

# **Phylogenetics for Predicting Virus Evolution**

---

**Anticipating next seasons influenza strains**

**Paul Wiesemeyer**

## Problem



HOUSTON,  
we have a — CHOO!

- Estimated 3 – 11% of the population catch symptomatic influenza each season [Tokars et al., 2018]
- Influenza evolves rapidly and evades immunity and vaccines
- Epidemics and pandemics caused by RNA recombination events
- Science is slow responding to health issues

# Problem

- $\sim 8$  months from submission to publication of a medical paper [AAMC, 2018]
- $\sim 6$  months from vaccine strain selection to distribution:



Figure 1: [Bedford, 2015]

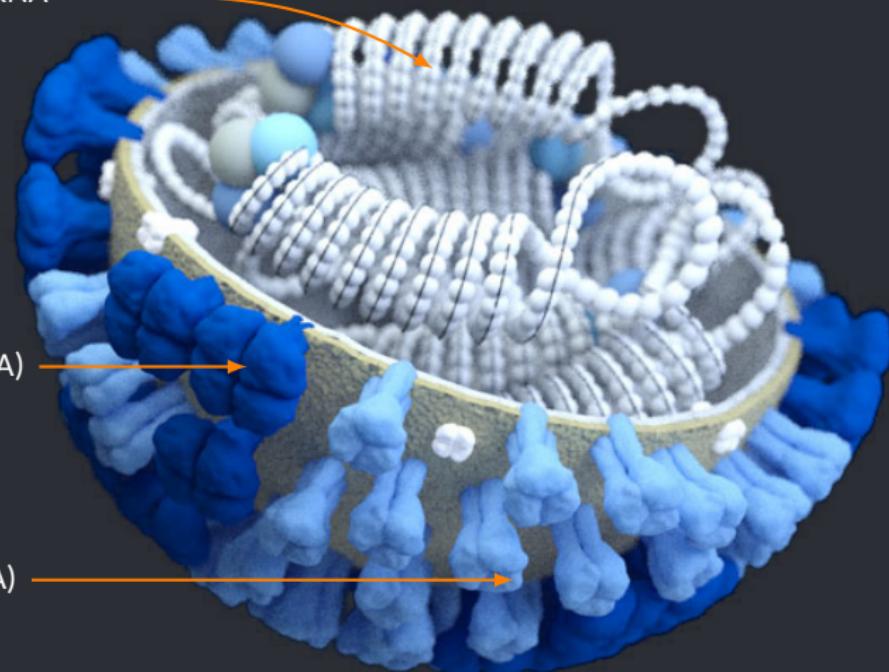
# Outline

1. Influenza
2. Phylogenetics
  - 2.1 The Molecular Clock
  - 2.2 Sequence Alignment
3. Predicting the next strain of influenza
  - 3.1 Predicting Virus Evolution
  - 3.2 The Hemagglutination Inhibition Assay
  - 3.3 Results
4. *Nextstrain*
  - 4.1 How to use the Framework
  - 4.2 The Powerful Meta Data
  - 4.3 Confidence Levels and Limitations

# Influenza

*a closer look*

8 single stranded RNA



Neuraminidase (NA)

Hemagglutinin (HA)

# Influenza

## *the recent history*

recent in sense of "blink of an eye" on evo scales

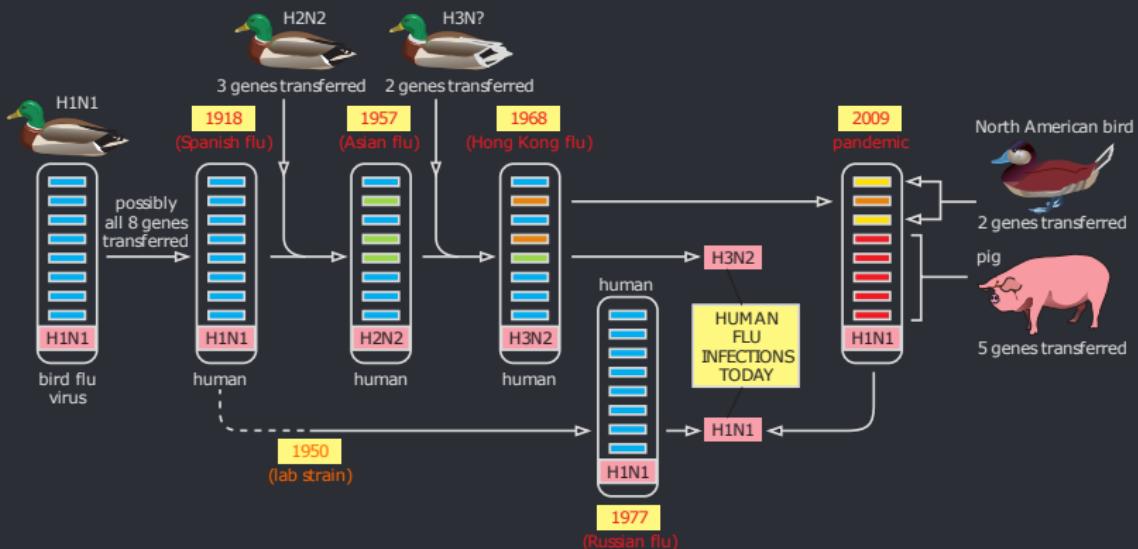


Figure 2: taken from [Alberts, 2015]

