bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Glycan Clock Forecasts Human Influenza A Virus Hemagglutinin Evolution

Meghan O. Altman[1]*, Matthew Angel[1], Ivan Košík[1], James S. Gibbs[1], Nídia S. Trovão[2,3], Seth J. Zost[4], Scott E. Hensley[4], Martha I. Nelson[2], Jonathan W. Yewdell[1]*

[1] Cellular Biology Section, Laboratory of Viral Diseases NIAID, NIH, Bethesda MD 20892

[2] Division of International Epidemiology and Population Studies, Fogarty International Center, NIH, Bethesda MD 20892

[3] Global Health and Emerging Pathogens Institute, Icahn School of Medicine at Mount Sinai, NY 11766

[4] Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

* Corresponding authors

Abstract

N-Linked glycosylation of the influenza A virus hemagglutinin (HA) globular domain greatly influences viral antigenicity, receptor binding, and innate immunity. Gel electrophoresis and bioinformatics analysis reveals that glycans are added to HA at regular intervals, until a glycan limit is reached, after which, at 2-fold longer intervals, glycans are either swapped to different glycan sites, or the HA is replaced by a new pandemic virus. These findings were used to predict a new glycan addition to pH1N1 HA over the 2015-2016 season. Together, these glycan patterns trace the timeline, trajectory, and eventual fate of HAs, from their pandemic introduction to eventual replacement. By deciphering the clock-like rhythm of glycan addition over the past century, we can forecast future influenza virus evolution.

Introduction

Seasonal influenza A virus (IAV) annually sickens millions and kills hundreds of thousands of people (WHO, 2003). IAV remains endemic despite high levels of adaptive immunity in nearly all humans after early childhood. IAV persists due to the rapid selection of mutants with amino acid substitutions in the viral hemagglutinin (HA) glycoprotein that enable escape from neutralizing antibodies (Abs). HA is a homotrimeric glycoprotein with a globular head domain resting atop a stem that anchors HA to the virion surface. The head has a receptor binding site that attaches virions to cell surface terminal sialic acid residues, initiating infection. Ab binding to the head neutralizes virus by blocking attachment or the conformational alterations required for HA-mediated membrane fusion. Understanding Ab-driven HA evolution is essential to improving influenza vaccination, which currently, offers only moderate protection from infection (CDC, 2017; Couch and Kasel, 1983).

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As newly synthesized HA enters the endoplasmic reticulum of IAV infected cells, oligosaccharides are attached to Asn residues present in the motif Asn-X-Ser/Thr-Y, where X/Y are all amino acids except Pro. Attached glycans can be high-mannose glycans or larger, more complex branched glycans (Keil et al., 1985; Khatri et al., 2016). HAs possess multiple highly conserved glycans on the stem domain crucial for HA folding and oligomerization (Daniels et al., 2003). Glycosylation of the head is much more variable, as evolving HAs add, retain, and occasionally lose potential head glycosylation sites (An et al., 2015; Sun et al., 2011). Head glycans can promote viral fitness by shielding virus from antibody binding and tuning receptor affinity and specificity (Job et al., 2013; Medina et al., 2013; Sun et al., 2013b; Vigerust et al., 2007). Conversely, glycans can deleteriously modulate receptor binding, enhance viral neutralization by innate immune lectins, or increase stress in the endoplasmic reticulum during translation (Alymova et al., 2016; Das et al., 2011; de Graaf and Fouchier, 2014; Hrincius et al., 2015; Park et al., 2016; Yang et al., 2015).

Glycosylation changes are generally not correlated with changes between antigenic clusters (Fonville et al., 2014), in part, because glycan addition typically lowers HA receptor avidity (Aytay and Schulze, 1991), complicating binary assays such as hemagglutination inhibition (HI) (Das et al., 2011). Further, antigenic clusters are defined by HI assays performed using ferret anti-sera, but HA head glycans are selected to escape neutralization by human antibodies specific for antigenic sites that are weakly targeted by ferret antibodies (Cobey and Hensley, 2017). Consequently, glycan evolution is typically unaccounted for in molecular evolution studies using HI-based analytic tools such as antigenic cartography (Blackburne et al., 2008; Sun et al., 2013a).

H1 and H3 HAs accumulate glycans at regular intervals before reaching an apparent glycan limit

Studies of HA glycan evolution rely nearly exclusively on bioinformatic predictions of glycan addition. NetNGlyc, the state of the art algorithm (Gupta, 2004), imperfectly predicts N-linked glycosylation (Khatri et al., 2016) and provides no information into the nature of added glycans. We therefore used gel electrophoresis to monitor glycan addition to HAs from 72 egg-grown H1N1 and H3N2 strains based on decreased HA migration (**Fig. 1**, **fig. S1-S4, file S1, table S1**). HA migration strongly correlates with the number of computationally predicted head glycosylation sites (Pearson coefficient $r = 0.96$, $P < 0.01$) (**Fig. 1A**). H1 adds an apparent 3.9 ± 0.1 kDa [$R^2 = 0.89$] per glycan, suggesting the majority of such glycans are highly branched. The average H3 glycan, by contrast, is smaller (3.0 ± 0.2 kDa) ([$R^2 = 0.90$]), consistent with greater addition of simpler glycans. The size difference between H1 and H3 HAs is consistent with mass spectrometry characterization showing that H1 glycans skew to larger, more complex structures, while H3 glycans at several sites are majority high-mannose (Khatri et al., 2016; Sun et al., 2013b).

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 1

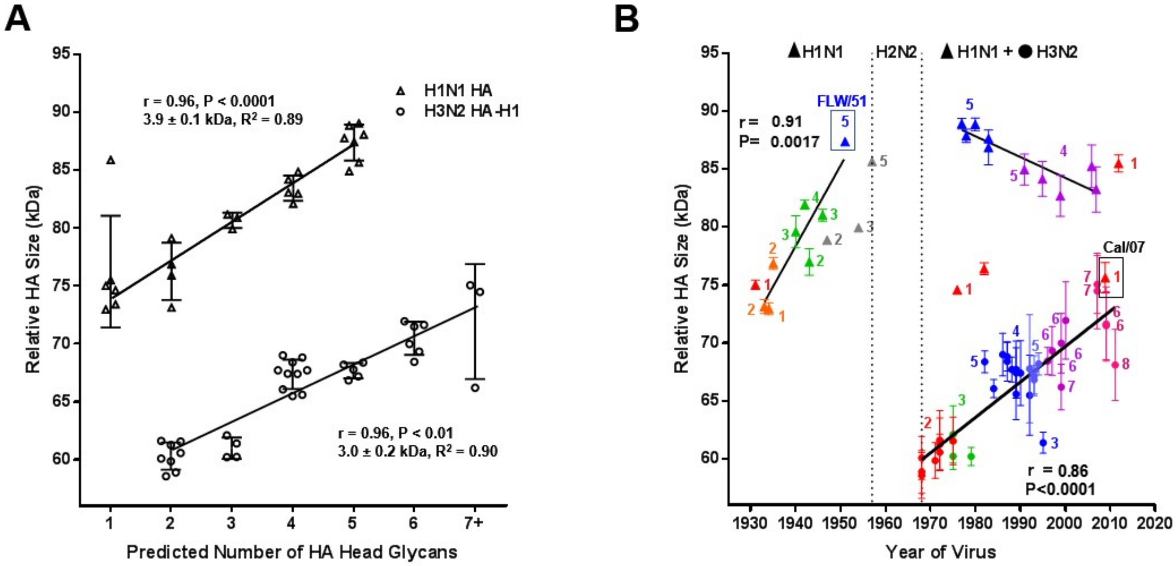


Fig. 1. Quantitative survey of glycan evolution in IAV. (A) Relative HA size, measured by migration rate through a gel, correlates with predicted number of head glycans. H1N1s (triangles) add 3.9 ± 0.1 kDa per glycan or 15.5 kDa total in increasing from 1 to 5 glycans. H3N2s (HA1, circles) add 3.0 ± 0.2 kDa per glycan or 15 kDa total in increasing from 2 to 7 glycans. **(B)** HA size over time for H1N1s (triangles) and H3N2s (HA1, circles) reveals regular glycan addition. Strains are color-coded to match glycan groups in **fig. S2.** Numbers of head glycans per strain are noted next to individual points. A glycan is added on average every 5 years for sH1N1 (3.9 kD) and every 8 years for H3N2 (3.0 kD). The strain most related to sH1N1 reintroduced in 1977, A/Fort Leonard Wood/1951, is boxed and labelled in blue. The pH1N1 strain, A/California/07/2009, is boxed and labelled in black. Pearson coefficient of correlation (r) and its P value are shown on both graphs. Error bars are standard error of the means from n = 3-5 blots.

Importantly, for H1 and H3 HAs, the total glycan mass added to the head during evolution in humans is, respectively, 15.5 kDa and 15 kDa. This is consistent with the idea of an upper limit of glycan shielding for H1 and H3 HA head domains totaling to ~20 kDa (5 glycans for H1, 7 for H3). This is not strictly due to HA functional constraints: H1 HAs can accommodate at least 8 head glycans while maintaining *in vitro* fitness (Eggink et al., 2014). Rather it points to glycan-based fitness costs *in vivo*, possibly due to innate immune mechanisms or alterations in virus binding to receptors found in human airways (Tate et al., 2014).

Plotting relative HA size versus time reveals the history of seasonal H1N1 (sH1N1) glycosylation (**Fig. 1B**, triangles). First-wave sH1N1 (1933-1957) HAs behaved differently than the second-wave sH1N1 representatives (1977-2008). During the first-wave, relative HA size increased by an average of 3.9 kDa (or one glycan) every 5.6 years (slope = 0.69 kDa/year, R^2 = 0.89). H1N1 strains that re-emerged after the 1977

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

introduction of H1N1 exhibited similar relative HA size to the 5-head glycan A/Fort Leonard Wood/1951 (inside blue box, **Fig. 1B**). Rather than increasing in number as occurred previously, electrophoresis confirms the NetNGlyc prediction that 5 head glycans were maintained until decreasing to 4 glycans after 1991.

These data indicate that circulating H1 HAs did not acquire more than 5 head glycans. Indeed, HA sequences predicted to have 6 head glycans are highly unusual over the century of H1 evolution (1918-2017), except in 1986 and 1987, when they competed among H1 strains (**fig S5A**). Rather than adding glycans, sH1N1 strains that reemerged after 1977 exhibited glycan adjustments spaced at 10.0 ± 0.9 year intervals (**Fig. 2**). In 1986, N71 pH1N1 (numbering convention is shown in **table S1**) replaced N172, as N144 swapped with N142, which though adjacent to 144, is spatially oriented to shield different epitopes. Around 1997, N286 was lost. Strains with the 4 remaining glycans circulated for 11 years before becoming extinct during the 2009 pH1N1 pandemic.

As with H1 HAs, H3 HAs exhibit a regular rhythm of glycan evolution (**Fig. 3**). Original (1968) pandemic H3 strains had two head glycans. Four times since 1968 glycans were added to H3 HA at regularly spaced 5.0 ± 0.6 year intervals (N126 by 1974, N246/N122 by 1980, N133 + N122 by 1998, and N144 by 2004) (**Fig. 4**). This timing is similar to the average rate of glycan addition (one every 5.6 years) seen by SDS-PAGE for sH1N1 from 1933 to 1951.

With the addition of the 7th head glycan in 2004, H3 HAs had increased 15 kDa from the original 1968 strain (**Fig 1B**). It is apparent from the global human influenzome (GHI) that H3 HAs with more than 7 head glycans lose in the competition for human circulation. While 8-glycan strains are present, they represent <1% of H3N2 sequences in any year (**fig. S5B**). As with first wave H1 viruses, ~10 years after reaching the glycan limit, a glycan swap was selected for, as the glycan at residue N144 was replaced by a glycan at residue N158 (**fig. S6A**).

There was one exception in the ~5 year interval timing for H3 glycan addition. Addition of a fifth glycan at residue 276, occurred 12.5 years after the previous glycan addition (**Fig. 3**, **orange**). In 1993 highly robust N276 strains swept the globe, dominating sH1N1 strains, before disappearing after only 3 seasons. N276 is unique because the Asn residue bearing the glycan is immediately adjacent to a disulfide-bonded cysteine (C277). We speculate that the structural rigidity conferred by the disulfide reduced the fitness of N276, requiring epistatic changes elsewhere in the HA to improve competitiveness, both delaying fixation of the glycan addition and also reducing its competitiveness against future strains lacking N276.

This longer time interval, associated with increased fitness costs in functionality of glycan addition mutants, can also be seen with human H2N2 strains, which never added a glycan during their 11-year era (1957-1967). While selection of H2N2 strains with additional head glycans readily occurs via *in vitro* antibody escape, these multi-glycan H2 HAs are

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

functionally constrained, as they exhibit less cell fusion and lower receptor binding (Tsuchiya et al., 2001).

