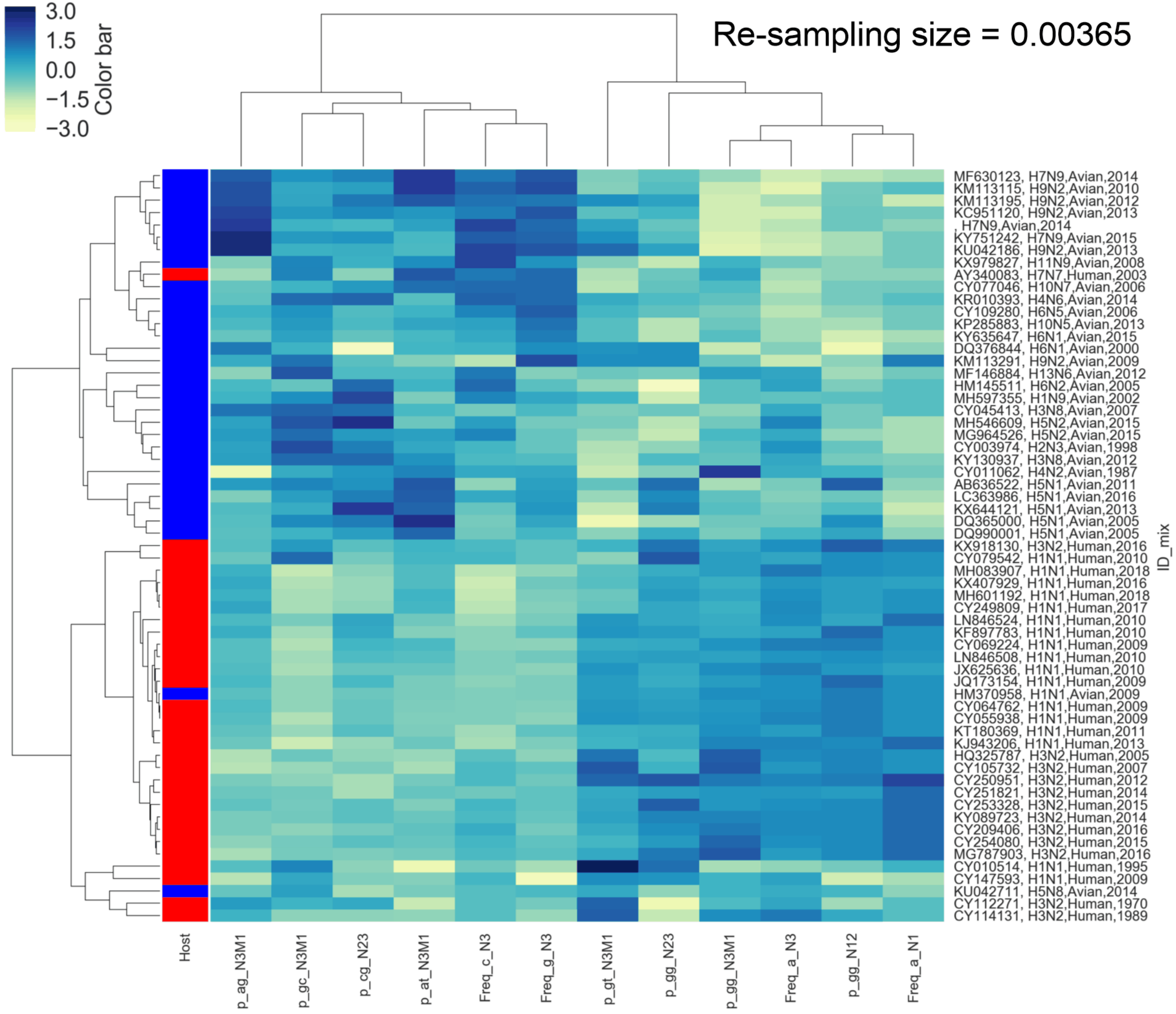
Supplemental Figure 9



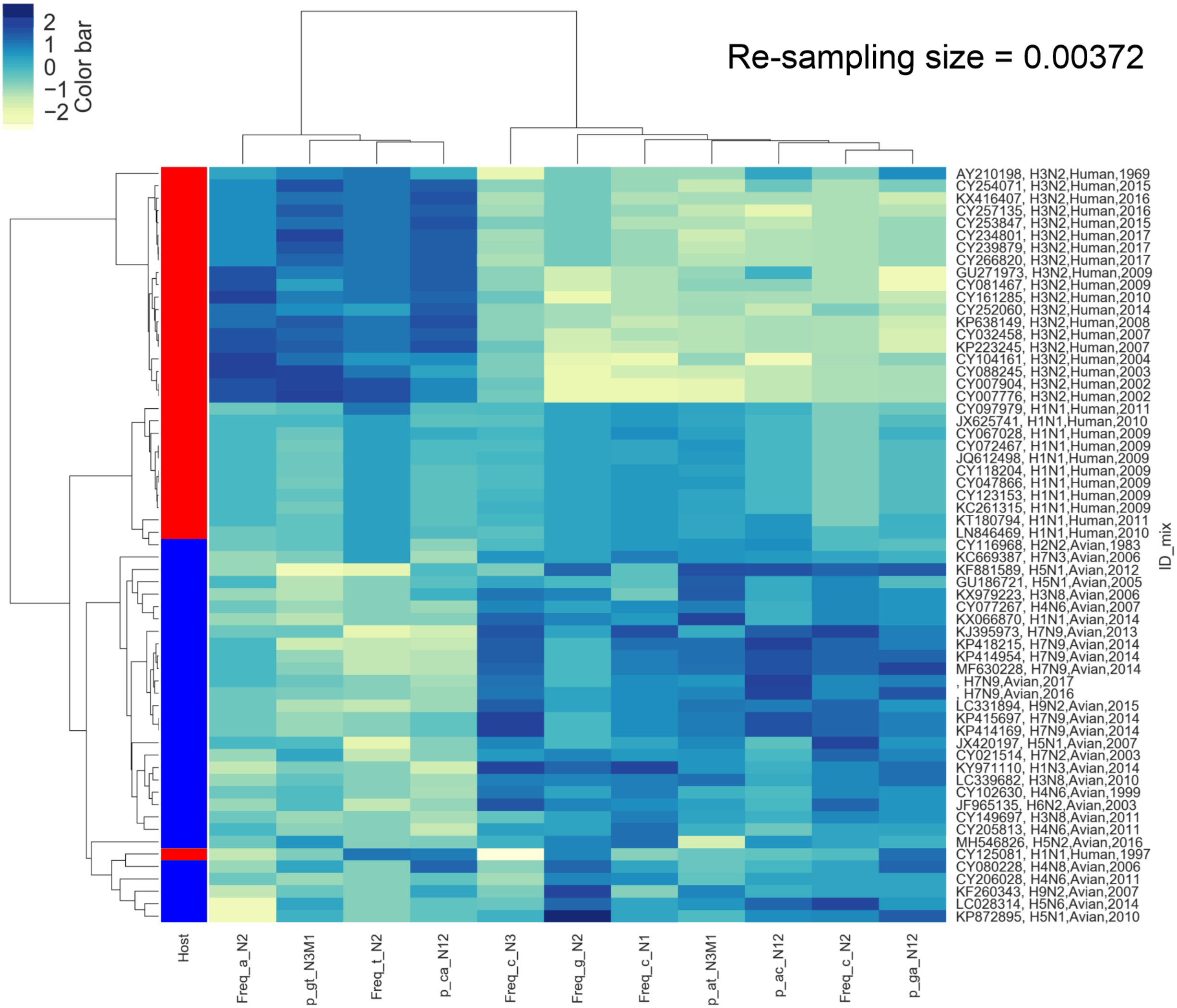
Supplemental Figure 10



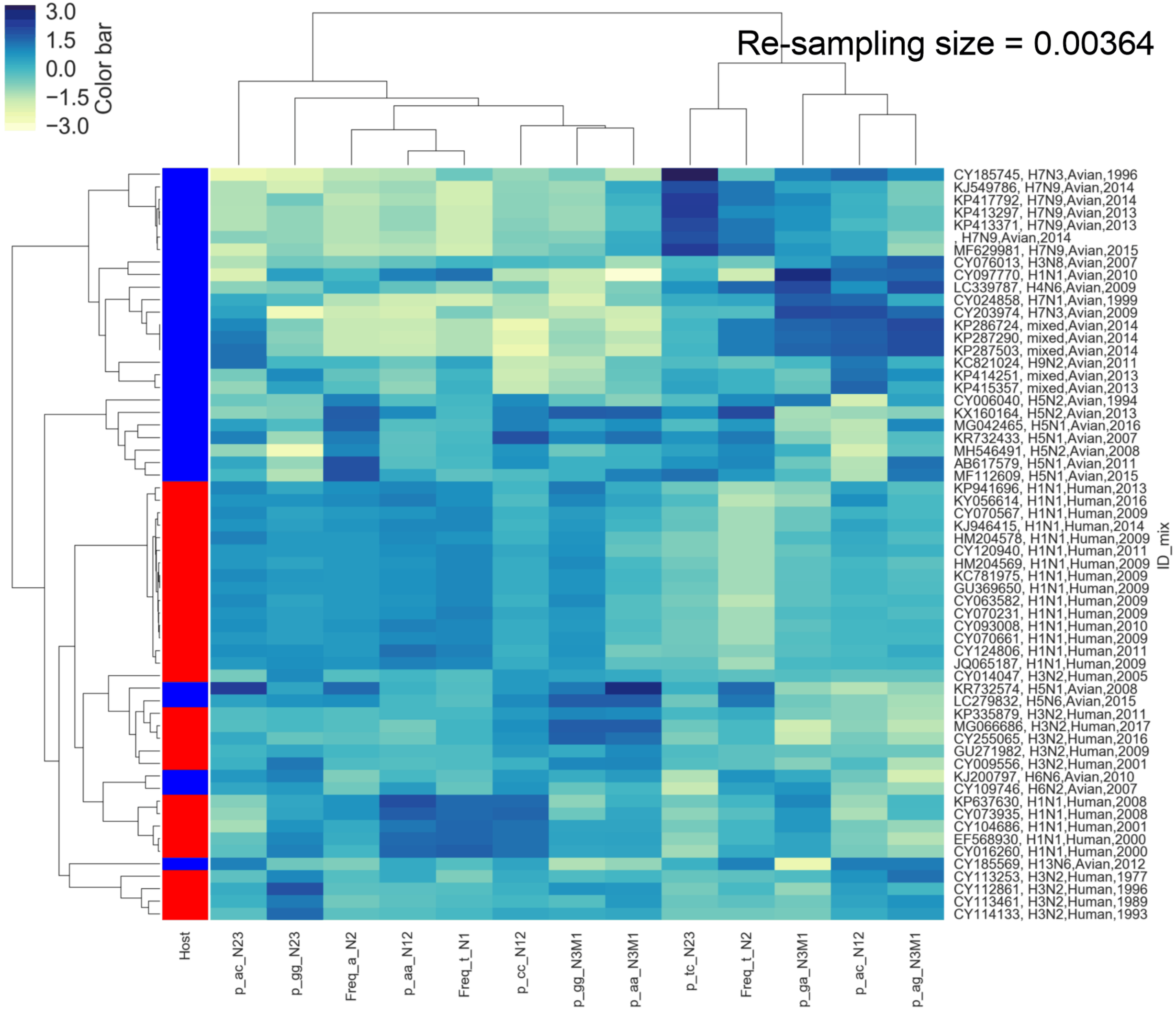