# Influenza

*the phylogeny*

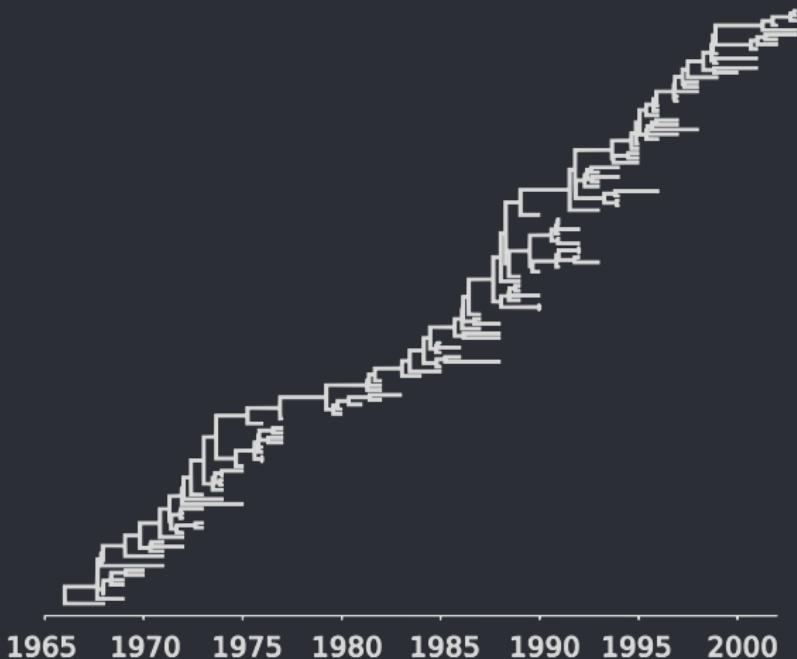


Figure 3:

# Influenza

*A virus that does not change escape vehicle will get caught*



Figure 4: It's like getting more and more stars, having to switch escape vehicle constantly

# Phylogenetics

*An old idea*

Haeckel, Darwin

Tree is simply connected graph, has parents and children, can be rooted/unrooted

clade means just any collection of branches, usually monophyletic:  
all children included

"leaves"

# Phylogenetics

*Concept: Parsimony*

Plot of tree

Ockhams razor

Cost function, sth. to minimize

# Phylogenetics

## Concept: The Molecular Clock

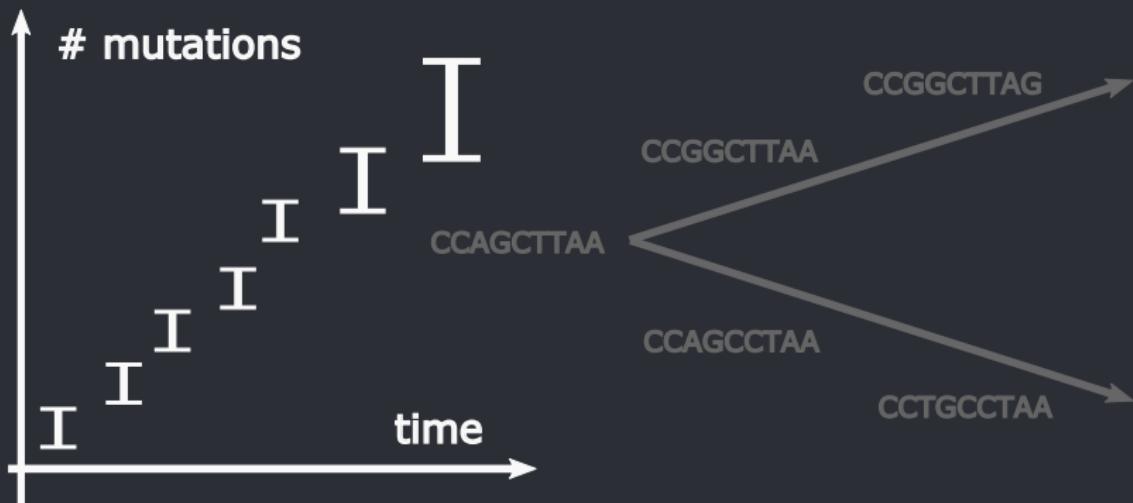


Figure 5: Linear time-mutation relationship

# Phylogenetics

Concept: The Molecular Clock

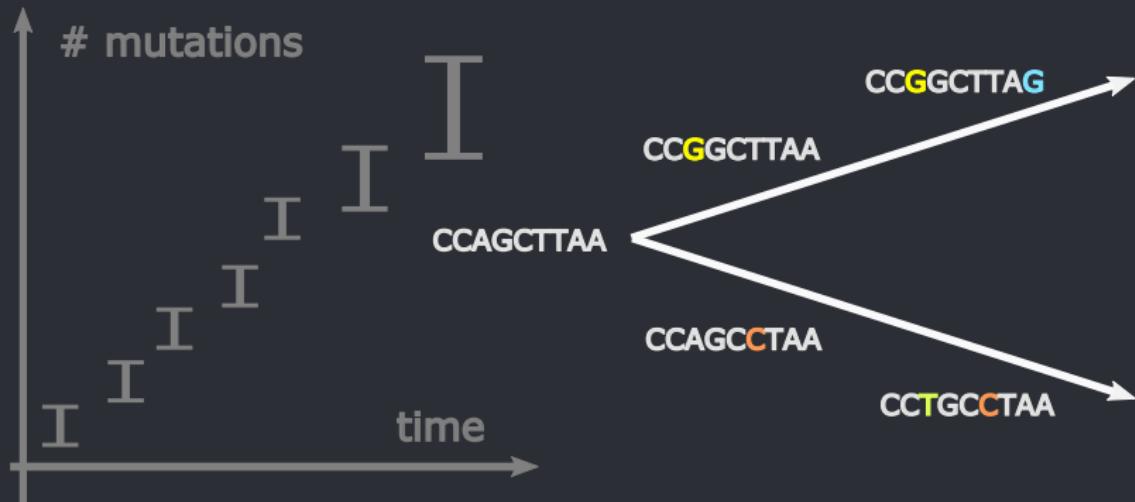


Figure 6: Linear time-mutation relationship

Corroborated by genetic equidistance.