Fig. 2

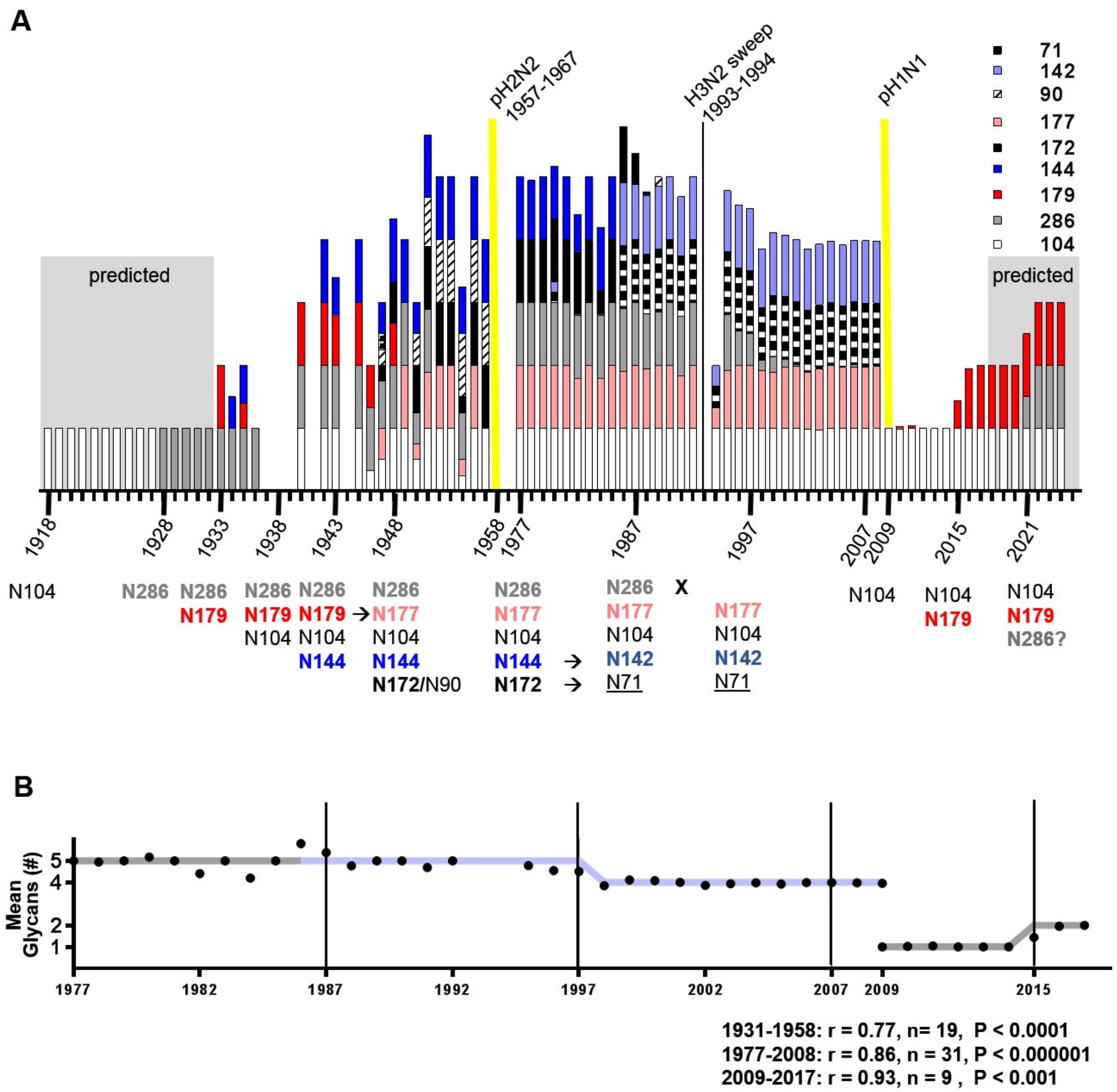


Fig. 2. H1 Glycan Evolution 1918-2024. (A) Stacked bars correspond to the percentage of all human H1N1 sequences from a given year containing a glycan listed in the legend. Dates of glycan transition are written on the timeline below along with the glycosylated residues. Projected values for strains from 1918-1933 and 2017-2024 are shown in shaded regions. Years with no data are left blank. Sequence counts for each year are shown in **fig. S5**. Pandemics are denoted by yellow lines. **(B)** The mean predicted number of glycans among all sequences from each year are shown as black circles. This data is laid atop the number of glycans predicted (grey line, light blue line indicates glycan swap from sequences containing N144/N172 to N142/N71). Black vertical lines denote predicted years for glycan evolution. Pearson's coefficient of correlation (r) with sample number (n) and P value between the model and the data is shown below.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 3

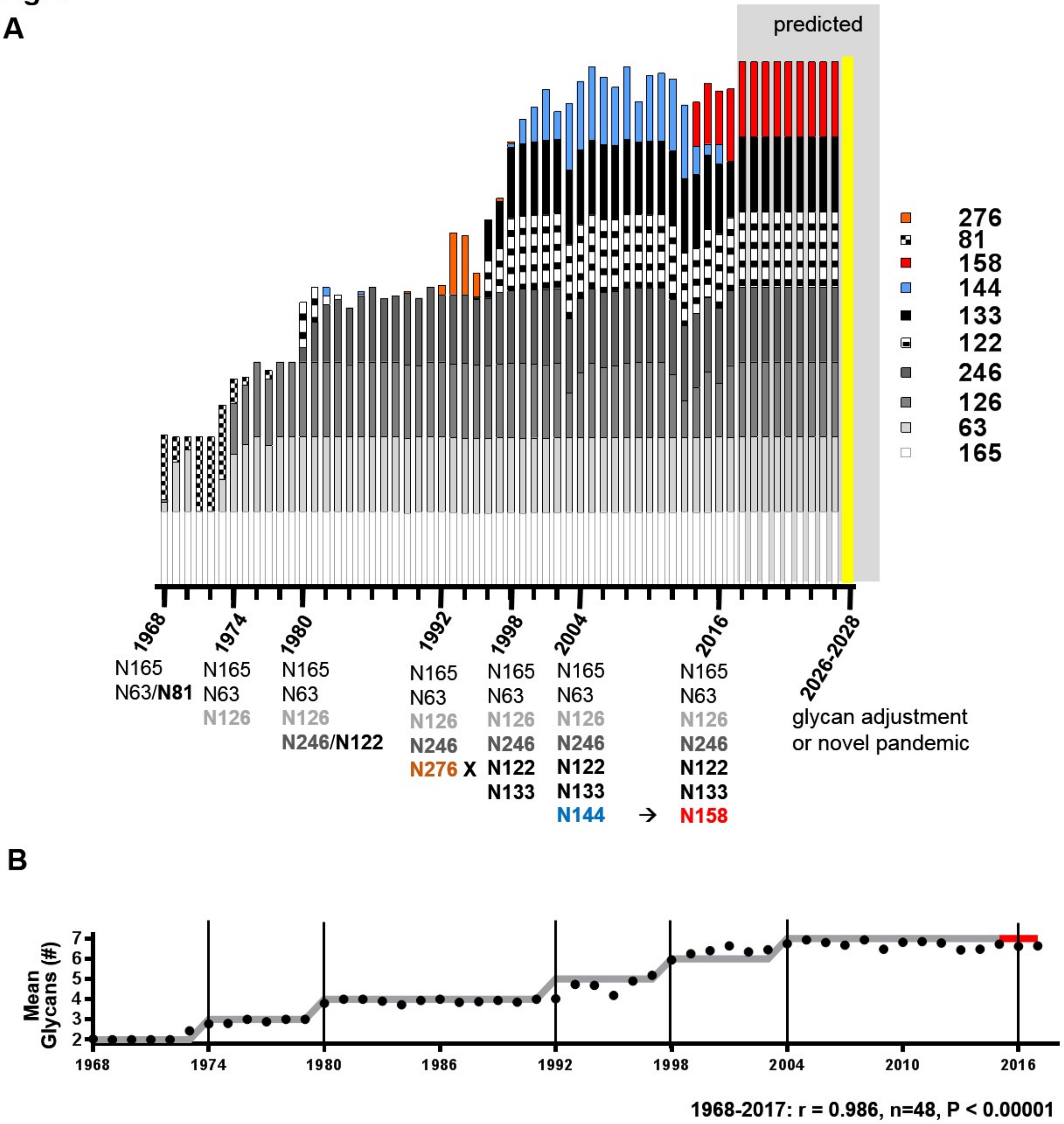


Fig. 3. H3 Glycan Clock 1968-2029. (A) Stacked bars correspond to the percentage of all human H3N2 sequences from a given year containing a glycan listed in the legend. Dates of glycan transition are written on the timeline below along with the glycosylated residues. Projected values for strains from 2017-2029 are shown in shaded regions. A potential pandemic is denoted by yellow line. **(B)** The mean predicted number of glycans among all sequences from each year are shown as black circles, which are displayed atop the number of glycans predicted (grey line, red line indicates glycan transition from sequences containing N144 to N158). Black vertical lines denote predicted years of glycan evolution. Pearson's coefficient of correlation (r) with sample number (n) and P value between the model and the data is shown below.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

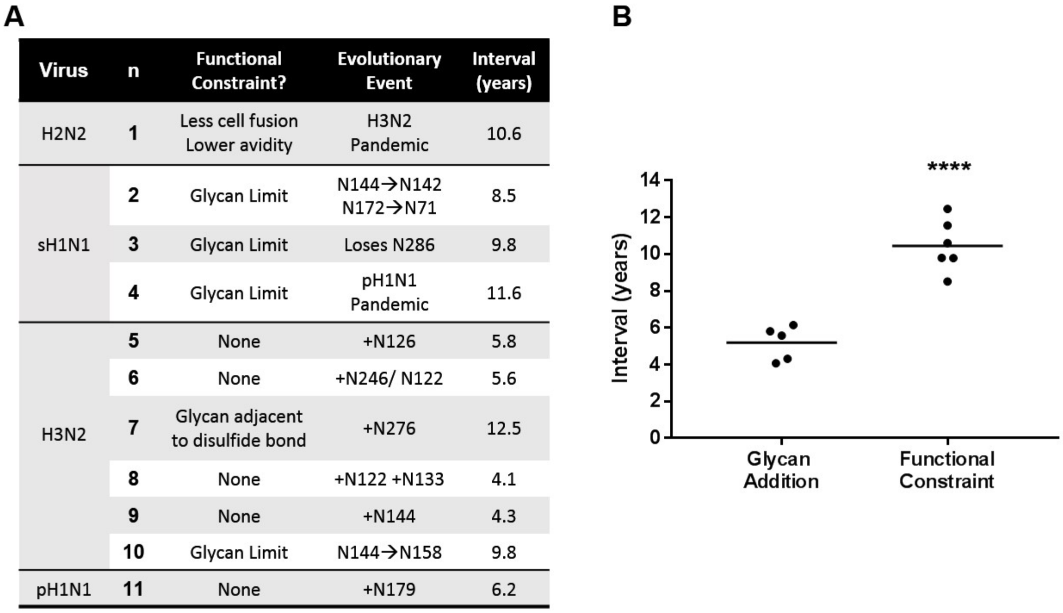
HA glycosylation patterns predict human influenza virus evolution

Based on these observations we can identify 3 patterns that consistently occur in human IAV evolution:

1. After introduction of a pandemic strain, glycans are typically added at intervals of ~5 ± 1 y.
2. Glycans are limited on the HA head to 5 and 7 glycans, respectively, for H1 and H3 HAs, with a total mass of ~20 kDa due to preferred use of complex (H1) *vs.* simple (H3) glycans.
3. If glycan addition is constrained by fitness loss, then glycan adjustment (i.e. swapping positions) or pandemic replacement occurs at intervals of ~10.5 ± 1 y.

Quantitative analyses supporting the statistical significance of patterns one and three are shown in **Fig. 4**.

