Phylogenetics for Predicting Virus Evolution

Including a brief guide to the open science tool *nextstrain.org*

A Contribution to the 2020 Seminar "Physics of Viruses" Conducted by Ulrich Schwarz and Frederik Graw at University Heidelberg

Paul Wiesemeyer

October 16, 2020

Aim of this write-up

After reading these pages—and potentially following the recommended material at the end of each section—the reader will be able to:

- Read phylogenetic trees
- Explore the framework nextstrain.org
- Have some in-detail knowledge about influenza
- Have gotten a close look at the phylogenetic approach at prediction of influenza evolution
- ... and hopefully end up with a sprawling interest in virus research!

As the research on this vast field proved cumbersome due to many terms and definitions unfamiliar to a physics graduate, this write-up comprises a glossary and a list of acronyms at the very end. Included terms are *italicized*.

Contents

1 Introduction

Old Introduction

Viruses, and the diseases they provoke, put a large burden on human society. Viruses are ...(TODO) . comparably simple structures are capable of modifying and possibly destroying vital functions in the human body, by infiltrating their genetic material into the reproductory apparatus of cells.

All viruses that persist in the human population over longer times share some common features:

- They need some form of protection from their surroundings, a hull
- They need to find a way to get into human cells
- They aim at reproducing quickly
- If there is no constant infection source, they need to find a path to get from one individual to another
- They will have to deal with human immune system response
- They will underlie some kind of evolutionary pressure (too general).

(This rough characterization is of course incomplete, and there may also be exceptions. The terms "need" and "aim at" are to be understood as: evolutionary processes strongly favor these characteristics.)

To give another rough image of what makes up a virus, here are the essential constituents:

- A piece of genetic code, most importantly categorized into: *Ribonucleic Acid* (*RNA*) or *Deoxyribonucleic Acid* (*DNA*), single or double stranded, positive or negative sense, length (usually some few kilo base (pairs) long)
- A hull: either only a capsid (of proteins) or an additional envelope (bilipid layer).

For most enveloped viruses such as members of the *Orthomyxoviridae* family (including influenza) or *Coronaviridae* (e.g. SARS-CoV-2), there is a variety of different surface proteins that populate the bi-lipid layer.

These surface proteins take up functions such as binding to a cell to infiltrate the viruses genetic sequence into it, or releasing a freshly assembled virus from the host cell's surface into the surrounding body fluids. This is usually done via interactions with specific host cell receptors, therefore defining a *Receptor Binding Domain* (*RBD*) as the part of the surface protein that fits onto the receptor like a key fits into a lock.

At the same time, the human immune system will also interact with mostly surface proteins. The immune system will eventually develop antibodies that are targeted to bind to a specific region of virus surface proteins, thereby rendering them innocuous.

The surface protein region targeted by the immune system is called *epitope site* and may have a large overlap with the *RBD*.

These two mechanisms put an evolutionary pressure on the virus, especially on the surface proteins and its epitope sites. While having to maintain essential functioning such as binding to receptors, the virus will draw large advantage from modifying its epitope site to such an extent that the antibodies cannot bind to it any longer. This way of disguising itself—by amino acid mutations in crucial places—will allow the virus to reinfect previously immune individuals of the host population.

An accelerated evolution therefore allows viruses to persist in the human population over long periods, but at the same time provides a large record to trace its spreading history, when the viral genome is sequenced and the information is curated.

In this write-up, we will look at how sequencing of the viral genome, along with phylogenetic tree inference, can provide useful insights on (i) the route that the virus takes to spread in the human population, (ii) the prevalence of strains within a virus type, and (iii) the (projected) evasion of the virus from immune system response and vaccines.

Vaccines need updates. Scientific publishing and vaccine manufacturing take in the order of half a year each. Two influenza seasons (at a given place)are usually only a few months apart.

The goal of the *nextstrain* research group is to make phylogenetic trees inferred from external sequencing labs quickly accessible and easily explorable for fellow scientists, health care officials and the public. This is urgently needed in highly dynamic situations such as the 2020 COVID-19 pandemic. Also, *nextstrain* provides an integrated visualization of pathogen spreading phenomena, as will be explained in section 4.

