# A new method for statistical clustering of influenza sequence data

**Alvin X. Han[1]* (hanxc@bii.a-star.edu.sg), Edyth Parker[2]*(ep484@cam.ac.uk), Sebastian Maurer-Stroh[1], Colin Russell[2]**

**[1]Bioinformatics Institute, A*STAR Singapore, [2]Department of Veterinary Medicine, University of Cambridge**

PhyCLIP

## Introduction

Frameworks for defining and delineating lineages and clades in phylogenetic trees are largely based on inconsistent and arbitrary criteria, with many phylogenetic clustering approaches requiring user-specified distance thresholds.

The most prominent nomenclature system for influenza viruses is the WHO/OIE/FAO nomenclature for the HA gene of the highly pathogenic A/goose/Guangdong/1/1996 (Gs/GD) lineage of H5 subtype viruses, which is updated periodically. The nomenclature of the HA gene of the low pathogenicity H9 subtype viruses is based on the H5 nomenclature principles. The diversity of H5 and H9 viruses has been extensively studied owing to their enzootic circulation in several countries as well as the frequency of sporadic human infection by these subtype viruses.

Clade designation in the WHO/OIE/FAO nomenclature is based on phylogenetic tree topology and nucleotide sequence divergence and have the following specific criteria:

**H5**

**1. Clades are required to be a monophyletic group, with a bootstrap support of ≥60% for the shared node in the phylogenetic tree.**

**2. The average pairwise genetic distance, calculated as p-distance, between sequences in the putative clade has to be <1.5%.**

**3.The average pairwise genetic distance to the closest clade has to be >1.5%.**

**4. The upper within-clade genetic distance threshold of putative clades can be relaxed for clades with highly evolved outliers**

**H9**

**1. The average pairwise genetic distance, calculated as p-distance, between sequences in the putative clade has to be <6%.**

**2.The C-value to the closest clade must be >1.1. The C-value is calculated as the ratio of the average pairwise genetic distance of the clade to its closest neighbouring clade divided by its within-clade average pairwise genetic distance.**

**3. Clade nomenclature is based on previous classification efforts and representative sequences**

Genetic distance based clustering approaches are convenient as they are rule-based and scalable for easy updates.

However, genetic distance based clustering is centrally limited by the arbitrariness of the thresholds defined for within- and between-cluster divergence. There is no clear measure of a ground-truth, well-delineated phylogenetic cluster for avian influenza viruses owing to its maintenance ecology in wild birds and the frequency of reassortment.

Identifying an appropriate distance threshold or resolution for partitioning phylogenetic trees into biologically meaningful clusters is therefore complex and is mostly defined *ad hoc* by exploratory analyses. Cluster membership is expectedly very sensitive to these thresholds, and therefore a robust statistical framework is needed to optimise phylogenetic clustering for influenza virus nomenclature.

In this poster we describe a statistically-principled phylogenetic clustering approach based on integer linear programming optimization, named PhyCLIP, which minimizes the arbitrariness of clustering approaches such as the WHO/OIE/FAO nomenclature and algorithms such as ClusterPicker and PhyloPart, which are based on user-specified distance thresholds.

## Methods:

### Phylogenetic clustering by integer linear programming optimization: PhyCLIP

For a rooted phylogenetic tree with $N$ internal nodes, let $\{n_1, n_2, ..., n_i, ..., n_N\}$ be the set of binary variables indicating if taxa subtended by internal node $i$ should be grouped as a cluster (i.e. $n_i$ = 1 if taxa subtended by internal node $i$ should be clustered; $n_i$ = 0 if otherwise).

The clustering problem is then formulated as an integer linear programming (ILP) problem with the objective function:

$$\max \sum_{i}^{N} n_i l_i$$

where

$$n_i \in [0,1]$$

$l_i$ = number of taxa subtended by internal node $i$

subject to the following constraints:

$$(n_i - 1)C \leq l_i - MC \quad \forall \ i \qquad (1)$$

where

$C$ = Large positive constant

$MC$ = *User-defined* minimum cluster size

$$n_i + \sum_k n_k \leq 1 \ \forall \ i,k \in \{ancestral\ nodes\ of\ i\} \qquad (2)$$

$$(n_i - 1)C \leq \mu_i - WCL \quad \forall \ i \qquad (3)$$

where

$\mu_i$ = Mean pairwise patristic distance of all taxa subtended by internal node $i$

