Computational Identification of Naturally-occurring Human Adaptation Sites in PB2 of H5N1 Avian Influenza Viruses

Alvin X. Han[1,2](hanxc@bii.a-star.edu.sg), Colin A. Rusell[3], Sebastian Maurer-Stroh[1,2]
[1]Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Singapore
[2]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore
[3]Department of Veterinary Medicine, University of Cambridge

Introduction

- Avian influenza viruses must adapt their polymerase complex for efficient replication in humans.
- Various host adaptive substitutions have been identified, mainly in the polymerase basic 2 (PB2) protein, with E627K being the most studied.
- However, many avian viruses that have infected humans do *not* possess 627K or any previously identified host adaptive mutations.
- The extent to which these known mutations form an exhaustive list of host adapting substitutions is also not known.
- We developed a sequence conservation analysis that incorporates phylogenetics, mutual information and the recently available polymerase crystal structure to reliably infer *lineage-specific* host switch substitutions.

Results

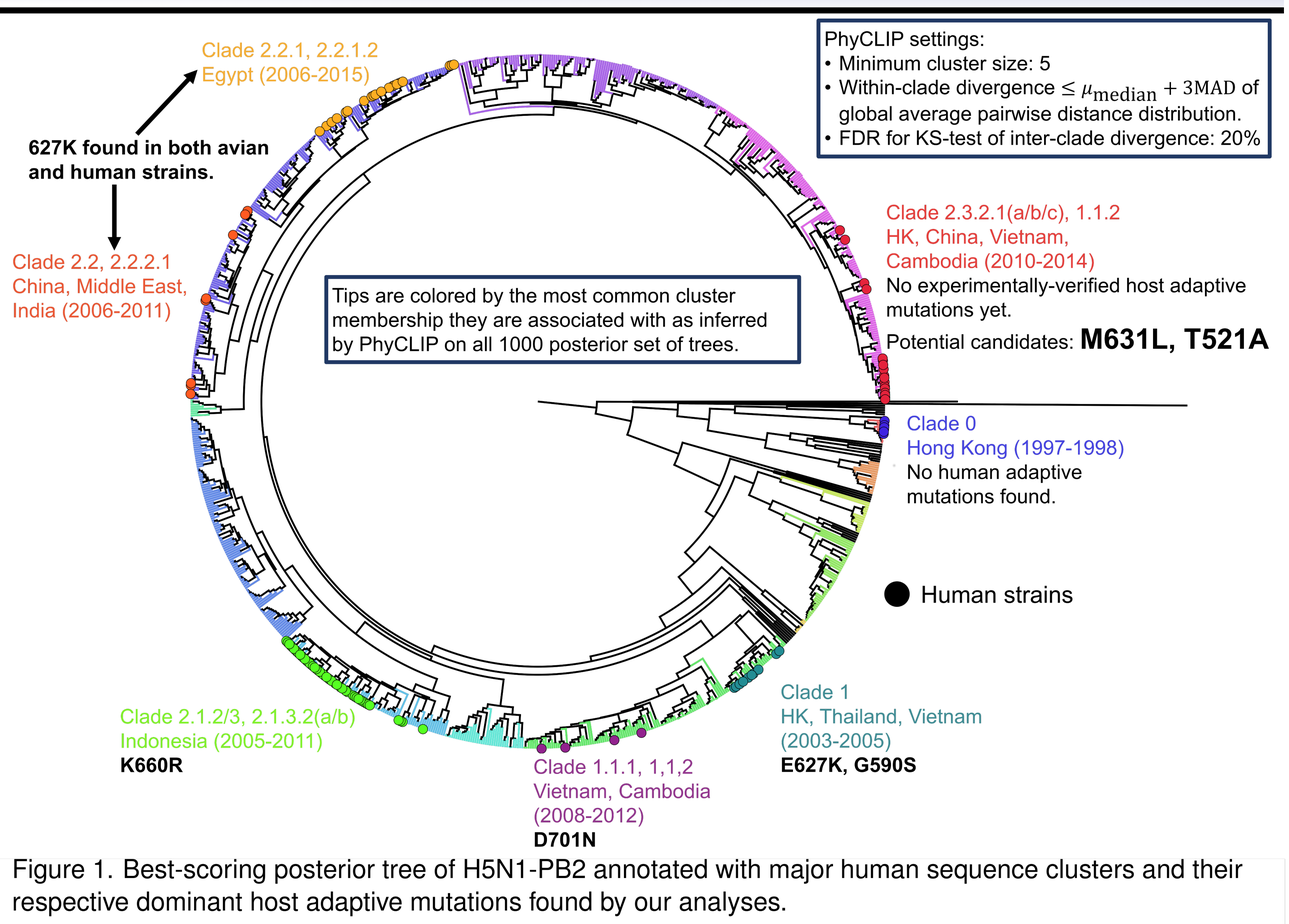
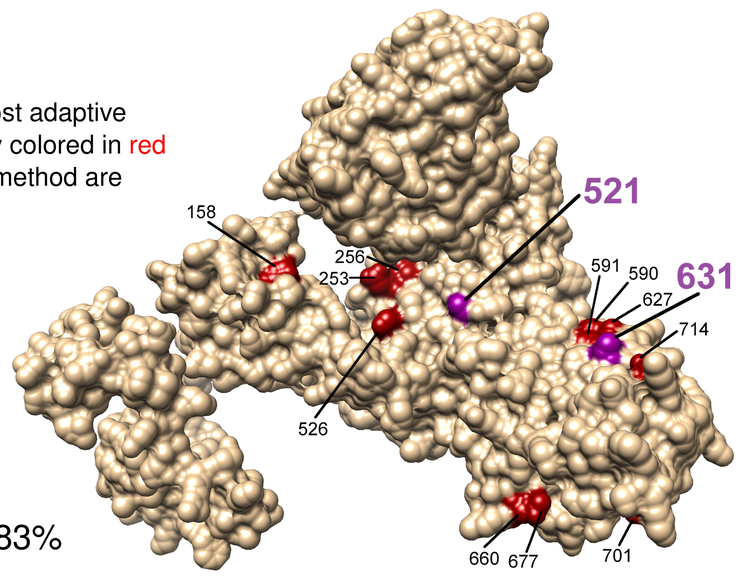