Fig. 4



Virus	n	Functional Constraint?	Evolutionary Event	Interval (years)
H2N2	1	Less cell fusion Lower avidity	H3N2 Pandemic	10.6
sH1N1	2	Glycan Limit	N144→N142 N172→N71	8.5
	3	Glycan Limit	Loses N286	9.8
	4	Glycan Limit	pH1N1 Pandemic	11.6
H3N2	5	None	+N126	5.8
	6	None	+N246/ N122	5.6
	7	Glycan adjacent to disulfide bond	+N276	12.5
	8	None	+N122 +N133	4.1
	9	None	+N144	4.3
	10	Glycan Limit	N144→N158	9.8
pH1N1	11	None	+N179	6.2

Fig. 4. Quantitative analysis of evolutionary intervals (A) The interval in years prior to the occurrence of glycan altering events was calculated by fitting the proportion of changed sequences to a 3-parameter logistic regression and finding the inflection point of the curve. Events were classified as either having (n=6) or not having (n=5) a functional constraint that limited glycan addition. Starting dates were determined by a fixed sliding window every 10 years for sH1N1 and 6 years for H3N2 (black vertical lines in **Figs. 2B + 3B**). **(B)** With no functional constraint, addition of a glycan occurs on average every 5.2 ± 0.9y. If the glycan is functionally constrained, evolution occurs at 2-fold longer intervals averaging 10.5 ± 0.9y (Student's t-test ****P<0.0001).Using these patterns, we can forecast future IAV glycan evolution (**Fig. 2A** and **3A**, grey areas). H3 HA reached its glycan limit in 2004, and consistent with pattern 3 above, swapped a glycan (N144 to N158) during the 2015-2016 season (**fig. S6**). In addition to this example, when we include all subtypes (H1, H2, and H3), there are 5 additional cases where functionally constrained viruses (i.e. at the glycan limit) adjust glycans after ~10.5 ± 1 y intervals. These data suggest H3 will either undergo a second glycan swap or be replaced by a new pandemic HA gene from the animal reservoir over the 2026-2028 seasons.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Evolution of pH1 HA over the past several years has been sufficiently well characterized to provide a clear example of clock-like glycan addition. Initial 2009 pH1N1 had a single predicted head glycan (N104), and as expected, the HA of a representative early strain, A/California/07/2009 (**Fig. 1B**, inside black box), co-migrated with strains from the 1930s and two other swine-origin human strains, A/New Jersey/8/1976 and A/Memphis/4/1982 with single predicted head glycan (**Fig. 1B**, red triangles).

According to pattern 1, we predicted that pH1 should add a glycan over the 2014-2016 seasons. Identifying all potential HA head glycosylation sites in pH1 sequences in the Flu Database (FluDB) revealed a predicted second glycan arising through an S to N change at residue 179 (**Fig. 5A**). N179 viruses were present at low frequencies in each season since pH1N1 introduction (**Fig. 5B**), demonstrating that selection for such viruses was not sufficient to overcome fitness limitations in the human population. Viruses with this mutation did not selectively sweep the human pH1N1 GHI until the 2015-2016 season (**fig. S6B**). All of these sweeper strains, unlike previous N179 strains, also had an I233T mutation.

Residue 179 is located in an antigenic site, called Sa, that is immunodominant in many contemporary human Ab responses. Two seasons before the N179 glycan sweep, over the 2013-2014 season, the GHI was swept by strains that were antigenically distinct due to a K180Q mutation (Huang et al., 2015; Linderman et al., 2014). Surprisingly, through reverse-genetic manipulation of the California/07/2009 strain, we found that despite having the amino acid sequence necessary for glycosylation, neither the N179 single- or N179/I233T-double mutants added a glycan to HA in virions. Rather all three mutations: N179, I233T, and crucially, K180Q were required for full glycosylation of the 179 site (**Fig. 5C**).

The pH1N1 HA structure provides a ready explanation for the epistatic interactions of the HA residues 179/180 and 233. Residues 233 and 179/180 are in close proximity across the trimer interface (**Fig. 5D**), with the change from bulkier and more hydrophobic Ile to Thr likely facilitating the addition of the large, polar glycan on the other side of the trimer interface. Notably, in H3 strains with a N179 glycan, the residue analogous to 233 in pH1N1 is also a small polar residue (Ser).

A phylogenetic analysis of all HA sequences from pH1N1 viruses collected globally during 2009-2017 revealed that N179 viruses emerged repeatedly during 2009-2014, but did not transmit onward (**file S2**). In 2015 a population of N179 viruses emerged and, within 15 months, became the exclusive global lineage in the sequence record. The earliest N179/I233T strains in our database are from Costa Rica, Turkey, and Iran in January 2015. Phylogeographic analysis determined that the N179/I233T clade likely emerged several months earlier, perhaps in the Middle East (**Fig. 6,** posterior probability 0.8). The virus rapidly disseminated to Asia and North America, and onward to Europe and other regions. North America had the largest number of N179 viruses, but the inference that

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

N179 strains may have first arose in the Middle East was found to be robust to sample bias (**fig. S8**).

Fig. 5

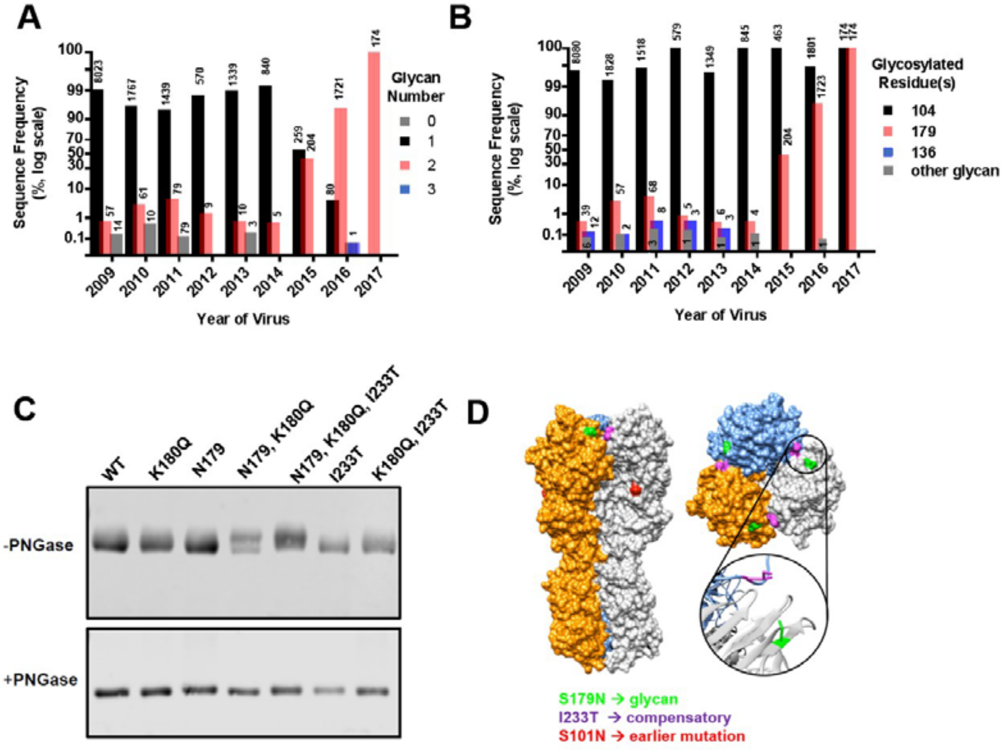


Fig. 5. Forecasted pH1N1 glycan addition. (A) Frequency of human pH1N1 sequences in FluDB by number of predicted glycans in each calendar year. The number of sequences represented by each bar is shown on top of bar. Until January 2015 >90% of sequences had 1 glycan (black). Starting in 2015, strains with 2 glycans swept (red). **(B)** Frequency of the individual residues glycosylated in human pH1N1 sequences. N104 predominates throughout (>95%, black). N179 (red) and N136 (blue) appear at low frequencies until 2015 when strains having both N179 and N104 extinguish N104-only strains. N179 strains are found in every year. Rarely, in addition to N104 a residue other than N179 or N136 has a potential glycosylation site (grey). Unlike N179, a single mutation with N136 alone is sufficient for glycosylation (Job et al., 2013). **(C)** HA Western blots of A/California/07/2009 mutants with permutations of N179, K180Q, and I233T with and without PNGase treatment to remove glycans. The single mutant N179 alone does not shift in size, indicating it is not detectably glycosylated. While a portion of the double mutant N179/K180Q shifts, a complete shift, indicating majority glycosylation, is only seen with the triple mutant N179/K180Q/I233T. **(D)** A/California/07/2009 (PDB: 3LZG) hemagglutinin crystal structure with N179 highlighted in green and the compensatory mutation, I233T, in purple adjacent across the trimer. A slightly earlier antigenic site mutation S101N is shown in red. Inset shows magnified view of the two mutated residues.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 6

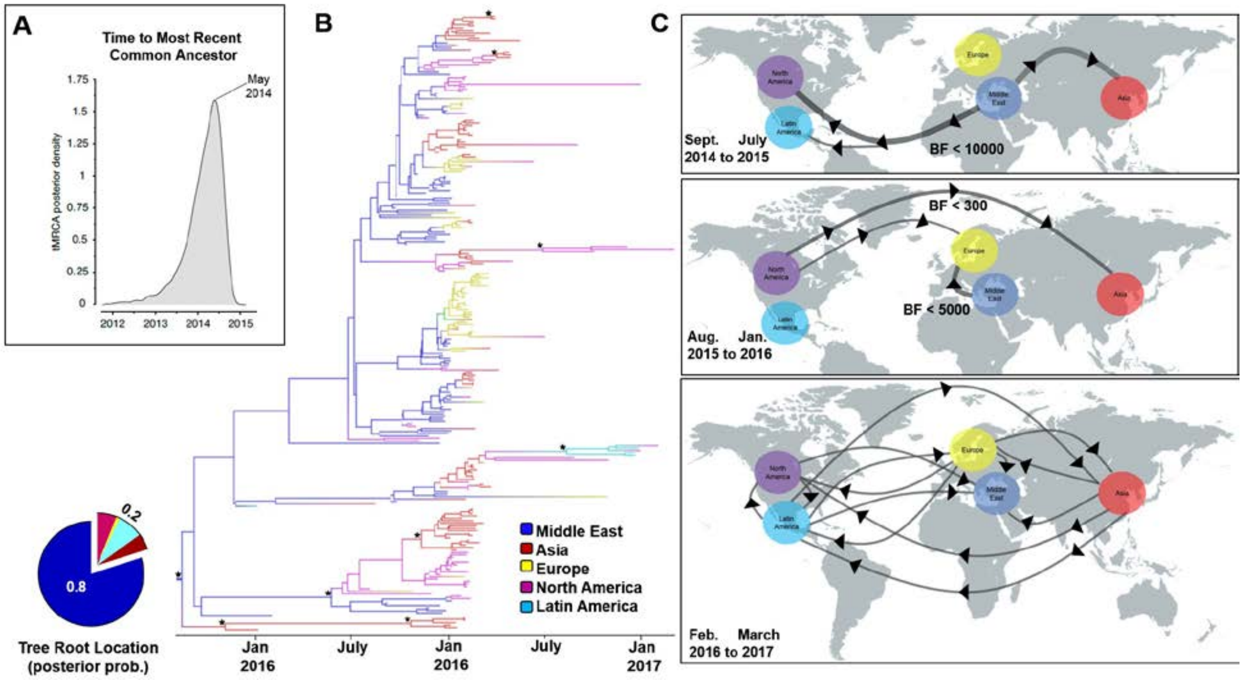


Fig. 6 Evolutionary relationships and reconstruction of the dispersal of N179/I233T HA strains. (A) Posterior density for the time to Most Recent Common Ancestor (tMRCA). Origin is estimated to have occurred in May 2014 (95% Highest Posterior Density: August 2013–November 2014). **(B)** Time-scaled MCC tree inferred for 255 N179/I233T HA sequences collected worldwide during 2015 – 2017 (dataset B). The color of each branch indicates the most probable location state. Asterisks indicate nodes with posterior probabilities >0.90. Pie chart indicates posterior probabilities for locations at the root of the phylogeny. The Middle East is estimated as the location where N179/I233T strains emerged with a posterior probability of 0.80. **(C)** A possible model for the spatial dispersal of N179/I233T strains was inferred from the MCC tree, based on the data available, and visualized using SpreaD3 at three time points (Bielejec et al., 2016). N179/I233T strains emerged around May 2014, potentially in the Middle East, and disseminated to Asia, Latin America and North America, and then globally from February 2016 on. Unless noted on the figure, the Bayes factors are < 100.

Based on biochemistry, epidemiology, and immunology we can reconstruct the selection of N179 glycan addition. Antibody pressure on the Sa antigenic site first selected for the K180Q escape mutation, which swept the GHI. Several years of immune pressure against the new site led to selection of the N179 glycosylation site in tandem with the I233T mutation, which enabled the trimer to accommodate the new glycan.

pH1N1 added a glycan during the 2015-2016 selective sweep. By pattern 1, another glycan addition should occur during the 2020-2022 seasons. The first 3 glycans added to sH1N1 were N104, N179, and N286 (see strain A/Hickox/JY2/1940). Pandemic H1N1 started with N104, and recently added N179, if pH1N1 recapitulates sH1N1, the next glycan added would be N286.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In conclusion, our findings show that HA glycan evolution in human strains follows predictable patterns. The consistency of the patterns remains to be explained: why does it typically require 5-6 year intervals for glycan addition mutants to become ascendant, and twice this time to swap between glycan sites when the glycan limit has been reached? Part of the answer, no doubt, is the time required for herd immunity to apply sufficient pressure. Another part of the answer appears to lie with the need for epistatic mutations to restore HA function. In any event, our findings can immediately be applied for vaccine selection, since they predict when strains with new glycans will become dominant and identify seasons with higher pandemic potential.

