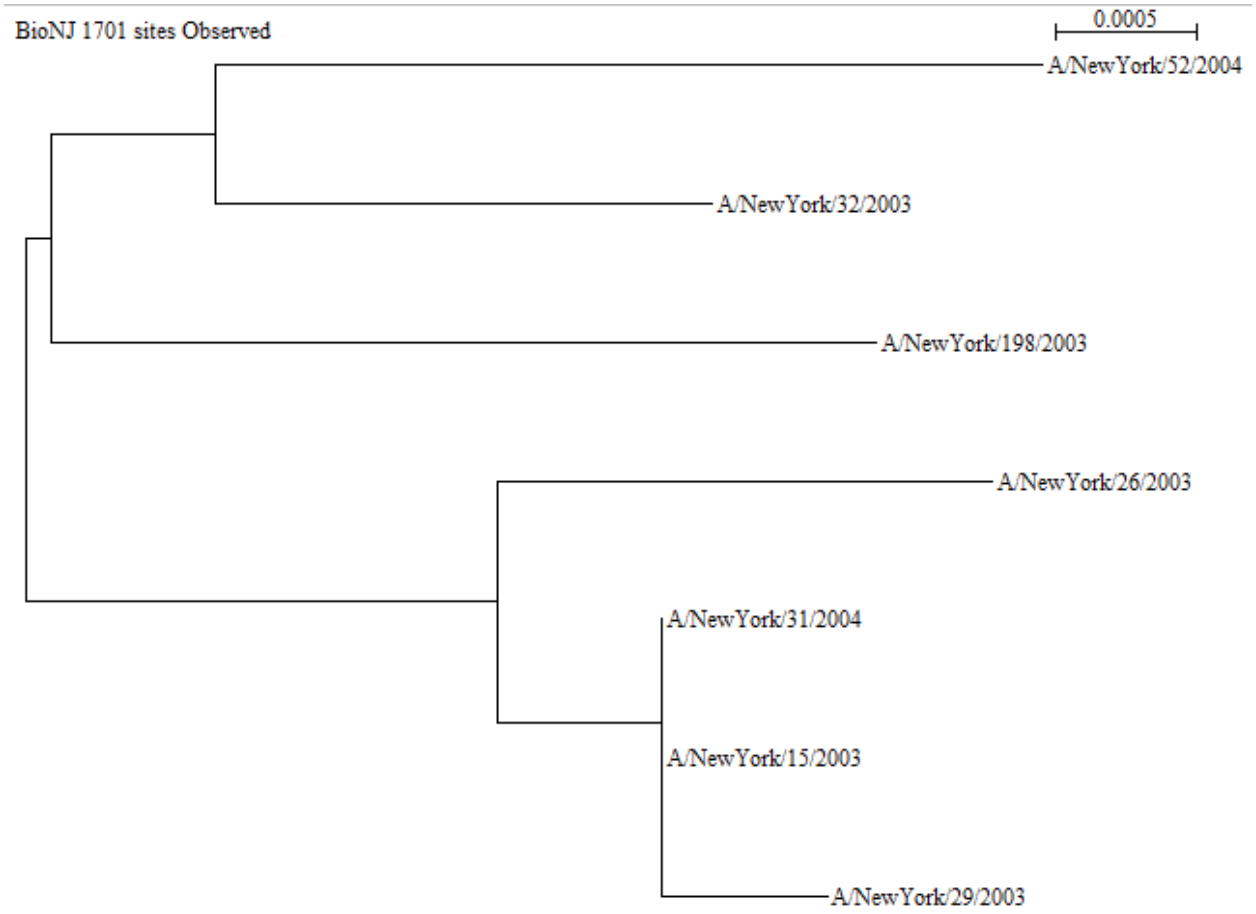
Khan Inan
Professor Truong
BI-GY 7683



1. The tree essentially varies with branches of different lengths representing different timeframes. One the anomalies observed is that one of the strains of the virus (A/New York52/2004) represents almost the entirety of the timeline and doesn't split off into a different strain while another one of the viruses (A/NewYork29/2003) is the result of a series of mutations/splits. My hypothesis is that the the more successful variations of the influenza virus (meaning most successful at spreading, and less affected by vaccines or treatments) are not subjected to as many selective pressures that induce mutations compared to the less successful variants. Based on reviewing numerous articles based on the coronavirus and its mutations, however, I don't think my hypothesis holds true. It seems that the most contagious and successful strains of the virus are subjected to more mutations and splits in the phylogenetic tree. The branches of the tree that are constant and never split are the strains of virus that lie dormant or have "latency" which delays the time for when it is activated in the host. This phenomena, however, may only apply to the coronavirus, and I need to do a bit more research to see if dormancy can also apply to influenza.

2. I can see the variants responsible for the clustering of the strains just by simply scrolling through the sequences at a glance. It seems that almost every single mutation that results in a split in the phylogenetic tree, is due to a single nucleotide substitution. How early along in the outbreak that the mutation occurs determines where the virus splits in the phylogenetic tree due to things like bottlenecks. The 4 different influenza variants that results in the cluster are A/NewYork/26/2003, A/NewYork/31/2004, A/NewYork/15/2003 and A/NewYork/29/2003, and I was able to determine this based on the close proximity of the splits in the tree.