$WCL$ = Upper within-cluster divergence limit (explanation follows)

$$(2 - n_i - n_j)C \geq q_{i,j} - FDR \quad \forall \ i,j \neq i \qquad (4)$$

where

$q_{i,j}$ = Adjusted p-value of the two-sample Kolmogorov-Smirnov test between cluster $i$ and $j$ (explanation follows)

$FDR$ = *User-defined* false discovery rate

Constraint (1) requires clusters to have at least $MC$ number of taxa.
Constraint (2) ensures that only 1 internal node along the ancestral trace of $i$ to the tree's root can be selected as a cluster.
Constraint (3) is the within-cluster divergence constraint. Should taxa subtended by node $i$ be clustered, the mean pairwise patristic distance ($\mu_i$) must be $\leq WCL$, which is defined as:

$$WCL = \mu_{med} + \gamma \cdot MAD$$

where

$\mu_{med}$ = Grand median of mean pairwise distances of all internal nodes

$\text{MAD} = \text{median}(\mu_i - \bar{\mu}) \quad \forall \quad i \text{ if } \mu_i \geq \bar{\mu}$

$\gamma$ = *User-defined* multiple of MAD

Constraint (4) is the inter-cluster divergence constraint. The pairwise distance distribution of cluster $i$ must be distinct from the distribution of a cluster combining $i$ and $j$, wherein $j$ denotes any other cluster apart from $i$. This was determined using the two-sample Kolmogorov-Smirnov (KS) test and the resulting p-value was adjusted for multiple testing using the Benjamini-Hochberg procedure.

As such, the user need only to define three analysis parameters:
1. $MC$ - minimum cluster size
2. $\gamma$ - multiple of MAD
3. $FDR$ - false discovery rate for KS-tests

### Selecting parameters for PhyCLIP

PhyCLIP was run with combinations of the different parameters ranging 2-5 for minimum cluster size, 5-20% for FDR and 1-3 for the MAD multiplier. The optimal clustering result was selected to maximise the number of sequences clustered (coverage), minimise the grand mean and standard deviation of the mean pairwise distance distribution for all nodes and minimise the variance in date and geographic proximity.

### Phylogenetic tree construction

**H5**

Retrieved all H5 HA nucleotide sequences from GISAID collected between 1997 and Aug-2017. Filtered for “high quality" (>90% full HA nucleotide length, <1% unknown residues) sequences and deleted duplicates (by strain names). CD-HIT was used to remove redundant sequences at amino acid level (based on translation of codon-alignment). 3,410 sequences were included in the final analyses, and annotated with H5 nomenclature using LABEL, H5v2015 module. Phylogenetic trees were constructed with RAxML, using the GTR+GAMMA substitution model.

**H9**

Retrieved all H9 HA nucleotide sequences used to build reduced reference tree in the development of H5-based H9 nomenclature (n=606). Sequences were annotated with H9 nomenclature using LABEL, H9v2011 module. Phylogenetic trees were constructed with BEAST, using the SRD06 substitution model, with a strict clock model and coalescent constant population prior. MCMC analysis was run for 200 million steps, sampled every 20,000 steps, with the first 10% removed as burn-in.

## Results

### H5 HA

The optimal clustering had a minimum clade size of 3, a FDR of 20% and median absolute deviation (MAD)-multiplier of 2 (Fig 1). In this tree, 87 clusters were resolved for the 2,872 sequences clustered (84.2% coverage, 538 outliers), as opposed to 43 clades in the WHO/OIE/FAO nomenclature.

