# 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" Held by Ulrich Schwarz and Frederik Graw at University Heidelberg

Paul Wiesemeyer

October 11, 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 viral 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 is augmented by a Glossary and List of Abbreviations (at the very end).

## Contents

# 1 Introduction

## 1.1 Problem at Hand

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.

# 2 Phylogenetics

Essentially the Wikipedia Article + Volz 2013 (TODO read)

## 2.1 Idea

*Phylogenetics* is the

## 2.2 How to Read a Phylogenetic Tree

## 2.3 Recommended Material

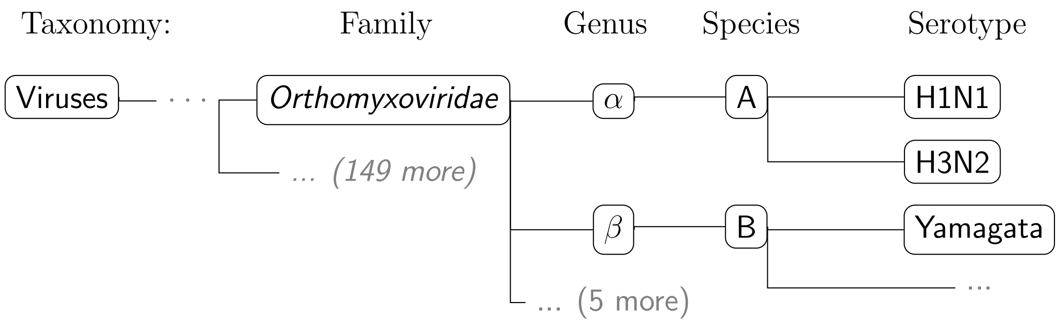
# 3 Influenza

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*



Recombination evo pathway. E.g. Avian H7N9 Flu CFR 15-60 % but no H2H

### 3.0.1 Basics

History diagram from [Alberts, 2015]
HA: epl
NA: expl
Phylo tree of A/H3N2 HA Volz+2013

### 3.0.2 Hemagglutination Inhibition Assay

Short description
Hirst 1943
mention cartography from Smith+2004 (?)

# 4 Nextstrain

## 4.1 Idea

## 4.2 Short Usage Instructions

## 4.3 Example

## 4.4 Recommended Material

# 5 Mapping Influenza Evolution

## 5.1 Mapping Titer to Tree

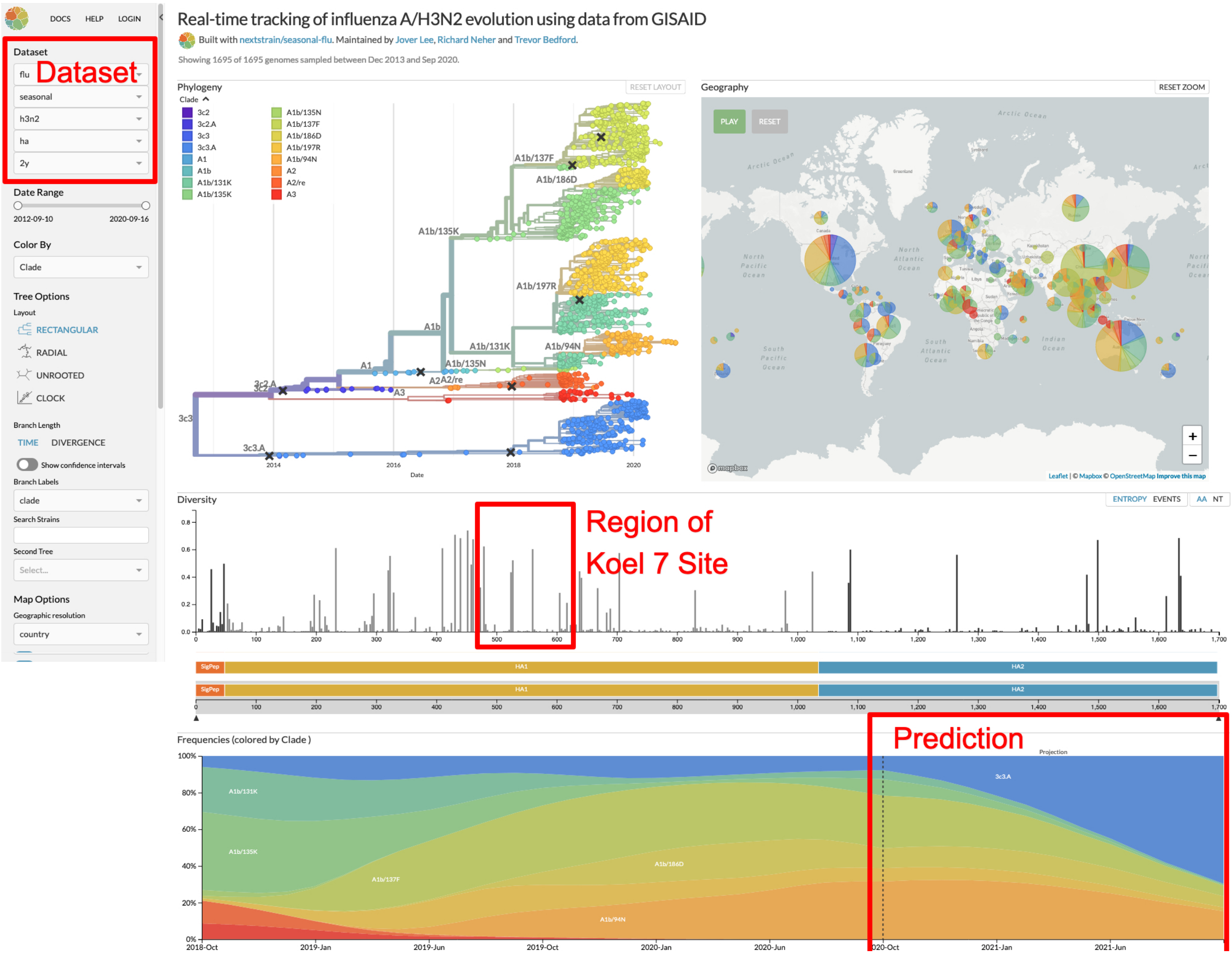
Minimizing a Cost function
Tree model vs. substitution model
Proving Treelikeness

