(1) Align sequences

- Command:

python3 -m augur align --sequences /path/to/FASTA2018-5.fa --reference-sequence /path/to/h3n2_ha_outgroup.gb --output /path/to/aligned.fasta --fill-gaps

- Input:

--sequences /path/to/FASTA2018-5.fa : Just the regular fasta file with the DNA sequences
--reference-sequence /path/to/h3n2_ha_outgroup.gb : Here I used *their* reference sequence, which can be downloaded at https://github.com/nextstrain/seasonal-flu/blob/master/config/h3n2_ha_outgroup.gb .
--output /path/to/aligned.fasta : all the DNA sequences from FASTA2018-5.fa aligned based on the sequence "h3n2_ha_outgroup.gb".

- Remarks:

- To run this command, it is necessary to install MAFFT in addition to Augur.
- If the sequences in the file "ASR_dna.fa" were correctly aligned with the reference sequence that they use, I could skip this step and the next step.

(2) Determine WHO clade designations

Priscila Biller wrote a Python script to assign clades. It receives as input the sequences (a FASTA file properly aligned as above using the specified outgroup), and the file "clades_h3n2_ha.tsv" (or the corresponding file for the other influenza virus variants), which defines the criteria used to determine if a sequence belongs to a certain clade or not. The sequence is assigned with a clade only if it satisfies all criteria of the clade. The output is a TSV file where each line corresponds to a sequence in the FASTA file and the clades that have all criteria satisfied by the sequence.

For a given sequence, the nucleotide sequence is cut in 3 pieces: "SigPep" (nucleotides 1..48), "HA1 protein" (nucleotides 49..1035), and "HA2 protein" (nucleotides 1036..1698), then each piece is translated into an AA sequence. During the translation of nucleotide sequences to amino-acids, the uncertain nucleotides are taken into account. For example, the codon "ATN" can be either "ATA", "ATC", "ATG", or "ATT". Then the constraints are processed. For example, "3b HA1 223 I" checks if the 223th amino-acid of "HA1 protein" is "I".

It is possible that a sequence satisfies the criteria of more than one clade. For instance, the sequence "KU289613_A/Corsica/33-02/2015_2015/02/20" satisfies the clades "3c","3c3", and "3c3.B". In this scenario, the most specific clade is assigned ("3c3.B" in the example). However, there are still some cases where there was more than one "specific clade". For instance, the sequence "KY273051_A/Xiamen/s175/2016_2016/03/14" satisfies "3c,3c2,A1,A1a". To resolve these we proceed visually (CC did this manually) using the trees at nextflu.org

The script is provided as identifyClades.py. BUT it will have to be customised because certain things are hard coded.

Input files from the augur/nextflu people

The relevant files for four influenza virus types are listed here.

Augur aligns to a distant reference / outgroup for H3N2, H1N1pdm, Vic and Yam. This doesn't have anything to do with what is chosen for seasonal vaccines.This reference sequence doesn't need updating from year-to-year, and sequences are available at:

* https://github.com/nextstrain/seasonal-flu/blob/master/config/reference_h3n2_ha.gb
* https://github.com/nextstrain/seasonal-flu/blob/master/config/reference_h1n1pdm_ha.gb
* https://github.com/nextstrain/seasonal-flu/blob/master/config/reference_vic_ha.gb
* https://github.com/nextstrain/seasonal-flu/blob/master/config/reference_yam_ha.gb

Clade definition files for all four lineages are available in the same directory:

* https://github.com/nextstrain/seasonal-flu/blob/master/config/clades_h3n2_ha.tsv
* https://github.com/nextstrain/seasonal-flu/blob/master/config/clades_h1n1pdm_ha.tsv
* https://github.com/nextstrain/seasonal-flu/blob/master/config/clades_vic_ha.tsv
* https://github.com/nextstrain/seasonal-flu/blob/master/config/clades_yam_ha.tsv

It is possible to annotate sequences with clades using "augur clades" as referenced here but we found that Priscila's custom script was effective and did not require the ancestral sequence reconstruction. The augur instructions are at:
https://github.com/nextstrain/seasonal-flu/blob/master/Snakefile_base#L711