Processing genomic signals for epidemic awareness

Lulu Han, Andrew Warren
This work is funded by CDC award no. 1 NU50CK000631-01-00

Introduction

One major limitation in understanding the COVID-19 pandemic is not knowing the true number of infections. Various models have agreed that the true number of infections surpass confirmed cases, but they disagree on how much.

Can we use analytical models in combination with genomic surveillance to estimate underlying aspects of the pandemic?

Wastewater Surveillance

Wastewater surveillance is commonly used as a tool for disease monitoring.

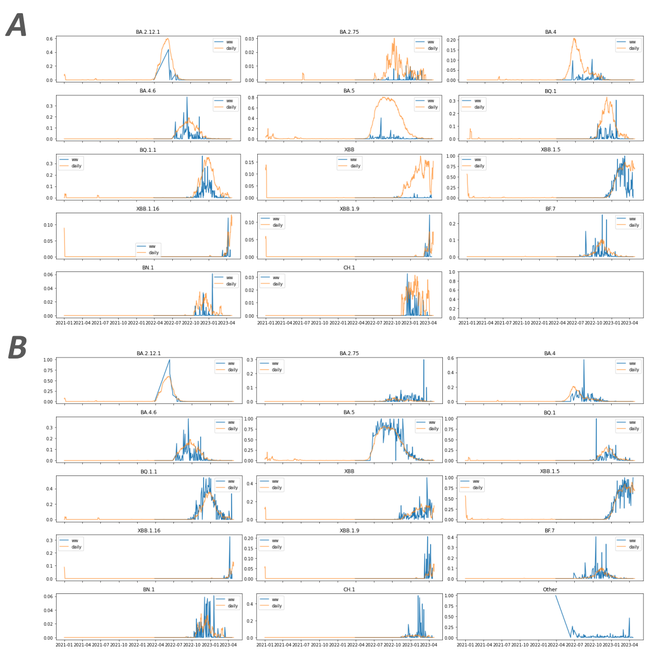


Figure #1

Figure 1A. Correlation between cases and variant prevalence in wastewater surveillance. **Figure 1B.** Using PangoAliasor to modify variant names of descendant variants, so they are accounted for in the variant prevalence.

Issues with Wastewater Surveillance

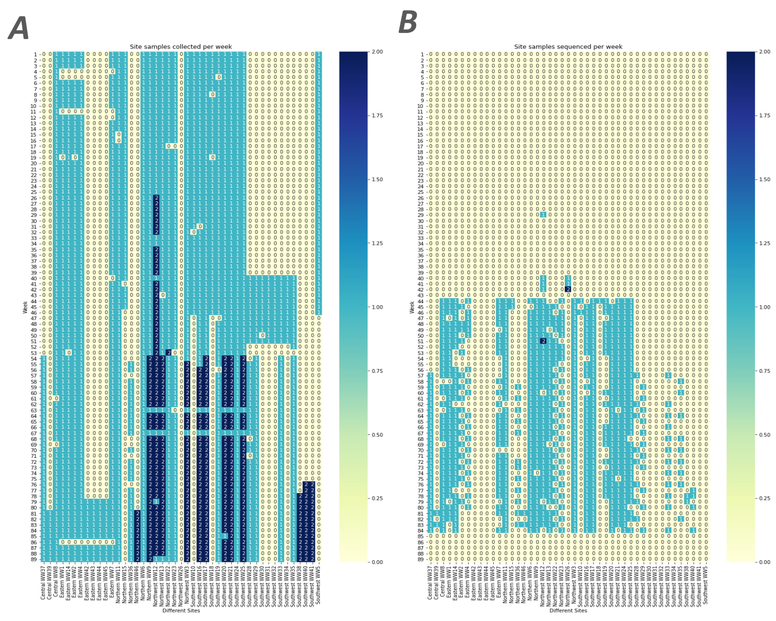


Figure #2

Figure 2A. Heatmap of sample collection frequency.
Figure 2B. Heatmap of sample sequence frequency

Not only was wastewater data not collected every week at all sites in Virginia, but the sequencing frequency was even less, with some sites having no samples sequenced.

Using Diversity As A Metric

What is the motivation behind using diversity?