## 5.2 Results

Figure 6 Interpretation

# 6 Conclusion

Veryasfsadffds interestidsfng! *mathematics Greatest Common Divisor* (*GCD*) sdfg *latex*



Real-time tracking of influenza A/H3N2 evolution using data from GISAID

Built with nextstrain/seasonal-flu. Maintained by Jover Lee, Richard Neher and Trevor Bedford.

Showing 1695 of 1695 genomes sampled between Dec 2013 and Sep 2020.

Dataset: flu, seasonal, h3n2, ha, 2y

Date Range: 2012-09-10 – 2020-09-16

Color By: Clade

Layout: RECTANGULAR

Branch Length: TIME

Branch Labels: clade

Geographic resolution: country

Figure 1: Taken from the 2 year seasonal H3N2 flu HA dataset visualization on **nextstrain.org**, Lee et al. [2020]

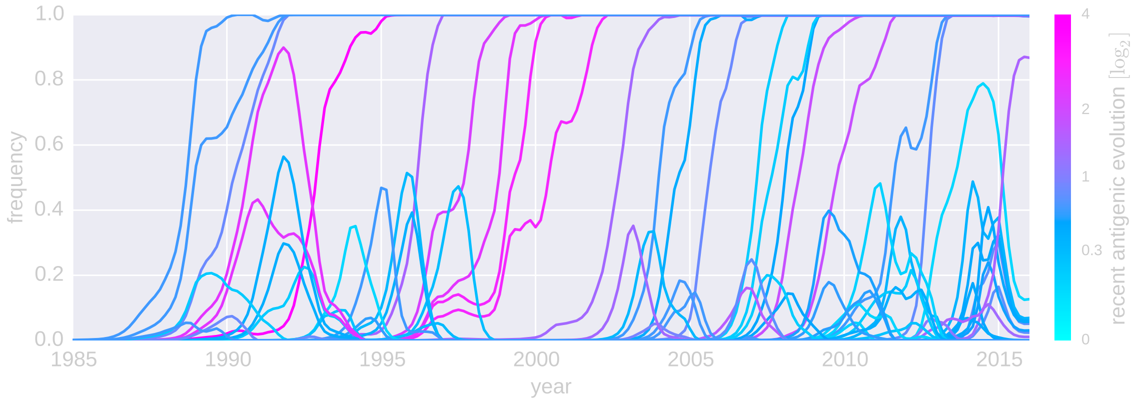


Figure 2: For high recent antigenic evolution traits, 25% prevalence directly entails 75% Neher et al. [2016]

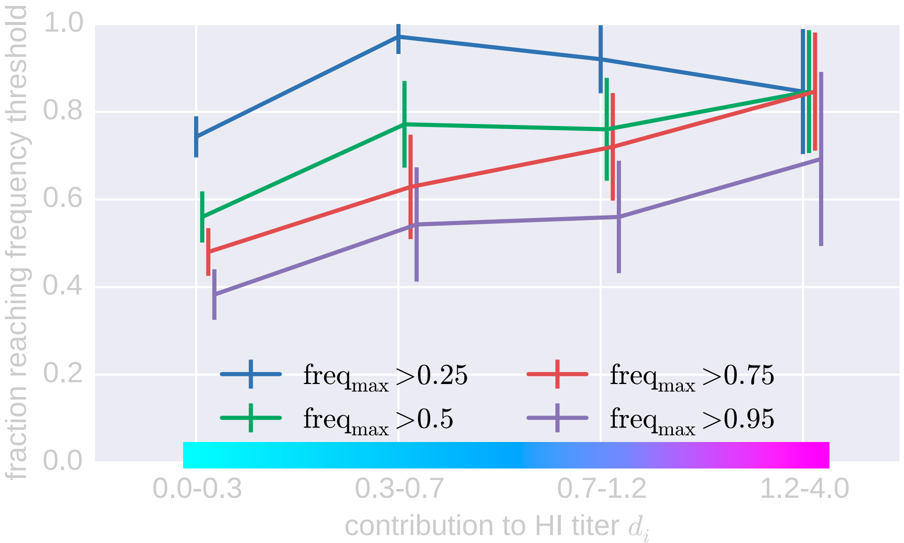


Figure 3: For high recent antigenic evolution traits, 25% prevalence directly entails 75% Neher et al. [2016]

# 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.

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., 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. 2

**GCD** Greatest Common Divisor. 4

**RBD** Receptor Binding Domain. 2

**RNA** Ribonucleic Acid. 2

## Glossary

**Hemagglutination Assay** A Lab experiment that classifies Influenza type A viruses according to their hemagglutinin surface protein. 3

**latex** Is a mark up language specially suited for scientific documents. 4

**mathematics** Mathematics is what mathematicians do. 4

**Phylogenetics** The science that extracts information from genetic sequences by classifying their inter-relatedness and inferring a *ML* tree.. 4