Rapid lineage assignment for the Influenza A Internal Genes

Andrew R. Dalby *

Department of Biomedical Sciences, University of Westminster, 115 New Cavendish Street, Westminster, W1W 6UW, UK

a.dalby@westminster.ac.uk

Sushant Bhat

The Pirbright Institute, Pirbright, Woking, Surrey, GU24 0NF, UK

Lorna Tinworth

Department of Biomedical Sciences, University of Westminster, 115 New Cavendish Street, Westminster, W1W 6UW, UK.

Munir Iqbal

The Pirbright Institute, Pirbright, Woking, Surrey, GU24 0NF, UK

*Corresponding author

Abstract

The hemagglutinin subtypes from Influenza A can be divided into distinct lineages. This is important for tracing the evolutionary history of the gene. It allows regional lineages to be identified and studied. The process of lineage identification depends on phylogenetic analysis to identify the distinct clades within the data.

Identification of lineages within the Influenza Internal genes would help to simplify the analysis of reassortment where these genes are transferred between subtypes. In this paper we show that a rapid clustering method can be used to assign lineages to the internal gene segments without the need for a full phylogenetic analysis.

Keywords

clustering, viral lineages, influenza, internal genes.

Acknowledgements

Funding: This work was funded in part by the BBSRC grant numbers BB/N002571/1 and BB/L018853/1.

Introduction

The influenza A virus genome is divided into 8 different RNA segments. Two of these segments encodes virus surface glycoproteins hemagglutinin (HA) and neuraminidase (NA) which characterise the viral subtypes H1-HA18 and N1-N11 respectively (1). The other six segments are commonly referred to as the internal genes and this includes the three polymerase segments PA, PB1 and PB2 a non-structural protein segment (NS) a Matrix protein M and nucleoprotein NP (2).

Each of the segments can be considered as evolving separately and there is exchange of the different segments between viral subtypes via the process of genetic reassortment (3). Determining if there has been exchange of the internal genes during reassortment depends on identifying the close homologues of the internal genes from the viral subtype under consideration (4). This is usually carried out by using a database search tool such as BLAST (5-7). Such analyses are time consuming and it is difficult to set an appropriate threshold for homology and commonly a cut-off of above 90% comparative identity might be used. These homology analyses also depend on the number of sequences returned by the search. If there are a large number of sequences a higher threshold will be applied to reduce the number, but if there are very few sequences the threshold would be increased.

The HA and NA genes are divided into subtypes based on their antigenic properties, but the internal genes do not have subtypes and are most often given the name of the glycolytic subtypes with which they are found. The HA subtypes have also been further divided into evolutionary lineages . A detailed classification has been developed by the WHO for the H5N1 highly pathogenic HA segment and similar nomenclatures also exist for the HA of H9N2 and H1N1 Swine influenza viruses (8-12). There is not currently a classification system available for the internal genes.

Recently a rapid sequence identity based clustering method, USEARCH has been used to filter sequences data (13). USEARCH applies a BLAST like approach but runs orders of magnitude faster than BLAST (14). In this case USEARCH was used to produce a non-redundant set of data where sequences have less than a specified degree of identity. A high identity threshold of 99% or even 100% can be used to remove identical sequences or nearly identical sequences. This is an example of an agglomerative clustering approach. In the study by Hurtado *et al.* the focus was on the representative sequences and not the clusters themselves (13).

An alternative approach is to use clustering to identify sub-groups which have sequence identity above a threshold. This will identify closely homologous sequences which form lineages. For lineage determination a lower identity threshold of between 90% and 95% should be used, as this is a divisive clustering approach. USEARCH generates two outputs, the representative sequences that

are at the centre of each cluster (the centroid sequences) and a set of files containing the sequences from each of the clusters (14).

H5N8 is a relatively rare highly pathogenic avian influenza subtype. H5N8 has occurred only sporadically until the 2014 outbreak in South Korea, and the 2015 outbreak in Taiwan (15). There are long gaps in the history of the subtype and it has been proposed that the virus has had multiple origins via reassortment (16-18). While there is strong evidence from analysis of the hemagglutinin and neuraminidase gene segments the additional evidence from the internal genes would give a more complete picture of reassortment. H5N8 is an ideal as a test case because there are a limited number of sequences (less than 250) and because of its wide spread geographical spread (it is present in both the USA and Eurasia).