Supplemental Figure 10



Supplemental Figure 11

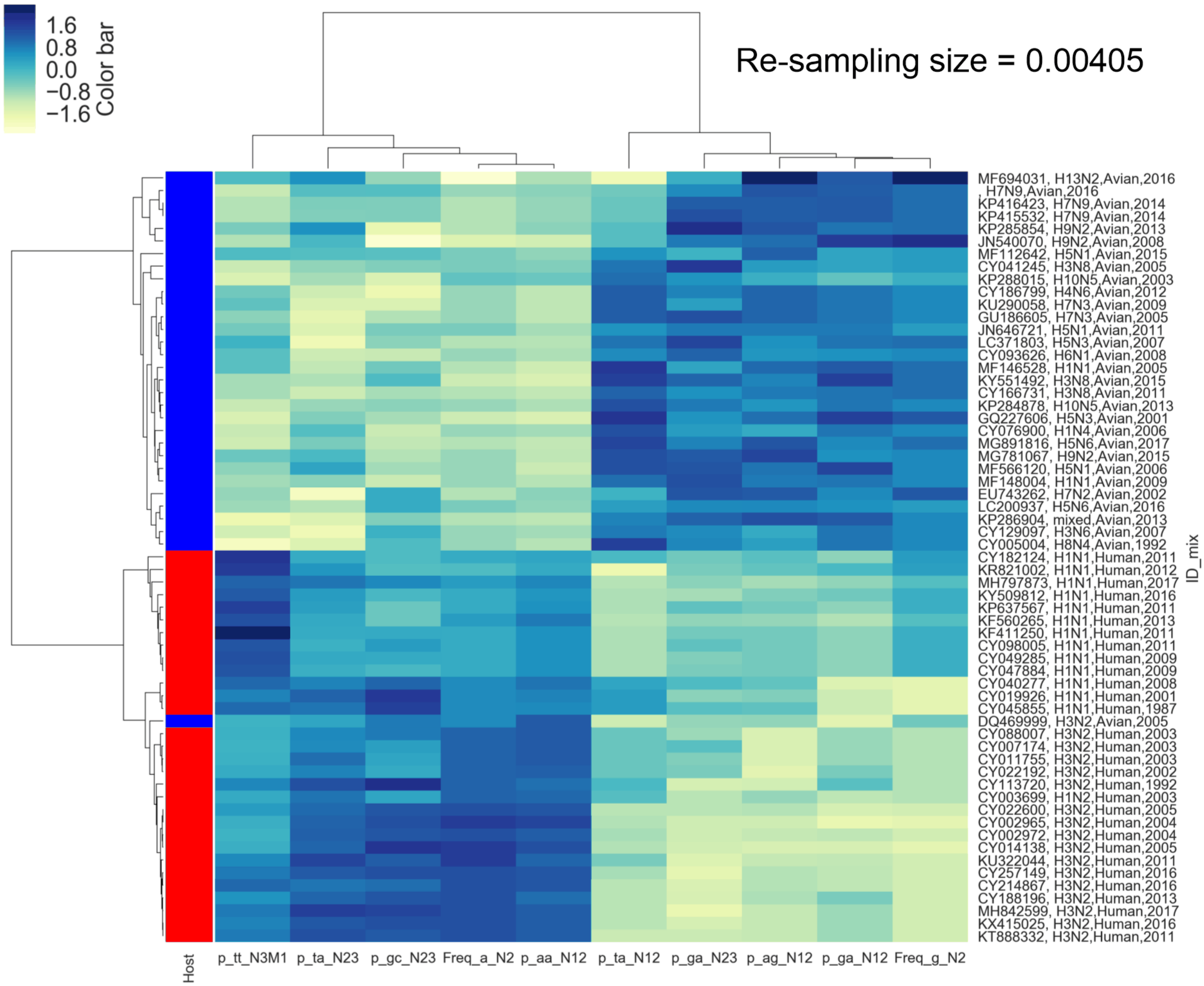


Supplemental Figure 13

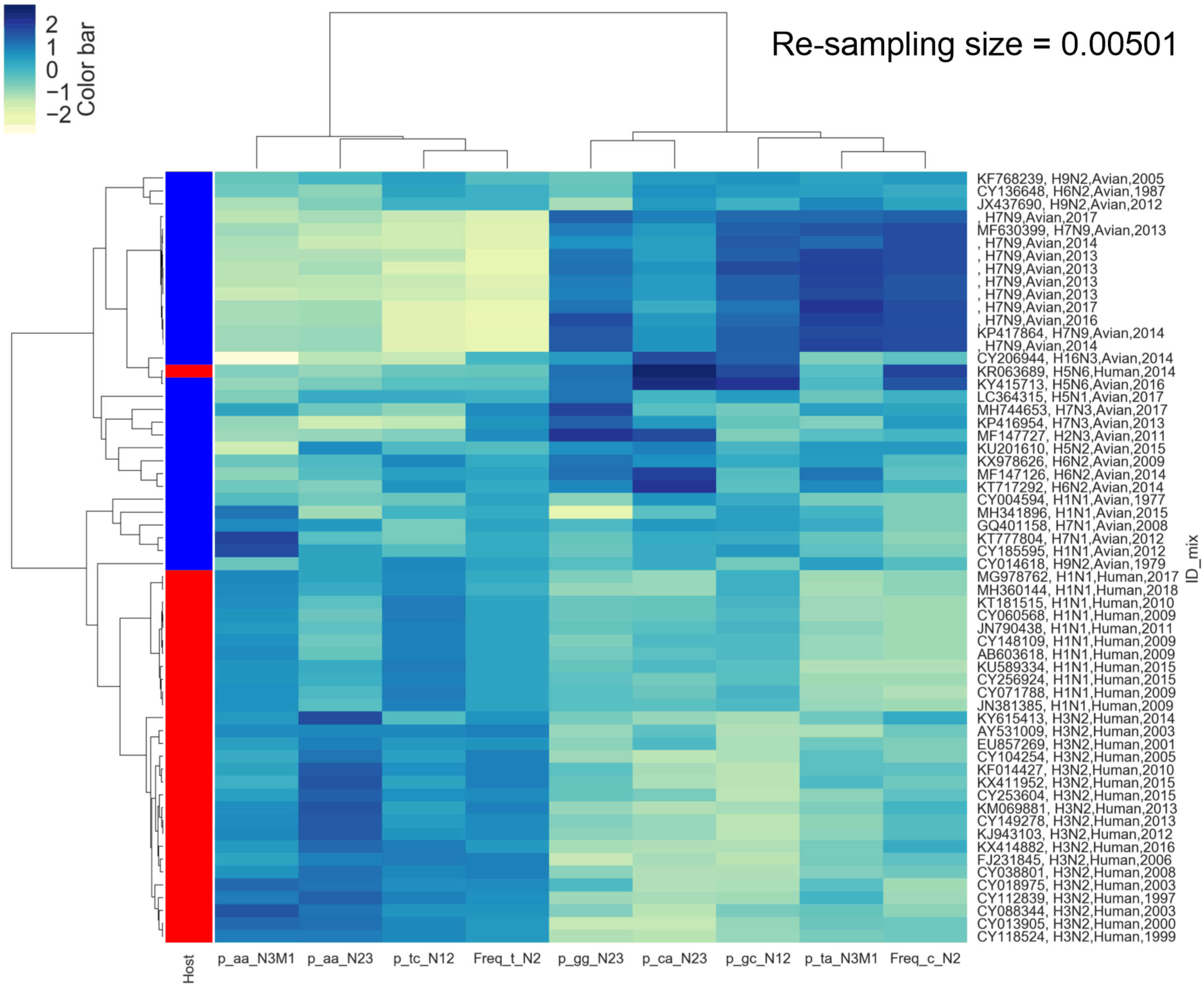