Limited by complete turnover time. (!?) synonymous/non-synonymous, point-mut.  
indels

# Phylogenetics

## *Example algorithm: Sequence Alignment*

N sequences of lengths  $n_1 \dots n_N$  (not same due to indels, seq. errors)

Pairwise matching, give a score, make N x N table

Best matches go together, form new "sequence"

Iterate. When comparing to sets of sequences, use mean.

# Predicting the next strain of influenza

## *Approaches*

Epidemiological: Is there a geographical region? PLOT/Picture?

Genealogical: Is there a certain type of Mutation? KOEL 7

Immunological: Hemagglutinin inhibition cartography

(Denote the scales here: Molecule/Cell/Population and that they couple)

Vision: Bring these levels together

# The Hemagglutination Inhibition Assay

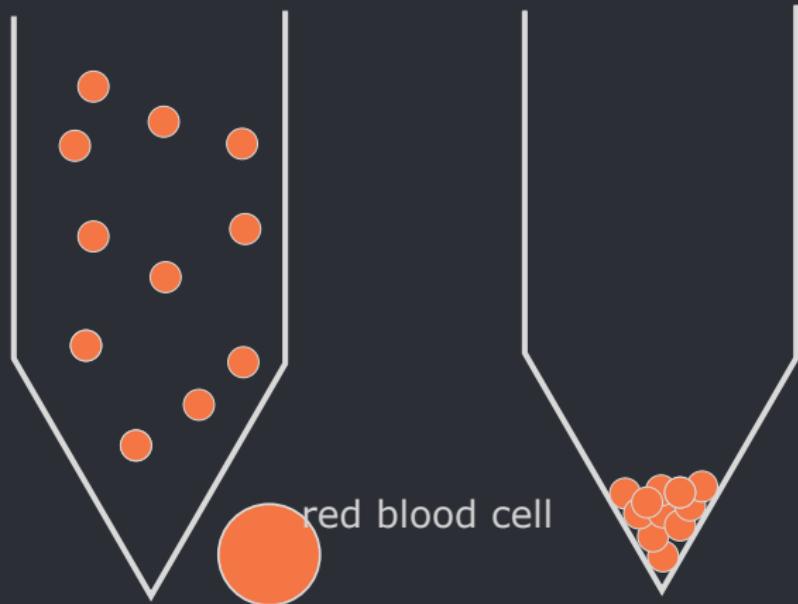


Figure 7: Red blood cells (RBC) precipitate.

## The Hemagglutination Inhibition Assay

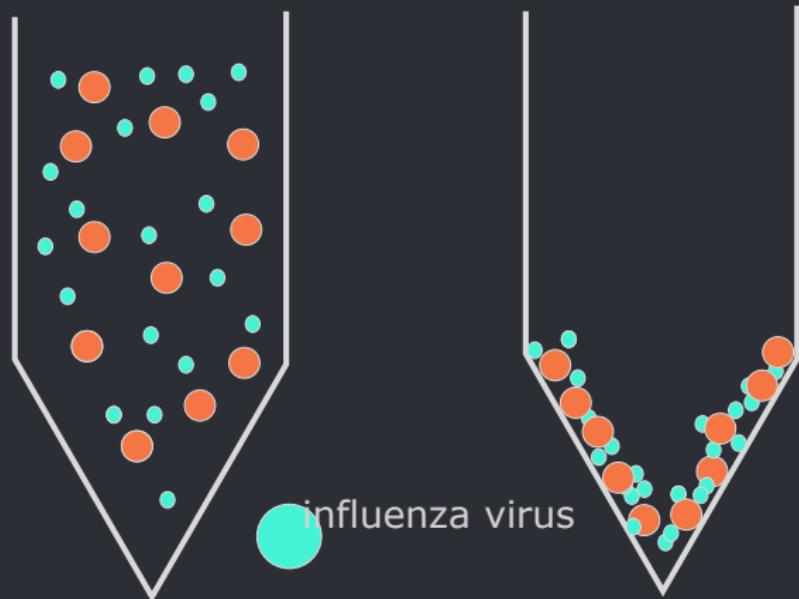


Figure 8: Influenza Hemagglutinin (HA) coagulates the RBC, forming a mat.

## The Hemagglutination Inhibition Assay

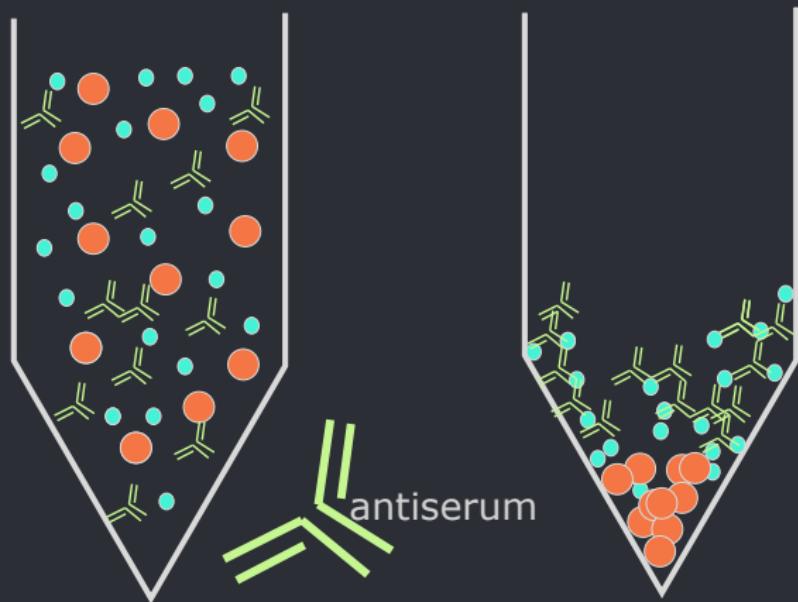


Figure 9: Antisera that fit the HA's epitope site bind it, letting the RBC sink to the bottom. Effect works up to a certain antigenic distance and .