Figure 1. Best-scoring posterior tree of H5N1-PB2 annotated with major human sequence clusters and their respective dominant host adaptive mutations found by our analyses.

- Human host adaptive mutations in H5N1-PB2 are lineage-specific, dependent on the polymerase genetic background.
- Our analyses were able to correctly identify ≥1 experimentally verified (>10x relative activity) human adaptive mutation for almost every major cluster with multiple (>3) unique human sequences.
  - –Experimental verification obtained from literature by Mänz *et al*.
- New likely human adaptive mutations found by our method - **M631L** & **T521A**. Proximity of 631 and 521 to experimentally verified host adaptive sites adds further support (Figure 2).

Figure 2. PB2 crystal structure. Host adaptive positions with >10x relative activity colored in red while new mutations found by our method are colored in purple.



- Predictive performance (Table 1):
  - –Recall (True positive rate): 80%
  - –Precision: 57%
  - –F1-score: 67%
  - –Specificity (True negative rate): 83%

H5 Clade	*Time period*	*Location*	*Posterior support*	*Inferred mutation*	*Host Z-scores*	*Neighbor Z-scores*	*Prediction (**Correct**)*	*Avg activity (*Alt mutation tested)*
1	2003-2005	HK, Thailand, Vietnam	92%	G590S	-3.3	-2.1	***Host***	0.65
				E627K	-6.7	-1.9	***Host***	195.55
1.1.1/2	2008-2012	Cambodia, Vietnam	99%	E191G	-6.1	-1.7	Host	2.22 (*E191K)
				D701N	-3.4	-2.1	***Host***	17.16
2.1.2/3, 2.1.3.2, 2.1.3.2(a/b)	2005-2011	Indonesia	75%	I64M	-5.5	-3.2	***Founder***	0.87