Here, we chose to look at influenza, as it is the evergreen. Recombination pathway for evo.

1.1 What are viruses?

History: infectious fluids w/o bacteria. Turns out the mere information can modify a living being, using its reproductory apparatus. Viruses also mark the border to what is considered the living world, being classified as below it.

Functional definition: Genetic material. Mechanism to enter "hijack" host cell. Often: protective hull.

Detailed definition (box?): Different types of genetic code. Different types of protection from the outer world. Different mechanisms to enter a cell.

Include: Spreading of viruses. Airborne, Waterborne etc. most importantly: directly host-to-host, or does it need a source or intermediary host or reservoir?

Organizational structure of human society makes up a big threat when it comes to human-to-human transmission. What used to be a spatial problem in past centuries (e.g. the wavelike patterns by the bubonic plague) has now turned into a highly interconnected world, that is the structure of the population has changed dramatically. This accelerates pathogen transmission, as it ... but it also open new opportunities for countermeasures.

1.2 What is problematic when viruses evolve?

Evade immune system. And even vaccines. Leads to a race and so called co-evolution. Mention the Red Queen here.

But one big advantage: We can now (since when?) sequence the genetic material and through the statistical nature of its modifications (mutations) extract information from it.

Most importantly we can interrelate the different probes. Where we just had 100 people with same symptoms, we can now make precise statements about who probably got the virus from whom. With a certain uncertainty.

So how can we make sense of this?

The theory behind this is called phylogenetics.

2 Phylogenetics

The classical example is that you have a number of different genetic sequences. Could be from different species or different individuals of the same species, as we will do here.

Then you need a model about how nature gets from one instance of a sequence to the other. The usually assumed manner is by point mutations and indels. And that they happen stochastically.

Picture of two different trees that explain three (or four?) different histories of a sequence.

Parsimony: Number of needed substitutions is to be reduced. This makes especially sense if single mutation probabilities are low. It is Ockham's razor applied.

Molecular clock: If the substitution rate was constant over all sites of the genome, one would expect the stochastically occurring mutations to yield one constant mutation rate for the whole genome. In reality, this is a strong simplification and each site has their own mutation rate. This is due to different evolutionary pressure laying on each site.

Synonymous mutations are such that do *not* alter the expressed amino acid. They leave the organism unchanged and are so to say just for our pleasure (??). Non-synonymous mutations change the amino acid and are therefore under selective pressure. That is: depending on whether they are beneficial or detrimental to the organism, they will establish in the population or vanish and be unlikely to show up in any probes. That is the nature of evolution, that—on large timescales—we only see in the data what is capable of surviving and reproducing.

2.1 What defines a Phylogenetic Tree?

Topological construct: Vertices linked by edges. To begin with, lengths and shapes have no meaning, only connectedness is relevant.

A construct of vertices and edges (called graph or network) that has no internal loops. Consequently, if you remove one edge of a tree, it decays into two disconnected graphs.

Vertices or nodes can be divided into external and internal, the external ones are sometimes called leaves.

Furthermore, trees are usually taken to be bifurcating, that is one internal vertex has exactly three edges. Trees that do not fullfill this criterion are called multifurcating or polytomous, which can then be looked at as a lack of resolution, as a multifurcating vertex can always be reduced to a set of bifurcating ones.

Now the idea is that you take a bunch of external nodes (sequences) and reconstruct their interrelatedness in form of such a tree by probabilistic considerations. This is called tree inference and more info can be found HERE.

In the next step, one would like to give the vertices an order, as in who are the descendants of whom. This is done by introducing a special vertex called root, which is then parent to all other vertices. Be it pointed out here, that it has nothing to do with time, so far we are talking only topology and hierarchy.

For N leaves there are $\frac{(2N-3)!}{2^{N-2}(N-2)!}$ possible bifurcating rooted trees (Wikipedia).

2.2 Visualization of especially sequence data?

One dimension could be mutation events or time

The other has no meaning.