# The Hemagglutination Inhibition Assay

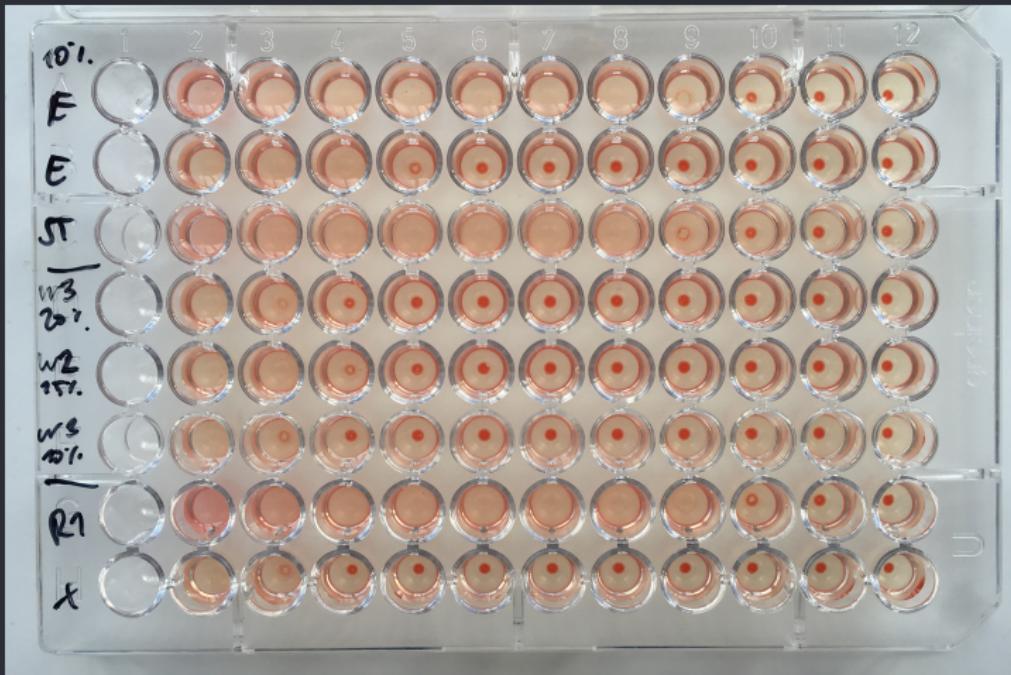
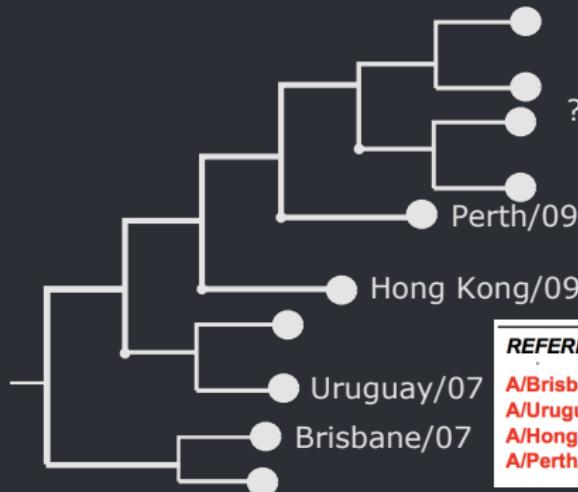


Figure 10: Here, one antiserum is tested in 12 different dilutions against 8 different virus strains. The highest dilution that prevents agglutination is called the titer.

# The Hemagglutination Inhibition Assay

*How this used to be looked at*

# The Hemagglutination Inhibition Assay



REFERENCE VIRUSES	A/Bris	A/Uru	A/HK	A/Perth
A/Brisbane/10/2007	2560	2560	80	<
A/Uruguay/7/16/2007	1280	2560	<	<
A/Hong Kong/1985/2009	80	160	1280	640
A/Perth/16/2009	<	40	640	640