ACKNOWLEDGEMENTS

This work was supported by the Division of Intramural Research, NIAID. Glennys Reynoso provided outstanding technical support. Jia Jie Wei, Mina Seedhom, and Davide Angeletti provided insightful conversations. Zhiping Ye at FDA generously supplied viruses. SEH is funded through NIAID grants 1R01AI113047 and 1R01AI108686. NST is supported by NIAID's Centers of Excellence in Influenza Virus Research and Surveillance (HHSN27772201400008C).

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

Alymova, I.V., York, I.A., Air, G.M., Cipollo, J.F., Gulati, S., Baranovich, T., Kumar, A., Zeng, H., Gansebom, S., McCullers, J.A., 2016. Glycosylation changes in the globular head of H3N2 influenza hemagglutinin modulate receptor binding without affecting virus virulence. Sci Rep 6, 36216.

An, Y., McCullers, J.A., Alymova, I., Parsons, L.M., Cipollo, J.F., 2015. Glycosylation Analysis of Engineered H3N2 Influenza A Virus Hemagglutinins with Sequentially Added Historically Relevant Glycosylation Sites. Journal of proteome research 14, 3957-3969.

Aytay, S., Schulze, I.T., 1991. Single amino acid substitutions in the hemagglutinin can alter the host range and receptor binding properties of H1 strains of influenza A virus. J. Virol. 65, 3022-3028.

Bielejec, F., Baele, G., Vrancken, B., Suchard, M.A., Rambaut, A., Lemey, P., 2016. SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. Mol Biol Evol 33, 2167-2169.

Blackburne, B.P., Hay, A.J., Goldstein, R.A., 2008. Changing selective pressure during antigenic changes in human influenza H3. PLoS Pathog 4, e1000058.

CDC, 2017. Vaccine Effectiveness - How Well Does the Flu Vaccine Work?

Cobey, S., Hensley, S.E., 2017. Immune history and influenza virus susceptibility. Current opinion in virology 22, 105-111.

Couch, R.B., Kasel, J.A., 1983. Immunity to influenza in man. Annu Rev Microbiol 37, 529-549.

Daniels, R., Kurowski, B., Johnson, A.E., Hebert, D.N., 2003. N-linked glycans direct the cotranslational folding pathway of influenza hemagglutinin. Mol Cell 11, 79-90.

Das, S.R., Hensley, S.E., David, A., Schmidt, L., Gibbs, J.S., Puigbo, P., Ince, W.L., Bennink, J.R., Yewdell, J.W., 2011. Fitness costs limit influenza A virus hemagglutinin glycosylation as an immune evasion strategy. Proc Natl Acad Sci U S A 108, E1417-1422.

de Graaf, M., Fouchier, R.A., 2014. Role of receptor binding specificity in influenza A virus transmission and pathogenesis. EMBO J 33, 823-841.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29, 1969-1973.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460-2461.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Eggink, D., Goff, P.H., Palese, P., 2014. Guiding the Immune Response against Influenza Virus Hemagglutinin toward the Conserved Stalk Domain by Hyperglycosylation of the Globular Head Domain. Journal of Virology 88, 699-704.

Fonville, J.M., Wilks, S.H., James, S.L., Fox, A., Ventresca, M., Aban, M., Xue, L., Jones, T.C., Le, N.M.H., Pham, Q.T., Tran, N.D., Wong, Y., Mosterin, A., Katzelnick, L.C., Labonte, D., Le, T.T., van der Net, G., Skepner, E., Russell, C.A., Kaplan, T.D., Rimmelzwaan, G.F., Masurel, N., de Jong, J.C., Palache, A., Beyer, W.E.P., Le, Q.M., Nguyen, T.H., Wertheim, H.F.L., Hurt, A.C., Osterhaus, A., Barr, I.G., Fouchier, R.A.M., Horby, P.W., Smith, D.J., 2014. Antibody landscapes after influenza virus infection or vaccination. Science 346, 996-1000.

Gupta, R.B.S., 2004. Prediction of N-glycosylation sites in human proteins.

Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22, 160-174.

Hrincius, E.R., Liedmann, S., Finkelstein, D., Vogel, P., Gansebom, S., Samarasinghe, A.E., You, D., Cormier, S.A., McCullers, J.A., 2015. Acute Lung Injury Results from Innate Sensing of Viruses by an ER Stress Pathway. Cell Rep 11, 1591-1603.

Huang, K.Y., Rijal, P., Schimanski, L., Powell, T.J., Lin, T.Y., McCauley, J.W., Daniels, R.S., Townsend, A.R., 2015. Focused antibody response to influenza linked to antigenic drift. J Clin Invest 125, 2631-2645.

Job, E.R., Deng, Y.M., Barfod, K.K., Tate, M.D., Caldwell, N., Reddiex, S., Maurer-Stroh, S., Brooks, A.G., Reading, P.C., 2013. Addition of glycosylation to influenza A virus hemagglutinin modulates antibody-mediated recognition of H1N1 2009 pandemic viruses. J Immunol 190, 2169-2177.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30, 772-780.

Keil, W., Geyer, R., Dabrowski, J., Dabrowski, U., Niemann, H., Stirm, S., Klenk, H.D., 1985. Carbohydrates of influenza virus. Structural elucidation of the individual glycans of the FPV hemagglutinin by two-dimensional 1H n.m.r. and methylation analysis. EMBO J 4, 2711-2720.

Khatri, K., Klein, J.A., White, M.R., Grant, O.C., Leymarie, N., Woods, R.J., Hartshorn, K.L., Zaia, J., 2016. Integrated Omics and Computational Glycobiology Reveal Structural Basis for Influenza A Virus Glycan Microheterogeneity and Host Interactions. Mol Cell Proteomics 15, 1895-1912.

Lemey, P., Rambaut, A., Drummond, A.J., Suchard, M.A., 2009. Bayesian phylogeography finds its roots. PLoS Comput Biol 5, e1000520.

Linderman, S.L., Chambers, B.S., Zost, S.J., Parkhouse, K., Li, Y., Herrmann, C., Ellebedy, A.H., Carter, D.M., Andrews, S.F., Zheng, N.Y., Huang, M., Huang, Y.,

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Strauss, D., Shaz, B.H., Hodinka, R.L., Reyes-Teran, G., Ross, T.M., Wilson, P.C., Ahmed, R., Bloom, J.D., Hensley, S.E., 2014. Potential antigenic explanation for atypical H1N1 infections among middle-aged adults during the 2013-2014 influenza season. Proc Natl Acad Sci U S A 111, 15798-15803.

Medina, R.A., Stertz, S., Manicassamy, B., Zimmermann, P., Sun, X., Albrecht, R.A., Uusi-Kerttula, H., Zagordi, O., Belshe, R.B., Frey, S.E., Tumpey, T.M., Garcia-Sastre, A., 2013. Glycosylations in the globular head of the hemagglutinin protein modulate the virulence and antigenic properties of the H1N1 influenza viruses. Sci Transl Med 5, 187ra170.

Pagel, M., Meade, A., Barker, D., 2004. Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53, 673-684.

Park, S., Lee, I., Kim, J.I., Bae, J.Y., Yoo, K., Kim, J., Nam, M., Park, M., Yun, S.H., Cho, W.I., Kim, Y.S., Ko, Y.Y., Park, M.S., 2016. Effects of HA and NA glycosylation pattern changes on the transmission of avian influenza A(H7N9) virus in guinea pigs. Biochem Biophys Res Commun 479, 192-197.

Rambaut, A., Drummond, A., 2009. FigTree v1. 3.1.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7, 539.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688-2690.

Suchard, M.A., Rambaut, A., 2009. Many-core algorithms for statistical phylogenetics. Bioinformatics 25, 1370-1376.

Sun, H., Yang, J., Zhang, T., Long, L.-P., Jia, K., Yang, G., Webby, R.J., Wan, X.-F., 2013a. Using Sequence Data To Infer the Antigenicity of Influenza Virus. mBio 4, e00230-00213.

Sun, S., Wang, Q., Zhao, F., Chen, W., Li, Z., 2011. Glycosylation site alteration in the evolution of influenza A (H1N1) viruses. PLoS One 6, e22844.

Sun, X., Jayaraman, A., Maniprasad, P., Raman, R., Houser, K.V., Pappas, C., Zeng, H., Sasisekharan, R., Katz, J.M., Tumpey, T.M., 2013b. N-linked glycosylation of the hemagglutinin protein influences virulence and antigenicity of the 1918 pandemic and seasonal H1N1 influenza A viruses. J Virol 87, 8756-8766.

Tate, M.D., Job, E.R., Deng, Y.M., Gunalan, V., Maurer-Stroh, S., Reading, P.C., 2014. Playing hide and seek: how glycosylation of the influenza virus hemagglutinin can modulate the immune response to infection. Viruses 6, 1294-1316.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tsuchiya, E., Sugawara, K., Hongo, S., Matsuzaki, Y., Muraki, Y., Li, Z.N., Nakamura, K., 2001. Antigenic structure of the haemagglutinin of human influenza A/H2N2 virus. The Journal of general virology 82, 2475-2484.

Vigerust, D.J., Ulett, K.B., Boyd, K.L., Madsen, J., Hawgood, S., McCullers, J.A., 2007. N-linked glycosylation attenuates H3N2 influenza viruses. J Virol 81, 8593-8600.

WHO, 2003. Influenza Fact Sheet No. 211.

Yang, H., Carney, P.J., Chang, J.C., Guo, Z., Villanueva, J.M., Stevens, J., 2015. Structure and receptor binding preferences of recombinant human A(H3N2) virus hemagglutinins. Virology 477, 18-31.

Zhang, Y., Aevermann, B.D., Anderson, T.K., Burke, D.F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C.N., Lee, A.J., Li, X., Macken, C., Mahaffey, C., Pickett, B.E., Reardon, B., Smith, T., Stewart, L., Suloway, C., Sun, G., Tong, L., Vincent, A.L., Walters, B., Zaremba, S., Zhao, H., Zhou, L., Zmasek, C., Klem, E.B., Scheuermann, R.H., 2017. Influenza Research Database: An integrated bioinformatics resource for influenza virus research. Nucleic Acids Res 45, D466-D474.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Glycan Clock Forecasts Human Influenza A Virus Hemagglutinin Evolution

Supplemental Material

Material and methods

fig. S1-S7

Table S1

Supplemental files S1-S8

Materials and Methods

Virus preparation and purification

We propagated each virus by infecting 5, 10-day old embryonated chicken eggs with 100µL of viral stock diluted 1:2000 in PBS. Infected eggs were incubated at 34.5˚C and 65% humidity for ~41h. We first clarified the allantoic fluid (AF) by centrifuging 4500 x g for 20min, then pelleted the virus by centrifuging for 2h at 26,000 x g. After incubating the viral pellet in 1.7mL PBS overnight at 4˚C, we layered the virus onto a discontinuous 15%-60% sucrose gradient and spun at 34,000 x g for 2h. We collected virus at the gradient interface and spun it down a final time in PBS at 34,000 x g for 2h. Virus was resuspended in 100µL PBS and stored as milky opalescent suspensions long term at 4˚C. Hemagglutination titers for each virus were measured from neat AF. A complete list of viruses and their origin is found in **file S1**. Total protein content of each virus prep was determined by Coomassie blue-based assay performed per manufacturer instructions [Bio-Rad, DC Assay].

Allantoic fluid (AF) from eggs infected with 93 out of the 97 viruses we attempted to grow showed hemagglutination activity (HAU), indicating viral presence. We collected pure virus from all viruses in amounts ranging from 20-6920 µg. Sham infection and purification of AF from control eggs did not produce enough protein for detection, and ran cleanly on a gel (10 µL of a 100 µL prep, Lane #29, **fig. S1**).

To determine the identity of individual bands on the Coomassie gel, we grew H1N1/H3N2 reassortant viruses A/HK/68 (HK), A/PR8/MCa (PR8), X31 (HK HA and NA, PR8 background), J1 (HK HA, PR8 background), and P50 (PR8 HA, HK background). These viruses were run together on the same gel and immunoblotted with appropriate antibodies to identify which bands corresponded to which proteins on via Coomassie blue staining (**fig. S2**).

