Using a fast clustering method for viral segment lineage determination, applied to the H9 influenza hemagglutinin.

Andrew R. Dalby
Department of Biomedical Sciences, University of Westminster, 115 New Cavendish Street, Westminster, W1W 6UW, UK
a.dalby@westminster.ac.uk

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

Joshua E. Sealy
The Pirbright Institute, Pirbright, Woking, Surrey, GU24 0NF, UK

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

Abstract

Lineage determination is an important part of the analysis of viral sequence data. Previously this has depended on phylogenetic analysis in order to identify distinct clades within the phylogenetic trees. This method is time consuming and dependent on a set of empirical rules for clade identification. An alternative approach is to use clustering. Clustering is commonly used to identify operational taxonomic units in next generation sequencing data. In this paper we use clustering in order to rapidly identify viral segment lineages and clades without the need for tree construction.

Keywords
clustering, viral lineages, influenza, hemagglutinin, H9

Background

It is well known in statistics that variance increases with sample size, as the rare cases within the population become more likely to be sampled. Biological sequence data is no exception. As the number of sequences included in a multiple sequence alignment increases so does the sequence diversity and as a consequence the number of totally conserved sequence positions decreases.

A multiple sequence alignment of 3659 influenza H9 hemagglutinin sequences from the Influenza Research Database reveals that less than 10% of the nucleotides are completely conserved [1]. Such a large amount of diversity and limited amount of conservation makes multiple sequence alignment challenging and the alignments can vary between different alignment programs and parameter settings. This degree of variation within H9 raises the question of population structure as the sequences have not diverged sufficiently in order to have produced a new subtype. Are there sub-groups which can be distinguished that have a smaller degree of variation? If this sub-structure exists then these would be much easier to align as separate groups.

These putative sub-populations would represent different lineages, and they are likely to have different antigenic properties. It is important to note that in the case of a segmented virus such as influenza these lineages need to be considered on a segment by segment basis. At the viral level there will be multiple different genotypes that contain the same gene segment lineage. For simplicity in this paper lineage will apply only at the segment level.

Segment lineages are sub-groups of sequences within a subtype that evolve independently of one another, but that may co-exist. These correspond to the highest level clades within the viral segment phylogenetic tree. The boundaries of the segment lineage will be carefully defined as all segment lineages will share a most recent common ancestor and so it can be difficult to determine when a new lineage begins.

Some segment lineages will still be circulating and are extant, but others will be extinct. The huge advantage of viral phylogenetics is that we have historical data and that we can observe viral evolution and the generation of new lineages. The assignment of segment lineages is biologically important for studies of viral reassortment, viral population genetics, phylodynamics and viral phylogeography.

Influenza A is already divided into subtypes dependent on the antigenic properties of the virus surface glycoproteins hemagglutinin (HA) and neuraminidase (NA). There are currently 18 known sub-types of HA and 11 subtypes of NA. In theory genetic reassortment events between these subtypes can give rise 198 different subtypes of virus (e.g. H9N1, H9N2, H9N3). However not all combinations have been detected in nature and there are preferential combinations for the viral packaging and fitness to cause productive infections in avian or mammalian species [2, 3].

Currently lineages have been identified for the highly pathogenic avian influenza form of the H5N1 hemagglutinin subtype descended from a goose in Guangdong (HPAI H5N1), H9N2,

global swine H1 and US swine H1 [4-8]. These lineage studies have focused on a single influenza subtype or in the case of the global swine H1 classification two subtypes H1N1 and H1N2 [8]. For the lineage analysis of HA it is logical to look at all the sequences from the same antigenic group as they form distinctive well defined groups, i.e. all of the H1 or all of the H5 sequences together. This has already begun to happen in the case of H5 where the nomenclature is no longer restricted to only the H5N1 subtype, but it still only contains sequences descended from the high pathogenicity avian influenza (HPAI) H5N1 and it does not include lineages for the low pathogenicity avian influenza (LPAI) lineages [9, 10].

The change in the HPAI H5N1 classification occurred in response to the recent spread of the HPAI H5 lineage to North America and the reassortment of the HA and NA segments into different subtypes (H5N2, H5N3, H5N6 and H5N8) demonstrates how a more comprehensive approach to lineage identification that crosses the boundaries of subtypes is required [11].

The current WHO clade nomenclature was produced from a multiple sequence alignment of all of the HPAI H5N1 hemagglutinin sequences using MAFFT and subsequent phylogenetic analysis using Fasttree [4]. Once the classification has been assigned new sequences can then be placed within the nomenclature using programmes such as pplacer, which is available through the Influenza Research Database [1, 12].

An alternative method which combines some of the elements of clade detection with classification is LABEL [7]. LABEL takes a supervised learning approach and uses HMMs to create patterns based on the existing classifications and subsequently SVM machine to assign sequences to the nomenclature. This has helped to clear up some of the discrepancies in the existing HPAI H5 classification as well as to develop a new classification scheme for H9N2 based on the existing literature.

Tree based approaches are very time consuming and often involve human intervention which prevents automation. The need to rebuild the tree when new sequences become available makes these methods particularly inefficient. Sequence alignment can be very slow for large numbers of sequences and although fast phylogenetic tree generation methods are available methods that use the non-parametric bootstrap are time consuming for large numbers of sequences.

Another problem of using tree based methods in order to identify the clade structure is that the quality of the tree depends on the quality of the multiple sequence alignment on which it is based. As alignments include a larger number of sequences with a higher variance the alignment becomes more ambiguous and the quality declines. A subset of sequences with less variation gives a more accurate alignment. For this reason an alternative approach to the tree based approach is presented here.

Clustering methods can be used to identify sub-groups in data and can be divided into two different approaches, divisive where the data is broken into ever smaller subsets and agglomerative where the subsets are built up from the individual members of the dataset [13]. All that is needed is a metric for assigning sequences to one group or another.

Clustering is an example of an unsupervised learning technique that will find groups within a dataset without prior classification. The data is divided into groups around a central "average" example. This is the centroid – the median example of the cluster.

USEARCH is an algorithm that was written for clustering next generation sequencing data in order to identify operational taxonomic units (OTUs) prior to multiple sequence alignment [14]. The task of identifying viral segment lineages (abbreviated from now on to lineages) is much the same except that the segments are longer that the reads produced by next generation sequencing and the numbers of sequences that need to be aligned are orders of magnitude smaller.

In this paper lineage is used to mean the deepest level of the clade structure. This means the first highest level sub-groups within the phylogenetic tree, having the lowest level of sequence identity. The terms lineage and clade will be used interchangeably.

Materials and Methods

A complete set of all of the influenza A HA sequences for all of the different subtypes was downloaded from the Influenza Research Database on 21st of May 2017. This dataset was labelled 2017_5_21_HA This provides a positive control for applying USEARCH to detecting groups within the viral sequence data as 18 hemagglutinin subtypes have already been assigned.

The complete sets of HA sequences for H1, H2, H3. H4, H5, H6, H7 and H9 were downloaded from the Influenza Research Database on 21st of May 2017. These datasets were used to test the statistical properties of the clustering method. USEARCH (version 9.2.64) was used to cluster the data at different levels of identity ranging between 80% and 99% [14].

An example of the command line for using USEARCH to calculate the cluster centroids at 90% is:

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

A complete clustering analysis at the 90% identity level was carried out for the H9 hemagglutinin sequences. This was verified against the clades that had been identified previously in the literature. The USEARCH command line for creating the set of clusters corresponding to the lineages was:

usearch –cluster_fast sequences.fasta –id 0.9 –clusters sequences_90_

The non-redundant set of sequences containing the centroid sequence from each of the lineages was also needed for 90% sequence identity. This was needed for a phylogenetic tree construction of the relationship between the lineages.

Cluster 2 is one of the two clades containing hemagglutinin sequences from the Middle East and contains over 400 sequences. To make the phylogenetic analysis more manageable it was partitioned further using a second clustering with USEARCH this time at the 93% level. This generated sub-clusters that were then used for subsequent phylogenetic tree analysis.

Cluster 1 is the other large clade with over 2300 sequences from the Far-East. This was also divided into sub-clades by the nested use of USEARCH. Classification was carried out at 93% identity and then 95% identity in order to create a set of nested clades for comparison with the results from LABEL [7].

Multiple sequence alignment was carried out for all of the clusters including the set of non-redundant sequences using Muscle in Mega v7.0.20 except for clusters1 and cluster2 that were aligned using the stand-alone version of Muscle v3.8.31 because of memory constraints in Mega [15, 16]. Intra-cluster distance statistics were calculated using the APE package in the R statistical environment using R-studio [17-19].

Where subsequent phylogenetic analysis of the lineages was required the analysis was carried out using BEAST2 [20]. All of the trees were generated using tip dates, a strict

evolutionary clock and the constant population coalescent tree prior. The general time reversible (GTR) substitution model with 50% invariant sites was used unless the simulations did not converge when the Tamura-Nei substitution model was used [21, 22]. All trees are the maximum clade credibility trees with the median node heights processed using Treeannotator v2.4.6. The trees were created and edited in FigTree v1.4.3 [23].

Results

USEARCH on the HA data

USEARCH identified 18 different clusters from the combined HA data at the 70% identity level. The representative sequence from each of the identified subgroups is given in table 1. Note that there are two sequences from H7 and no sequences from the H15 subtype were classified in their own cluster.

USEARCH correctly classified 17 out of the 18 hemagglutinin subtypes.

USEARCH on the Hemagglutinin Data

All of the different HA subtypes exhibited the same statistical properties for the clustering as the percentage identity was increased (table 2). Initially there is a single cluster that breaks down into a large cluster with a small number of much smaller clusters. These smaller clusters continue to fragment as the percentage identity increases until finally this large cluster breaks up. This can be seen clearly in the graphs of maximum cluster size and the number of clusters (figures 1 and 2).

USEARCH on H9 Hemagglutinin

USEARCH identified 19 clusters at the level of 90% identity. A summary of the clusters is given in table 3. Of these clades five are monophyletic for subtype and five are monophyletic for geographical location (if China and Hong Kong are considered together). The subtype originated in Wisconsin in 1966 and this clade continues to be in circulation.

A phylogenetic tree of the lineages is given in figure 3. The clades are numbered according to the dates in which the earliest members were sampled. This numbering is likely to remain consistent as most sampling occurs in real-time and the additional of historical samples occurs very rarely. The nodes are labelled with the posterior probabilities which are mostly close to one.