Figure 11: Mapping the chart to the tree constructed from sequences

# The Hemagglutination Inhibition Assay

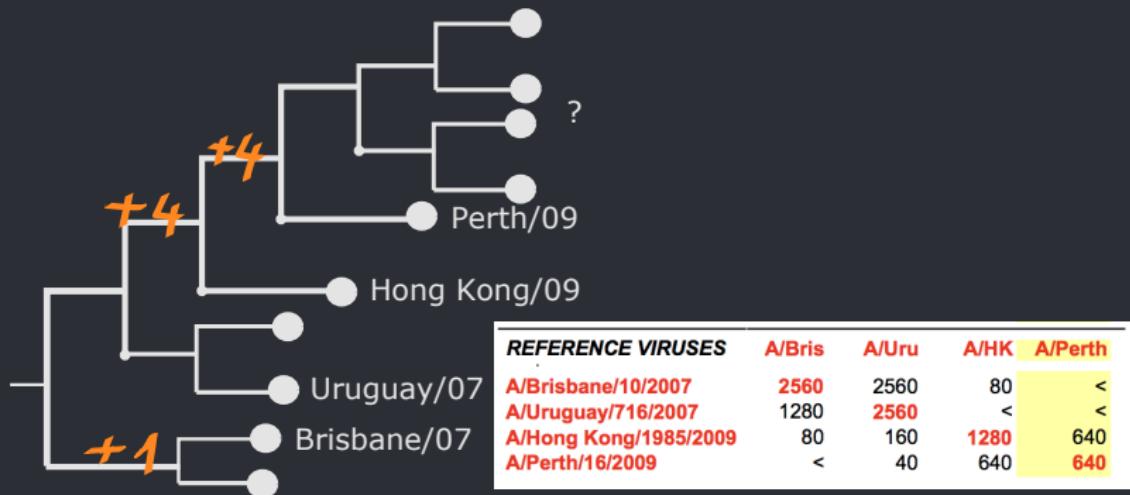
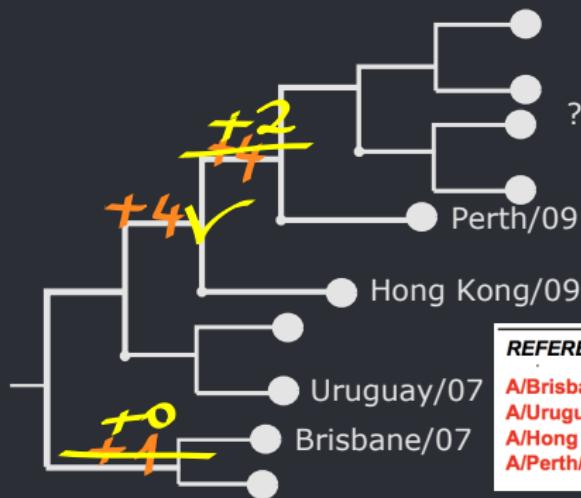


Figure 12: Here, one antiserum is tested in 12 different dilutions against 8 different virus strains. The highest dilution that prevents agglutination is called the titer.

# The Hemagglutination Inhibition Assay



REFERENCE VIRUSES	A/Bris	A/Uru	A/HK	A/Perth
A/Brisbane/10/2007	<b>2560</b>	2560	80	<
A/Uruguay/716/2007	1280	<b>2560</b>	<	<
A/Hong Kong/1985/2009	80	160	<b>1280</b>	640
A/Perth/16/2009	<	40	640	<b>640</b>

Figure 13: Here, one antiserum is tested in 12 different dilutions against 8 different virus strains. The highest dilution that prevents agglutination is called the titer.

# The Hemagglutination Inhibition Assay

*The asymmetry makes sense, think of it like this but with more dimensions*



Figure 14: [rosipaw, 2010]

## Formulas I

$T_{a\beta}$  ..... HI titer of virus  $a$  against antiserum  $\beta$  (virus  $b$ )

$H_{a\beta}$  .....  $\log_2$  relative titer (we'll use this one)

$$H_{a\beta} = \log_2(T_{b\beta}) - \log_2(T_{a\beta}) \quad (1)$$

$\hat{H}_{a\beta}$  ..... predicted  $\log_2$  relative titer

$v_a$  ..... avidity of virus  $a$  (=greediness)

$p_\beta$  ..... potency of antiserum  $\beta$  (=effectiveness)

$D_{a\beta}$  ..... genetic component of titer drop

$$\hat{H}_{a\beta} = v_a + p_\beta + D_{ab} \quad D \text{ is between two viruses} \quad (2)$$

## Formulas II

What remains is split up into a sum over individual mutation contributions:

$$D_{ab} = \sum_{i \in (a \dots b)} d_i \quad (3)$$

Where the sum is over the path connecting virus  $a$  and virus  $b =$  antiserum  $\beta$ . Now we want

$$\hat{H}_{a\beta} \stackrel{!}{=} H_{a\beta} \quad (4)$$

so we minimize a cost function  $C$  of the whole tree:

$$C := \sum_{a,\beta} (\hat{H}_{a\beta} - H_{a\beta})^2 + \lambda \sum_i d_i + \gamma \sum_a v_a^2 + \delta \sum_\alpha p_\alpha^2 \quad (5)$$

# Results

*Omitting 10% of the HI titer data, to make a test*

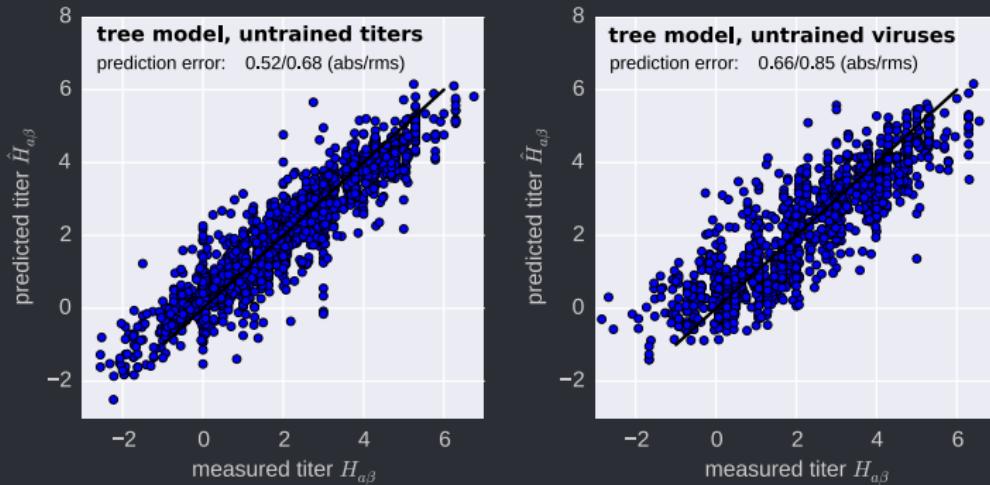


Figure 15: On left, 10% randomly picked measurements were omitted, on right, 10 % of the titer columns were held back, as if it was a new virus