Radial trees are only a way of clamming more information into a 2D piece of paper, as opposed to Cartesian visualizations.

PICTURE?

2.3 Phylodynamics and what we can read from a tree

Phylodynamics—as opposed to phylogenetics—is not satisfied with a tree but also asks: What processes shape a viral phylogeny?

First: Data availability. As long as scientific field does not abound with data, or as data content is dominated by the means available to gather it, a phylogenetic tree will always rather be the skewed little window through which we look at a problem, than an accurate image of any of the processes below.

Second: Epidemiological processes. For example growth vs. constant population PICTURE.

Third: Immunological processes. For example imbalanced vs. balanced. Maybe HIV as a PICTURE here? Also included here: Evolutionary feedbacks.

Fourth: Host population structure.

Note, that we are increasingly adding information to the vertices, leaving the theoretical realm of tree construction and entering into the matter of interpreting trees that have been constructed by appropriate algorithms.

As a last note, the mapping between these processes and the observed ML tree is many-to-one REF VOLZ+2013, that is, there are usually many possible explanations for one observation, and phylodynamics deals with this interplay.

2.4 Recommended Material

Good Phylogenetics page
Anders Gorm Petersen
Volz+2013

But—to continue our journey through phylodynamic methods—let us now take a closer look at one most interesting virus family, the influenza viruses.

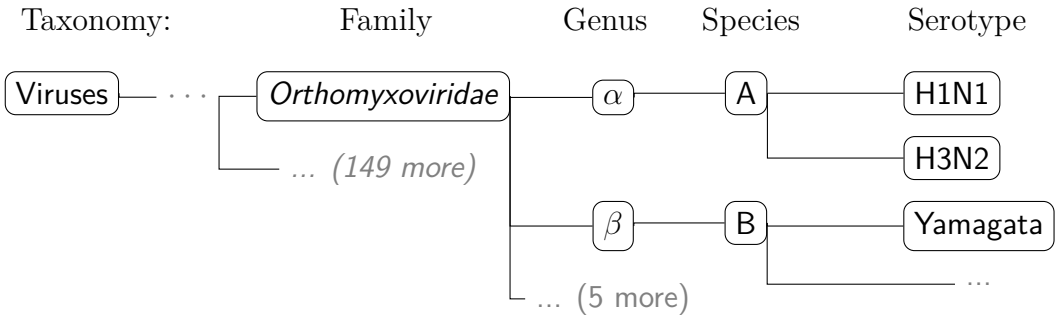
3 Influenza

Influenza is one of the most common viral diseases [Duda and Menna, 2020] and puts a large burden on human health. The disease is estimated to cause symptomatic illness in about $3-11\%$ of the human population (in the U.S.) [Tokars et al., 2018] every season. Globally, up to $650,000$ individuals die from influenza illness every season [see Iuliano et al., 2018].

Influenza refers to a disease caused by a virus of the influenza family *Orthomyxoviridae*. All *orthomyxo* viruses have an envelope that carries surface proteins and their genome is a negative sense RNA.

Most relevant for human infectious diseases are the genera *Alpha-* and *Betainfluenzavirus* that contain the species *Influenza A* and *Influenza B* respectively.

They are further subdivided into their so called *Serotype*, a classification by coagulation behavior in the *Hemagglutination Assay*



Where they come from (mention important pandemics).
INFLUVENNZA PICTURE

3.1 Hemagglutination Inhibition Assay

Hemagglutinin is the surface protein for entering host cells. It is intended to dock at the XXX receptor, but can also bind to red blood cell's YYY which is where it got its name from.

When a virus is added to blood serum, the virus and RBCs will form a matrix structure, the blood will agglutinate. This is particularly useful since it is an effect visible with bare eye.

In Hemagglutination Inhibition Assays, a third component is added: Antibodies to a particular influenza subtype that were incubated in for example eggs. These antibodies bind to the virus epitope sites, preventing the blood from coagulating. This process depends on the antibody concentration and how closely related they are

Hirst 1943

mention cartography from Smith+2004 (?)