While many of the clades are monophyletic to Eurasia or the Americas they are not clearly distinguishable on the phylogenetic tree. Clades 1,2,3,4,6,9, 16 and 17 contain sequences from the Americas most of these form part of a consistent group highlighted in orange along with clades 10, 11 and 19 which are monophyletic to South Korea, Malaysia and Vietnam respectively (clade 10 has a single Egyptian sequence from 2016).

Clades 3 and 4 have a particularly widespread distribution and can be considered as circulating globally. The oldest sequence from clade 3 is AY206674 from a Hong Kong duck in 1976. Clade 3 has three subclades, 1976-1984, Canada/USA 1978-1991 and USA/China/South Korea 2005-2014. The phylogenetic tree for clade 3 is shown in figure 4. Clade 4 does not show any clear sub-structure and is shown in figure 5.

Clades 12 and 15 are mostly Middle Eastern and South Asian sequences and these are highlighted in blue in figure 3. The evolution of H9N2 in the Middle East has been the subject of previous phylogenetic studies and clades 12 and 15 were chosen for further analysis. A second round of clustering using USEARCH was used on Clade 15 to make the cluster size more manageable and generated 13 sub-clades. A summary of these sub-clades is given in table 4.

The phylogenetic trees for these two clades and their sub-clades are shown in figures 6 and 7a-h. The Guinea Fowl and Ferret sequences had to be removed from Clade 12 as they are undated. Clade 15.4 has an unusually large median distance between sequences and so it was removed as an outlier from the clade 15 trees which improved the convergence of the BEAST phylogenetic analysis.

The Nested Application of USEARCH to Cluster One

At the level of 93% identity there were 3 sub-clades Clades14.1, Clades14.2, Clade14.3, Clade14.4. Clades14.3 contains 2280 sequences and was subdivided again. This time at 95% identity there were 10 sub-clusters but 2 contained 10 or less sequences and one sub-subclade 14.3.2 contained 1278 sequences. If classification continues at 97% there are 31 clusters but half of these contain less than five sequences and so it was decided to end subdividing the clades at the 95% level.

Discussion

The clustering of the influenza viral hemagglutinins using USEARCH proved that clustering can correctly identify the viral subtypes from the sequence data. USEARCH has been used previously in order to classify influenza genomes but this is the first time that it has been used for lineage detection in viral gene segments [24, 25].

These subtypes have been assigned by antigenic properties and the differences between the hemagglutinin subtypes is clear at the sequence level. The reason for the failure to detect the H15 subtype is likely to be because there are such a limited number of H15 sequences available (there are currently only 21 complete sequences for H15 hemagglutinin in the Influenza Research Database). Why two H7 clusters were detected remains unclear.

The same properties are observed in the clustering statistics for all of the different hemagglutinin subtypes (table 1, figures 1 and 2). This is consistent with there being a population structure within the gene segments and that is not an artefact of the clustering process. In all subtypes the clusters are centred around a large main grouping while the other smaller clusters are much more diverse.

The best choice of percentage identity for clustering is when this large cluster first becomes clearly defined but when the smaller clusters have not been fragmented too much. From the graph of maximum cluster size this is beginning of the plateau region. For H9 this is at 90% but it varies between the different subtypes and for example it is 93% for H7.

The median sequence to sequence distances are given in table 3 for all of the clusters. This compares with a minimum distance between the centroids of 11% and a median difference between the centroids of 18%. This shows that the clusters are well separated from one another. However, the posterior probability for the split between clades 10 and 17 is only 0.48 which suggests that these two clades might need to be merged.

Where USEARCH produces very large clusters a nested approach can be used [14]. Each of the large clusters can be divided by performing further runs of USEARCH classification at a higher degree of sequence identity. For the phylogenetic analysis of clade 15 the resulting classification at 93% identity produced 13 clusters with a maximum cluster size of 196, which was considered manageable for tree analysis. This process of nested clustering can be repeated as often as required by increasing the identity threshold until either an acceptable maximum cluster size is obtained or the clustering reaches 98.5% identity which is threshold for the WHO nomenclature framework.

A summary of the second round of classification for clade 15 is given in table 4, this includes the median sequence to sequence distances for the clade. Apart from subclade 15.4 these values are much smaller than the minimum distance between centroids od 7.7%. Clade 15.4 only contains 3 sequences and it was considered an outlier. It could be that this is just a result of poor sampling from within that sub-clade. Care must be taken over how you deal with the small clusters produced by successive runs. In this case the small sub-clades of four

or less sequence were merged for phylogenetic analysis as otherwise BEAST performance is impaired.

There has been a previous study that used a phylogenetic tree approach to analyse the complete set of H9N2 hemagglutinins [26]. That study highlighted fours main clades labelled A, B and C and a very large Chinese Clade. Clade A corresponds to the clades coloured in orange in figure 3 and clade B corresponds to the Middle-Eastern clades highlighted in blue. Clade C corresponds to clade 13 in the current paper and the large Chinese clade corresponds to clade 14. That paper attempted to bring together all of the different clade names that had been used in the literature but this was carried out more systematically in LABEL [7].

LABEL uses a hierarchical approach which is analogous to the nested approach used in clustering here. The prototype sequences used by LABEL and their equivalents in the current dataset are given in table 5. The current method does not distinguish as many lineages within cluster 4 as are present in the LABEL classification, but it is important to note that LABEL was constrained by the literature definitions of the lineages and may be subjective.

The origin of clade 12 is the G1 sequence AF156378/KY785896 from a quail in Hong Kong in 1997 and clades 12 and 15 match with the previously identified G1 lineage from the paper of Fusaro *et al.* [27] Clade 15.7 corresponds exactly to the G1_Mideast_B clade from Shanmuganathan *et al.* 2014 [28]. This clade was not available when LABEL was developed and has been added as G1_Mideast_B2. The other G1 clades are classified as separate clusters in the current analysis except that A and D are both part of clade 12. However from figure 6 it is clear that clade 12 has three identifiable sub-groups and that A and D are separate at the sub-clade level in this analysis.

The H9 Y280 clade and its relatives make up a substantial portion of the sequences classified by LABEL. This H9 Y280 sequence itself is a partial sequence and so it is not present in the current dataset. The nearest related sequence that is present is AF461530 a chicken from Beijing in 1997, this is found in the large Far-Eastern clade 14 in sub-clade 14.1. LABEL does not have the nested substructure of these mostly Chinese clades that was reported previously by Dalby and Iqbal and that is found here in the sub-clades of clade 14 [26].

The novel clades that have been identified in this study are clades 2,3,7,10,11 and 13. It should be noted that most of these clades have bifurcated from a clade that is present in the LABEL classification and that these are small changes in the classification system. In general the agreement between the two classifiers is strong.

This method goes beyond what LABEL can do by identifying new clades without dependence on the literature. Clustering methods quickly identify lineages within the dataset. By the successive use of identify cut-offs large clusters can be broken into more manageable subsets in a nested algorithmic way that is consistent with the phylogenetic analysis. Multiple sequence alignment of the subsets is rapid and phylogenetic tree generation using Bayesian coalescence based methods becomes possible for all of the clades.

Previous clade classification systems have tended to filter on influenza sub-type such as H5N1, H1N1 or H9N2. This study shows that clades can often be polyphyletic with regard to subtype because of reassortment of the hemagglutinin and neuraminidase genes in order to move between subtypes. By taking an inclusive and comprehensive approach a more complete picture of viral reassortment can be achieved.

Using clustering a complete phylogenetic analysis for the H9 HA sequences becomes computationally much more efficient. There is no need for users to subset the data or to apply filters depending on subtype or location which might introduce bias into the subsequent analysis. Visualisation of the resulting trees is also improved as well as the quality of the multiple sequence alignments. It is easier to see patterns in the data and to determine evolutionary constraints and processes.

Conclusion

Clustering also allows for the possibility of a comprehensive investigation of the influenza internal genes. The internal genes are not classified into subtypes like the glycoproteins HA and NA. This makes any phylogenetic analysis involving the internal genes complex and subject to selection criteria. Making it particularly difficult to analyse reassortment events where internal genes transfer between viral subtypes. By applying the clustering approach presented in this paper the analysis becomes much more manageable and it is possible to quickly identify clades within the internal genes where there is evidence for reassortment. These analyses will be much more objective because they can include all of the available data.

Table 1: The representative sequence from clustering the influenza A hemagglutinin sequences.

HA Subtype	Strain	Accession no.	Host Location and Date
H1	H1N1	CY0202085	A/Human/USA/1943
H2	H2N2	AB432937	A/Human/Japan/ 1957
H3	H3N2	HM628694	A/Human/Brazil/Human/2010
H4	H4N6	CY138045	A/American Black Duck /Canada /2010
H5	H5N1	KP739421	A/Green Winged Teal/USA/2014
H6	H6N8	CY190587	A/American Black Duck/USA/2007
H7	H7N7	K00429	A/Sea Mammal /USA/1980
H7	H7N3	AB558258	A/Chicken/Chile /2002
H8	H8N4	CY128894	A/American Black Duck /Canada/2007
H9	H9N2	KX185901	A/Oystercatcher//Chile/2015
H10	H10N6	CY139489	A/ American Black Duck /Canada/ ?/2010
H11	H11N2	KJ729359	A/Adelie Penguin /Antarctica//2013
H12	H12N6	CY139550	A/American Black Duck /Canada/2010
H13	H13N6	KX979504	A/Gull/Netherlands/?/2012
H14	H14N3	KY644423	A /Blue Winged Teal / Guatemala/ 2011
H15			-
H16	H16N3	GQ907294	A/Black Headed Gull/ Mongolia/2006
H17	H17N10	CY103892	A/Yellow Shouldered Bat /Guatemala/2010
H18	H18N11	KR077932	A/ Bat/Bolivia/ 2011

Table 2: The statistical properties of the clusters produced by USEARCH for the H1, H2, H3, H4, H5, H6, H7 and H9 influenza A hemagglutinins.