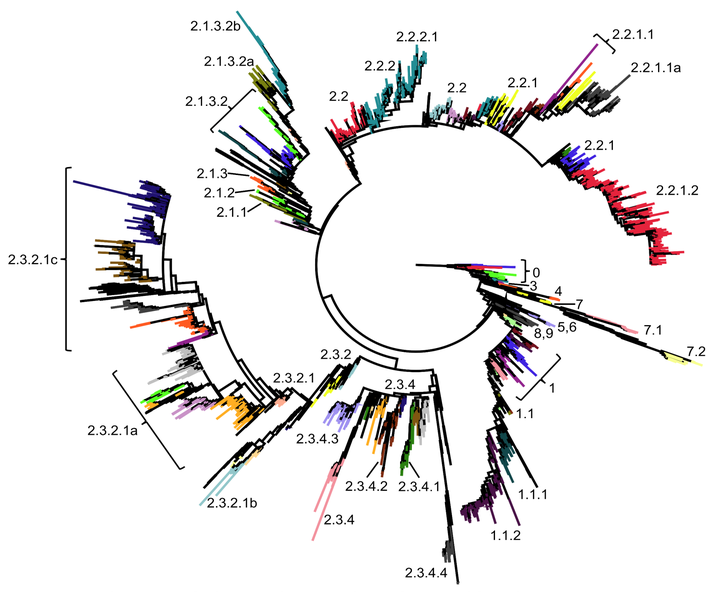


**Figure 1: Phylogenetic tree of H5 clustered by PhyCLIP**

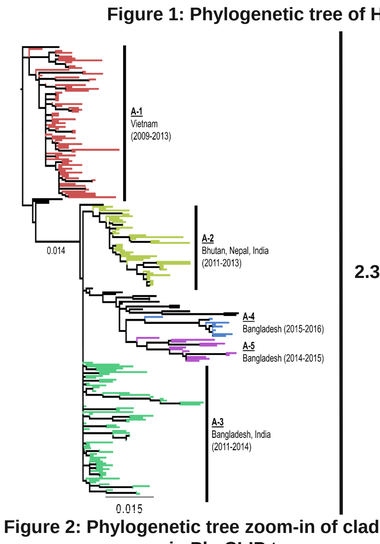


**Figure 2: Phylogenetic tree zoom-in of clade 2.3.2.1a in PhyCLIP tree**

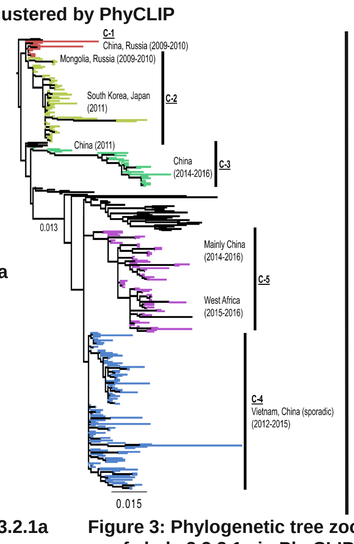


**Figure 3: Phylogenetic tree zoom-in of clade 2.3.2.1c in PhyCLIP tree**

The WHO/OIE/FAO nomenclature system follows a "greedy" strategy, clustering as many strains within each cluster as possible that fulfils the internal node support and pairwise distance cut-offs. This approach loses critical information on the genetic diversity of a virus that has yet been subjected to much selective pressure, as was the case for Clade 1 (Fig 4).

While PhyCLIP also follow a "greedy" strategy, it assigns clades based on how the sequence divergence of each cluster compares to each other and to the entire evolutionary trajectory of H5. This, in turn, allows for the characterisation of Clade 1's genetic diversity (Fig 4).

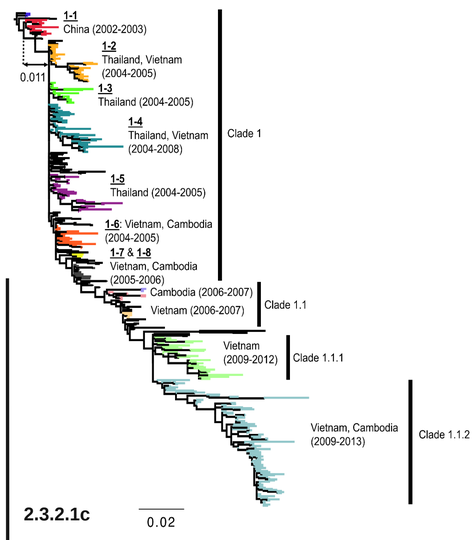


**Figure 4: Phylogenetic tree zoom-in of clade 1 in ILP-clustered tree**

There is ample evidence by the WHO/OIE/FAO H5 nomenclature criteria that these clades should have been further delineated in the latest nomenclature update (conducted for sequences up to 2014), as their within-clade average pairwise distance is > 1.5% by both p-distance (2.3.2.1a: 1.93%; 2.3.2.1c: 1.75%) and patristic distance.
PhyCLIP captured clear lineage distinctions between viruses from different geographic regions:
- In 2.3.2.1a, there is a clear delineation between viruses found in Vietnam and South Asia (Fig 2)
- In 2.3.2.1c, there is a clear delineation between viruses found in Vietnam, Japan and China (Fig 3)

PhyCLIP could also suggest potential delineations for the more recent 2.3.2.1a cases in Bangladesh (Cluster A-4 and 5 in Fig 2) and 2.3.2.1c viruses reported in China and West Africa (C-3, 4 and 5 in Fig 3).