				I292T	-14.1	-41.2	***Founder***	0.71
				P465S	-22.1	-14.6	***Founder***	0.50
				T524I	-38.6	-29.3	***Founder***	0.65
				K526R	-11.8	-37.9	Founder	10.60
				K660R	-4.5	-2.1	***Host***	12.32
				E677G	-6.6	-3.7	***Founder***	1.19
2.2, 2.2.2.1	2006-2011	China, Middle East, Bangladesh	68%	M47I	-3.3	-1.1	Host	2.29
				M570V	-9.1	-74.7	***Founder***	0.36 (*M570I)
2.2.1, 2.2.1.2	2006-2015	Egypt	92%	D195Y	-11.9	-45.9	***Founder***	0.53
				K197R	-4.2	-63.8	***Founder***	0.02 (*K197N)
				I292M	-5.8	-32.3	***Founder***	0.71 (*I292T)
				R389K	-6.6	-3.0	***Founder***	0.42
				T559A	-9.5	-40.9	***Founder***	1.20 (*T559N)
2.3.2.1a/b/c, 1.1.2	2010-2014	HK, China, Vietnam, Cambodia	100%	I411V	-16.9	-6.1	***Founder***	0.21 (*I411M)
				N456D	-6.7	0.6	Host	0.99
				T683K	-11.1	-12.7	***Founder***	1.61 (*T683A)
				A684S	-3.7	-4.5	***Founder***	2.54
				K702R	-3.1	-28.0	***Founder***	1.69

Table 1. Prediction performance for substitutions experimentally tested by Mänz et al. (2016).

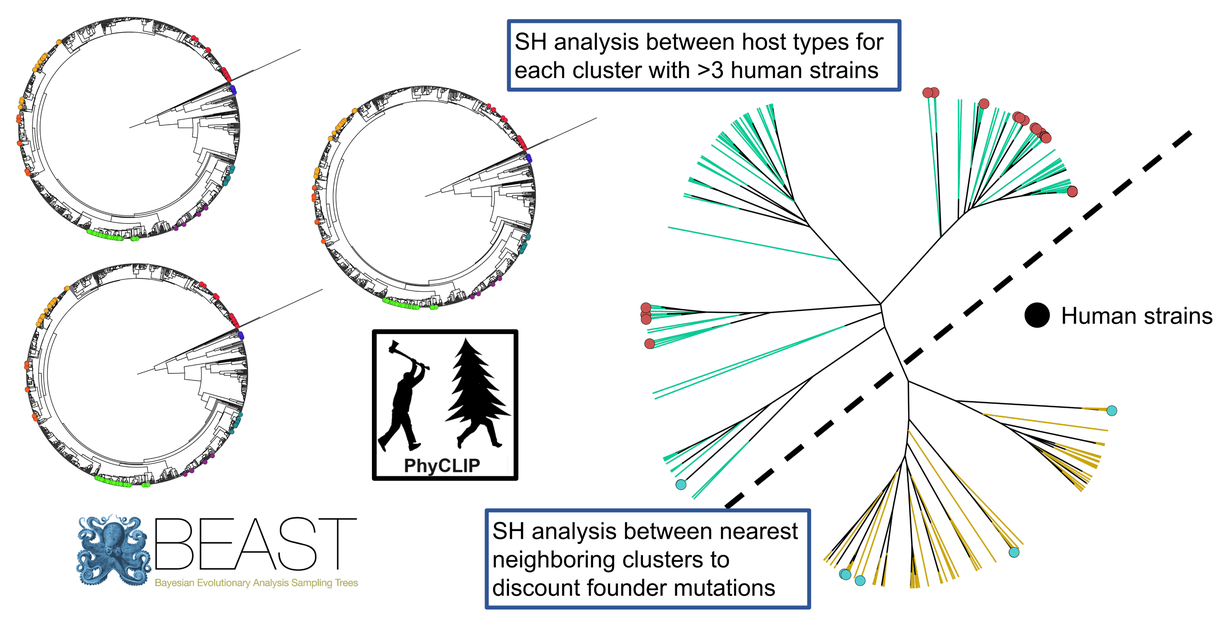
Methods

Sequence curation & Bayesian posterior set of phylogenetic trees (PST)

- Downloaded 1908 H5N1-PB2 nucleotide sequences collected between 1997-Jul 2017 (>90% full length, <1% unknown residues, GsGd-lineage, non-redundant at amino acid level) from GISAID.
- Random down-sampling based on H5-clade and host type ⇒ 1011 sequences.
- 1000 Bayesian posterior trees using BEAST (GTR+GAMMA, 150 million steps, 50 million burn-in, ESS>700).

