



Understanding the spread of SARS-CoV-2 clusters through an integrated pipeline using UShER, Cluster Tracker and StrainHub



Adriano de Bernardi Schneider¹, Colby T Ford^{2,3}, Jakob McBroome¹, Jennifer Martin¹, Daniel Janies², Yatish Turakhia⁴, Russel Corbett-Detig¹

1. Background

Response by the scientific community to the SARS-CoV-2 pandemic has created an unprecedented amount of genomic data that has to be processed and analyzed in a timely manner to have public health impact.

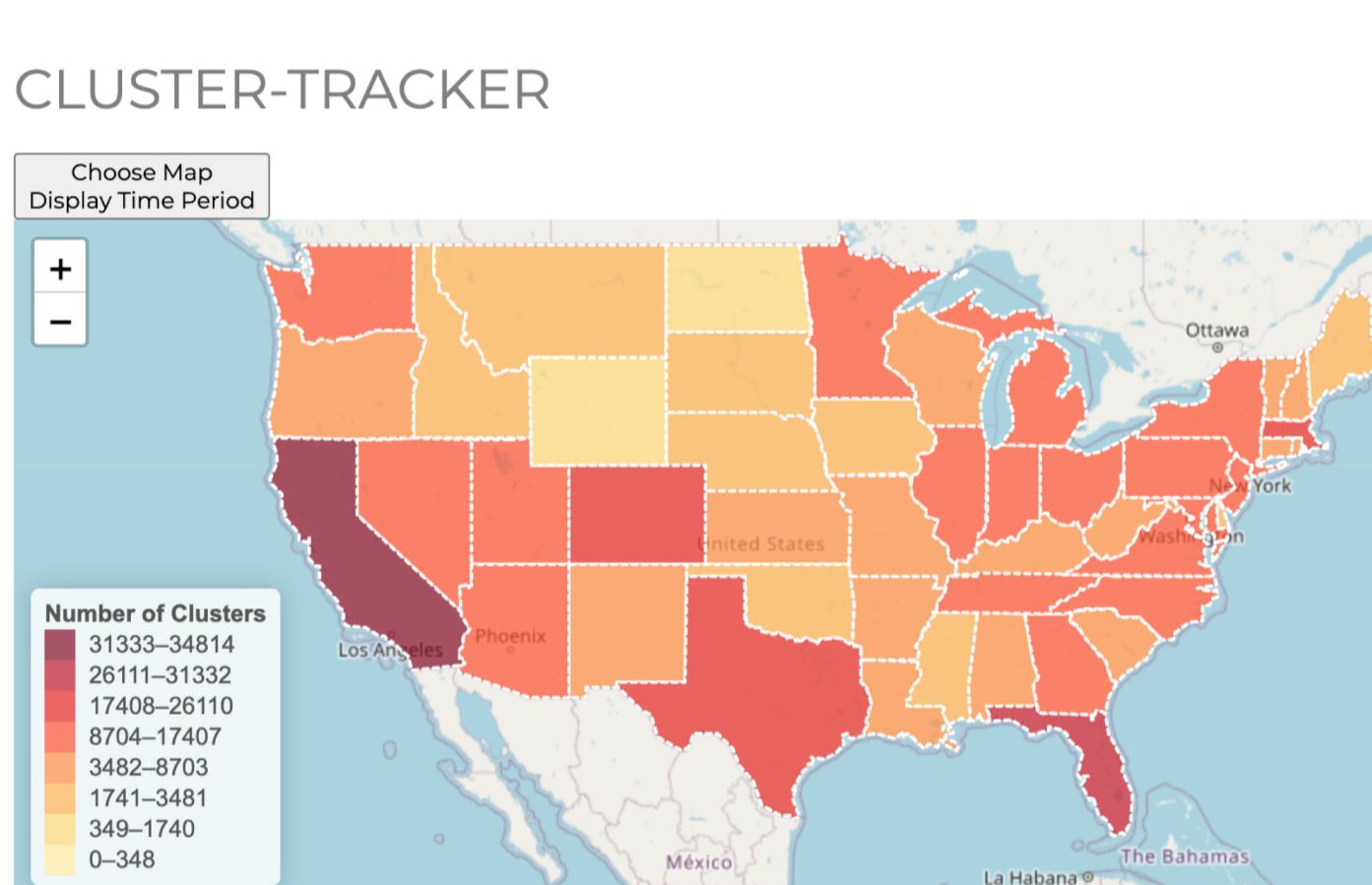
These circumstances push for a change in the current way genomic data has been evaluated in order to assist epidemiologists and public health officials make effective policy changes.

We integrate a pipeline to bring together three tools currently available and supported by our research groups to evaluate select SARS-CoV-2 clusters identified to understand the spread of these clusters in the United States.

2. Method



UShER rapidly places new SARS-CoV-2 genomic sequences to an existing phylogeny. Using the matUtils introduce tool, we identify clusters to be further studied using Cluster-Tracker.