3.2 What is special about influenza?

It has a relatively high mutation rate (EXACT COMPARISON, PLOT?), that is X substitutions per kilo base per reproductory cycle or per year (distinguish here to evolutionary rate?). Recombination evo pathway. E.g. Avian H7N9 Flu CFR 15-60 % but no H2H

3.3 Basics

History diagram from [Alberts, 2015]

HA: epl

NA: expl

Phylo tree of A/H3N2 HA Volz+2013

4 Nextstrain

In 2015, a group of scientists from Fred Hutchinson Cancer Research Center and University of Basel published a web tool to track and visualize the evolution of influenza viruses. Under the name *nextflu*, they had automatized sequence alignment, tree inference and visualization into two data processing pipelines, allowing for a quick analysis of the genomic data at hand [Neher and Bedford, 2015].

Shortly thereafter, phylodynamic analyses of other viruses like *zika* and *ebola* were included in the platform, now running under the name nextstrain.

Realizing that a limiting factor of phylodynamics is data availability, the aim was to make the sharing of sequence data and phylodynamic analyses easier between scientists. At the same time nextstrain allowed to directly present the phylogenetic results to the scientific community and the public.

In February 2017 the team around nextstrain won the first *open science* prize for their work [Hudson, 2017].

While bringing the scientific community and the public closer together, the ingenuity of nextstrain can as well be found in another domain: The integration of the multiple scales of phylodynamic processes into one visualization. So far, evolution of the viral genome had been studied by looking at the phylogeny while epidemic spreading patterns were visualized for example with geographic maps of the affected regions. Nextstrain provides an integrated view through these two different windows being in sync, allowing for a picture much closer to the process found in nature.

The following section will give a short hands-on introduction to the platform and its most relevant features.

4.1 Using nextstrain

On *nextstrain.org*, the visitor can first chose between a wide range of available pathogens. Beside viruses, even a bacterial disease—tuberculosis—can be selected. Contributions by external scientific groups are included further down the main page.

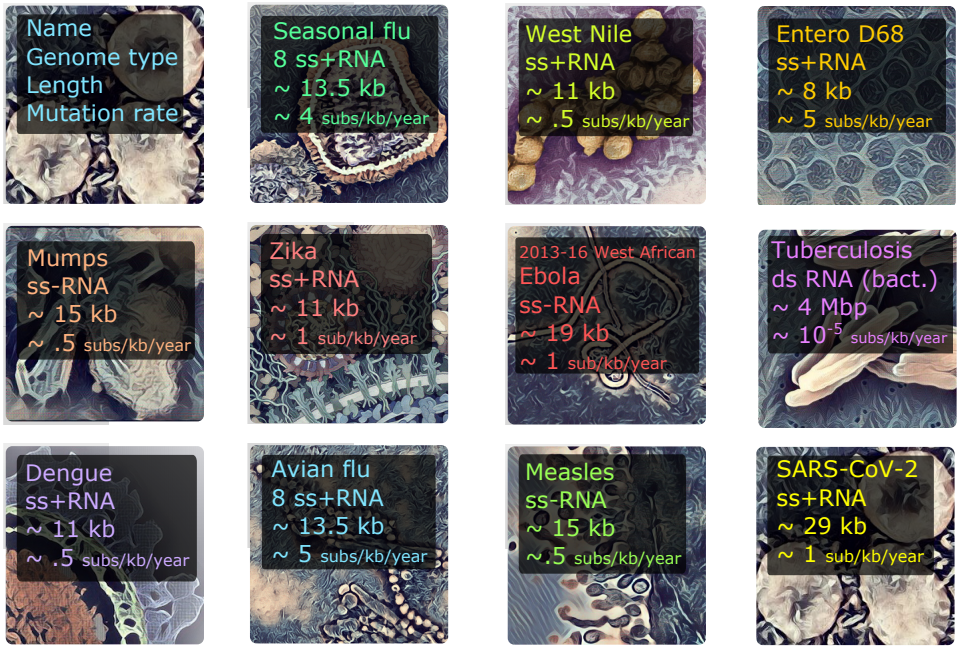


Figure 1: Augmented pathogen visualization, modified from **nextstrain.org**, Lee et al. [2020]

Selecting for instance seasonal flu, the visitor gets to what we will call the *phylodynamic dashboard*. In the following, we will go through its most basic features and discuss some of its possible uses.