### H9 HA

The optimal clustering had a minimum clade size of 3, a FDR of 20% and median absolute deviation (MAD)-multiplier of 2 (Fig 5). In this tree, 33 clusters were resolved for the 535 sequences clustered (88.28% coverage, 71 outliers), as opposed to 23 clades in the H5-based nomenclature.

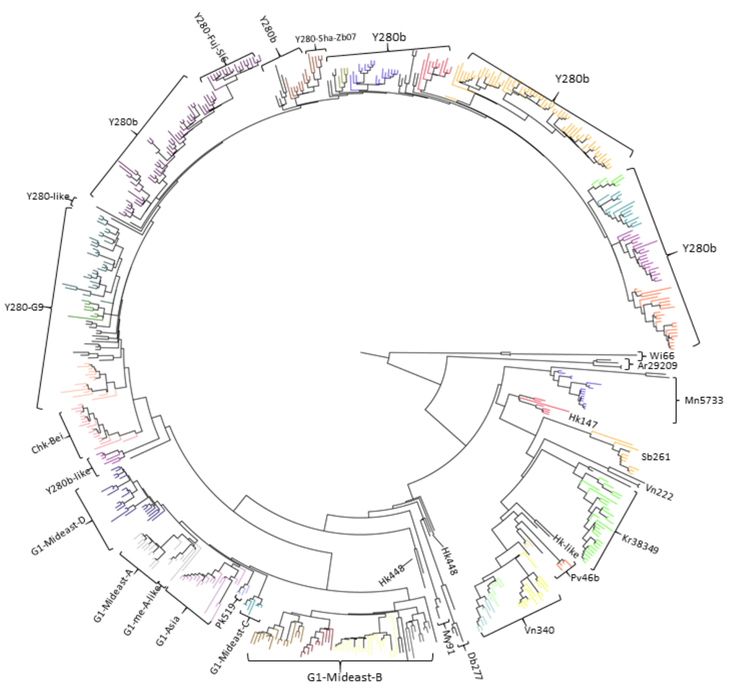


**Figure 5: Phylogenetic tree of H9 clustered by ILP strategy.**

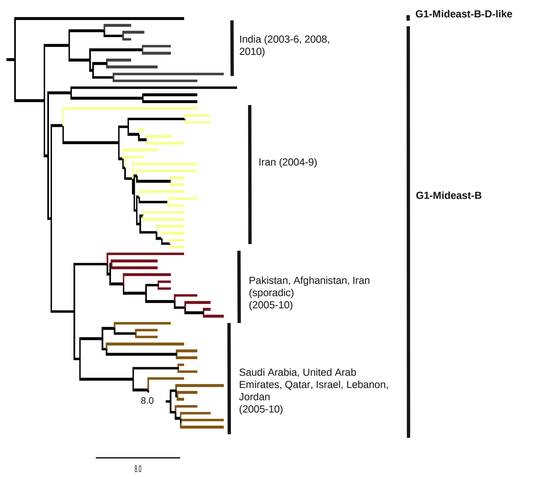


**Figure 6: Phylogenetic tree zoom in of Clade G1-MidEast-B in PhyCLIP tree**

There is only one H5 principle-based H9 nomenclature, with no updates (includes viruses up to 2011).

The WHO/OIE/FAO nomenclature system defined a high within-clade average pairwise p-distance threshold at 6%, to capture under-represented and potentially non-circulating clades. The WHO/OIE/FAO nomenclature defined clades with ranging sampling (MY91:3 members; Y280b:824 members) and within-clade average pairwise distance (My91:0.8%; Y280b:5.2%).

The G1-Mideast-B clade has a within-clade average pairwise distance of 5% in the WHO/OIE/FAO nomenclature. PhyCLIP captures clear lineage distinctions between viruses from different geographic regions to resolve four clades instead of one (Fig 6), including India- and Iran-specific clades.

The Y280b clade has a within-clade average pairwise distance of 5.2% in the WHO/OIE/FAO nomenclature. PhyCLIP delineates this large clade into ten clades (Fig 7).