In this paper we apply a clustering methodology in order to classify the influenza internal gene segments and we demonstrate how this classification can be used in order to identify potential reassortment events in H5N8.

Materials and Methods

All of the influenza A internal gene segments were downloaded from the Influenza Research Database (IRD) on the 25th of May 2017 (19). The H5 and H9 hemagglutinin segment sequences were downloaded from the IRD on the 21st of May 2017. A summary of these datasets is given in supplementary table 1.

Cluster identification using USEARCH was applied at a range of different identity levels to evaluate the statistical properties of the clustering (14). At this point only the centroids, the representative set of sequences for each cluster was determined. This was also carried out for the H5 and H9 hemagglutinin which act as a reference.

From the statistical properties of the clustering an optimal clustering identities for each of the internal genes and classification can be chosen and clustering can then be carried out using USEARCH for each of the segments. A level of identity of 90% was chosen for clustering all the internal gene segments using the following command.

usearch –cluster_fast sequences.fasta –id 0.9 –centroids sequences_nr.fasta

A multiple sequence alignment of all the clusters was carried out using Muscle v3.8.31 or for very large clusters containing over 2000 sequences using Mafft v7.271 (20, 21). Intra-cluster and inter-cluster distances were calculated using the APE module within the R statistical programming environment and R-studio (22-24). Inter-cluster distances are the distances between centroids. Scripts were used to automate the analysis (supplementary file S1).

A simple shell grep command was used to search for clusters containing H5N8 sequences for each of the internal gene segments.

grep H5N8 *.fas

Results

The clustering statistics for the initial classifications using USEARCH are given in supplementary table 2. From these data an optimum clustering identity of 90% was chosen for clustering all of the internal gene segments. This is the level of identity where the main cluster appears and where there are still aa manageable number of clusters.

USEARCH generated 25 clusters for the M gene segment, 35 for the NP, 54 for the NS segment, 46 for PA, 62 for PB1 and 48 for PB2. The number of clusters for the PB1 segment is significantly larger than for the others. In total there are 270 cluster across the 6 gene segments.

Summaries of the cluster properties for each of the internal gene segments are given in table 1-6.

The H5N8 containing clusters are summarised in table 7. The H5N8 sequences can be broken down into 13 representative groups. This includes 2 from the origin of the subtype in Ireland in 1983 and 5 groups from the USA that contain either 1 or two sequences. The South Korean groups that were identified previously based on the H5 hemagglutinin trees are the Buan and Gochang groups (25). However, these two groups are not easily separable based on the internal gene segment clustering.

Discussion

The HA sequences of H5 and H9 viruses were included as a reference because of the extensive literature on lineage assignment for these two hemagglutinin subtypes. The statistical properties for the clustering of the hemagglutinins can be compared to that of the internal genes to establish if they exhibit the same properties.

The existence of a plateau in the maximum cluster sizes in case of the hemagglutinin and all the internal genes shows that there is a genuine structure to the sequence data and that the main cluster is stable over a range of identities. This indicates that there are multiple groups of sequences which have small numbers of changes between them that are separated by larger distances to other groups in a multi-modal population. This is strong evidence to support the use of classification based on clustering as an appropriate method of analysing the internal genes. By using clustering we can identify groups that share a close ancestry and we also reduce the variability in the dataset which improves alignments.

The sequences for the internal genes are often incomplete as they are not sequenced as often and in as much detail as the HA and NA segments. This is particularly an issue with older sequences and also with the PB1 segment. USEARCH was developed for use in next generation sequencing projects to assign sequences to operational taxonomic units (14). USEARCH is based around a BLAST like algorithm which detects words in common between the sequences and so clustering does not require the full-length sequences. For this reason, USEARCH seems to be robust to partial sequences, although the ambiguity introduced by missing sections might explain the large number of clusters for the PB1 segment. Some of the PB1 clusters such as cluster 4 contain very short stretches of sequence (only 173 bases) and it is possible that some clusters result from the truncation.

In almost all cases the median of the distances between the cluster centroids is smaller than the maximum distance within the clusters and the maximum distances between centroids are significantly larger. This indicates that the clusters are well separated and that only a small number of sequences will lie at the border between clusters.