# Results

*Asymmetry not interesting, but maybe quartet rule?*

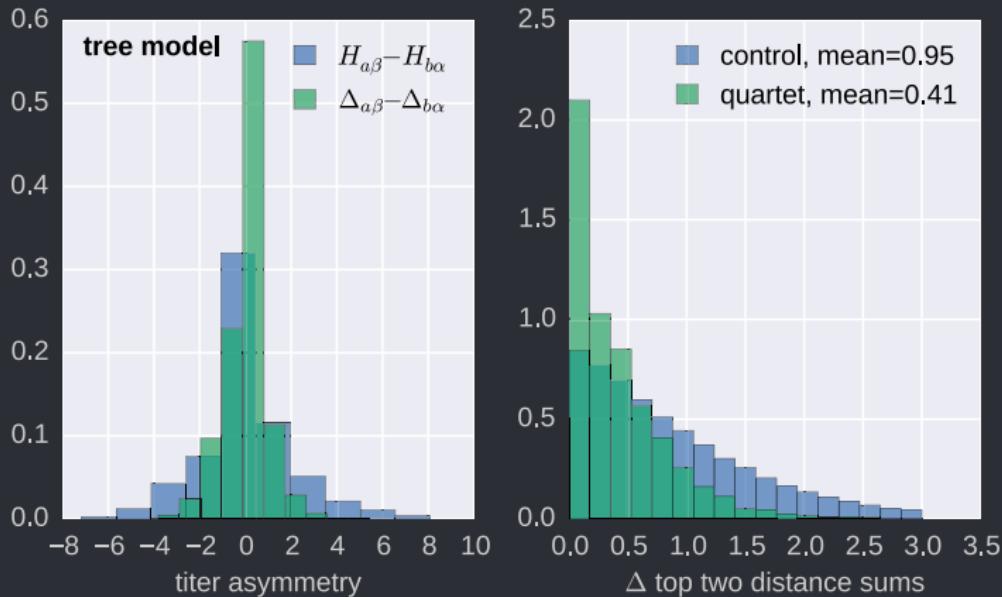


Figure 16: a

# Results

*Does recent antigenic evolution have fixation implications?*

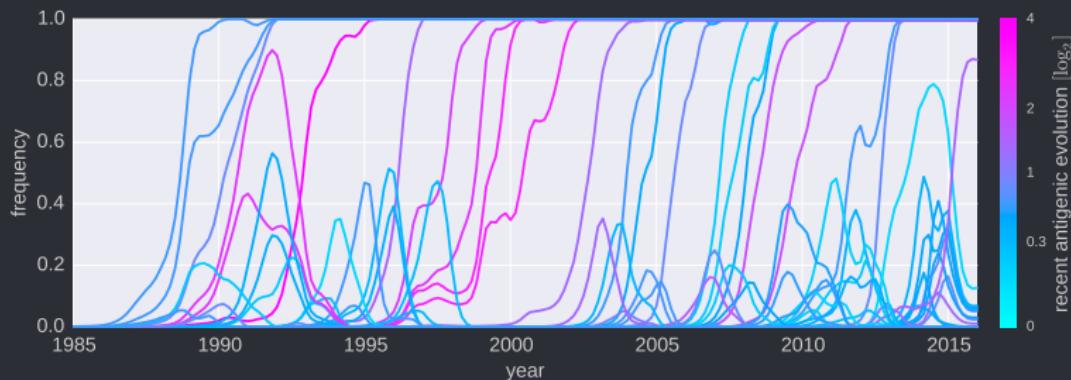


Figure 17: Fraction of samples having a certain mutation plotted vs. time  
Strains with frequencies smaller than 0.01 were omitted.