	Number of Clusters							
Identity (%)	H1	H2	H3	H4	H5	H6	H7	H9
80	3	1	8	2	3	3	5	2
85	13	2	16	6	5	5	7	6
90	45	6	41	11	20	14	12	19
91	61	8	48	15	22	17	14	26
92	96	11	69	19	24	23	18	36
93	128	13	90	23	35	28	28	46
94	166	16	126	30	55	45	31	64
95	271	23	149	42	86	64	46	88
96	357	34	257	64	133	98	67	150
97	550	59	362	106	328	173	105	249
98	1091	99	678	281	511	298	237	600
99	3499	213	2731	618	1624	622	470	1582

	Maximum Cluster Size							
Identity (%)	H1	H2	H3	H4	H5	H6	H7	H9
80	17945	610	18001	1386	4296	1401	1496	3508
85	17851	331	17596	1294	3945	928	1469	2906
90	17141	327	15227	1009	2975	656	904	2398
91	16170	282	15286	961	2254	646	1346	2328
92	15453	255	14844	934	2495	588	1273	2371
93	15405	235	14487	890	998	456	880	2319
94	15378	214	14256	832	715	404	910	1627
95	15372	193	8111	653	600	381	815	1290
96	15082	187	6909	380	916	294	829	1639
97	15294	135	6466	260	726	232	783	1268
98	9631	93	5821	182	436	84	768	671
99	3325	43	3604	72	217	44	596	109

	Average Cluster Size							
Identity (%)	H1	H2	H3	H4	H5	H6	H7	H9
80	7899	610	2531.4	909	1656	556.3	508.4	1829.5
85	1822.3	305	1265.7	303	993.6	333.8	363.1	609.8
90	526.6	101.7	493.9	165.3	248.4	119.2	211.8	192.6
91	388.5	76.2	421.9	121.2	225.8	98.2	181.6	140.7
92	246.8	55.5	293.5	95.7	207	72.6	141.2	101.6
93	185.1	46.9	225	79	141.9	59.6	90.8	79.5
94	142.8	38.1	160.7	60.6	90.3	37.1	82	57.2
95	87.4	26.5	135.9	45.3	57.8	26.1	55.3	41.6
96	66.4	17.9	78.8	28.4	37.4	17	37.9	24.4
97	43.1	10.3	55.9	17.2	21.8	9.6	24.2	14.7
98	21.7	6.2	29.9	6.5	9.7	5.6	10.7	6.1
99	6.8	2.9	7.4	2.9	3.1	2.7	5.4	2.3

Table 3: Summary of the H9 lineages identified by USEARCH at the 90% identity level.

USEARCH Cluster	No. of Sequences	Median Distance	Subtypes	Geographical Location	Dates	Clade Number	Host Species	Representative Sequence, Accession Number
0	6	10%	H9N2, H9N7, H9N9, mixed	Argentina, Chile, USA, Delaware	1996-2015	9	Oystercatcher, Pochard, Plover, Ruddy Turnstone, Shorebirds	KX185901
1	2398	1.2%	H9N2, H9N6, H9N9, mixed	China, Hong Kong, Viet Nam, Japan, Myanmar	1997-2016	14	Crane, Chicken, Dog, Duck, Goose, Human, Quail, Mink, Pigeon, Sparrow, Swine, Weasel	KM245331
2	407	8.4%	H9N2	Bangladesh, Egypt, Hong Kong, India, Iran, Israel, Jordan, Kuwait, Lebanon, Libya, Nepal, Pakistan, Saudi Arabia, Tunisia, UAE, USA	1999-2016	15	Chicken, Curlew, Duck, Falcon, Pheasant, Pigeon, Quail, Turkey, Bustard, Ferret	FJ464718
3	62	5.5%	H9N1, H9N2, H9N5, H9N6, H9N9, mixed	Canada, China, Hong Kong, Italy, New Zealand, South Korea, USA	1976-2014	3	Goose, Teal, Gull, Mallard, Pintail, Shoveler, Ruddy Turnstone, Turkey, Waterfowl	KJ013297

4	84	6.7%	H9N1, H9N2, H9N3	Austria, Bangladesh, Belgium, China, Finland, Germany, Hong Kong, Iran, Ireland, Italy, Japan, Netherlands, Norway, Portugal, Russia, South Africa, South Korea, Sweden, Switzerland, Thailand, UK, USA, Viet Nam, Zambia	1977-2014	4	Mallard, Teal, Swan, Chicken, Duck, Wigeon, Goose, Ostrich, Pelican, Ruddy Turnstone, Sanderling, Turkey	CY041274
5	222	5.4%	H9N2	China, Hong Kong	1994-2013	8	Chicken, Magpie, Duck, Human, Swine, Goose, Guinea Fowl, Quail, Sparrow	EU086265
6	36	2.3%	H9N1, H9N2, H9N3, H9N5, H9N7, H9N8, H9N9, mixed	Georgia, Singapore, UK, USA	2003-2015	17	Chicken, Common Murre, Gull, Ruddy Turnstone, Shorebirds, Sanderling,	CY185545
7	11	6.7%	H9N2, H9	China, Hong Kong	1979-2014	7	Chicken, Duck	KP766620
8	207	2.4%	H9N2, H9	China, Hong Kong, Japan, Viet Nam	2007-2014	18	Chicken, Duck, Guinea Fowl, Pheasant, Pigeon, Quail, Sparrow	KP186947

9	73	5.9%	H9N1, H9N2, H9N8	China, Hong Kong, Iran, Israel, Japan, Lebanon, UAE, Viet Nam	1997-2011	12	Babbler, Chicken, Duck, Garganey, Guinea Fowl, Human, Ostrich, Parakeet, Pigeon, Quail, Turkey	KF259135
10	19	8.0%	H9N1, H9N2, H9N5, mixed	Canada, Hong Kong, Hungary, USA	1979-2001	6	Chicken, Duck, Goose, Gull, Knot, Ruddy Turnstone	JX273541
11	28	4.8%	H9N2, H9	Egypt, South Korea	1996-2016	10	Chicken, Duck, Swine	KT157796
12	1	NA	H9N6	Viet Nam	2010	19	Duck	AB569975
13	20	3.9%	H9N1, H9N2, H9N7, H9N9	USA, Delaware, California, Arkansas, Alaska, Maryland, Texas, Massachusetts, New Jersey, Missouri	2002-2013	16	Shorebirds, Mallard, Pintail, Shoveler, Ruddy Turnstone, Teal, Gull,	CY146513
14	2	2.1%	H9N1, H9N2	USA, Wisconsin	1974-1976	2	Mallard, Teal	CY181337
15	12	8.3%	H9N2	China, Hong Kong, Iran, Japan	1978-2008	5	Chicken, Duck, Parakeet.	AY206675
16	9	4.1%	H9N2	USA, Wisconsin, China.	1966-2012	1	Turkey, Chicken, Swine	CY087824
17	59	4.3%	H9N2	China, Japan	1997-2013	13	Chicken, Duck, Goose, Human, Pigeon, Swine	FJ190112
18	3	1.0%	H9N2	Malaysia	1997-2001	11	Duck	JQ344326

Table 4: Summary of the sub-clades from clade 15

USEARCH Cluster	No. of Sequences	Median Distance	Subtypes	Geographical Location	Dates	Sub-Clade Number	Host Species	Representative Sequence, Accession Number
0	196	4.3%	H9N2	Bangladesh, Egypt, India, Iran, Israel, Lebanon, Pakistan, Saudi Arabia, UAE	2006-2015	15.8	Avian, Chicken, Dove, Duck, Falcon, Pigeon, Quail, Turkey	FJ464718
1	20	5.4%	H9N2	Iran, Israel, Jordan, Pakistan, Saudi Arabia, UAE	1999-2005	15.1	Avian, Chicken, Quail, Turkey	EF492228
2	99	3.8%	H9N2	Bangladesh (97), India, Iran	2005-2015	15.7	Chicken, Duck, Environment, Pigeon, Quail	KC757937
3	2	0	H9N2	Hong Kong	2011	15.12	Chicken	KF188335
4	5	4.4%	H9, H9N2	India, Nepal	2010-2013	15.11	Chicken	KT285333
5	6	3.8%	H9, H9N2	India, Nepal	2009-2010	15.10	Chicken	JX310065
6	24	3.5%	H9N2	Iran, Pakistan	2005-2013	15.6	Chicken	JX294920
7	6	3.6%	H9N2	Iran, Pakistan	2008-2014	15.9	Chicken	KJ69534
8	17	3.7%	H9N2	Iran, Tunisia, UAE	2004-2010	15.3	Bustard, Chicken	GU071981
9	22	5.0%	H9, H9N2	India, Israel, Libya, Pakistan, Saudi Arabia, UAE, USA(lab)	2003-2011 2012 (lab)	15.2	Avian, Bustard, Chicken, Ferret (lab), Pheasant, Quail	JX273557

10	3	6.9%	H9N2	Bangladesh, Kuwait	2004-2010	15.4	Chicken, Environment	JX273545
11	3	0.5%	H9N2	Pakistan	2015	15.13	Chicken, Pigeon	KU042910
12	4	2.6%	H9N2	UAE	2005	15.5	Bustard, Curlew, Quail.	KF188337

Table 5: The clade equivalences between the LABEL study and the current clustering analysis.

Representative Virus	Equivalent Accession	LABEL Clade [7]	Current Clade
A/quail/Hong_Kong/G1/97	KY785896	G1_Asia	12
A/chicken/Middle_East/ED-1/1999	GQ120553	G1_Mideast_A	12
A/chicken/Iran/B102/2005	EF063733	G1_Mideast_B1	15.8
		G1_Mideast_B2	15.7
A/quail/Dubai/202/2000	EF063512	G1_Mideast_C	15.1
A/chicken/Saudi_Arabia	AB049160	G1_Mideast_D	12
A/chicken/Pakistan/AG519/98	JX465626	G1_Pakistan	5
A/chicken/Beijing/1/94	KF188294	Chk_Bei	5
A/chicken/Hong_Kong/G9/97	KF188366	Y280_G9	8
A/duck/Hong_Kong/Y280/97	AF461530	Y280B	14.1
A/chicken/Fujian/SL6/2011	JF715052	Y280_Fuj_SL6	18
A/chicken/Shandong/ZB/2007	KF746843	Y280_Sha_ZB07	8
A/quail/Arkansas/29209-1/93	CY101224	AR29209	9
A/shorebird/Delaware_Bay/277/2000	CY102633	DB277	16
A/duck/Hong_Kong/448/78	AY206673	HK448	5
A/mallard/Ireland/PV46B	AB303077	PV46B	4
A/shorebird/DE/261/2003	CY005992	SB261	17
A/duck/Viet_Nam/340/2001	AB262463	VN340	4
A/Chicken/Korea/38349-p96323/96	KF188387	KR38349	4
A/duck/Hong_Kong/147/77	AY206671	HK147	4
A/goose/Minnesota/5733/80	KF188339	MN5733	6
A/turkey/Wisconsin/1/1966	CY014663	WI66	1

Figure 1: A plot of maximum cluster size against sequence identity for the H2, H4, H6, H7 and H9 influenza A hemagglutinin clustering with USEARCH.