Re-sampling size = 0.00364

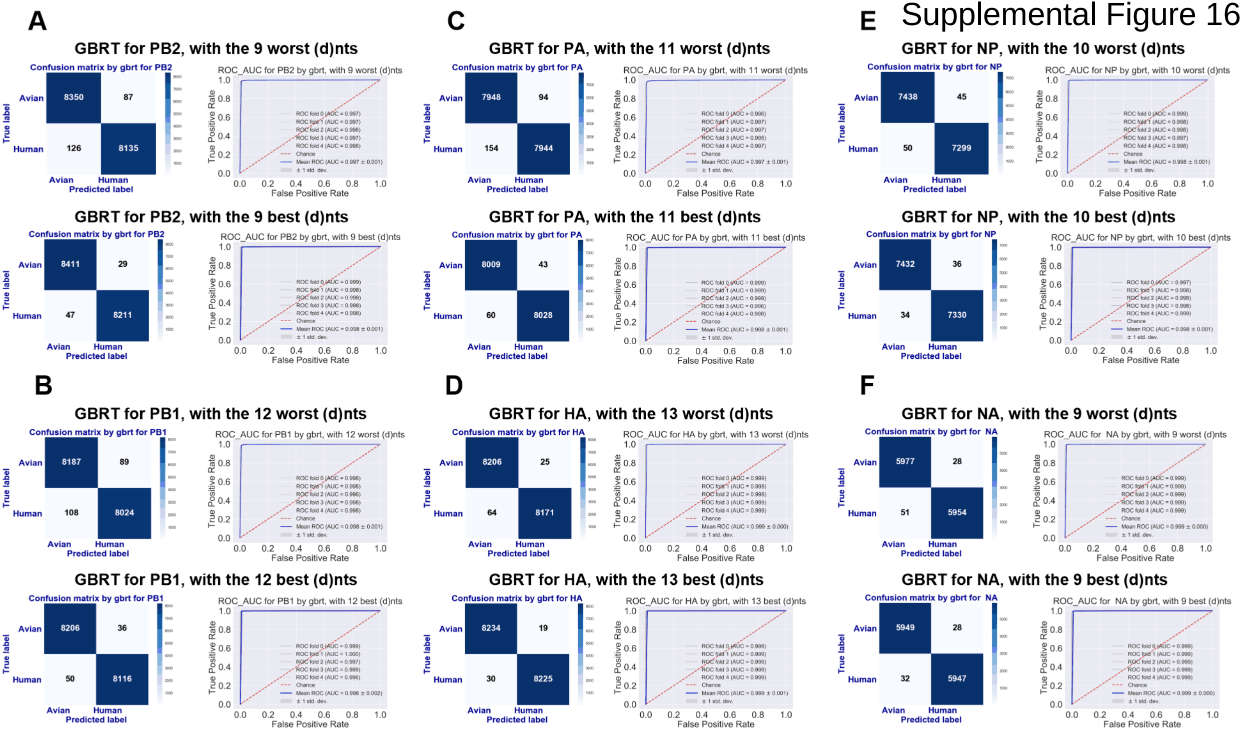
Supplemental Figure 14



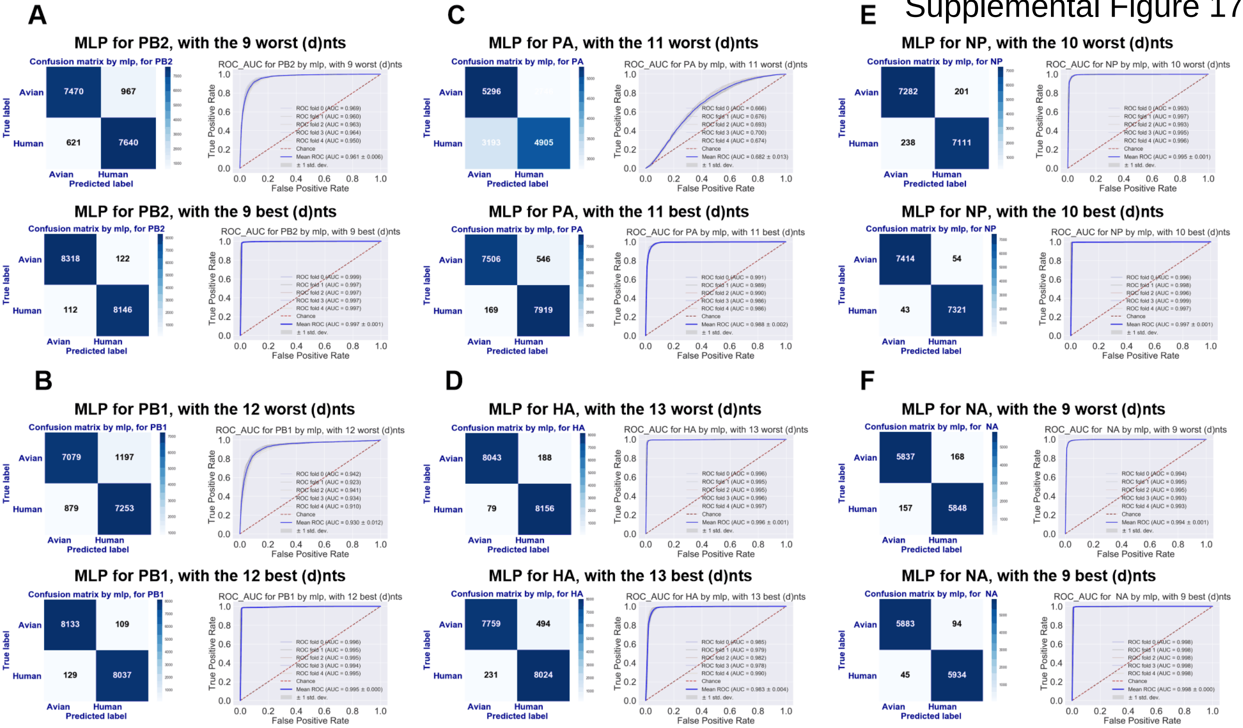
Supplemental Figure 15



