Kyu Eun Lee <kyueunl@uw.edu>

Project description

Huddleston, John L <jhuddles@fredhutch.org> Fri, Nov 14, 2025 at 3:15 PM
To: Kyu Eun Lee <kyueunl@uw.edu>

Hi Kyu,

This is really cool! Now that I'm seeing these map views of the data, I'm realizing I didn't have a specific hypothesis beforehand for what the distributions would look like across states. To make sure I understand how you produced these plots, I'll try to summarize how I think it's working:

1. Download a recent 12-year H3N2 HA tree from nextstrain.org
2. For each season between 2010-2011 through 2019-2020
    a. Find the vaccine strain that was available that season (based on "selection date" details in the tree or WHO recommendations website?)
    b. Find the clades that were circulating in the USA that season
    c. Count the number of sequences per clade and state from GISAID for virus isolates collected during the season
    d. Calculate clade frequencies per state from the count data
    e. For each clade, find the amino acid substitutions on the path between the clade's most ancestral node in the 12-year tree and the vaccine strain from that season.

       (I'm realizing now as I type this that your substitutions between vaccine strain and inferred clade sequence must reflect the same direction from the vaccine to the clade. In other words, you are trying to recreate from the tree the mutations that occur as you move from the vaccine sequence to the clade sequence. In your earlier example with A/Perth/16/2009, you found HA1 mutations E62K, N114K, and I214S on the branch leading to that Perth vaccine strain. But when you consider the distance from Perth to clade 3C, those mutations will be flipped such that Perth has 62K but 3C has 62E. So, the mutations that occur on branches between the vaccine strain and the MRCA of the vaccine strain and clade need to be flipped from E62K, for example, to K62E when you're calculating the antigenic distance between the sequences. The mutation that occur on branches between the clade and its MRCA with the vaccine can stay in the same orientation. In the 3C example earlier, you'd keep HA1 S45N and T48I as they are.)
    f. In the table of antigenic effects per season ending in October or April, find the closest timepoint to your current season and load the effects per mutation from that timepoint.

       (As I type this, I'm also realizing that you could reasonably use weights from both October and April of a given season. Those weights will reflect antigenic effects going into the Northern Hemisphere season in October and at the end of the season in April. If you had to pick between the two, the April timepoints would be most appropriate since they reflect the retrospective antigenic values from the season after it happened. It could be interesting to compare the results you find if you ran the analysis with October timepoints vs. April timepoints, though.)

g. For each vaccine/clade pair, sum the antigenic effects associated with the mutations found on the path between the vaccine and clade nodes in the tree. This gives you the antigenic distance between vaccine and clade for the season.
h. For each state, calculate the weighted antigenic distance from the vaccine strain weighting the antigenic distance per clade by the frequency of the clade in the state.

Does that seem more or less correct? I'm sorry I didn't catch the issue with the directionality of the amino acid substitutions between the vaccine strain and MRCA of the vaccine and each clade, but that's one big change to make before going farther with the analysis.

The other big change is the one you mentioned about which timepoints to use for your season's weights. If I had to pick one timepoint of antigenic weights per season in your analysis, I'd pick the April timepoints (so, for 2010-2011, use the 2011-04-01 weights) because these reflect the best antigenic data that were available at the end of the season. I would not use weights from different seasons (e.g., weights from 2019 in the 2010 season), since the specific HA positions that matter for immune escape change so much between seasons. The global H3N2 population usually turns over completely within 6 years such that the most recent common ancestor of any H3N2 virus circulating today dates back to a successful ancestor 6 years ago. Even on shorter time periods than that, the specific positions associate with antigenic drift change frequently. If you used weights from 2019 in the 2010 season, you'd be effectively time traveling to see what immunity in 2019 would have looked like in 2010.

I'm sorry to have written another essay-length response! We can always Zoom chat again, if that would be easier...

[Quoted text hidden]