The within cluster distances are NA when there is a single sequence because there is no distance to calculate but there are also cases where the intra-cluster distances cannot be calculated for non-singleton clusters. These occur for the PA and PB1 sequences (tables 4 and 5) and they are caused by a large number of sequences with long stretches of unknown nucleotides which are represented as the string N. In one cluster the number of Ns is so large that the sequence is 100% different to the other sequences. The only way of dealing with this challenge is manual editing of the alignments

which with such large numbers of sequences is resource intensive and time consuming. Manual examination and editing of the polymerase cluster 1 showed that the cluster was well defined but given that there are over 5000 sequences it was not possible to edit the alignment to remove all the problem sequences. Automated methods for quality control need to be developed for the PA and PB1 segments where there are very large variations in the data collected.

One of the feature we would expect to see if the data has been appropriately clustered into lineages is that some of the clusters should be homogeneous for subtype or location. This is equivalent to having a monophyletic clade in a phylogenetic analysis.

Most of the small clusters containing less than 10 sequences are likely to be homogeneous (table 1-6). These demonstrate that the cluster membership is much more than a random assignment and that clades do have biological significance. For example, two bat sequences from South America and the H18N11 subtype can only be found in cluster 23 of the matrix segment, cluster 16 of the NP segment and cluster 29 of the PB2 segment. There are only single H18N11 sequences available in the other gene segments. Another example from the PB2 internal gene clusters, cluster 26 is monophyletic for China and nearly monophyletic for H9N2 as there are only a few H3N8, H5N1 and H5N3 sequences within the clade.

Examination of most of the clades reveals some degree of homogeneity in either subtype, location or host species, but there are some clades which are highly polyphyletic and require further investigation. One of these is clade 24 from the PB2 internal gene clusters. Clade 24 contains a mixture of subtypes and locations as well as hosts. There are two distinct groups, one is Murre from Alaska in 1976 of the H1N6 subtype the others are Eurasian sequences from the 1980s and 1990s from a wide range of subtypes. It is possible that this lineage spread from sea-birds to other water fowl in the 1980s via bird migration, but based on the available evidence we cannot suggest more than a plausible hypothesis in this case. This is supported by the PB1 cluster 6 which contains the same Alaskan Murre samples from 1976 along with a wide variation of other subtypes including water-fowl from the 1980s. In these highly heterogeneous cases we are more confident about the cluster assignment if it has been identified in other segment lineages.

The H5N8 subtype was used as a test case to see if the clustering agreed with previous phylogenetic studies. There is good agreement with the previous studies of the sporadic US sequences but the results are less clear for the recent Eurasian outbreak and its sub-sequent spread to North America (16, 26). Sequences from that outbreak have been classified into two groups based on the hemagglutinin gene segment, Gochang and Buan (25). The Gochang sequences also form a cluster

with the sequences from China and this is the suggested point of origin. The Buan clade spread more widely including to Taiwan in 2015.

However, from the internal genes the clustering suggests a much more complex pattern of reassortment (table 7). These results are consistent with the current phylogenies but they indicate that there have been numerous reassortment events during the outbreak including a change in the matrix gene segment when the virus spread to Taiwan. The Viet Nam sequence also corresponds to a divergent viral genome even though it has some features in common with the Chinese, South Korean and Taiwanese viruses. There is evidence in a recent paper that the Viet Nam sequence is the source of another H5N8 lineage (27).

The North American groups are a much easier to analyse because of the small number of available sequences. While they all share common features they are nevertheless all unique in their patterns of internal genes. This lends further support to the hypothesis that the H5N8 subtype has arisen multiple times in North America from H5 and N8 containing subtypes, but that this subtype was short-lived and did not spread widely until the 2014 outbreak.

This study is only the beginning of analysing the massive amounts of internal gene segment sequence data that are available. In this study we have classified nearly a quarter of a million sequences classified into 270 clusters across the six internal gene segments. All of the original data, the aligned clustered sequences and the script used for analysis are available for download via Zenodo (https://doi.org/10.5281/zenodo.832431). Unravelling the reassortment history for influenza is going to be an enormous task beyond any single group and we welcome everyone to use the data for their own analysis.

The next step is a complete statistical analysis of the distribution of hemagglutinin and neuraminidase subtypes amongst the different lineages, as well as their geographical and host distributions. From this we can start to build a better understanding of the extent of reassortment in the influenza A viral genome.