The dashboard consists of primarily a phylogeny and a geographic map, as well as a diversity and frequency panel further down.

The **Phylogeny** may be considered the core of the visualization. It displays sequences as colored dots, showing the collected meta data when hovering or clicking. The branches are of the inferred *ML* tree, and come with a confidence interval that is displayed when hovering. The black crosses are exclusive to the influenza datasets, they mark the location of a vaccine strain.

The **Geographic Map**, shows the same data as the phylogeny, but mapped out according to the sequencing location (from meta data). The coloring is the same as the phylogeny. Some datasets display transmission lines between the geographic regions, these are a reflection of the branches of the phylogeny.

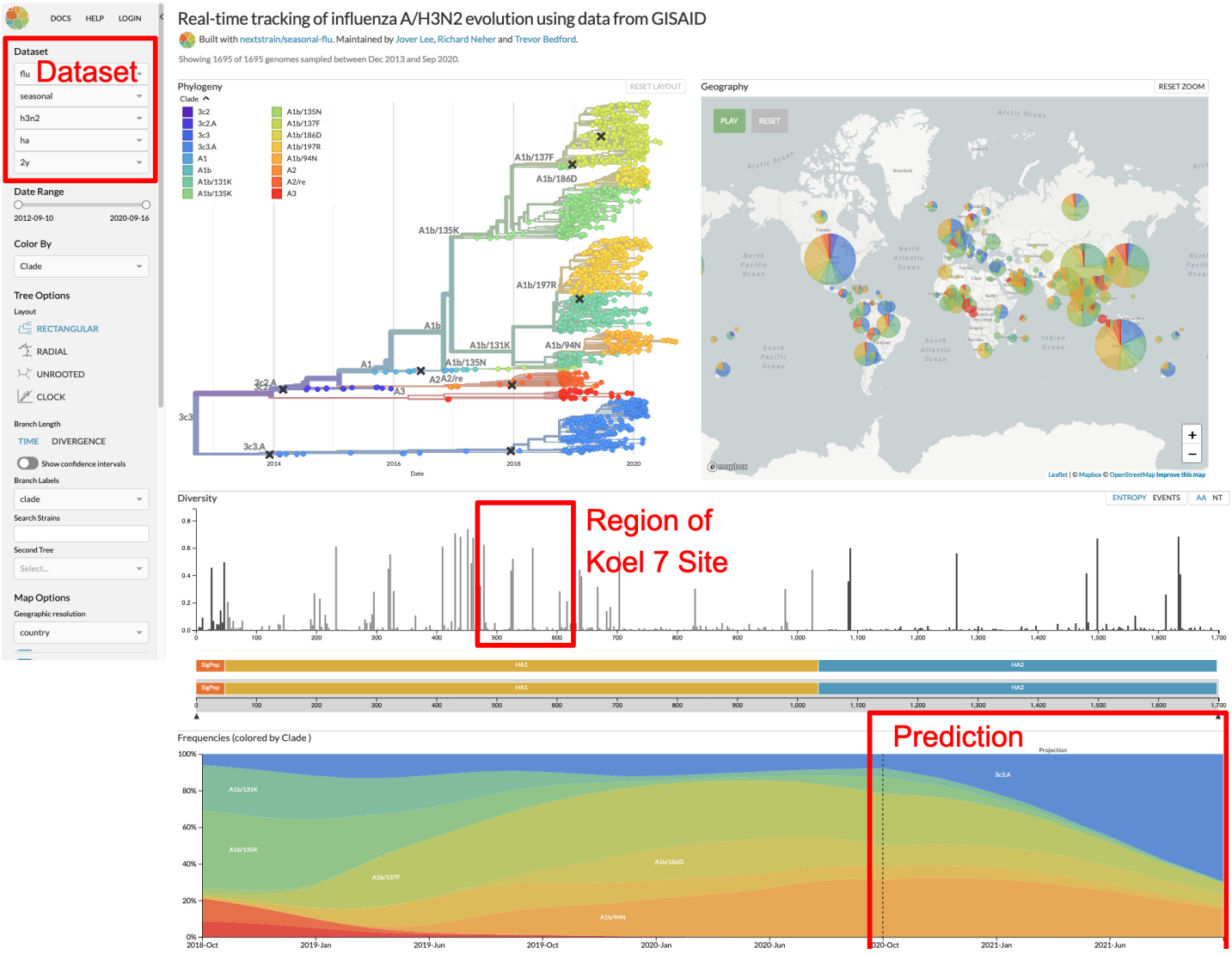


Figure 2: Taken from the 2 year seasonal H3N2 flu HA dataset visualization on **nextstrain.org**, Lee et al. [2020]

The **play button** allows to watch the disease unfolding as an animation, where the phylogeny and the geography is in sync.

Below, the **Diversity Panel** shows the sequence under investigation, with an identification of specific regions such as the signal peptide. The histogram displays either the number of mutation events or the entropy of each site in the genome. Entropy is calculated from the presence of different mutations $S = -\sum_{\text{mutations } i}(p_i \log(p_i))$, where p_i is the probability that a randomly picked sequence shows the mutation i at the site, such that exactly two equally abundant mutations would yield an entropy of about .69. The histogram can be binned by *amino acid* (AA) or by *nucleotide* (NT). If a genome site is clicked, the entire dashboard coloring will be set to discriminate between different mutations of that very position.