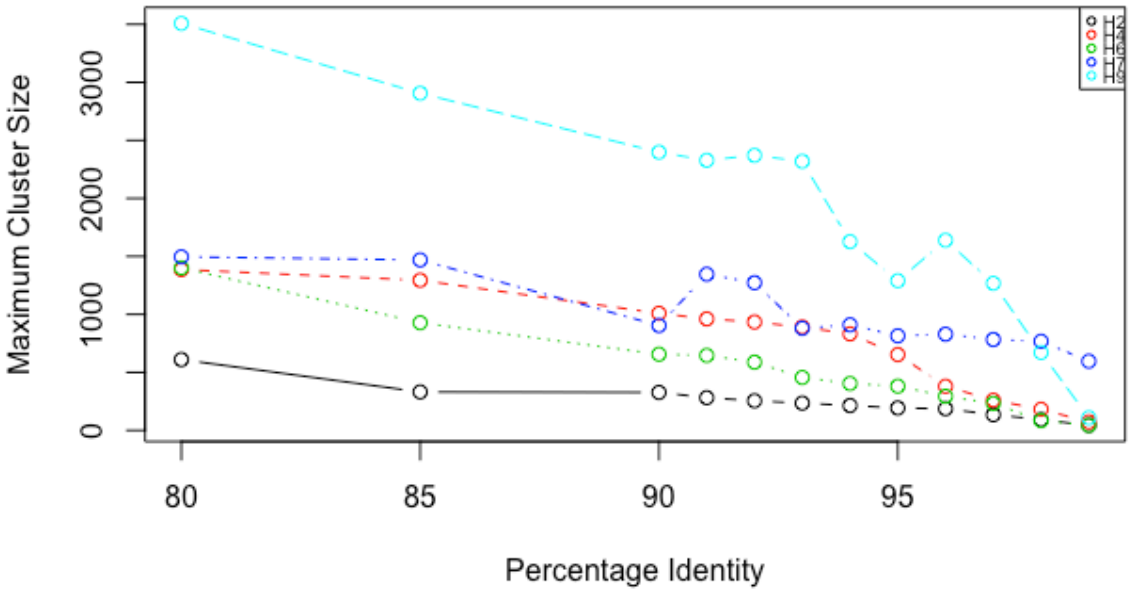


Figure 2: A plot of the number of clusters against sequence identity for the H2, H4, H6, H7 and H9 influenza A hemagglutinin clustering with USEARCH.

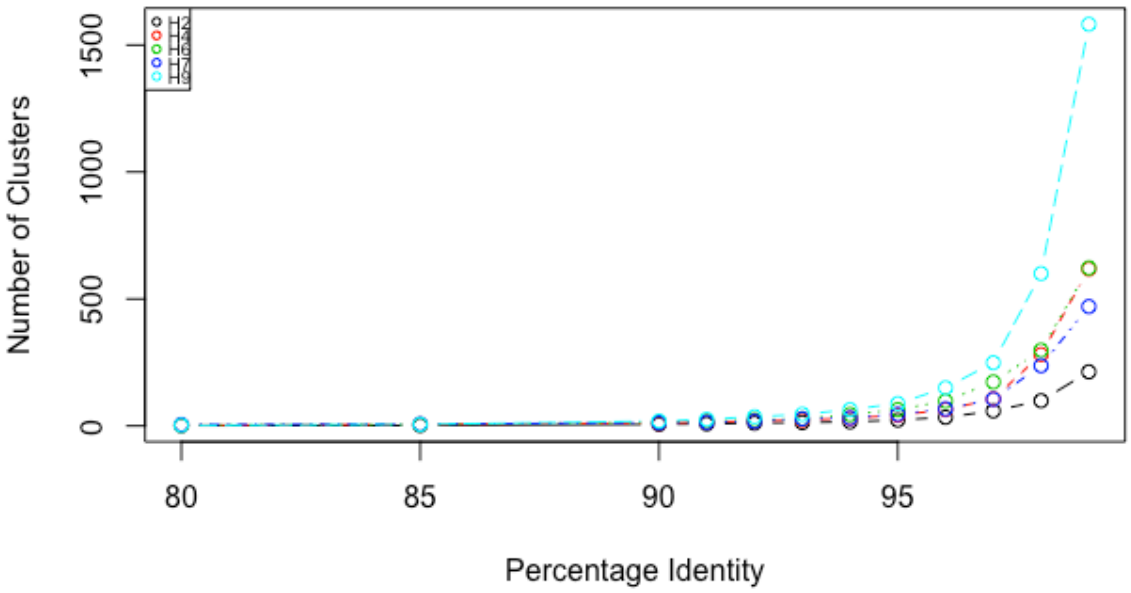


Figure 3: A BEAST phylogenetic tree of the representative sequences (centroids) from the H9 hemagglutinin clades.

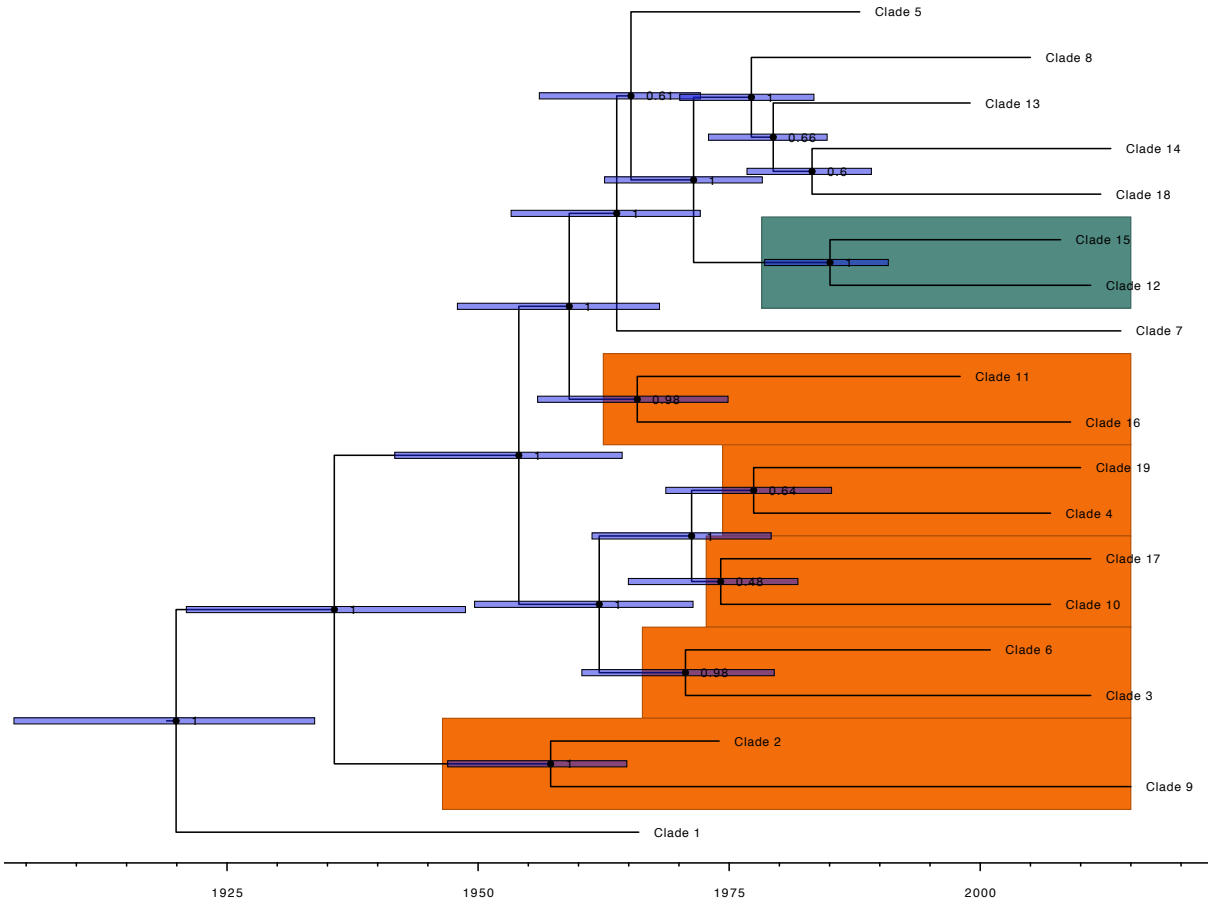


Figure 4: The BEAST2 maximum clade credibility coalescent phylogenetic tree for Clade 3. Nodes are labelled with the posterior probabilities and the bars show the 95% highest posterior density for the node age.