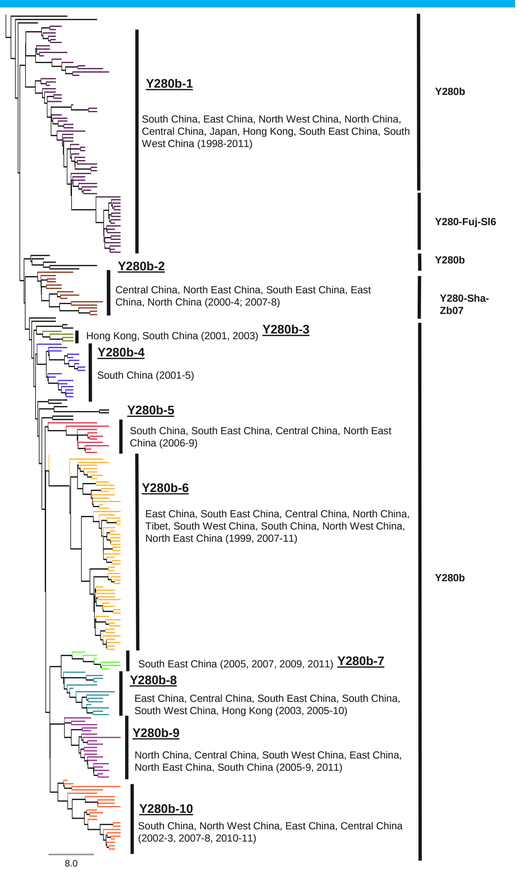


**Figure 7: Phylogenetic tree zoom in of Clade Y280b in PhyCLIP tree.**

## Conclusion

- Clade delineation in recent WHO/OIE/FAO updates for H5 are typically only performed for clades with new strains added after the previous update, with new clades only assigned to new viruses. While this minimises expansion of the nomenclature and prevents reassignment of older viruses to new clades as new information about the evolutionary trajectory is known, it results in violations to the stipulated bootstrap/pairwise distance cut-offs as the H5 phylogeny is based on all viruses collected to date.

- The WHO/OIE/FAO nomenclature proposes *ad-hoc* rules to resolve such violations. However, these rules change for different updates and are typically disconnected from the divergence between the different evolutionary lineages. This results in unreproducible inconsistencies as well as incongruence between the different clades.

- PhyCLIP provides a statistically-principled framework that can recapitulate the current H5 and H9 nomenclature developed on sequence divergence alone (e.g. differentiation between 1.1.1 and 1.1.2 in H5 nomenclature) without the introduction of arbitrary distance thresholds for clade designation. It also provides the foundation to an alternative nomenclature based on phylogenetics and statistics that minimises the limitations of the current system.

- PhyClip can be generalised to a variety of research questions concerning the identification of evolutionarily relevant clusters in phylogenies. This includes application to other influenza subtypes or gene segments, where the evolutionary continuum is not as well captured as H5 viruses owing to sampling bias.

## Literature cited

1. WHO/OIE/FAO H5N1 Evolution Working Group. Toward a unified nomenclature system for highly pathogenic avian influenza virus (H5N1). Emerg Infect Dis 14(7) (2008).
2. WHO/OIE/FAO H5N1 Evolution Working Group. Continuing progress towards a unified nomenclature for the highly pathogenic H5N1 avian influenza viruses: divergence of clade 2·2 viruses. Influenza Other Respi. Viruses 3, 59–62 (2009).
3. WHO/OIE/FAO H5N1 Evolution Working Group. Revised and updated nomenclature for highly pathogenic avian influenza A (H5N1) viruses. Influenza Other Respi. Viruses 8, 384–8 (2014).
4. Smith, G. J. D., Donis, R. O. & WHO/OIE/FAO H5 Evolution Working Group. Nomenclature updates resulting from the evolution of avian influenza A (H5) virus clades 2.1.3.2a, 2.2.1, and 2.3.4 during 2013-2014. Influenza Other Respi. Viruses 9, 271–6 (2015).
5. Shepard, S. S. *et al*. LABEL: Fast and Accurate Lineage Assignment with Assessment of H5N1 and H9N2 Influenza A Hemagglutinins. PLoS One 9, e86921 (2014).
6. Ragonnet-Cronin, M. *et al*. Automated analysis of phylogenetic clusters. BMC Bioinformatics 14, 317 (2013).
7. Prosperi, M. C. F. *et al*. A novel methodology for large-scale phylogeny partition. Nat. Commun. 2, 321 (2011).