To determine the effect of S179N, K180Q, and I233T mutations on glycosylation of pH1N1 HA, we made HA plasmids with different mutation combinations in A/California/07/2009 HA background, which also included an additional D225G mutation to increase yield in eggs (**Fig. 5D**). Reassortant viruses with mutant HA and NA from A/California/07/2009 and internal segments from PR8 were rescued and grown in eggs. Virus was enriched from AF by centrifugation, standardized by ELISA using a stem reactive antibody (C179), and visualized by SDS-PAGE using the anti-HA CM-1 antibody. A portion of each virus was treated with PNGase F.

Protein gels and immunoblotting

We mixed purified virus (1 µg protein) with 4× NuPage loading buffer (Invitrogen), and boiled for 10 min at 96˚C. We electrophoresed samples with Chameleon Duo Li-Cor ladder on 4–12% Bis-Tris Gels (Invitrogen) at 200 V for 55 min. To visualize proteins, we fixed gels for 10 min with 10 ml 10% acetic acid and 50% methanol, shaking at RT. After removing fixative, we added 10 ml

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

GelCode stain (Pierce) and shook for 30 min at RT, then destained the gels with water overnight. For immunoblotting, we transferred proteins from gels to PVDF membranes with the iBLOT at P3 setting for 7 min. We blocked membranes for 30 min at room temperature (RT) StartingBlock (Thermo). After incubating with primary Ig or sera for 1 hr at RT, and washing 5x for 5 min each in TBST (10 mM Tris, 150 mM NaCl, 0.1% Tween-20), we added secondary Ig, repeating the washing step incubation. We imaged blots and Coomassie gels on a Li-Cor Odyssey. Band signals and molecular weights were calculated using Li-Cor's ImageStudio software. Glycoprotein molecular weights were normalized for gel distortion, commonly called "smiling," by adjusting each glycoprotein molecular weight by the average amount that the M1 and NP molecular weight for each virus varied from the mean molecular weight of M1 and NP proteins on each gel (**fig. S3**). Linear regression, Pearson coefficient, and P-values were calculated using PRISM software.

Two data points in **Fig. 1** are outliers. The first, A/Georgia/M5081/2012 (#28), an pH1N1 is predicted to have 1 head glycan, but migrates much slower than the other swine origin viruses. We made a new viral stock, and deep-sequenced it to confirm its identity. There were no additional predicted N-linked glycans. This HA alone in our panel of H1N1s (which includes another contemporary pH1N1 A/California/07/2009, #27), does not immunoblot with the anti-H1 HA monoclonal CM1 antibody even through the CM-1 peptide epitope used to generate the mAb is present. We suspect the conformation of this protein under SDS-PAGE conditions is somehow altered, changing the HA migration through the gel, and altering its interaction with CM-1. These interesting and perplexing findings will be pursued in future experiments. The second outlier, A/Victoria/361/2011 (#72) is the only H3N2 predicted to have 8 head glycans in our panel. As indicated in the text, H3N2s with more than 7 glycans are not common, indicating possible functional constraints on H3 HA. A possible constraint might be incomplete glycosylation of the virions.

Influenza sequences and N-glycosylation prediction

Human HA protein sequences were retrieved from the NIAID Influenza Research Database (IRD) (Zhang et al., 2017) through the web site at http://www.fludb.org (accessed April 20, 2017). Sequences with identical strain names were retained only if the majority of members had identical sequences. Sequences were aligned to A/California/04/2009 (H1N1) (FJ966082) with MAFFT v7.305b, and removed if indels existed in relation to the reference or if any ambiguities existed within the HA head. The amino acid numbering of sites important for this work with both H1 and H3 numbering conventions is shown in **table S1**. Potential glycosylation sites were identified by searching for the NX[S/T]Y motif where X/Y were any amino acid other than proline or by using NetNGlyc 1.0 artificial neural network-based prediction software (Gupta, 2004). The glycan frequencies used for Fig. 2-5 and fig. S5 are found in Supplementary files S3-8. Stacked glycan plots (Fig. 2A + 2B) are similar to a previous publication {Lee, 2014 #59}. For the quantitative analysis of SDS PAGE in **Fig. 1** and **fig. S4**, residues with positive NetNGlyc scores were considered glycosylated "N+", and residues with only one negative mark were recorded as "N-". Both conditions were treated as having a glycan present. Glycan predictions with "N--" were counted as negative.

pH1N1 phylogenetic analysis

Sequences glycosylated at N179 from 2009-2015 were aligned by ClustalΩ (Sievers et al., 2011). An unrooted phylogenetic tree of all 365, N179-glycosylated sequences was made using FigTree (Rambaut and Drummond, 2009). The earliest strains in the same clade as the SWSs were

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

identified as: A/Shiraz/4/2015, A/Shiraz/3/2015, A/Taipei/0021/2015, A/Adana/08/2015, A/Shiraz/6/2015.

To identify any additional mutations co-occurring with N179 we identified all sequences with >99% sequence similarity to the 6 earliest sweeper strains (SWSs) using USEARCH (max ~17nt differences) (Edgar, 2010). This represented 15% of all FluDB sequences. An unrooted phylogenetic tree of a downsampled set of 370 sequences representing pre-SWS and SWSs (as well as outgroup A/California/07/2009) was aligned in ClustalΩ is shown to the right. From this alignment, the two conserved amino acid changes that define SWSs, S179N and I233T were clearly seen.

To determine time and location of SWSs emergence, all pH1N1 in FluDB from 2009 to present day were collected and annotated with collection country, date, and glycosylation status. We also screened GISAID EpiFlu database for N179/I233T viruses. We down-sampled more than 16000 sequences to ~200 sequences per year (dataset A), taking into account location diversity and maximum temporal representation per year (total number of sequences = 1853). Dataset A was aligned using MAFFT v7.310 (Katoh and Standley, 2013). We inferred a phylogenetic tree using the maximum likelihood (ML) method available in the program RAxML v7.2.6 (Stamatakis, 2006) incorporating a general time-reversible (GTR) model of nucleotide substitution with a gamma-distributed (Γ) rate variation among sites. To assess the robustness of each node, a bootstrap resampling process was performed (1000 replicates), again using the ML method available in RAxML v7.2.6.

A smaller dataset B was assembled from the dominant clade with N179/I233T sequences to reconstruct its evolutionary and dispersal dynamics. Dataset B has 255 sequences from 2015 to 2017 with appropriate geographical and temporal representation.

Phylogenetic relationships were inferred using the time-scaled Bayesian approach using MCMC available via the BEAST v1.8.4 package (Drummond et al., 2012) and the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov). A relaxed molecular clock was used, with a constant population size, and a Hasegawa, Kishino and Yano (HKY) 85 model (Hasegawa et al., 1985) of nucleotide substitution with gamma-distributed rate variation among sites. For viruses for which only the year of viral collection was available, the lack of tip date precision was accommodated by sampling uniformly across a one-year window from January 1st to December 31st. The MCMC chain was run separately at least three times for each of the data sets and for at least 100 million iterations with sub-sampling every 10,000 iterations, using the BEAGLE library to improve computational performance (Suchard and Rambaut, 2009). All parameters reached convergence, as assessed visually using Tracer v.1.6, with statistical uncertainty reflected in values of the 95% highest posterior density (HPD). At least 10% of the chain was removed as burn-in, and runs for the same lineage and segment were combined using LogCombiner v1.8.4 and downsampled to generate a final posterior distribution of 500 trees that was used in the subsequent spatial analysis (Pagel et al., 2004).

The phylogeographic analysis considered 5 locations. Four Asian countries: China, Japan, South Korea and Taiwan; Two North American countries: Canada and the United States of America (USA); Two European countries: Czech Republic and Russia; Three Middle Eastern countries:

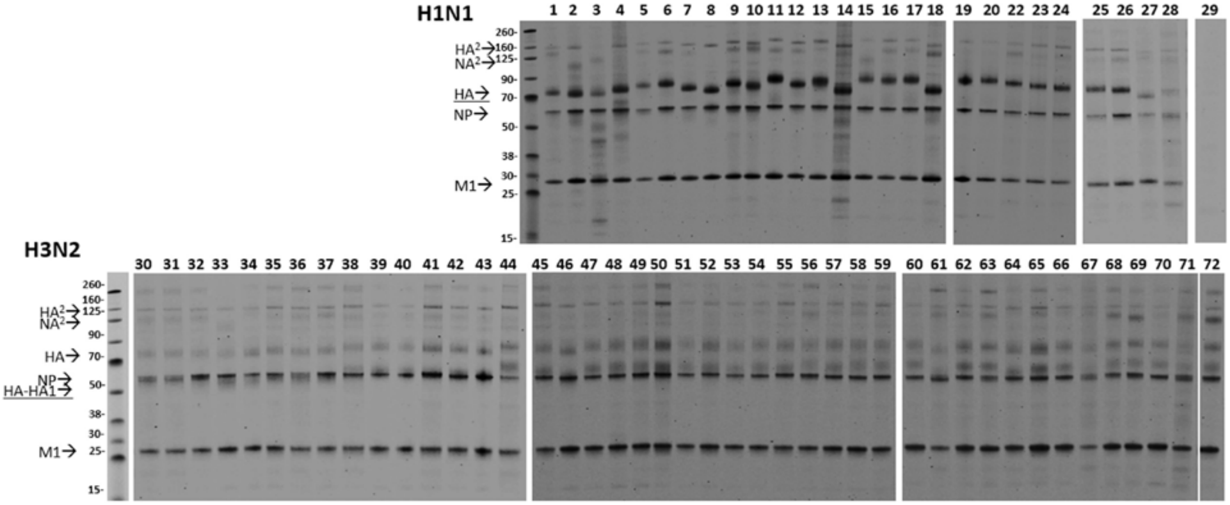
bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

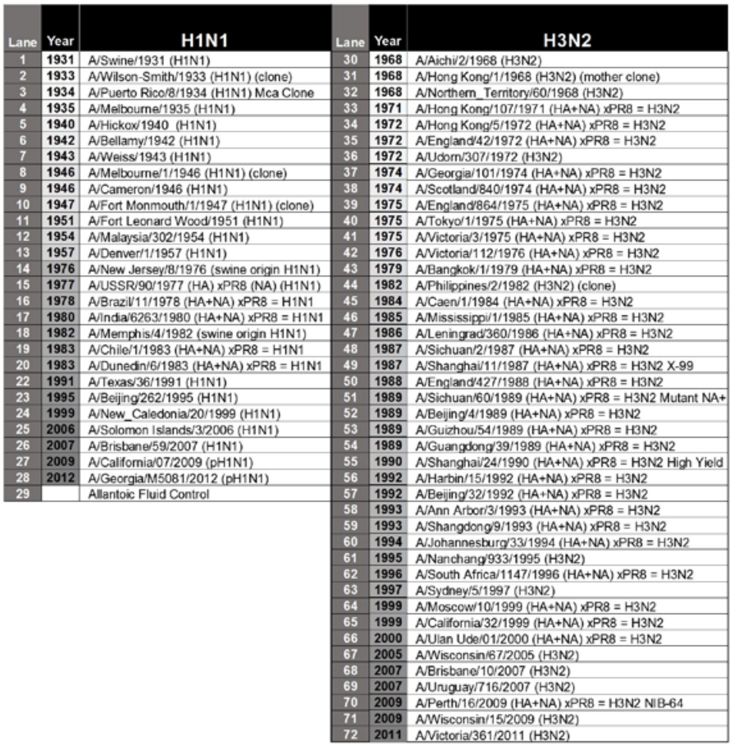
Iran, Saudi Arabia and Turkey; Two Latin American countries: Costa Rica and Mexico. A non-reversible discrete model was used to infer the rate of location transitions, along with Bayesian Stochastic Search Variable Selection (BSSVS) to identify those highly significant transitions while improving statistical efficiency (Lemey et al., 2009). For computational efficiency the phylogeographic analysis was run using an empirical distribution of 500 trees (Pagel et al., 2004), allowing the MCMC chain to be run for 50 million iterations, sampling every 5,000. Maximum clade credibility (MCC) trees were summarized using TreeAnnotator v1.8.4 and the trees were visualized in FigTree v1.4.3. The worldwide dispersal of N 179 - glycosylated viruses was visualized using SpreaD3 (Bielejec et al., 2016).

Additionally, we performed a sensitivity analyses by down-sampling all oversampled locations in dataset B to 45 sequences per location (dataset C), in order to address if the phylogegraphic estimates described above, were robust to any sampling bias.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. S1.