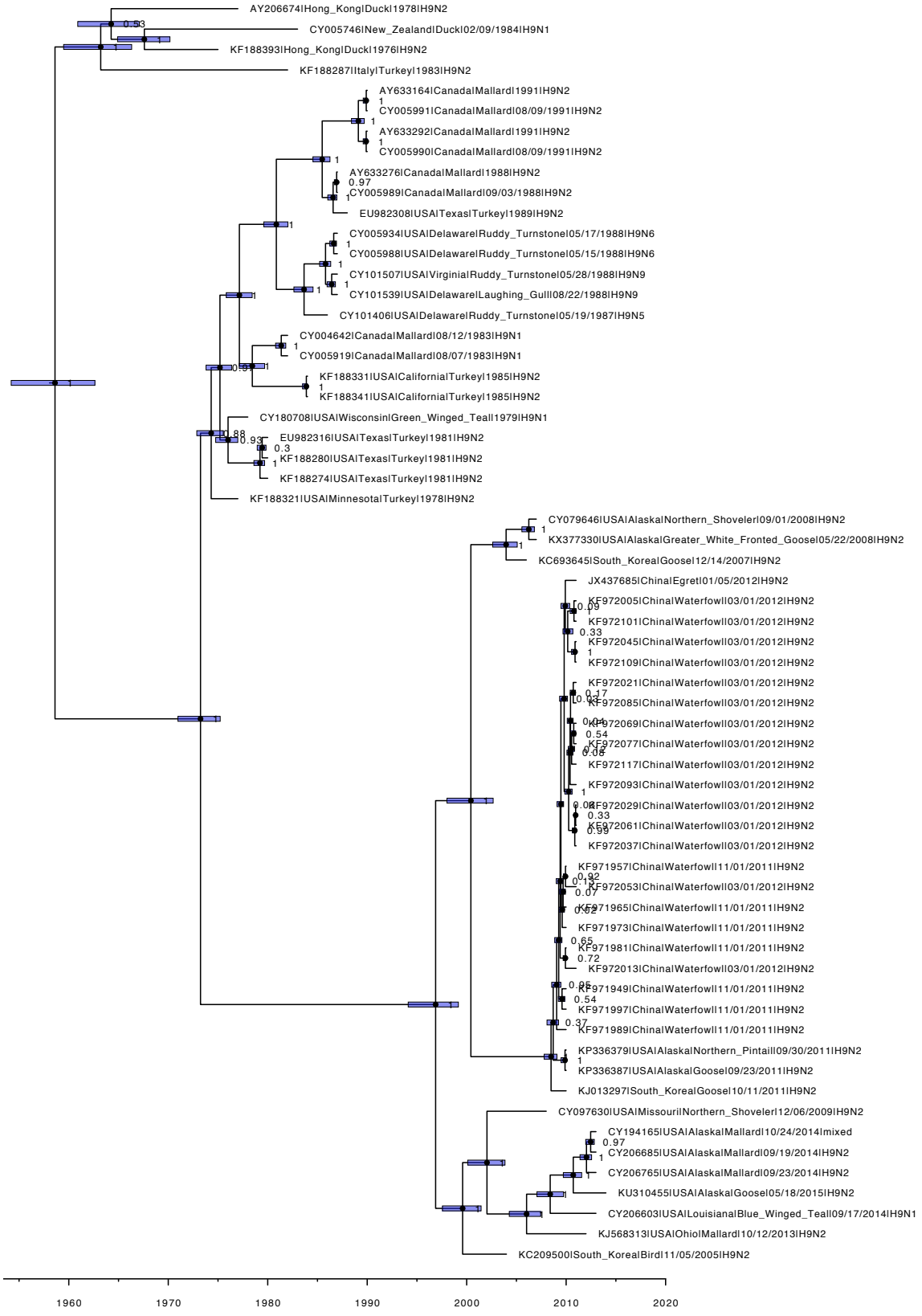


Figure 5: The BEAST2 maximum clade credibility coalescent phylogenetic tree for Clade 4. Nodes are labelled with the posterior probabilities and the bars show the 95% highest posterior density for the node age.

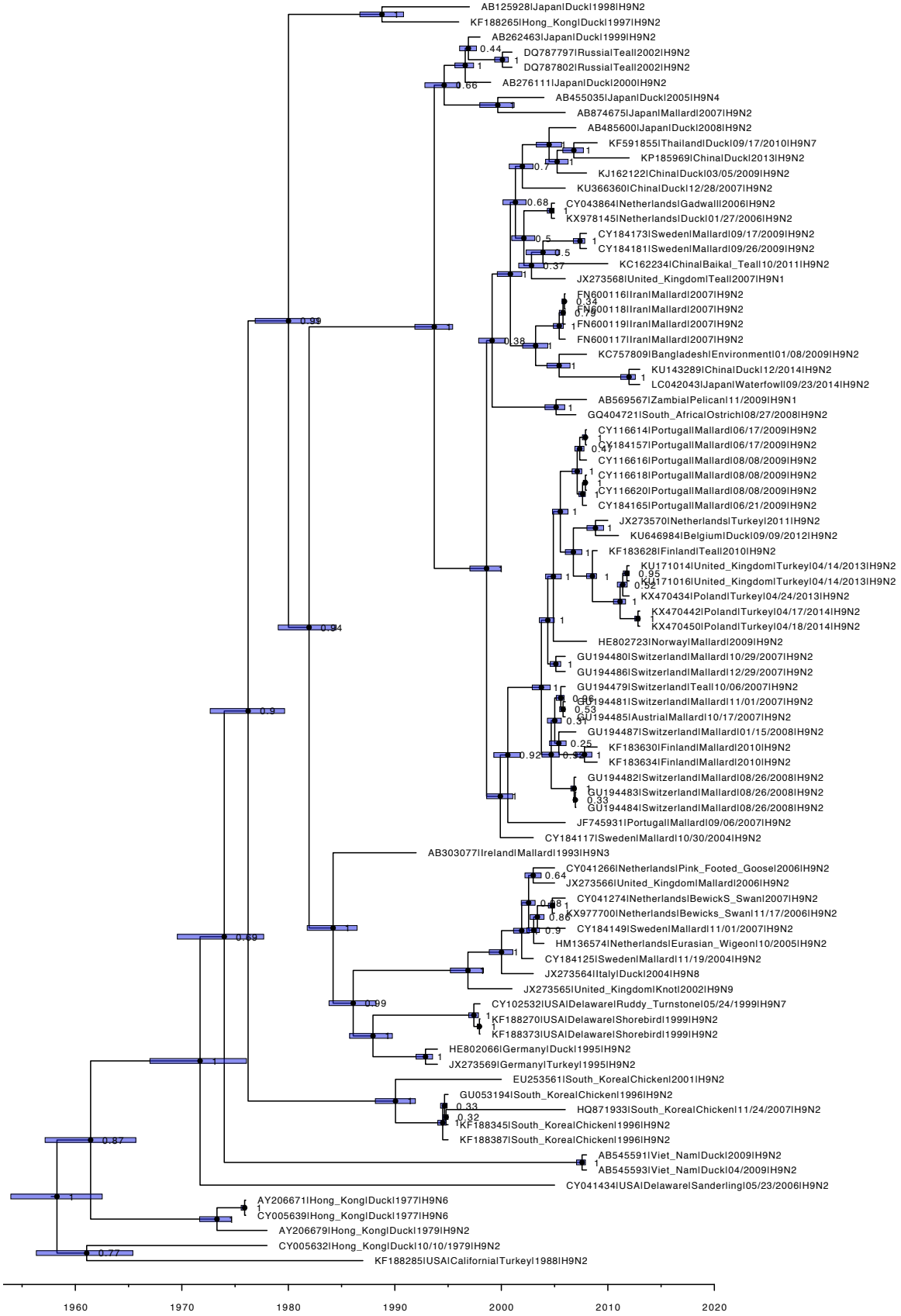


Figure 6: The BEAST2 maximum clade credibility coalescent phylogenetic tree for Clade 12. Nodes are labelled with the posterior probabilities and the bars show the 95% highest posterior density for the node age.

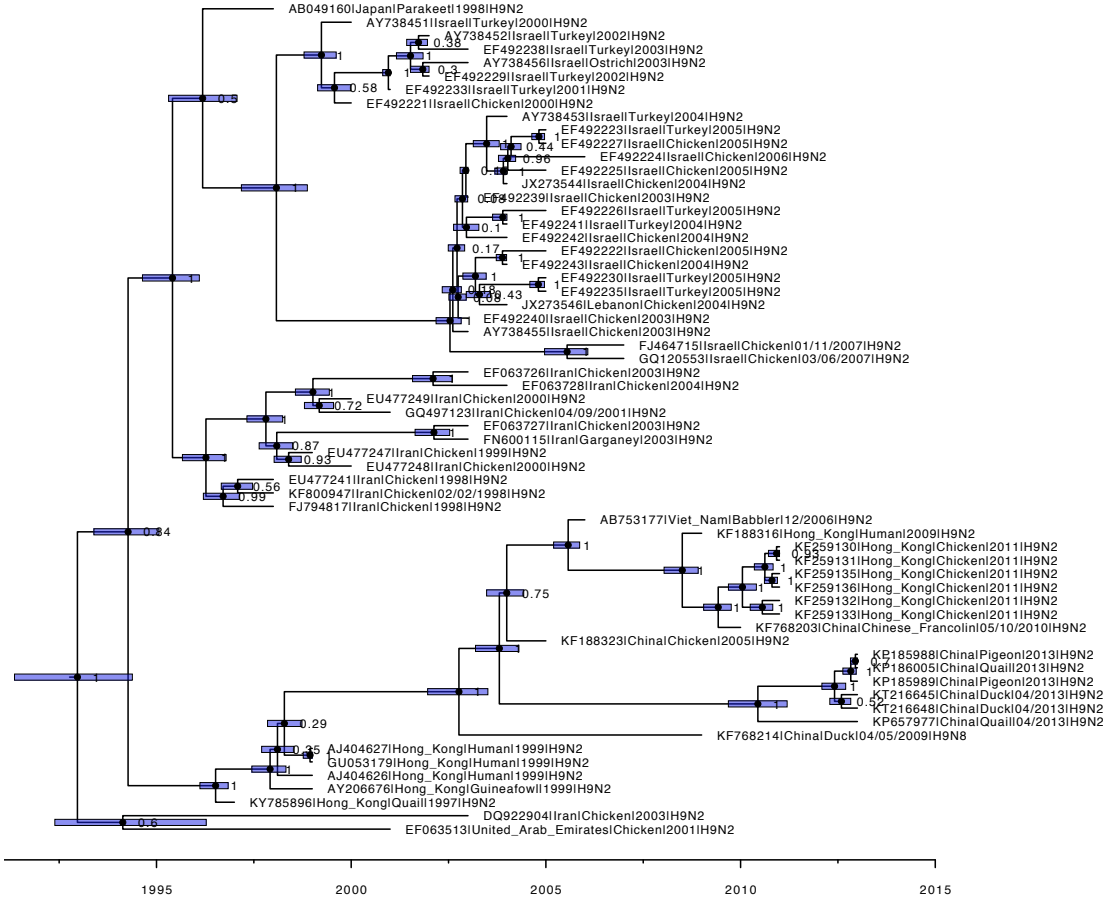


Figure 7a: The BEAST2 maximum clade credibility coalescent phylogenetic tree for Clade 15. Nodes are labelled with the posterior probabilities and the bars show the 95% highest posterior density for the node age.