Cluster	Number of Sequences	Alignment	Median Cluster Distance (%)	Maximum Cluster Distance (%)
0	14952	MAFFT	0	8.2
1	12309	MAFFT	1.1	12
2	2908	MAFFT	0	14
3	5539	MAFFT	3.2	15
4	2607	MAFFT	2.5	14
5	153	Muscle	8.2	14
6	378	Muscle	4	11
7	2417	MAFFT	5	14
8	480	Muscle	1.9	7.8
9	388	Muscle	6.8	14
10	902	Muscle	8.1	15
11	15	Muscle	1.5	2.7
12	219	Muscle	6.3	14
13	1230	Muscle	1.1	8.5
14	220	Muscle	4.4	13
15	1685	Muscle	3.4	13
16	3	Muscle	1.7	1.7
17	137	Muscle	4.9	14
18	24	Muscle	6.3	7
19	494	Muscle	4.4	12
20	769	Muscle	5.1	16
21	206	Muscle	3.9	11
22	92	Muscle	7.5	13
23	2	Muscle	9.4	9.4
24	77	Muscle	6.9	14
Centroids			14	39

Table 1: A summary of the within cluster statistics for the M gene segment. The summary statistics for the centroids (between cluster distances) are also given for comparison.

Cluster	Number of Sequences	Alignment	Median Cluster Distance (%)	Maximum Cluster Distance (%)
0	10344	MAFFT	0.97	10
1	11761	MAFFT	1.1	12
2	826	Muscle	6.2	12
3	1768	Muscle	3.5	9.5
4	1648	Muscle	2.1	11
5	36	Muscle	1.4	7.1
6	6534	MAFFT	4.6	12
7	2	Muscle	9	9
8	4087	MAFFT	1.1	8.8
9	322	Muscle	5.6	10
10	53	Muscle	7.3	12
11	383	Muscle	6.8	13
12	268	Muscle	6	13
13	50	Muscle	6.9	9.8
14	197	Muscle	2.8	9.6
15	103	Muscle	7.2	12
16	2	Muscle	4.7	4.7
17	14	Muscle	3.1	7.3
18	6	Muscle	0.34	0.4
19	1237	Muscle	6	12
20	2	Muscle	11	11
21	3	Muscle	10	11
22	146	Muscle	0.13	11
23	7	Muscle	1.7	4.7
24	37	Muscle	4.3	10
25	3	Muscle	4.8	4.8
26	20	Muscle	6.5	7.8
27	1		NA	NA
28	63	Muscle	3.9	11
29	572	Muscle	5.4	12
30	17	Muscle	5	10
31	106	Muscle	1.4	11
32	1		NA	NA
33	1		NA	NA
Centroids			19	46

Table 2: A summary of the within cluster statistics for the NP gene segment. The summary statistics for the centroids (between cluster distances) are also given for comparison.

Cluster	Number of Sequences	Alignment	Median Cluster Distance (%)	Maximum Cluster Distance (%)
0	4392	MAFFT	7	17
1	10955	MAFFT	1.6	14
2	2336	MAFFT	0.69	6.5
3	10319	MAFFT	0.83	10
4	2806	MAFFT	3.1	15
5	38	Muscle	1.5	5.1
6	798	Muscle	2.9	14
7	3318	MAFFT	4.3	15
8	97	Muscle	2.5	12
9	234	Muscle	3.8	9.3
10	339	Muscle	2.1	8.5
11	116	Muscle	9.3	14
12	321	Muscle	6.7	14
13	18	Muscle	6.1	9.4
14	1690	Muscle	3.5	11
15	464	Muscle	1.6	10
16	45	Muscle	3.3	8.7
17	153	Muscle	4.6	11
18	44	Muscle	7	12
19	2	Muscle	9.6	9.6
20	913	Muscle	2.2	12
21	5	Muscle	1.7	2.2
22	217	Muscle	3.8	11
23	8	Muscle	0.43	3.4
24	36	Muscle	2.7	7.7
25	1		NA	NA
26	329	Muscle	4.7	13
27	3	Muscle	3.7	3.7
28	2	Muscle	0	0
29	2	Muscle	3.8	3.8
30	1170	Muscle	4.8	18
31	67	Muscle	2.8	8.3
32	46	Muscle	5.7	12
33	14	Muscle	0.58	9.5
34	9	Muscle	2.4	5.3
35	344	Muscle	2.2	8.8
36	64	Muscle	4.7	12
37	179	Muscle	6	13
38	13	Muscle	8.2	10
39	209	Muscle	6.2	15
40	266	Muscle	3.9	11
41	1		NA	NA
42	27	Muscle	3.4	10
43	111	Muscle	3.4	11
44	93	Muscle	4.3	10
45	74	Muscle	2.1	17