# Results

## thresholds

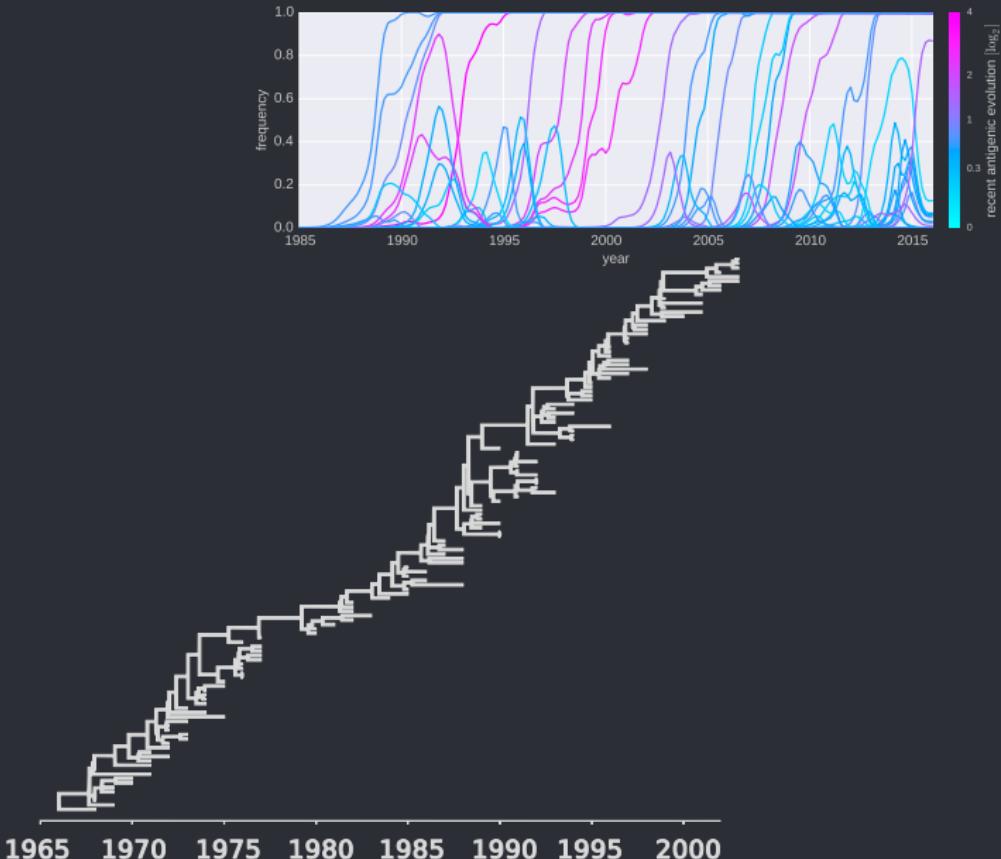


Figure 18: GET nextstrain tree with trunk colored according to antigenic advance 28/39

# Results

## thresholds

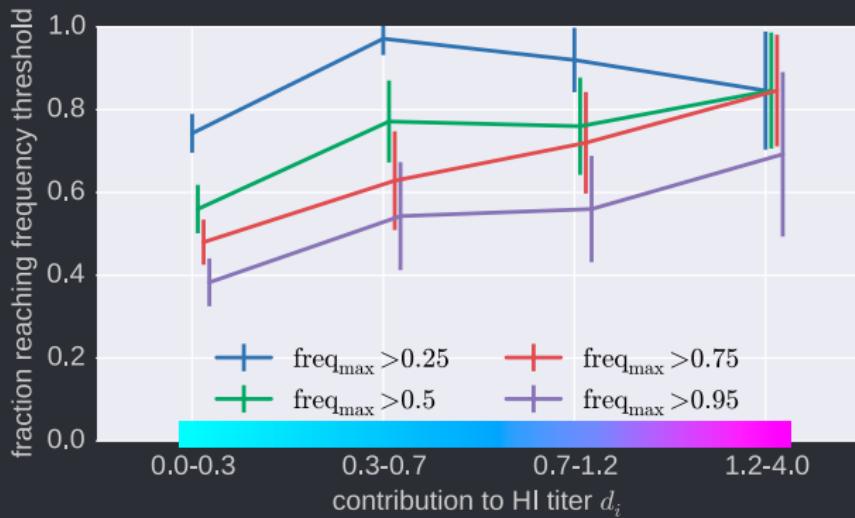
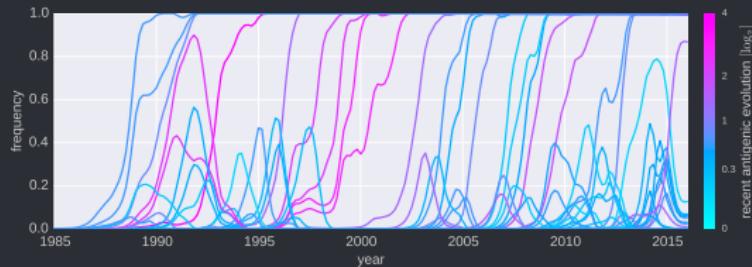