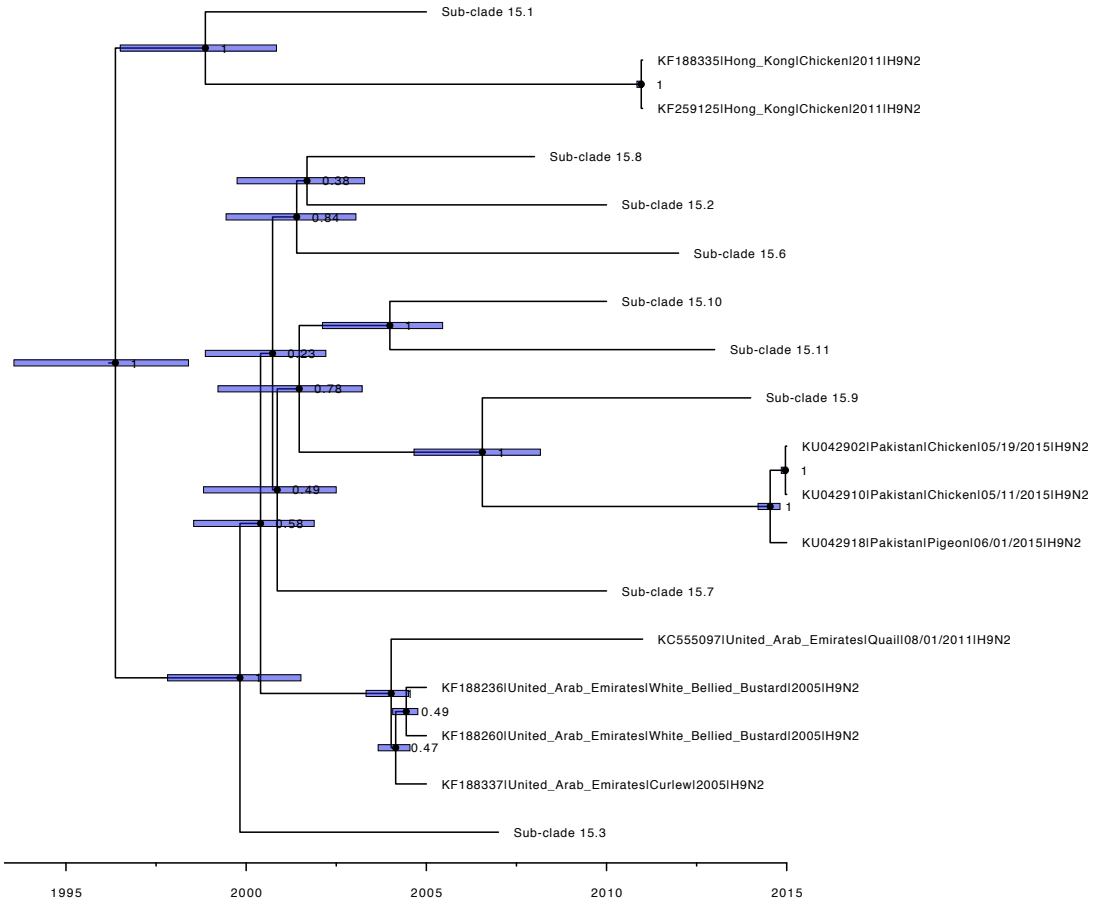
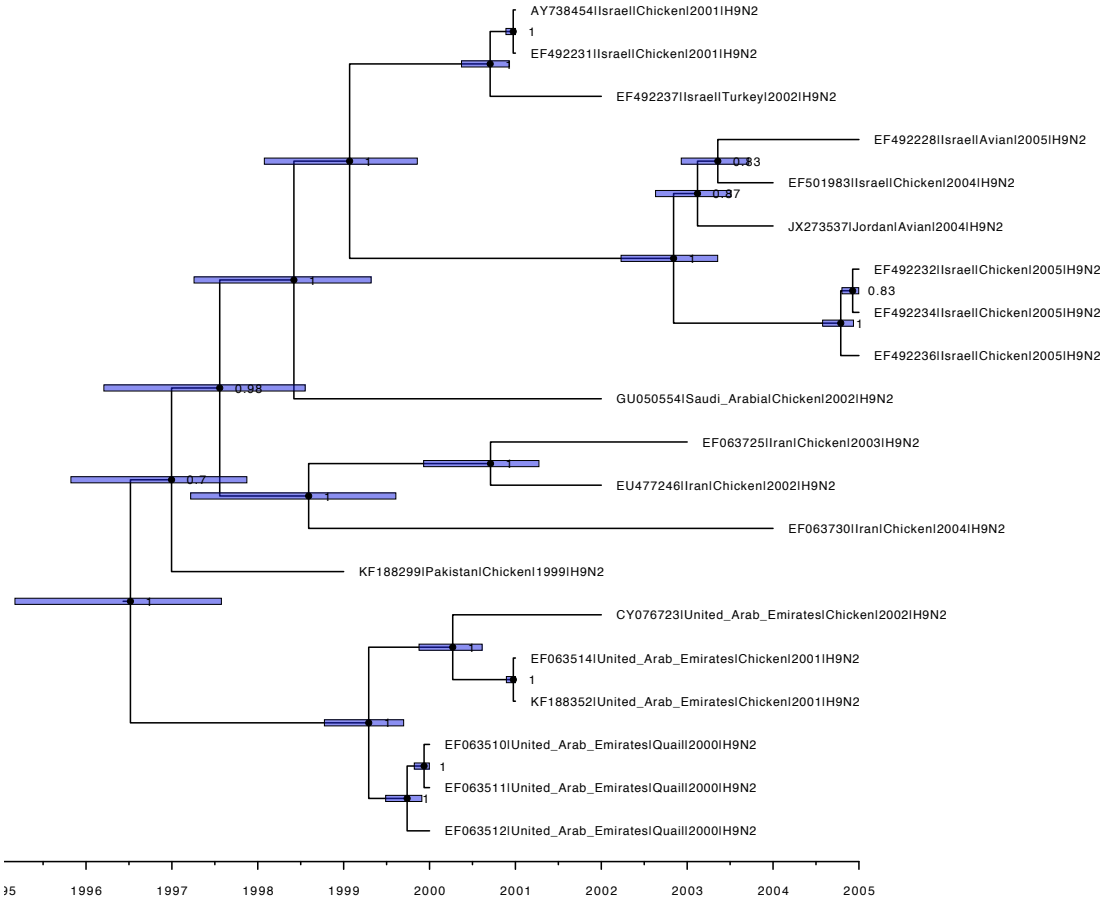


Figure 7b: The BEAST2 maximum clade credibility coalescent phylogenetic tree for Clade 15.1. Nodes are labelled with the posterior probabilities and the bars show the 95% highest posterior density for the node age.



Clade 7c: The BEAST2 maximum clade credibility coalescent phylogenetic tree for Clade 15.2. Nodes are labelled with the posterior probabilities and the bars show the 95% highest posterior density for the node age.

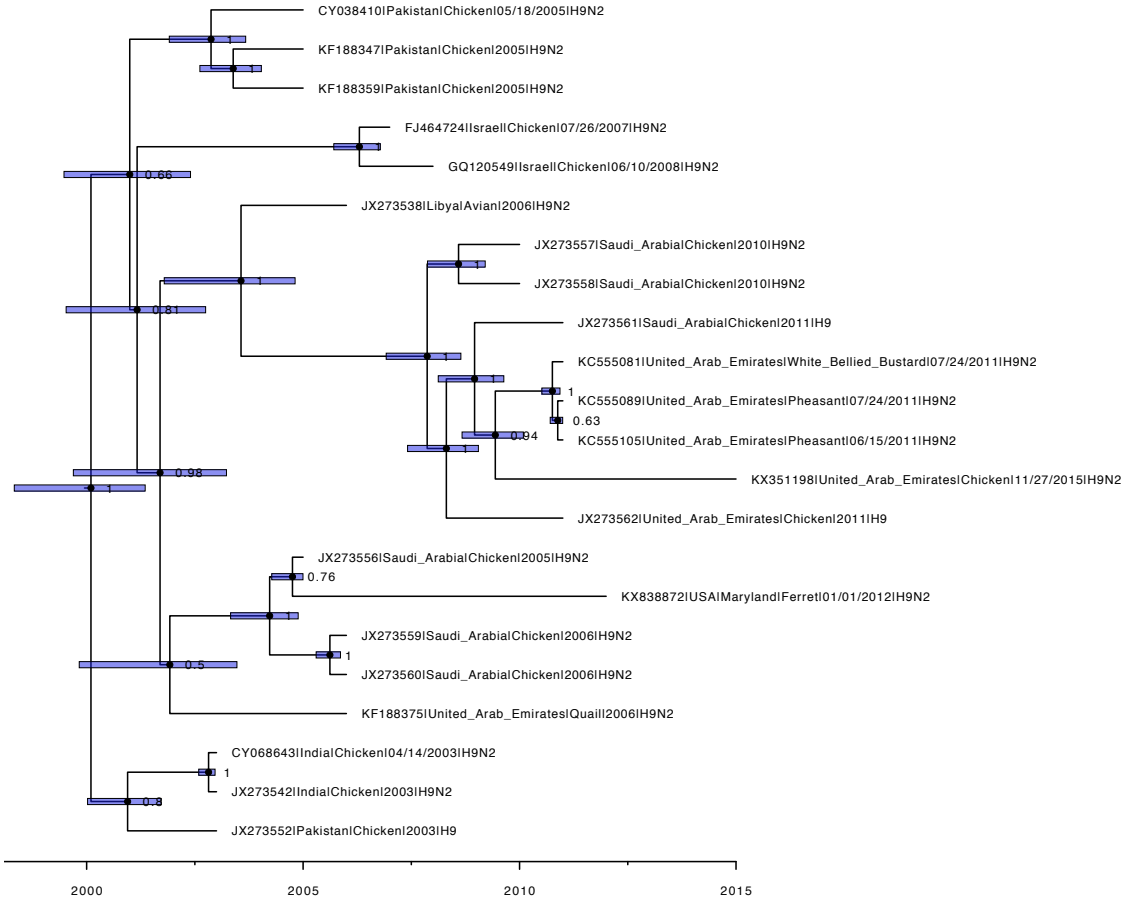


Figure 7d: The BEAST2 maximum clade credibility coalescent phylogenetic tree for Clade 15.3. Nodes are labelled with the posterior probabilities and the bars show the 95% highest posterior density for the node age.