The **Frequency Panel** displays the abundance of the selected variation over the time period of the dataset and a projection beyond. Here, *"selected variation"* refers to any information that the data can be colored by, be it the strain (as is the default) or a specific mutation or anything else. For example when coloring by a high entropy site, the frequency panel should display at least two abundant different colors. The **Projection** of the frequencies is a feature specific to the influenza dataset, and its scientific origin will be discussed in section 5.

In the leftmost **navigation bar**, the main visualization adjustments can be made. Primarily, the dataset can be narrowed down to a specific part of the genome, or to a time range over which the data was collected. The data usually originates from many laboratories throughout the world submitting their sequences

4.2 Example

WILL I DO THIS?

4.3 Recommended Material

The platform is self-explanatory to some degree, but for a deeper introduction to the topic it provides plenty of material:

- The Documentation and Help. It should answer the most frequent questions and it gives a detailed guide on how to build an entire phylodynamic analysis on your own machine with the pipelines `augur` and `auspice`.
- The Discussion forum
- Guide on Narratives

The video series with Colin Megill

5 Mapping Influenza Evolution

5.1 Mapping Titer to Tree

Minimizing a Cost function
Tree model vs. substitution model
Proving Treelikeness

5.2 Results

Figure 6 Interpretation

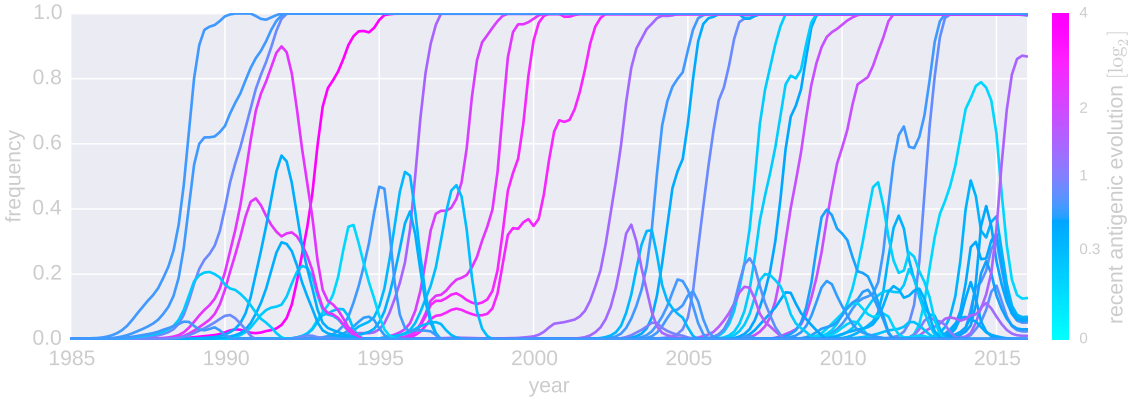


Figure 3: For high recent antigenic evolution traits, 25% prevalence directly entails 75% Neher et al. [2016]