Phylogenetic Clustering by Linear Integer Programming (PhyCLIP)

- **Group taxa in *each* posterior tree into statistically principled clusters using PhyCLIP, a newly developed clustering algorithm.**
- *Objective*: Assign cluster membership to as many taxa as possible.
- *Within-cluster constraint*: mean pairwise taxa patristic distance of each cluster ≤ upper limit based on global distribution of pairwise distances.
- *Inter-cluster constraint*: Using the two-sample Kolmogorov–Smirnov test, the pairwise taxa patristic distance distribution of cluster i must be distinct from that should clusters i and j form a cluster together ($j \neq i$) .
- **More details on PhyCLIP can be found in our poster entitled "A New Method for Statistical Clustering of Influenza Sequence Data".**



Sequence Harmony (SH)

- Mutual information based metric that identifies residues specific to user-defined groups for a given protein alignment (Pirovano *et al*., 2006).
- SH analysis performed for groups that differentiate:
  - –**Host types (avian vs. human) in *each* cluster to identify *host-specific* mutations**
  - –**Nearest neighbouring clusters to identify *lineage-specific* founder mutations (i.e. substitutions enriched in geographic region and time period without conferring host adaptive phenotype)**.
- From the proportions of sequences with amino acid x in position i associated with group A and that for sequences in group B:

$$SH_i^{A/AB} = \sum_x p_{i,x}^A \log_2 \frac{p_{i,x}^A}{p_{i,x}^A + p_{i,x}^B} \quad SH_i = \frac{1}{2}\left(SH_i^{A/AB} + SH_i^{B/AB}\right) \in [0,1]$$

- Symmetric; $SH_i = 0 \Rightarrow$ completely non-overlapping residue composition; $SH_i = 1 \Rightarrow$ identical compositions.
- Predictability can be improved by incorporating scores of spatial neighbours when overall conservation is high (Panchenko *et al*., 2004). Using PB2 crystal structure (PDB: 4WSB, Pflug *et al*., 2014), we calculated:

$$SH_i' = \frac{1}{2}SH_i + \frac{1}{2}\left(\frac{\sum_{j\in\text{window}} SH_j}{|\text{window}|}\right), \quad \text{window} \in \{j \quad \forall \quad \text{d}(\mathbf{i},\mathbf{j}) \leq 8\text{Å}\}$$

- Reliability of SH' inferred from empirical *Z*-scores; 1000 randomization of alignment's group labels to obtain null (random group labels) distribution of SH_i' scores.
- **Lower SH_i' score implies specificity ⇒ Lower (negative) *Z*-score indicates greater significance.**
- **Separate SH analyses performed for all posterior trees to account for phylogenetic uncertainty.**

Conclusions

- We have developed a computational method that can reliably identify both known and potential lineage-specific molecular signatures of host adaptation.
- Without training against any experimental data sets, our method has already demonstrated high specificity (83%), obviating redundant testing of founder mutations that do not confer host adaptation phenotype.
- The tools described here can be applied to any influenza subtypes and proteins for making similar inferences.

Literature cited:
1. Pirovano, W., K.A. Feenstra, and J. Heringa, *Sequence comparison by sequence harmony identifies subtype-specific functional sites.* Nucleic Acids Res, 2006. 34(22): p. 6540-8.
2. Panchenko, A.R., F. Kondrashov, and S. Bryant, *Prediction of functional sites by analysis of sequence and structure conservation.* Protein Science, 2004. 13(4): p. 884-892.
3. Lee, R.T., et al., *All that glitters is not gold - founder effects complicate associations of flu mutations to disease severity.* Virology Journal, 2010. 7(1): p. 297.
4. Mänz, B., et al., *Multiple Natural Substitutions in Avian Influenza A Virus PB2 Facilitate Efficient Replication in Human Cells.* Journal of Virology, 2016. 90(13): p. 5928-5938.