46	36	Muscle	2.3	15
47	1		NA	NA
48	14	Muscle	5.1	8.3
49	9	Muscle	8	10
50	2	Muscle	9.9	9.9
51	3	Muscle	2.1	2.1
52	13	Muscle	8.1	11
53	1		NA	NA
Centroid			17	59

Table 3: A summary of the within cluster statistics for the NS gene segment. The summary statistics for the centroids (between cluster distances) are also given for comparison.

Cluster	Number of Sequences	Alignment	Median Cluster Distance (%)	Maximum Cluster Distance (%)
0	1852	Muscle	0	100
1	5812	MAFFT	NA	NA
2	11	Muscle	10	13
3	2052	MAFFT	1.5	13
4	6924	MAFFT	NA	NA
5	515	Muscle	4.5	17
6	1215	Muscle	1.8	16
7	1935	Muscle	0	84
8	676	Muscle	4.9	18
9	549	Muscle	0.17	4.3
10	877	Muscle	0.95	10
11	2528	MAFFT	0	NA
12	164	Muscle	3.3	11
13	9	Muscle	3.6	9.3
14	90	Muscle	3.6	7.5
15	123	Muscle	3.5	9.7
16	10	Muscle	2.9	4.5
17	46	Muscle	2.9	12
18	88	Muscle	1.9	6.3
19	60	Muscle	2.8	10
20	520	Muscle	NA	NA
21	10	Muscle	1.7	2.7
22	65	Muscle	1.3	11
23	441	Muscle	3.6	14
24	233	Muscle	NA	NA
25	24	Muscle	7.1	12
26	104	Muscle	6.9	73
27	1107	Muscle	3.5	15
28	5	Muscle	0.097	4.6
29	1592	Muscle	NA	NA
30	3	Muscle	3.1	3.1
31	32	Muscle	3.3	7.8
32	23	Muscle	3.3	5.8
33	13	Muscle	6.1	8.4
34	3	Muscle	0.25	0.31
35	328	Muscle	6.7	11
36	65	Muscle	6.8	11
37	131	Muscle	NA	NA
38	1031	Muscle	3	22
39	147	Muscle	5.5	12
40	10	Muscle	4.7	10
41	3	Muscle	3.5	3.7
42	276	Muscle	2.8	7.9
43	1		NA	NA
44	2	Muscle	10	10

45	29	Muscle	1.4	4.4
Centroid			9.7	96

Table 4: A summary of the within cluster statistics for the PA gene segment. The summary statistics for the centroids (between cluster distances) are also given for comparison.

Cluster	Number of Sequences	Alignment	Median Cluster Distance (%)	Maximum Cluster Distance (%)
0	1076	Muscle	5.4	25
1	5281	MAFFT	0	28
2	29	Muscle	2.5	13
3	4019	Muscle	0	16
4	448	Muscle	2.1	8.7
5	473	Muscle	8.1	25
6	6786	Muscle	5.8	28
7	97	Muscle	4	15
8	11306	MAFFT	NA	NA
9	1811	Muscle	NA	NA
10	1767	Muscle	0	16
11	44	Muscle	10	21
12	59	Muscle	1.3	13
13	165	Muscle	NA	NA
14	1971	Muscle	2	28
15	174	Muscle	7.2	20
16	23	Muscle	6.7	15
17	38	Muscle	4.6	16
18	174	Muscle	2.6	10
19	225	Muscle	7.1	23
20	22	Muscle	8.9	16
21	1		NA	NA
22	9	Muscle	8.9	13
23	144	Muscle	3.3	15
24	193	Muscle	6.5	35
25	220	Muscle	7.9	18
26	178	Muscle	3.9	18
27	26	Muscle	7.6	9.7
28	261	Muscle	7.1	22
29	2	Muscle	6.1	6.1
30	1		NA	NA
31	69	Muscle	2.6	14
32	20	Muscle	9.7	16
33	422	Muscle	5.3	27
34	123	Muscle	6.5	36
35	356	Muscle	7.6	36
36	66	Muscle	10	24
37	4	Muscle	1.3	2.6
38	40	Muscle	1.5	8.2
39	5	Muscle	2	7.5
40	448	Muscle	2.1	8.7
41	442	Muscle	6.7	19
42	252	Muscle	1.7	18
43	6	Muscle	2.6	4.7
44	5	Muscle	3.4	4.1