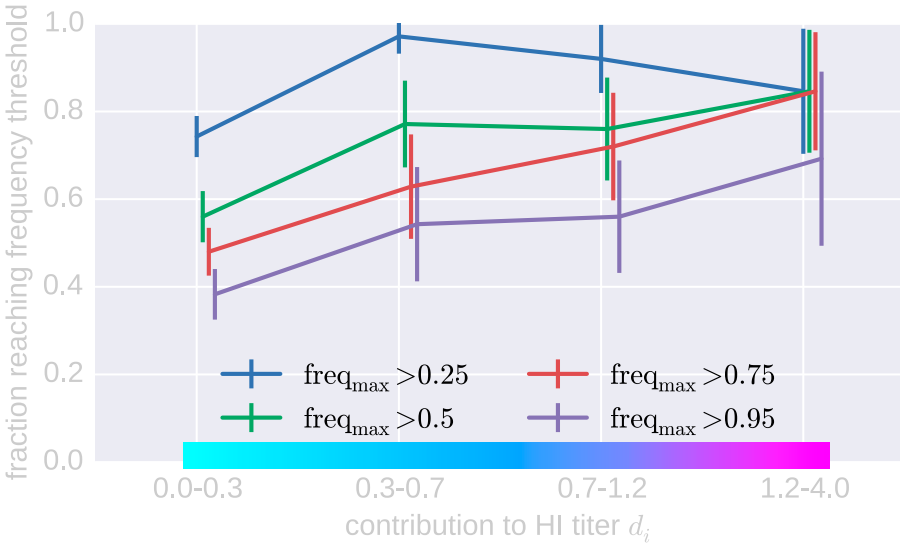


Figure 4: For high recent antigenic evolution traits, 25% prevalence directly entails 75% Neher et al. [2016]

6 Conclusion

Veryasfsadffds interestidsfng!

7 References, Acronyms, and Glossary

References

Alberts, B. (2015). *Molecular Biology of the Cell*.

Duda, K. and Menna, M. (2020). 7 Remarkably Common Viral Infections Explained. https://www.verywellhealth.com/common-viral-infections-770660.

Hudson, C. (2017). Open Science Prize Goes to Software Tool for Tracking Viral Outbreaks.

Iuliano, A. D., Roguski, K. M., Chang, H. H., Muscatello, D. J., Palekar, R., Tempia, S., Cohen, C., Gran, J. M., Schanzer, D., Cowling, B. J., Wu, P., Kyncl, J., Ang, L. W., Park, M., Redlberger-Fritz, M., Yu, H., Espenhain, L., Krishnan, A., Emukule, G., van Asten, L., Pereira da Silva, S., Aungkulanon, S., Buchholz, U., Widdowson, M.-A., Bresee, J. S., and Global Seasonal Influenza-associated Mortality Collaborator Network (2018). Estimates of global seasonal influenza-associated respiratory mortality: A modelling study. *Lancet (London, England)*, 391(10127):1285–1300.

Lee, J., Moncla, L., Anderson, B., Black, A., Sagulenko, P., Huddleston, J., Neher, R. A., Bedford, T., Hadfield, J., Megill, C., Callender, C., Fay, K., Potter, B., Bell, S., Hodcroft, E., Sibley, T., Ilcisin, M., and Grubaugh, N. (2020). Nextstrain.org. https://nextstrain.org/.

Neher, R. A. and Bedford, T. (2015). Nextflu: Real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, 31(21):3546–3548.