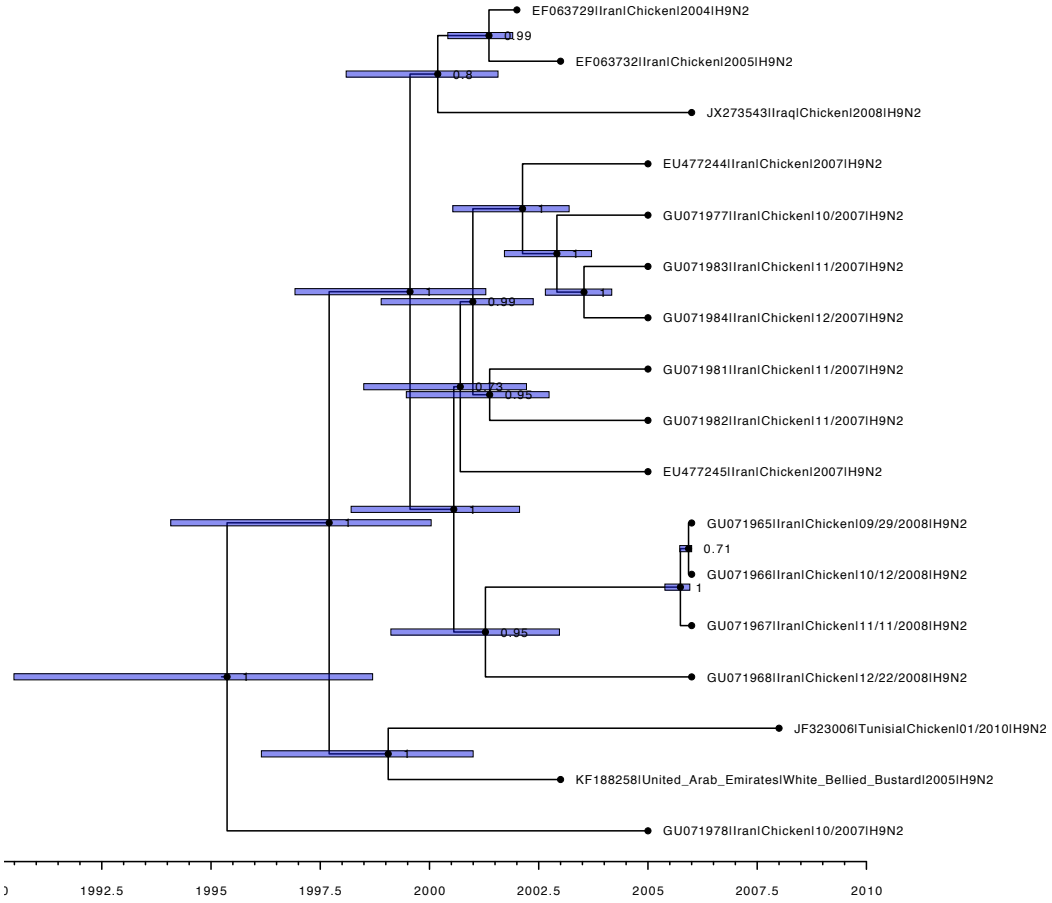


Figure 7e: The BEAST2 maximum clade credibility coalescent phylogenetic tree for Clade 15.6. Nodes are labelled with the posterior probabilities and the bars show the 95% highest posterior density for the node age.

.

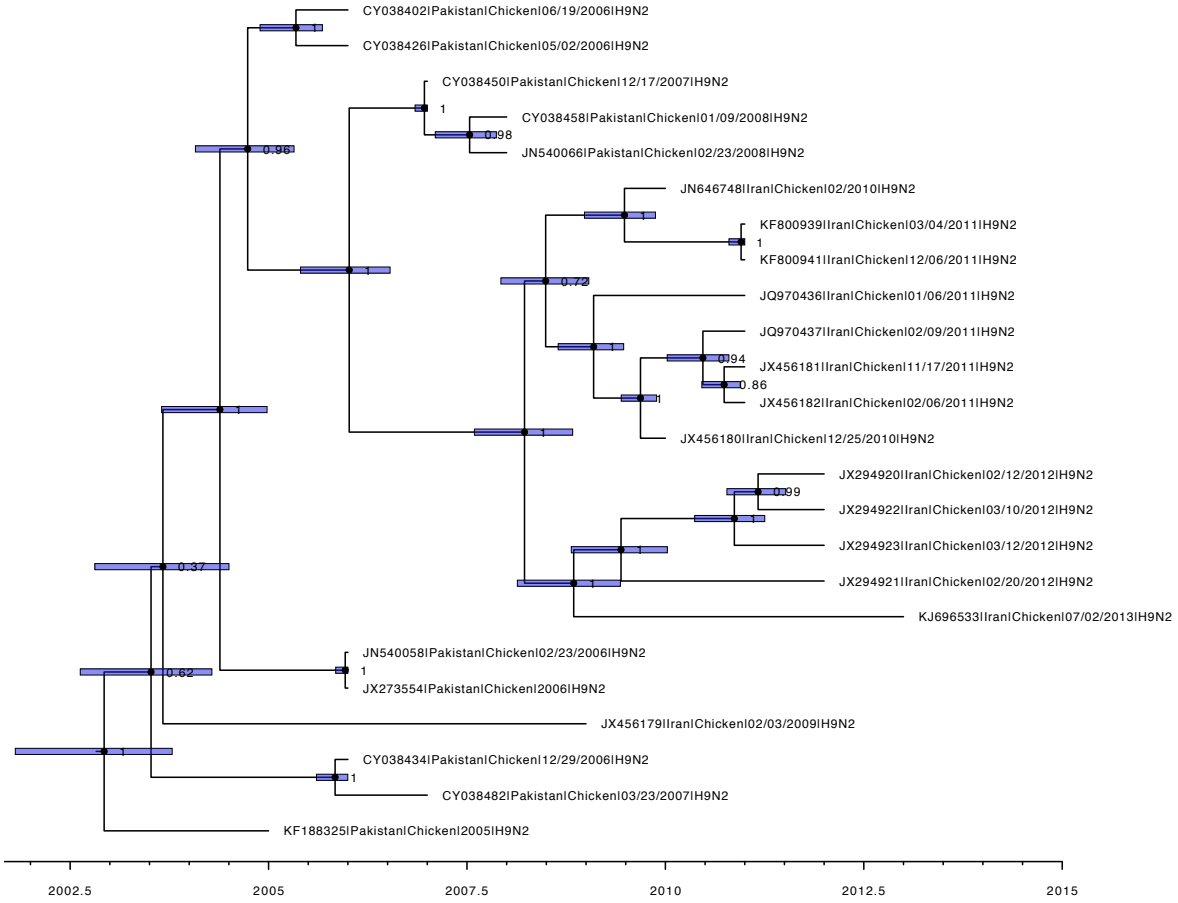


Figure 7f: The BEAST2 maximum clade credibility coalescent phylogenetic tree for Clade 15.7. Nodes are labelled with the posterior probabilities and the bars show the 95% highest posterior density for the node age.

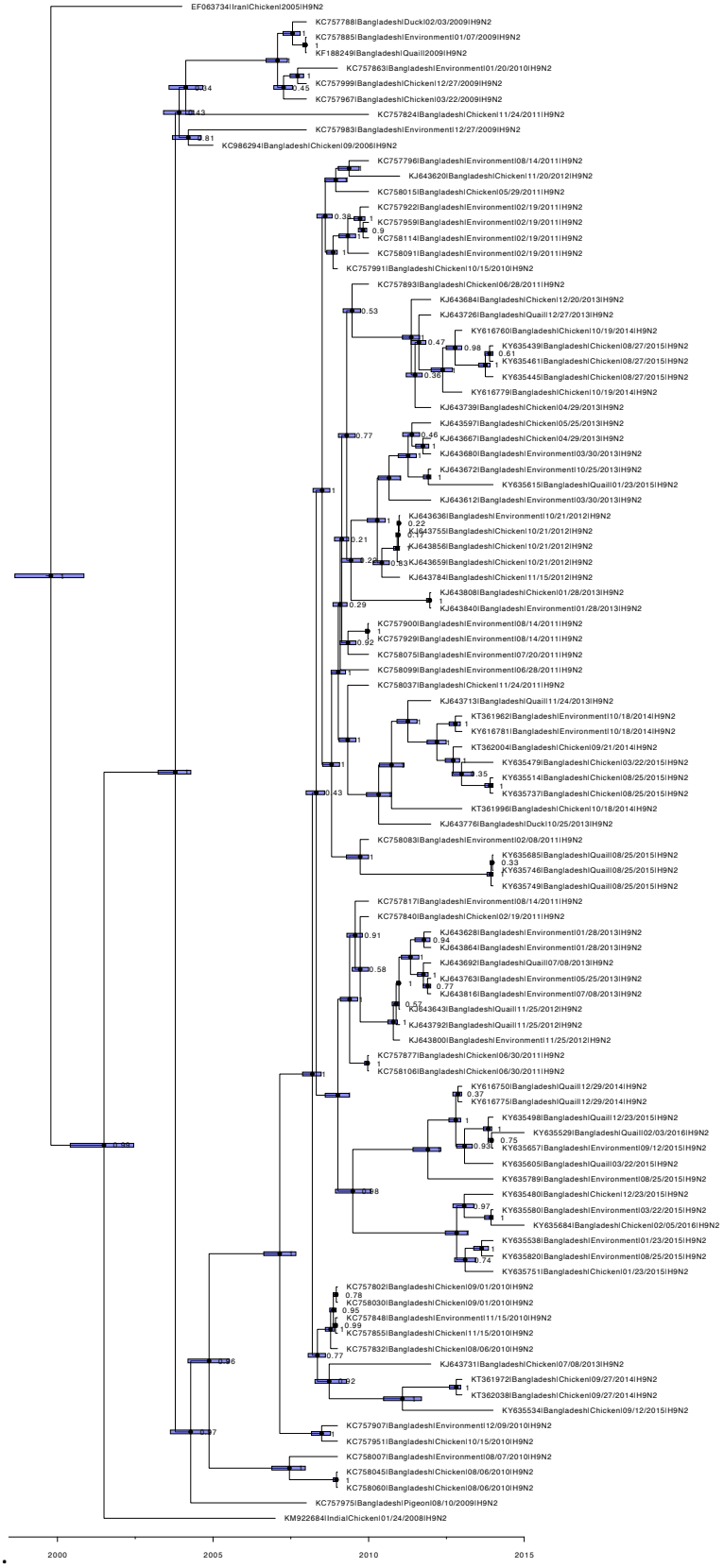


Figure 7g: The BEAST2 maximum clade credibility coalescent phylogenetic tree for Clade 15.8. Nodes are labelled with the posterior probabilities and the bars show the 95% highest posterior density for the node age.