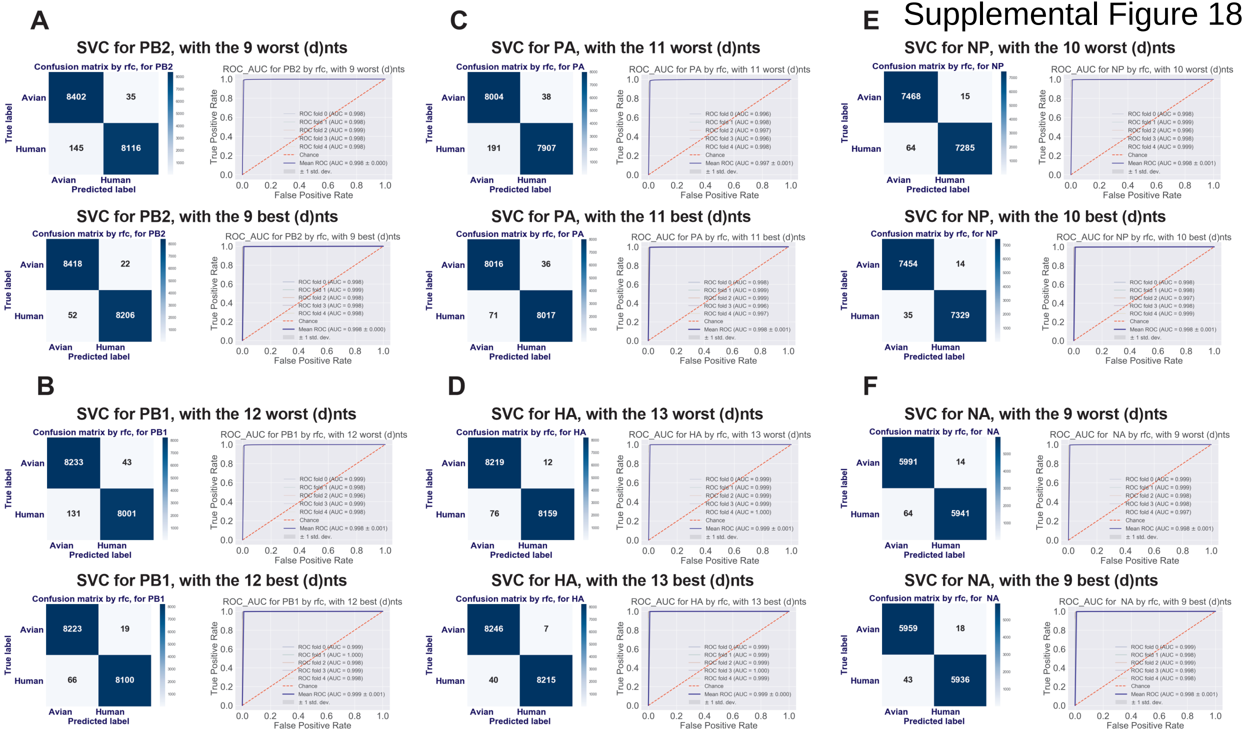
Supplemental Figure 16



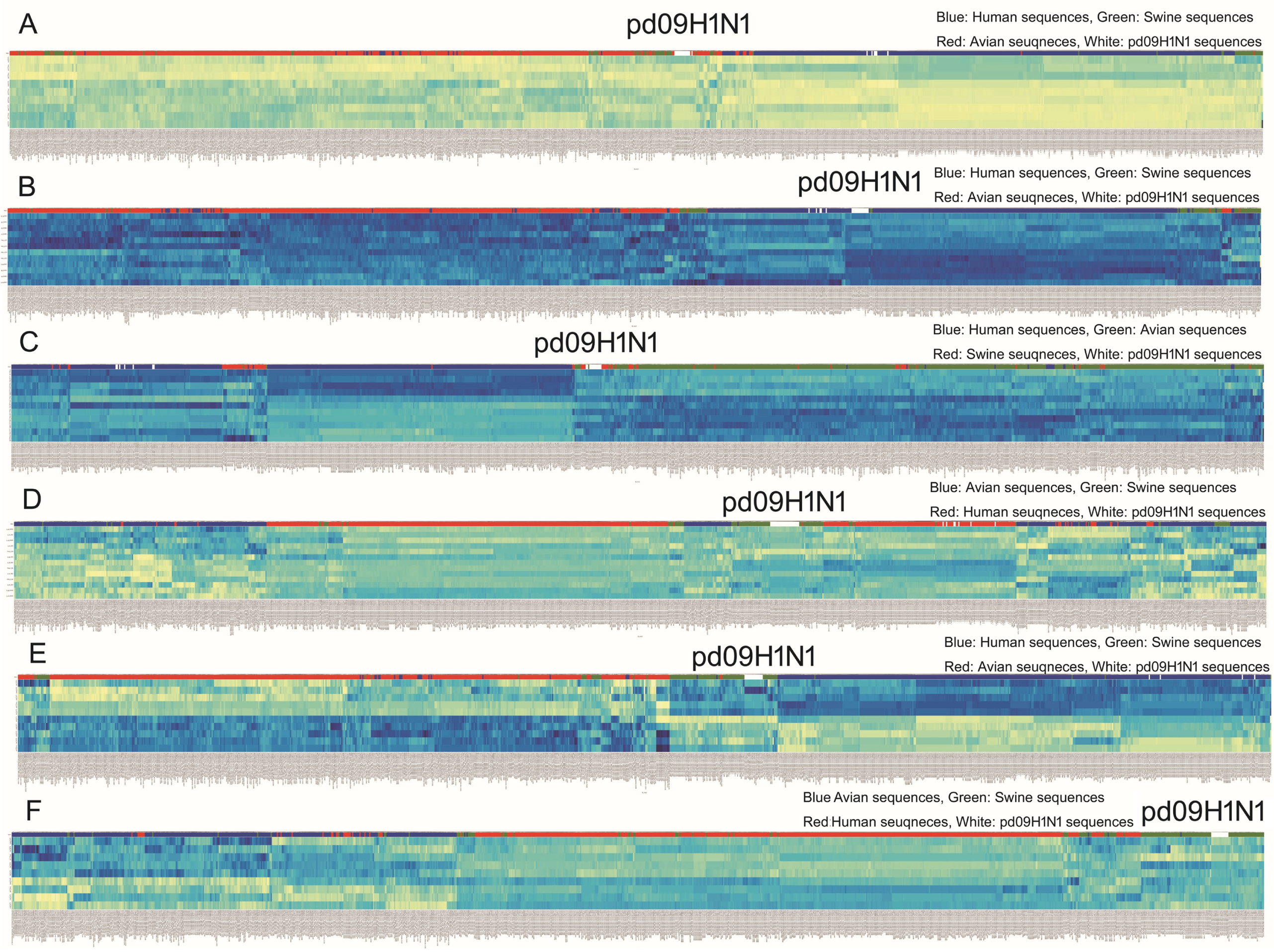
Supplemental Figure 17



Supplemental Figure 18



Supplemental Figure 19


A
pd09H1N1
Blue: Human sequences, Green: Swine sequences
Red: Avian seuqneces, White: pd09H1N1 sequences
B
pd09H1N1
Blue: Human sequences, Green: Swine sequences
Red: Avian seuqneces, White: pd09H1N1 sequences
C
pd09H1N1
Blue: Human sequences, Green: Avian sequences
Red: Swine seuqneces, White: pd09H1N1 sequences
D
pd09H1N1
Blue: Avian sequences, Green: Swine sequences
Red: Human seuqneces, White: pd09H1N1 sequences
E
pd09H1N1
Blue: Human sequences, Green: Swine sequences
Red: Avian seuqneces, White: pd09H1N1 sequences
F
Blue Avian sequences, Green: Swine sequences
Red:Human seuqneces, White: pd09H1N1 sequences
pd09H1N1