- RNA is bad at replicating itself, meaning that as the virus passes from person to person, the viral genome would change
- If we can produce a way to calculate this change, then we can use this rate to determine how many people have been infected

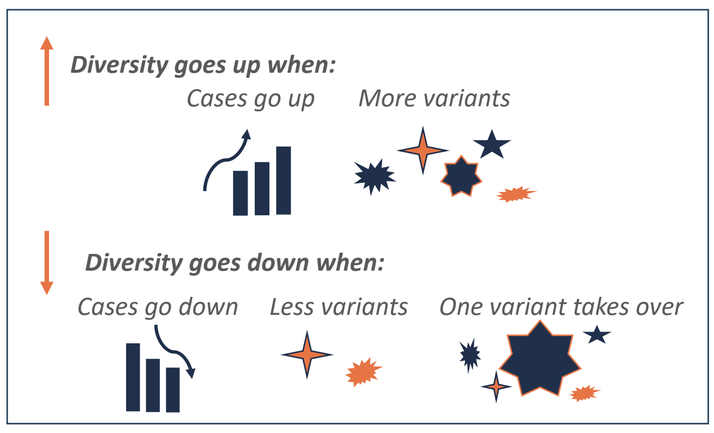


Figure #3

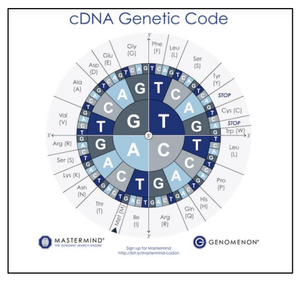
Since case data is more structured than wastewater, and we have an abundance of sequences from cases, we explored the diversity of sequences in cases first.

Analysis of Case Surveillance

A synonymous change has a much smaller effect on the viral genome than a change in other positions.

We hypothesize that the virus does not care about the 3rd position, where synonymous changes occur

- If so, then the change must be due to replication errors going from person to person than effects of positive selection



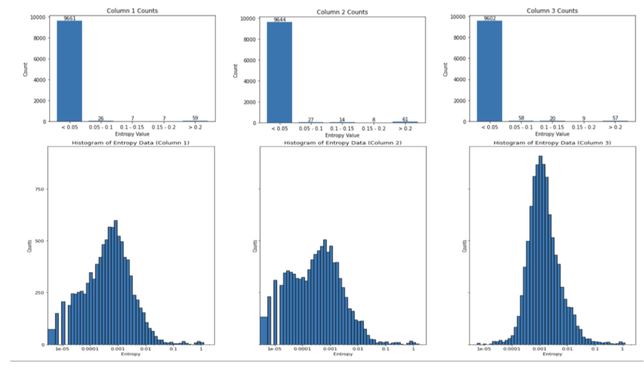


Figure #4

Figure 4. Distribution of column entropy. More positions in column three exhibit higher entropies than those in the other two columns on a logarithmic scale.

Column three has a higher correlation with cases than any other metric. The correlation varies because of how variants affect diversity. The first dip in correlation is from one variant taking over, decreasing diversity.

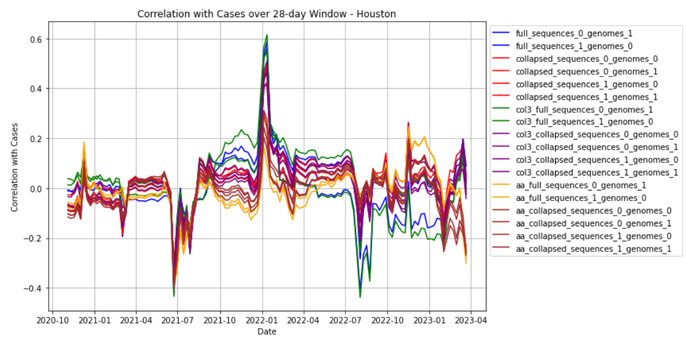


Figure #5

Future Work

We want to form an expectation of the entropy value based on the number of samples drawn.

Instead of using sequences grouped by location and time, we want to use phylogenetic tree connections to turn samples into a contact network.

Using this contact network, we can then estimate how many people were infected using the expected entropy