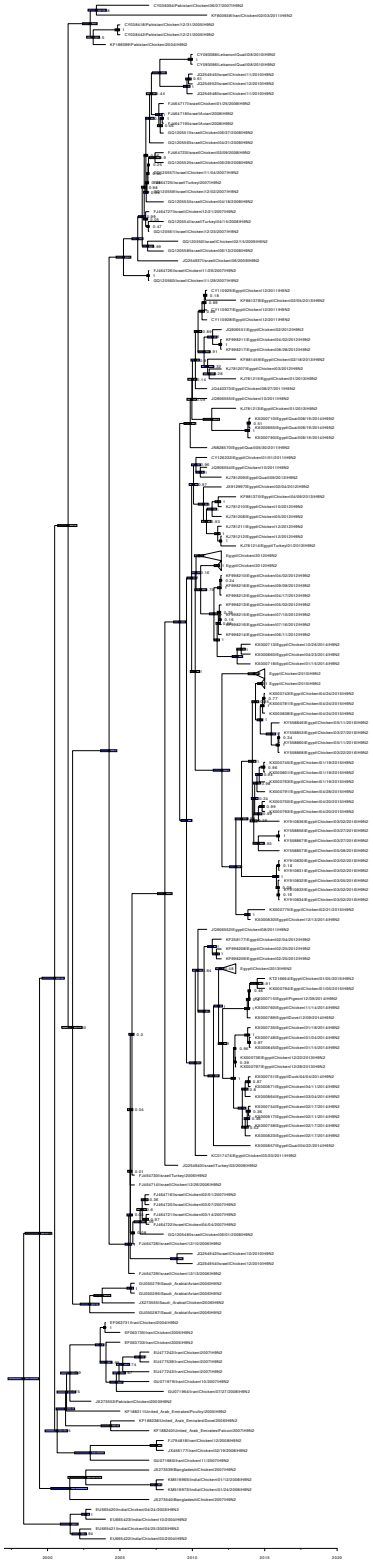
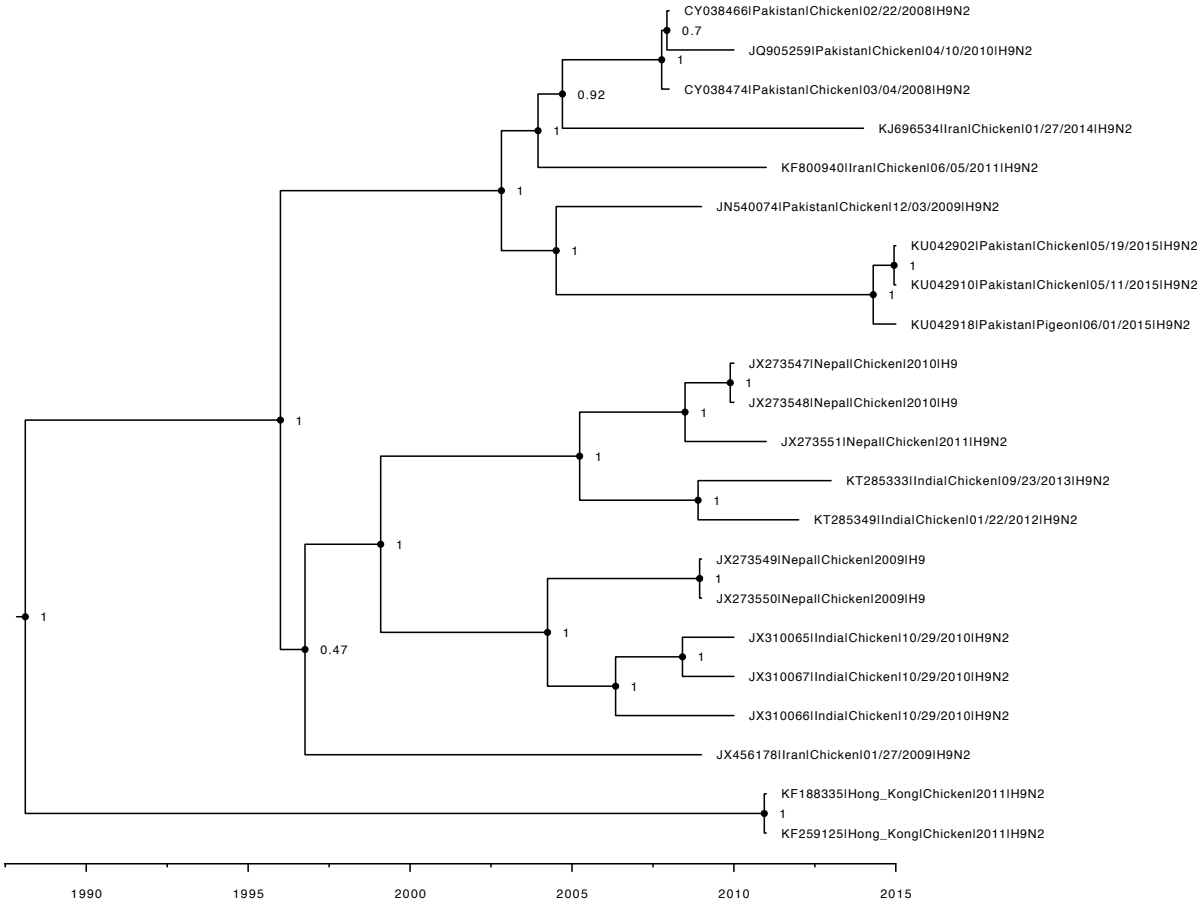


Figure 7h: The BEAST2 maximum clade credibility coalescent phylogenetic tree for Clade 15.9-15.13. Nodes are labelled with the posterior probabilities.



References

1. Squires RB, Noronha J, Hunt V, García-Sastre A, Macken C, Baumgarth N, Suarez D, Pickett BE, Zhang Y, Larsen CN: **Influenza research database: an integrated bioinformatics resource for influenza research and surveillance**. *Influenza and other respiratory viruses* 2012, **6**(6):404-416.
2. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y: **Evolution and ecology of influenza A viruses**. *Microbiological reviews* 1992, **56**(1):152-179.
3. Krauss S, Walker D, Pryor SP, Niles L, Chenghong L, Hinshaw VS, Webster RG: **Influenza A viruses of migrating wild aquatic birds in North America**. *Vector-Borne & Zoonotic Diseases* 2004, **4**(3):177-189.
4. Smith GJ, Donis RO: **Nomenclature updates resulting from the evolution of avian influenza A (H5) virus clades 2.1. 3.2 a, 2.2. 1, and 2.3. 4 during 2013–2014**. *Influenza and other respiratory viruses* 2015, **9**(5):271-276.
5. Anderson TK, Macken CA, Lewis NS, Scheuermann RH, Van Reeth K, Brown IH, Swenson SL, Simon G, Saito T, Berhane Y: **A Phylogeny-Based Global Nomenclature System and Automated Annotation Tool for H1 Hemagglutinin Genes from Swine Influenza A Viruses**. *mSphere* 2016, **1**(6):e00275-00216.
6. WHO O: **Toward a unified nomenclature system for highly pathogenic avian influenza virus (H5N1)**. *Emerging infectious diseases* 2008, **14**(7):e1.
7. Shepard SS, Davis CT, Bahl J, Rivailler P, York IA, Donis RO: **LABEL: fast and accurate lineage assignment with assessment of H5N1 and H9N2 influenza A hemagglutinins**. *PloS one* 2014, **9**(1):e86921.
8. Anderson TK, Campbell BA, Nelson MI, Lewis NS, Janas-Martindale A, Killian ML, Vincent AL: **Characterization of co-circulating swine influenza A viruses in North America and the identification of a novel H1 genetic clade with antigenic significance**. *Virus research* 2015, **201**:24-31.
9. Ip HS, Dusek RJ, Bodenstein B, Torchetti MK, DeBruyn P, Mansfield KG, DeLiberto T, Sleeman JM: **High Rates of Detection of Clade 2.3. 4.4 Highly Pathogenic Avian Influenza H5 Viruses in Wild Birds in the Pacific Northwest During the Winter of 2014-15**. *Avian diseases* 2016, **60**(1s):354-358.
10. Donis RO, Smith GJ: **Nomenclature updates resulting from the evolution of avian influenza A (H5) virus clades 2.1. 3.2 a, 2.2. 1, and 2.3. 4 during 2013-2014**. *Influenza and other respiratory viruses* 2015.
11. Lee M-S, Chen L-H, Chen Y-P, Liu Y-P, Li W-C, Lin Y-L, Lee F: **Highly pathogenic avian influenza viruses H5N2, H5N3, and H5N8 in Taiwan in 2015**. *Veterinary microbiology* 2016, **187**:50-57.
12. Matsen FA, Kodner RB, Armbrust EV: **pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree**. *BMC bioinformatics* 2010, **11**(1):538.
13. Rousseeuw PJ, Kaufman L: **Finding Groups in Data**: Wiley Online Library; 1990.
14. Edgar RC: **Search and clustering orders of magnitude faster than BLAST**. *Bioinformatics* 2010, **26**(19):2460-2461.
15. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic acids research* 2004, **32**(5):1792-1797.

16. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets**. *Molecular biology and evolution* 2016, **33**(7):1870-1874.
17. R-core Team: **R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014**. In.; 2014.
18. R Studio: **RStudio: integrated development environment for R**. *RStudio Inc, Boston, Massachusetts* 2012.
19. Paradis E, Claude J, Strimmer K: **APE: analyses of phylogenetics and evolution in R language**. *Bioinformatics* 2004, **20**(2):289-290.
20. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ: **BEAST 2: a software platform for Bayesian evolutionary analysis**. *PLoS computational biology* 2014, **10**(4):e1003537.
21. Waddell PJ, Steel M: **General time-reversible distances with unequal rates across sites: mixing Γ and inverse Gaussian distributions with invariant sites**. *Molecular phylogenetics and evolution* 1997, **8**(3):398-414.
22. Tamura K, Nei M: **Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees**. *Molecular biology and evolution* 1993, **10**(3):512-526.
23. Rambaut A: **FigTree v1. 4**. *Molecular evolution, phylogenetics and epidemiology Edinburgh, UK: University of Edinburgh, Institute of Evolutionary Biology* 2012.
24. Koçer ZA, Carter R, Wu G, Zhang J, Webster RG: **The genomic contributions of avian H1N1 influenza A viruses to the evolution of mammalian strains**. *PloS one* 2015, **10**(7):e0133795.
25. Hurtado R, Fabrizio T, Vanstreels RET, Krauss S, Webby RJ, Webster RG, Durigon EL: **Molecular characterization of subtype H11N9 avian influenza virus isolated from shorebirds in Brazil**. *PloS one* 2015, **10**(12):e0145627.
26. Dalby AR, Iqbal M: **A global phylogenetic analysis in order to determine the host species and geography dependent features present in the evolution of avian H9N2 influenza hemagglutinin**. *PeerJ* 2014, **2**:e655.
27. Fusaro A, Monne I, Salviato A, Valastro V, Schivo A, Amarin NM, Gonzalez C, Ismail MM, Al-Ankari A-R, Al-Blowi MH: **Phylogeography and evolutionary history of reassortant H9N2 viruses with potential human health implications**. *Journal of virology* 2011, **85**(16):8413-8421.
28. Shanmuganatham K, Feeroz MM, Jones-Engel L, Walker D, Alam S, Hasan M, McKenzie P, Krauss S, Webby RJ, Webster RG: **Genesis of avian influenza H9N2 in Bangladesh**. *Emerging microbes & infections* 2014, **3**(12):e88.