Lane	Year	H1N1	Lane	Year	H3N2
1	1931	A/Swine/1931 (H1N1)	30	1968	A/Aichi/2/1968 (H3N2)
2	1933	A/Wilson-Smith/1933 (H1N1) (clone)	31	1968	A/Hong Kong/1/1968 (H3N2) (mother clone)
3	1934	A/Puerto Rico/8/1934 (H1N1) Mca Clone	32	1968	A/Northern_Territory/60/1968 (H3N2)
4	1935	A/Melbourne/1935 (H1N1)	33	1971	A/Hong Kong/107/1971 (HA+NA) xPR8 = H3N2
5	1940	A/Hickox/1940 (H1N1)	34	1972	A/Hong Kong/5/1972 (HA+NA) xPR8 = H3N2
6	1942	A/Bellamy/1942 (H1N1)	35	1972	A/England/42/1972 (HA+NA) xPR8 = H3N2
7	1943	A/Weiss/1943 (H1N1)	36	1972	A/Udorn/307/1972 (H3N2)
8	1946	A/Melbourne/1/1946 (H1N1) (clone)	37	1974	A/Georgia/101/1974 (HA+NA) xPR8 = H3N2
9	1946	A/Cameron/1946 (H1N1)	38	1974	A/Scotland/840/1974 (HA+NA) xPR8 = H3N2
10	1947	A/Fort Monmouth/1/1947 (H1N1) (clone)	39	1975	A/England/864/1975 (HA+NA) xPR8 = H3N2
11	1951	A/Fort Leonard Wood/1951 (H1N1)	40	1975	A/Tokyo/1/1975 (HA+NA) xPR8 = H3N2
12	1954	A/Malaysia/302/1954 (H1N1)	41	1975	A/Victoria/3/1975 (HA+NA) xPR8 = H3N2
13	1957	A/Denver/1/1957 (H1N1)	42	1976	A/Victoria/112/1976 (HA+NA) xPR8 = H3N2
14	1976	A/New Jersey/8/1976 (swine origin H1N1)	43	1979	A/Bangkok/1/1979 (HA+NA) xPR8 = H3N2
15	1977	A/USSR/90/1977 (HA) xPR8 (NA) (H1N1)	44	1982	A/Philippines/2/1982 (H3N2) (clone)
16	1978	A/Brazil/11/1978 (HA+NA) xPR8 = H1N1	45	1984	A/Caen/1/1984 (HA+NA) xPR8 = H3N2
17	1980	A/India/6263/1980 (HA+NA) xPR8 = H1N1	46	1985	A/Mississippi/1/1985 (HA+NA) xPR8 = H3N2
18	1982	A/Memphis/4/1982 (swine origin H1N1)	47	1986	A/Leningrad/360/1986 (HA+NA) xPR8 = H3N2
19	1983	A/Chile/1/1983 (HA+NA) xPR8 = H1N1	48	1987	A/Sichuan/2/1987 (HA+NA) xPR8 = H3N2
20	1983	A/Dunedin/6/1983 (HA+NA) xPR8 = H1N1	49	1987	A/Shanghai/11/1987 (HA+NA) xPR8 = H3N2 X-99
22	1991	A/Texas/36/1991 (H1N1)	50	1988	A/England/427/1988 (HA+NA) xPR8 = H3N2
23	1995	A/Beijing/262/1995 (H1N1)	51	1989	A/Sichuan/60/1989 (HA+NA) xPR8 = H3N2 Mutant NA+
24	1999	A/New_Caledonia/20/1999 (H1N1)	52	1989	A/Beijing/4/1989 (HA+NA) xPR8 = H3N2
25	2006	A/Solomon Islands/3/2006 (H1N1)	53	1989	A/Guizhou/54/1989 (HA+NA) xPR8 = H3N2
26	2007	A/Brisbane/59/2007 (H1N1)	54	1989	A/Guangdong/39/1989 (HA+NA) xPR8 = H3N2
27	2009	A/California/07/2009 (pH1N1)	55	1990	A/Shanghai/24/1990 (HA+NA) xPR8 = H3N2 High Yield
28	2012	A/Georgia/M5081/2012 (pH1N1)	56	1992	A/Harbin/15/1992 (HA+NA) xPR8 = H3N2
29		Allantoic Fluid Control	57	1992	A/Beijing/32/1992 (HA+NA) xPR8 = H3N2
			58	1993	A/Ann Arbor/3/1993 (HA+NA) xPR8 = H3N2
			59	1993	A/Shangdong/9/1993 (HA+NA) xPR8 = H3N2
			60	1994	A/Johannesburg/33/1994 (HA+NA) xPR8 = H3N2
			61	1995	A/Nanchang/933/1995 (H3N2)
			62	1996	A/South Africa/1147/1996 (HA+NA) xPR8 = H3N2
			63	1997	A/Sydney/5/1997 (H3N2)
			64	1999	A/Moscow/10/1999 (HA+NA) xPR8 = H3N2
			65	1999	A/California/32/1999 (HA+NA) xPR8 = H3N2
			66	2000	A/Ulan Ude/01/2000 (HA+NA) xPR8 = H3N2
			67	2005	A/Wisconsin/67/2005 (H3N2)
			68	2007	A/Brisbane/10/2007 (H3N2)
			69	2007	A/Uruguay/716/2007 (H3N2)
			70	2009	A/Perth/16/2009 (HA+NA) xPR8 = H3N2 NIB-64
			71	2009	A/Wisconsin/15/2009 (H3N2)
			72	2011	A/Victoria/361/2011 (H3N2)

Fig. S1. SDS-PAGE of purified virus. Molecular weight standards are labelled. Representative blots are shown. Lane identifications with reassortant status are shown below blots. Source of strain, yield and hemagglutinin units are shown for each lane in file S1.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. S2.

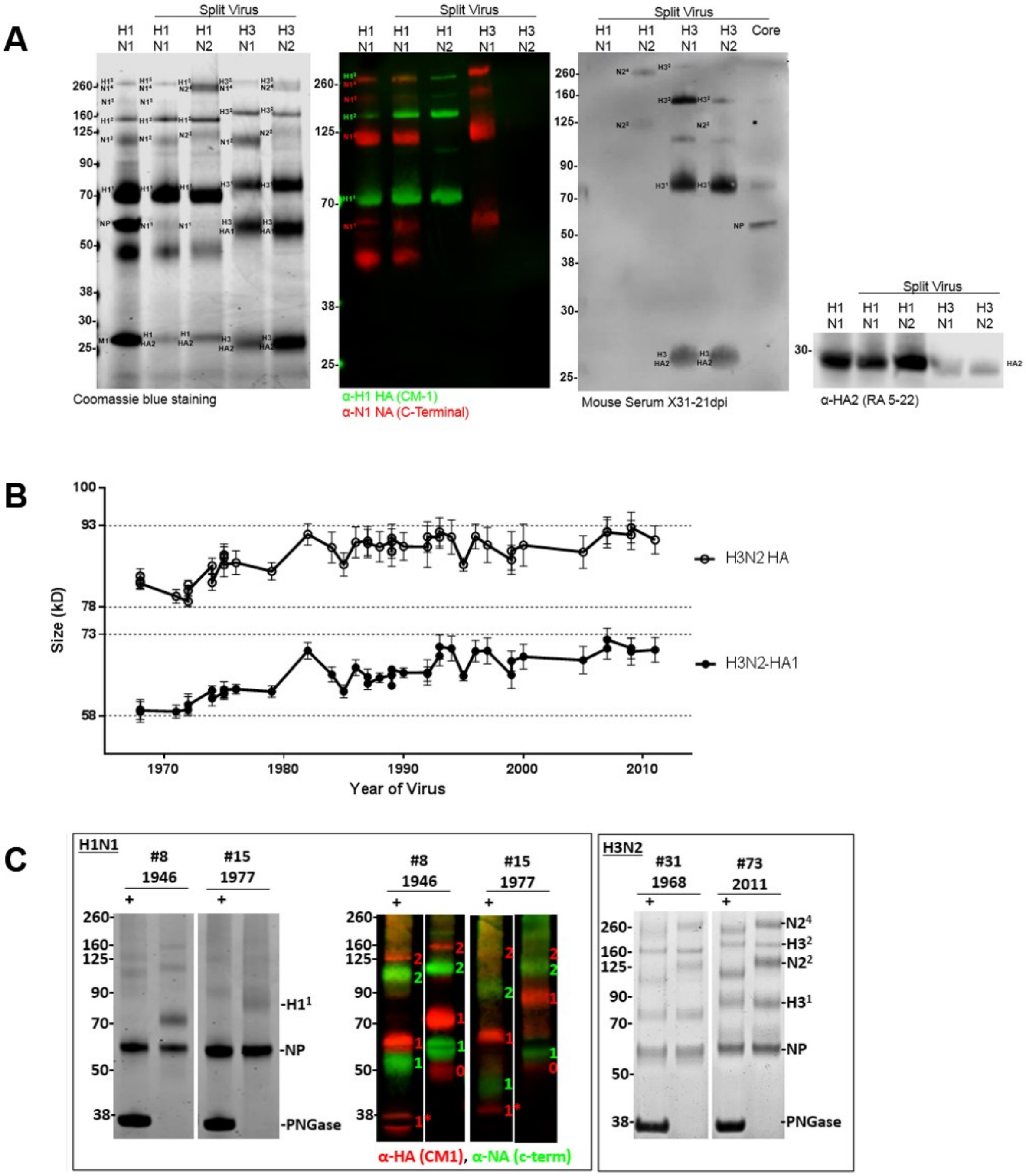


Fig. S2. Identification of viral proteins in Coomassie blue stained gels. **(A)** Purified whole PR8 or reassortant, detergent split viruses were stained with Coomassie blue; probed with monoclonal antibodies against HA1 (CM1), NA N1 (C-terminal), and HA2 (RA-5 22); or convalescent sera from a H3N2 reassortant infected mouse. **(B)** H3N2 HA monomer and H3N2-HA1 band sizes correlate almost perfectly. As the resolution was better for H3N2-HA1, these were used for analysis. **(C)** PNGase treatment of viruses shows deglycosylation of glycoproteins for representative viruses. Coomassie stained gels and monoclonal antibodies against H1 (CM1) and N1 (C-terminal) for H1N1.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. S3.

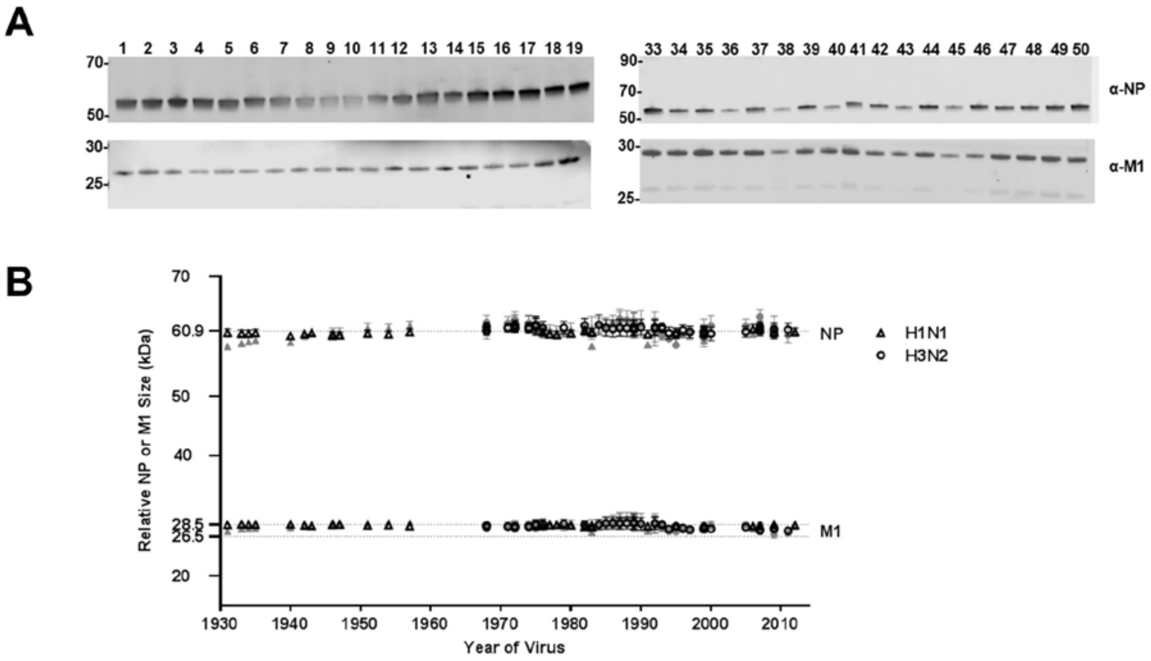
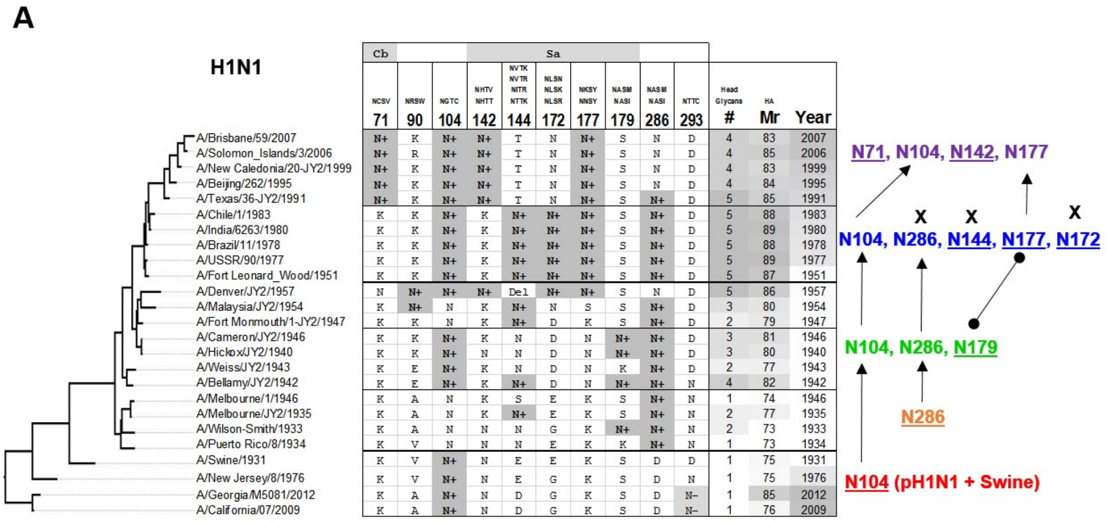


Fig. S3. NP and M1 proteins serve as internal ruler. (A) Western blots of NP and M1 for representative viruses. **(B)** Quantitation of NP and M1 relative size from SDS-PAGE. Raw data (grey) was normalized to account for gel distortion (open black symbols). Average size of each protein is shown on the Y-axis. Error bars are standard error from n = 3-6 blots.