45	25	Muscle	7.4	10
46	12	Muscle	8.4	14
47	653	Muscle	8.2	39
48	1		NA	NA
49	12	Muscle	4.8	8.2
50	34	Muscle	5.2	16
51	1		NA	NA
52	3	Muscle	0.19	0.23
53	67	Muscle	5.6	13
54	2	Muscle	5.2	5.2
55	34	Muscle	11	22
56	1		NA	NA
57	11	Muscle	6.7	11
58	1		NA	NA
59	74	Muscle	3.4	15
60	2	Muscle	11	11
61	1		NA	NA
Centroid			11	81

Table 5: A summary of the within cluster statistics for the PB1 gene segment. The summary statistics for the centroids (between cluster distances) are also given for comparison.

Cluster	Number of Sequences	Alignment	Median Cluster Distance (%)	Maximum Cluster Distance (%)
0	1622	Muscle	2.8	16
1	11288	MAFFT	0.96	11
2	9959	MAFFT	1	9.7
3	125	Muscle	6.5	13
4	4444	Muscle	4.6	12
5	1160	Muscle	4.1	9.6
6	3346	MAFFT	2.4	8.9
7	69	Muscle	6.2	13
8	101	Muscle	5.9	17
9	485	Muscle	5.2	12
10	138	Muscle	7.1	13
11	107	Muscle	3.1	14
12	875	Muscle	4	15
13	288	Muscle	5.1	11
14	43	Muscle	6.8	10
15	60	Muscle	5.7	11
16	166	Muscle	2.8	9.7
17	82	Muscle	6	11
18	436	Muscle	4.2	11
19	218	Muscle	7.6	13
20	10	Muscle	8.2	8.6
21	57	Muscle	3	10
22	407	Muscle	4.7	15
23	330	Muscle	2.3	9.8
24	60	Muscle	9.1	13
25	288	Muscle	3.3	9.1
26	61	Muscle	3.7	12
27	37	Muscle	9.8	13
28	3	Muscle	4.1	4.8
29	2	Muscle	2.1	2.1
30	2	Muscle	2.4	2.4
31	11	Muscle	5.9	11
32	5	Muscle	0.13	3.7
33	24	Muscle	5.1	13
34	549	Muscle	6.9	12
35	3	Muscle	4.9	4.9
36	1535	Muscle	2.9	10
37	11	Muscle	4.8	14
38	11	Muscle	9.6	10
39	19	Muscle	8.4	10
40	11	Muscle	0.13	0.62
41	76	Muscle	3.5	5.4
42	111	Muscle	2	11
43	233	Muscle	4	7.9
44	132	Muscle	5.1	12

45	1		NA	NA
46	2	Muscle	0.75	0.75
47	1		NA	NA
Centroid			18	49

Table 6: A summary of the cluster statistics for the PB2 gene segment. The summary statistics for the centroids (between cluster distances) are also given for comparison.

Sub-clade	M	NP	NS	PA	PB1	PB2
CY0*** Ireland 1983	10	3	20	10	15	27
GU0*** Ireland 1983	8	-	17	12	53	-
New Jersey Ruddy Turnstone 2001	3	2	14	38	6	4
Maryland 2007	7	6	14	29	6	36
Colorado 2006 New Jersey 2006	3	6	3	7	6	36
China 2010-14	4	3/19	7	11/42	35	5
California 2011	7	6	15	29	6	4
China 2012	4	3/19	4	8	7/14/35	5/6
Viet Nam 2013	2	19	4	8	10	6
South Korea 2014	0/4	3/19	0/7/20	8/11/27/42	7/9/10/35	5/6
California Quail 2014	3	6	14	-	6	4
Taiwan Duck 2015	4	15	15	11	10	6
Taiwan Goose 2015	22	19	7/20	8	7/10/35	5

Table 7: Summary of the internal gene cluster assignments for the sequences from the H5N8 subtype for the H5 hemagglutinin sub-clades.

Compliance with Ethical Standards

Conflict of Interest: The authors declare that they have no conflict of interest.

This article does not contain any studies with human participants or animals performed by any of the authors. The research does not require informed consent.

.