Neher, R. A., Bedford, T., Daniels, R. S., Russell, C. A., and Shraiman, B. I. (2016). Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proceedings of the National Academy of Sciences*, 113(12):E1701–E1709.

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.

Acronyms

DNA Deoxyribonucleic Acid. 1, 4

HA Hemagglutinin. 1

HI Hemagglutination Inhibition. 1

ML Maximum Likelihood. 1, 14

NA Neuraminidase. 1

RBD Receptor Binding Domain. 1, 4, 5

RNA Ribonucleic Acid. 1, 4, 15

Glossary

allele SOMETHING SOMETHING. 1

antigen Any kind of structure, that provokes a reaction by the immune system. See *epitope*. 1, 13

antigenic advancement SOMETHING SOMETHING. 1

antigenic drift SOMETHING SOMETHING. 1

antigenic shift SOMETHING SOMETHING. 1

bilipid layer The most ubiquitous membrane found in nature that consists of a double layer of lipid molecules. 1

capsid The capsid of a virus is the protein shell that encloses the genetic material. It usually consists of many so called protomers that form an almost spherical structure (platonic body). 1

epidemiology Branch of science that studies the spreading and distribution of diseases and other health factors in populations. This can range from the abstract formulation of governing equations all the way to the exact description of a particular epidemic such as COVID-19. 1

epitope The part of an *antigen*, such as a viral surface protein, that is recognized by the immune system and that antibodies bind to. 1, 13

Evolution Systems that have a sense of succession and contain some form of *heredity*, *variation* and *selection* show evolution. This can be as simple as "survival of the fittest" but also—depending on system complexity—a wide plethora of other dynamics. 1, 14, 15

Hemagglutination Assay A Lab experiment that classifies Influenza type A viruses according to their hemagglutinin surface protein. 1, 8

heredity In *Evolution*, heredity means that one generation inherits traits from its preceding generation. 1, 14

immunology Branch of science that deals with the immune system of organisms. 1

indel SOMETHING SOMETHING. 1

Influenza The disease that is caused by one of the influenza viruses. Also called flu. 1

open science The development to make scientific results, data and alike easily accessible to the public. 1, 9

Phylodynamics The branch of science that analyzes the various processes shaping phylogenies. These usually comprise evolutionary feedbacks as well as shorter timescale processes such as an extinction event. 1

Phylogenetics The science that extracts information from genetic sequences by classifying their inter-relatedness and inferring a *ML* tree. 1

phylogeny another word for phylogenetic tree. 1

point mutation SOMETHING SOMETHING. 1

. 1

selection In *Evolution*, selection is the feedback between fitness and probability to produce a next generation. 1, 14

sense In genetics, the sense of a nucleic acid, indicates the roles of the strand (and its complement) in specifying a sequence of amino acids. In virology, positive-sense viral RNA means that the viral RNA sequence may be directly translated into viral proteins (5'-to-3'). 1, 15

serotype In biology, "species" is the most widely used classification term with a range of definitions. Traditionally defined as "largest group of organisms that are capable of producing fertile offspring" it is used even beyond the realms of sexual reproduction with more sophisticated methods—e.g. based on genomic data—to distinguish between species. 1

species In biology, "species" is the most widely used classification term with a range of definitions. Traditionally defined as "largest group of organisms that are capable of producing fertile offspring" it is used even beyond the realms of sexual reproduction with more sophisticated methods—e.g. based on genomic data—to distinguish between species. 1

strain In biology, a strain refers to a subtype or a variant of some species.. 1

strandedness In genetics, strandedness refers to whether a nucleic acid consists of one or two strands. The latter usually provides more stability, i.e. higher fidelity when copied. 1, 15

titer SOMETHING SOMETHING. 1

variation In *Evolution*, variation refers to the stochasticity of the properties that one generation inherits from its precursor. 1, 14

Zika virus A *single stranded positive sense*, enveloped *RNA* virus of the Flavoviridae family, that is transmitted by mosquito bites and can cause zika fever and micro encephalitis in newborns. 1