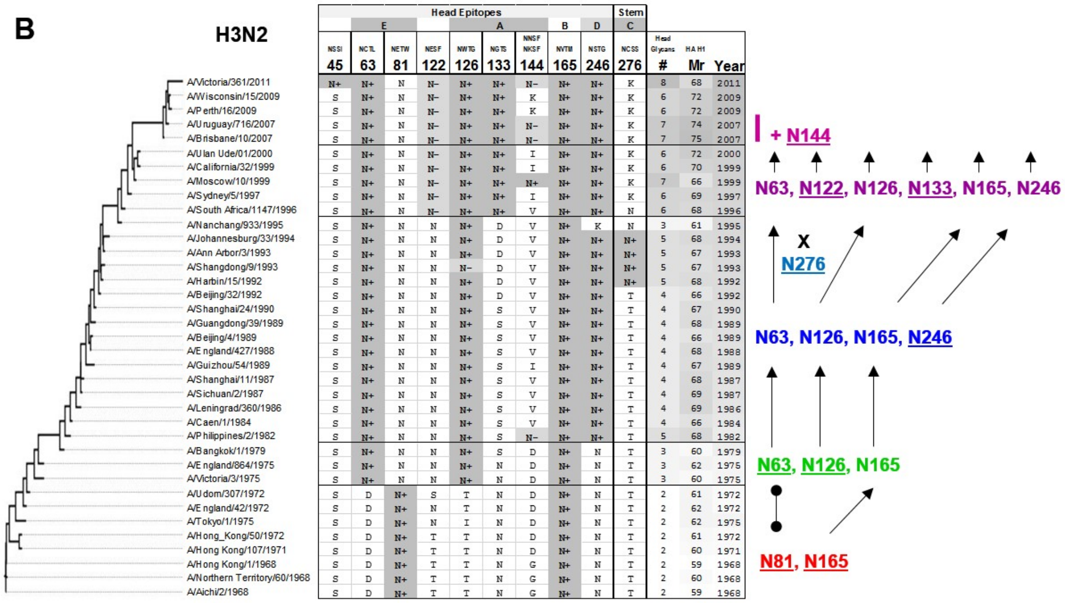
bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. S4.



	Cb			Sa									
	NCSV 71	NRSW 90	NGTC 104	NHTV NHTT 142	NVTK NVTR NITR NTTK 144	NLSN NLSK NLSR 172	NKSY NNSY 177	NASM NASI 179	NASM NASI 286	NTTC 293	Head Glycans #	HA Mr	Year
A/Brisbane/59/2007	N+	K	N+	N+	T	N	N+	S	N	D	4	83	2007
A/Solomon_Islands/3/2006	N+	R	N+	N+	T	N	N+	S	N	D	4	85	2006
A/New Caledonia/20-JY2/1999	N+	K	N+	N+	T	N	N+	S	N	D	4	83	1999
A/Beijing/262/1995	N+	K	N+	N+	T	N	N+	S	N	D	4	84	1995
A/Texas/36-JY2/1991	N+	K	N+	N+	T	N	N+	S	N+	D	5	85	1991
A/Chile/1/1983	K	K	N+	K	N+	N+	N+	S	N+	D	5	88	1983
A/India/6263/1980	K	K	N+	K	N+	N+	N+	S	N+	D	5	89	1980
A/Brazil/11/1978	K	K	N+	K	N+	N+	N+	S	N+	D	5	88	1978
A/USSR/90/1977	K	K	N+	K	N+	N+	N+	S	N+	D	5	89	1977
A/Fort Leonard_Wood/1951	K	K	N+	K	N+	N+	N+	S	N+	D	5	87	1951
A/Denver/JY2/1957	N	N+	N+	N+	Del	N+	N+	S	N	D	5	86	1957
A/Malaysia/JY2/1954	K	N+	N	K	N+	N	S	S	N+	D	3	80	1954
A/Fort Monmouth/1-JY2/1947	K	K	N	K	N+	D	K	S	N+	D	2	79	1947
A/Cameron/JY2/1946	K	K	N+	K	N	D	N	N+	N+	D	3	81	1946
A/Hickox/JY2/1940	K	K	N+	N	N	D	N	N+	N+	D	3	80	1940
A/Weiss/JY2/1943	K	E	N+	K	N	D	N	K	N+	D	2	77	1943
A/Bellamy/JY2/1942	K	E	N+	K	N+	D	N	N+	N+	N	4	82	1942
A/Melbourne/1/1946	K	A	N	K	S	E	K	S	N+	N	1	74	1946
A/Melbourne/JY2/1935	K	A	N	K	N+	E	K	S	N+	N	2	77	1935
A/Wilson-Smith/1933	K	A	N	N	N	G	K	N+	N+	N	2	73	1933
A/Puerto Rico/8/1934	K	V	N	N	N	E	K	K	N+	N	1	73	1934
A/Swine/1931	K	V	N+	N	E	E	K	S	D	D	1	75	1931
A/New Jersey/8/1976	K	V	N+	N	E	G	K	S	D	N	1	75	1976
A/Georgia/M5081/2012	K	A	N+	N	D	G	K	S	D	N–	1	85	2012
A/California/07/2009	K	A	N+	N	D	G	K	S	D	N–	1	76	2009

*Plus conserved N27, N28, N40, N304, N498, N557



	Head Epitopes									Stem			
		E			A			B	D	C			
	NSSI 45	NCTL 63	NETW 81	NESF 122	NWTG 126	NGTS 133	NNSF NKSF 144	NVTM 165	NSTG 246	NCSS 276	Head Glycans #	HA H1 Mr	Year
A/Victoria/361/2011	N+	N+	N	N–	N+	N+	N–	N+	N+	K	8	68	2011
A/Wisconsin/15/2009	S	N+	N	N–	N+	N+	K	N+	N+	K	6	72	2009
A/Perth/16/2009	S	N+	N	N–	N+	N+	K	N+	N+	K	6	72	2009
A/Uruguay/716/2007	S	N+	N	N–	N+	N+	N–	N+	N+	K	7	74	2007
A/Brisbane/10/2007	S	N+	N	N–	N+	N+	N–	N+	N+	K	7	75	2007
A/Ulan Ude/01/2000	S	N+	N	N–	N+	N+	I	N+	N+	K	6	72	2000
A/California/32/1999	S	N+	N	N–	N+	N+	I	N+	N+	K	6	70	1999
A/Moscow/10/1999	S	N+	N	N–	N+	N+	N+	N+	N+	K	7	66	1999
A/Sydney/5/1997	S	N+	N	N–	N+	N+	I	N+	N+	K	6	69	1997
A/South Africa/1147/1996	S	N+	N	N–	N+	N+	V	N+	N+	N	6	68	1996
A/Nanchang/933/1995	S	N+	N	N	N+	D	V	N+	K	N	3	61	1995
A/Johannesburg/33/1994	S	N+	N	N	N+	D	V	N+	N+	N+	5	68	1994
A/Ann Arbor/3/1993	S	N+	N	N	N+	D	V	N+	N+	N+	5	67	1993
A/Shangdong/9/1993	S	N+	N	N	N–	D	V	N+	N+	N+	5	67	1993
A/Harbin/15/1992	S	N+	N	N	N+	D	V	N+	N+	N+	5	68	1992
A/Beijing/32/1992	S	N+	N	N	N+	D	V	N+	N+	T	4	66	1992
A/Shanghai/24/1990	S	N+	N	N	N+	S	V	N+	N+	T	4	67	1990
A/Guangdong/39/1989	S	N+	N	N	N+	S	V	N+	N+	T	4	68	1989
A/Beijing/4/1989	S	N+	N	N	N+	S	V	N+	N+	T	4	66	1989
A/England/427/1988	S	N+	N	N	N+	S	V	N+	N+	T	4	68	1988
A/Guizhou/54/1989	S	N+	N	N	N+	S	I	N+	N+	T	4	67	1989
A/Shanghai/11/1987	S	N+	N	N	N+	S	V	N+	N+	T	4	68	1987
A/Sichuan/2/1987	S	N+	N	N	N+	S	V	N+	N+	T	4	69	1987
A/Leningrad/360/1986	S	N+	N	N	N+	S	V	N+	N+	T	4	69	1986
A/Caen/1/1984	S	N+	N	N	N+	S	V	N+	N+	T	4	66	1984
A/Philippines/2/1982	S	N+	N	N	N+	S	N–	N+	N+	T	5	68	1982
A/Bangkok/1/1979	S	N+	N	N	N+	S	D	N+	N	T	3	60	1979
A/England/864/1975	S	N+	N	N	N+	N	D	N+	N	T	3	62	1975
A/Victoria/3/1975	S	N+	N	N	N+	N	D	N+	N	T	3	60	1975
A/Udom/307/1972	S	D	N+	S	T	N	D	N+	N	T	2	61	1972
A/England/42/1972	S	D	N+	N	T	N	D	N+	N	T	2	62	1972
A/Tokyo/1/1975	S	D	N+	N	I	N	D	N+	N	T	2	62	1975
A/Hong_Kong/50/1972	S	D	N+	T	T	N	D	N+	N	T	2	61	1972
A/Hong Kong/107/1971	S	D	N+	T	T	N	D	N+	N	T	2	60	1971
A/Hong Kong/1/1968	S	D	N+	T	T	N	G	N+	N	T	2	59	1968
A/Northern Territory/60/1968	S	D	N+	T	T	N	G	N+	N	T	2	60	1968
A/Aichi/2/1968	S	D	N+	T	T	N	G	N+	N	T	2	59	1968

*Plus conserved N8 (except Hong Kong/50/1972 + Hong Kong/107/1971), N22, N38, N285, N483

Fig. S4. Glycan groups of IAV viral panel. (A) H1N1 or **(B)** H3N2 viruses and NetNGlyc predicted glycosylation status of HA Asn (N+ or N-), the substituted residue present, or a deletion (Del). The total number of head glycans, relative HA size as migration rate through gel (Mr), and year of virus collection is shown for each virus. The glycosylation progression is shown color coded to match individual viruses in **Fig. 1B**. New glycosylation sites are underlined. Arrows connect conserved glycosylation sites and barbells connect glycans in the same epitope, but not same site. Lost glycosylation sites are noted with an X. Conserved glycans among all viruses are listed below the table.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. S5.