Supplementary Materials

Supplementary Table 1: Summary of the downloaded Influenza A virus sequence data.

Supplementary Table 2: The statistical properties of the clusters produced by USEARCH for the internal gene segments and H5 and H9 influenza A hemagglutinins.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3396v1 | CC BY 4.0 Open Access | rec: 7 Nov 2017, publ: 7 Nov 2017

References

1. Alexander D.J., Vaccine 25, 5637-5644, 2007.
2. Hiromoto Y., Yamazaki Y., Fukushima T., Saito T., Lindstrom S.E., Omoe K., Nerome R., Lim W., Sugita S., and Nerome K., Journal of General Virology 81, 1293-1303, 2000.
3. Webster R.G., Bean W.J., Gorman O.T., Chambers T.M., and Kawaoka Y., Microbiological Reviews 56, 152-179, 1992.
4. Zhou N.N., Senne D.A., Landgraf J.S., Swenson S.L., Erickson G., Rossow K., Liu L., Yoon K.-j., Krauss S., and Webster R.G., Journal of virology 73, 8851-8856, 1999.
5. McGinnis S., and Madden T.L., Nucleic acids research 32, W20-W25, 2004.
6. Liu Q., Ma J., Liu H., Qi W., Anderson J., Henry S.C., Hesse R.A., Richt J.A., and Ma W., Archives of virology 157, 555-562, 2012.
7. Guo Y.J., Wen L.Y., Zhang Y., Wang M., Guo J.F., Li Z., and Shu Y.L., Zhonghua shi yan he lin chuang bing du xue za zhi = Zhonghua shiyan he linchuang bingduxue zazhi = Chinese journal of experimental and clinical virology 19, 358-361, 2005.
8. WHO O., Emerging infectious diseases 14, e1, 2008.
9. Donis R.O., and Smith G.J., Influenza and other respiratory viruses, 2015.
10. Smith G.J., and Donis R.O., Influenza and other respiratory viruses 9, 271-276, 2015.
11. Anderson T.K., Campbell B.A., Nelson M.I., Lewis N.S., Janas-Martindale A., Killian M.L., and Vincent A.L., Virus research 201, 24-31, 2015.
12. Anderson T.K., Macken C.A., Lewis N.S., Scheuermann R.H., Van Reeth K., Brown I.H., Swenson S.L., Simon G., Saito T., and Berhane Y., mSphere 1, e00275-00216, 2016.
13. Hurtado R., Fabrizio T., Vanstreels R.E.T., Krauss S., Webby R.J., Webster R.G., and Durigon E.L., PloS one 10, e0145627, 2015.
14. Edgar R.C., Bioinformatics 26, 2460-2461, 2010.
15. Dalby A.R., and Iqbal M., PeerJ 3, e934, 2015.
16. Dalby A.R., F1000Research 5, 2016.
17. Dalby A., PeerJ PrePrints 3, e1489v1481, 2015.
18. Dalby A., PeerJ PrePrints 3, e1250v1251, 2015.
19. Squires R.B., Noronha J., Hunt V., García-Sastre A., Macken C., Baumgarth N., Suarez D., Pickett B.E., Zhang Y., and Larsen C.N., Influenza and other respiratory viruses 6, 404-416, 2012.
20. Edgar R.C., Nucleic acids research 32, 1792-1797, 2004.
21. Katoh K., and Standley D.M., Molecular biology and evolution 30, 772-780, 2013.
22. Paradis E., Claude J., and Strimmer K., Bioinformatics 20, 289-290, 2004.
23. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. 2014.
24. R Studio., RStudio Inc, Boston, Massachusetts, 2012.
25. Jeong J., Kang H.-M., Lee E.-K., Song B.-M., Kwon Y.-K., Kim H.-R., Choi K.-S., Kim J.-Y., Lee H.-J., and Moon O.-K., Veterinary microbiology 173, 249-257, 2014.
26. Ip H.S., Torchetti M.K., Crespo R., Kohrs P., DeBruyn P., Mansfield K.G., Baszler T., Badcoe L., Bodenstein B., and Shearn-Bochsler V., Emerging infectious diseases 21, 886, 2015.
27. Selim A.A., Erfan A.M., Hagag N., Zanaty A., Samir A.-H., Samy M., Abdelhalim A., Arafa A.-S.A., Soliman M.A., and Shaheen M., Emerging infectious diseases 23, 1048, 2017.