Figure 19: a

# Results

## *cumulative antigenic change*

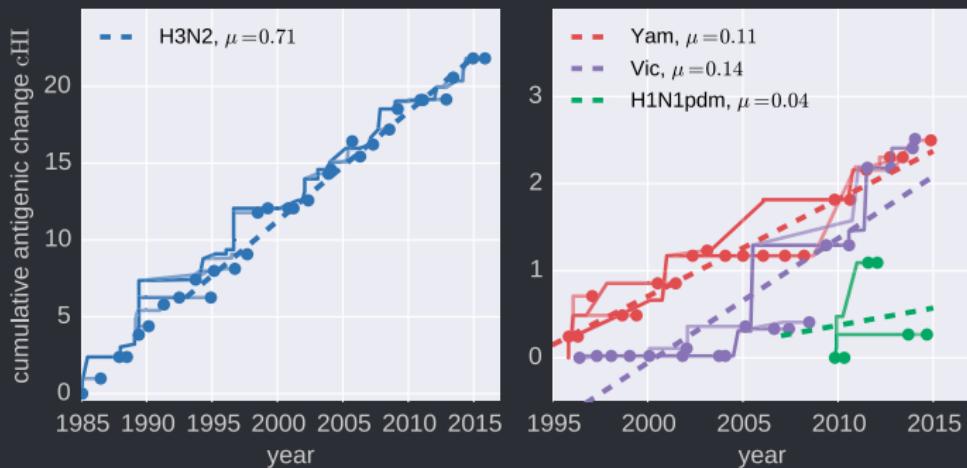


Figure 20: a

## Results

*distance to season year*

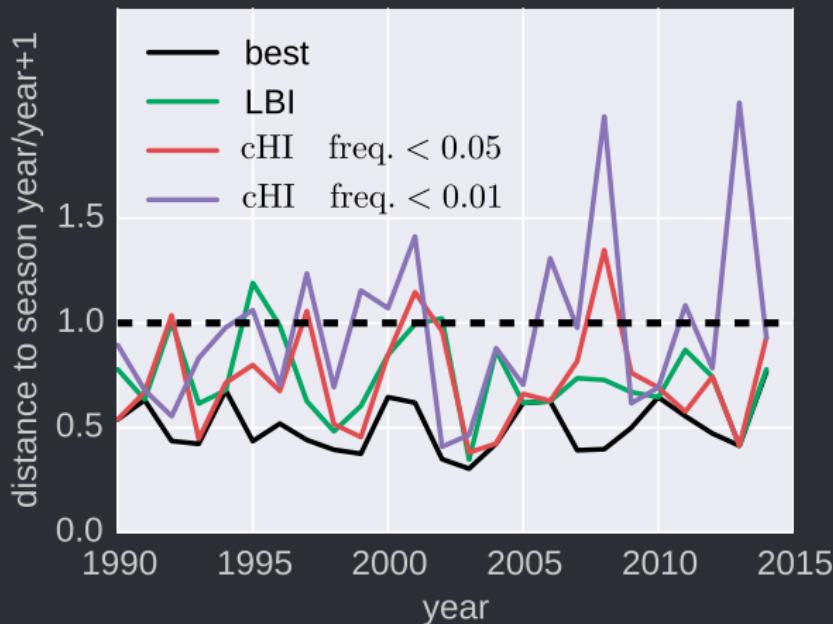


Figure 21: a

## Results

*years antigenic advance*

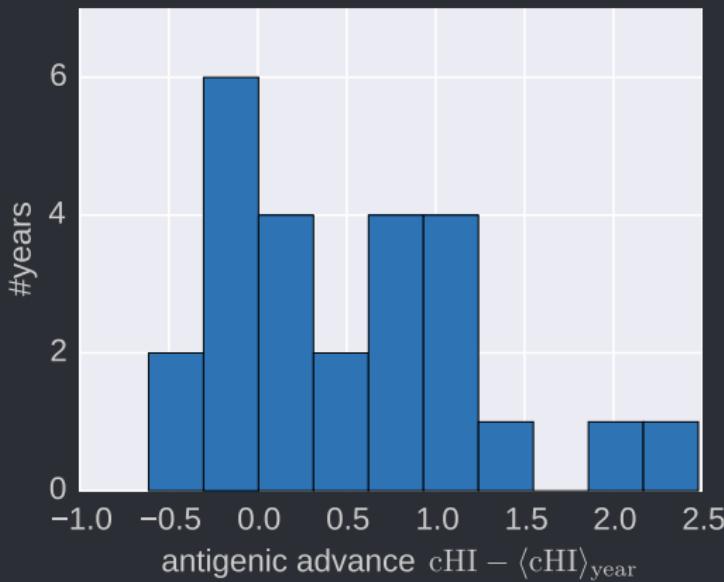


Figure 22: a

# *Nextstrain*

*Mending pieces together*

Please visit [nextstrain.org/narratives/.....](https://nextstrain.org/narratives/)

Ideas: COVID-Line? H3N2-Line?

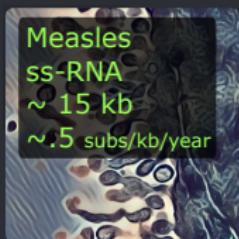
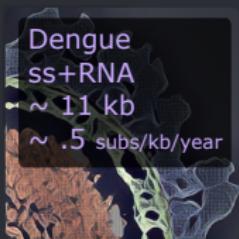
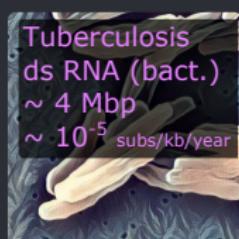
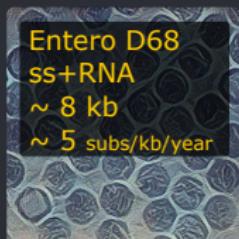
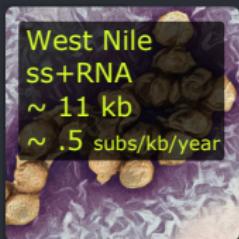
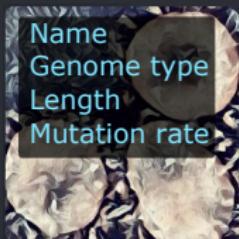
Multiple data sets?

de Vries Epitope sites

Koel 7 sites

D614G — Epidemiologic turning point

# Nextstrain: How to use the Framework (this is mirrored in the narrative)



## **Nextstrain: The Powerful Meta Data**

*(this is mirrored in the narrative)*

## *Nextstrain: Confidence Levels and Limitations*

*(this is mirrored in the narrative)*

## Multi Scale Evolution

If a single event mutation occurs, say **D 186 G** in the HA genome, it is subject to multi scale evolutionary selection:

- this RNA instance vs. the other RNA strands in the same cell
- this cell's mutated viruses vs. other viruses inside the host
- this host's viruses vs. viruses in rest of the population
- this population vs. other populations

These scales are difficult to separate. At the population level *epidemiological* processes may dominate.

# Outlook *and closing remarks*

- Difficult to disentangle levels, therefore integrate visualization
- Include more meta data (symptoms, severity, & c.)
- Interesting: larger evolutionary timescale
- Include other pathogens, like bacteria, CAVEAT: HGT
- My question: what happens if epidemic threshold is crossed even further, and zoonotic forcing is stronger?
- nice to see instance of Red Queen Hypothesis
- Transparency and Community building very important!

# References I

[AAMC, 2018] AAMC (2018).

Journal of the Association of American Medical Colleges Frequently Asked Questions.

[Alberts, 2015] Alberts, B. (2015).

*Molecular biology of the cell.*

[Bedford, 2015] Bedford, T. (2015).

Real-time evolutionary forecasting for influenza vaccine strain selection.

[rosipaw, 2010] rosipaw (2010).

146/365 square peg into a round hole.

[Tokars et al., 2018] Tokars, J. I., Olsen, S. J., and Reed, C. (2018).

Seasonal Incidence of Symptomatic Influenza in the United States.

*Clinical Infectious Diseases*, 66(10):1511–1518.

Publisher: Oxford Academic.