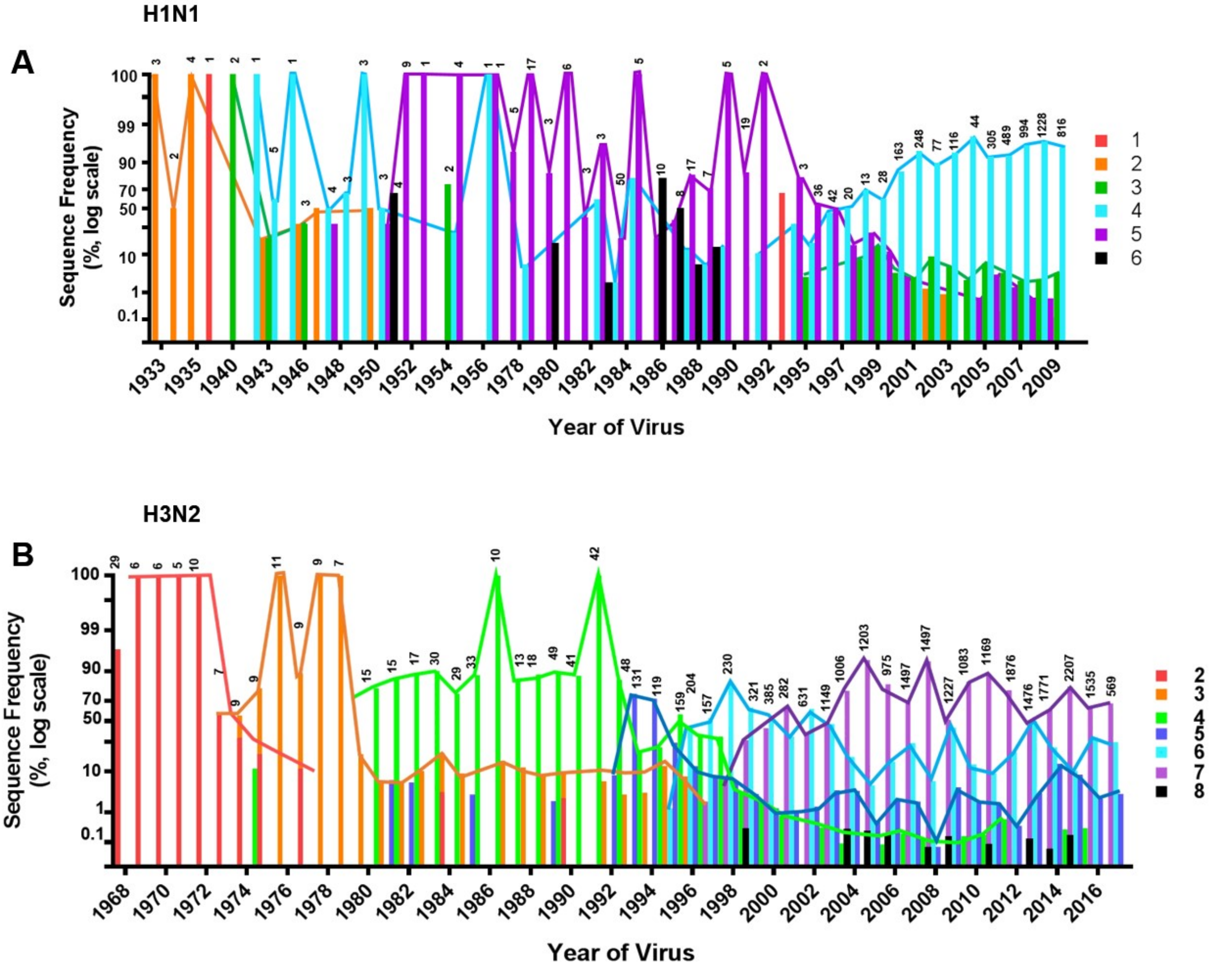


Fig. S5. Frequency of total sequences by number of glycosylation sites. (A) Percentage of H1N1 or **(B)** H3N2 sequences in FluDB with indicated number of predicted head glycans. Above each lane are total number of sequences in FluDB from that year. Y-Axis is log scale. Sequences above the glycan limit 6 for H1N1 and 8 for H3N2 are shown in black.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. S6.

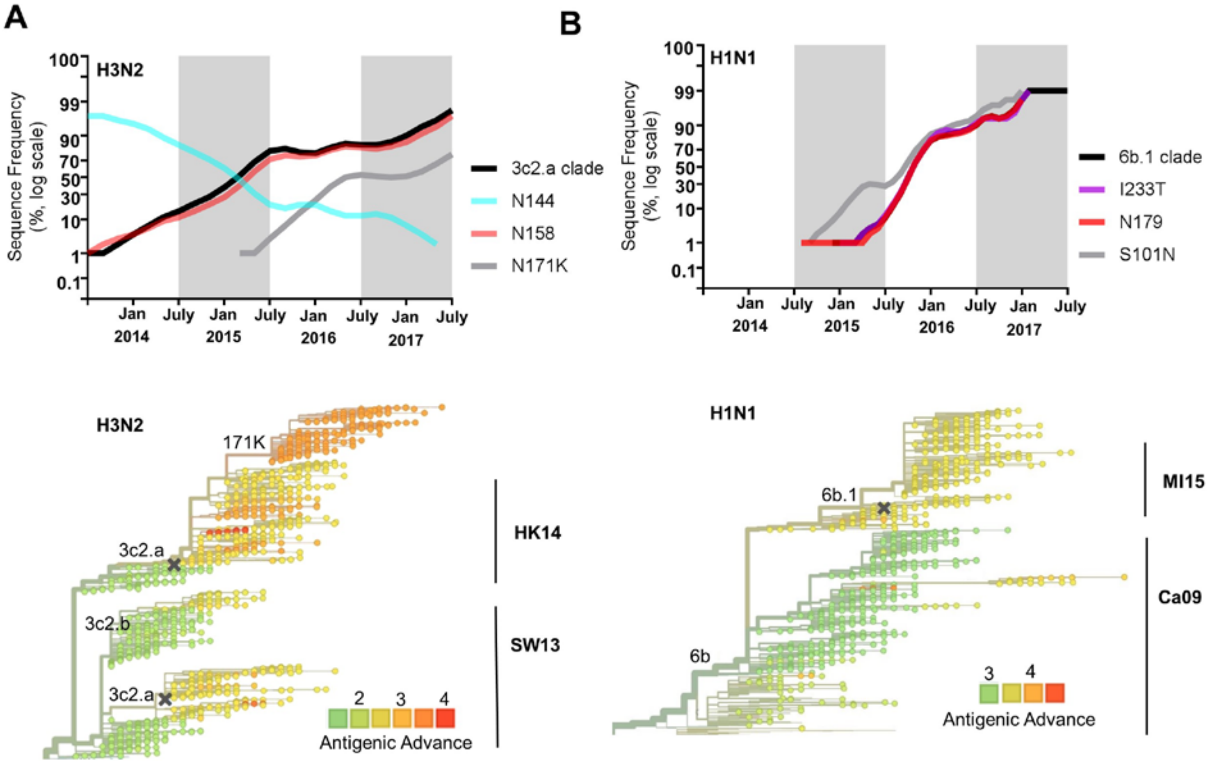


Fig. S6. Major IAV clades track with glycan evolution. Using nextflu.org we tracked the most recent glycan changes for **(A)** H1N1 and **(B)** H3N2 viruses in terms of frequency of the major clade (black) and other mutations of interest. On bottom, phylogenic trees are colored according to antigenic advance measured as $\log_2$ HI titer distance from the root of the tree. Vaccine strains are marked with an X. Antigenic clusters are labelled to the right.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. S7

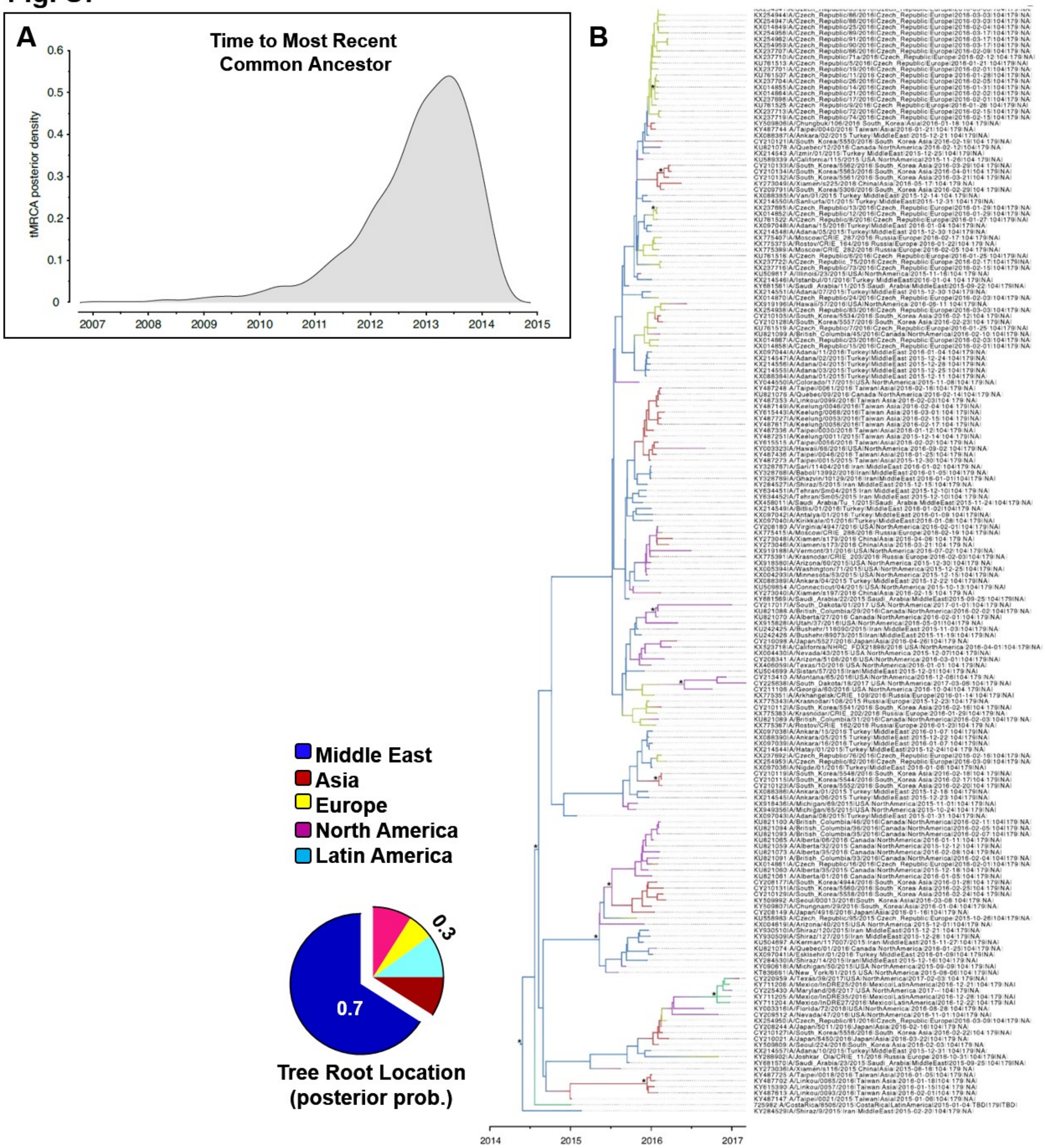


Fig. S7. Sensitivity analyses for the evolutionary relationships between N179+ HA sequences. (A) Posterior density for the time to Most Recent Common Ancestor (tMRCA). Origin is estimated to have occurred in March 19th, 2013 (95% Highest Posterior Density: May 1st, 2011 – September 28th, 2014). **(**B) Time-scaled Bayesian MCC tree inferred for 184 N179/I233T HA sequences collected worldwide during 2015 – 2017 (dataset C). The color of each branch indicates the most probable location state. Asterisks indicate nodes with posterior probabilities higher than 0.90. Pie Graph depicts the posterior probabilities for locations at the root of the phylogeny. Middle East is estimated as the location where N179+ viruses emerged with a posterior probability of 0.66.

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S1. Relevant amino acid residues with H1 and H3 numbering. Calculated using the "HA subtype numbering conversion tool (beta)" on FluDB. Numbering in our text starts from the initial HA methionine (M=1).

In Paper	M=1	H1	H3
N27	27	10	20
N28	28	11	21
N40	40	23	33
N71	71	54	63
N90	90	73	82
S101N	101	84	92
N104	104	87	94
N136	136	119	125.1
N142	142	125	129
N144	144	127	131
N172	172	155	158
N177	177	160	163
N179	179	162	165
K180Q	180	163	166
I233T	233	216	219
N276	276	259	261
C277	277	275	292
N286	286	269	271
N304	304	287	289
N498	498	481	483
N557	557	HA2 213	541

bioRxiv preprint doi: https://doi.org/10.1101/163345; this version posted August 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Files

File S1. Viruses used for SDS PAGE. Viral source, yield and hemagglutinin units for each strain for in Fig. S1.

File S2. Maximum likelihood tree of dataset A. The phylogenetic tree was inferred using the maximum likelihood method for the dataset A sequences, entailing 1853 H1 sequences from around the globe. The tree is midpoint rooted, all branch lengths are drawn to scale, and bootstrap values >70 are provided for key nodes. N179+ strains are shaded blue.

File S3. Frequency of pH1N1 sequences by glycan count.

File S4. Frequency of pH1N1 sequences by glycan residue.

File S5. Frequency of sH1N1 sequences by glycan count.

File S6. Frequency of sH1N1 sequences by glycan residue.

File S7. Frequency of H3N2 sequences by glycan count.

File S8. Frequency of H3N2 sequences by glycan residue.