Cluster-Tracker displays pre-calculated clusters with their inferred origin for US states and Puerto Rico. The tool also displays metrics such as inferred origin confidence and growth score.



StrainHub takes in the inferred origin of the clusters of interest selected in Cluster-Tracker and plots a network that includes the cluster of interest and surrounding clusters.

3. Clusters of interest

We selected four large clusters to display the capabilities of these tools.

Cluster 1: California_node_48404 / N = 11900 / Lineage AY.44 (Delta) / Inferred origin: indeterminate

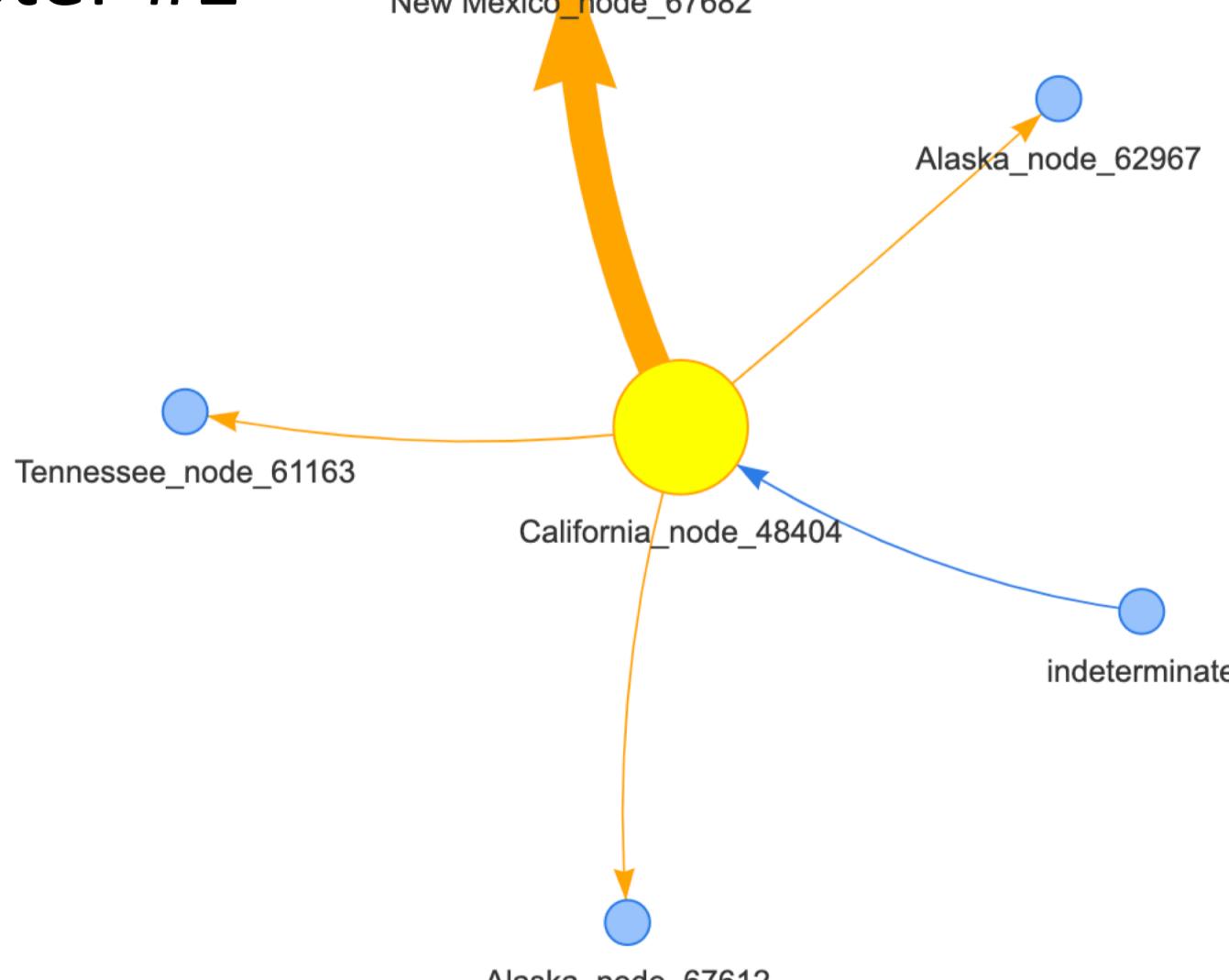
Cluster 2: Utah_node_31820 / N = 6005 / Lineage AY.44 (Delta) / Inferred origin: Colorado

Cluster 3: New York_node_776355 / N = 1998 / Lineage BA.2.12.1 (Omicron) / Inferred origin: Washington

Cluster 4: Minnesota_node_138024 / N = 475 / Lineage AY.103 (Delta) / Inferred origin: Michigan

4. Cluster networks

Cluster #1



Network of California Node 48404. Network shows that California received transmissions from an indeterminate source (non-US), while it seems to have originated four clusters in three distinct states. Nodes are scaled by Degree centrality (larger nodes = larger number of transmissions).

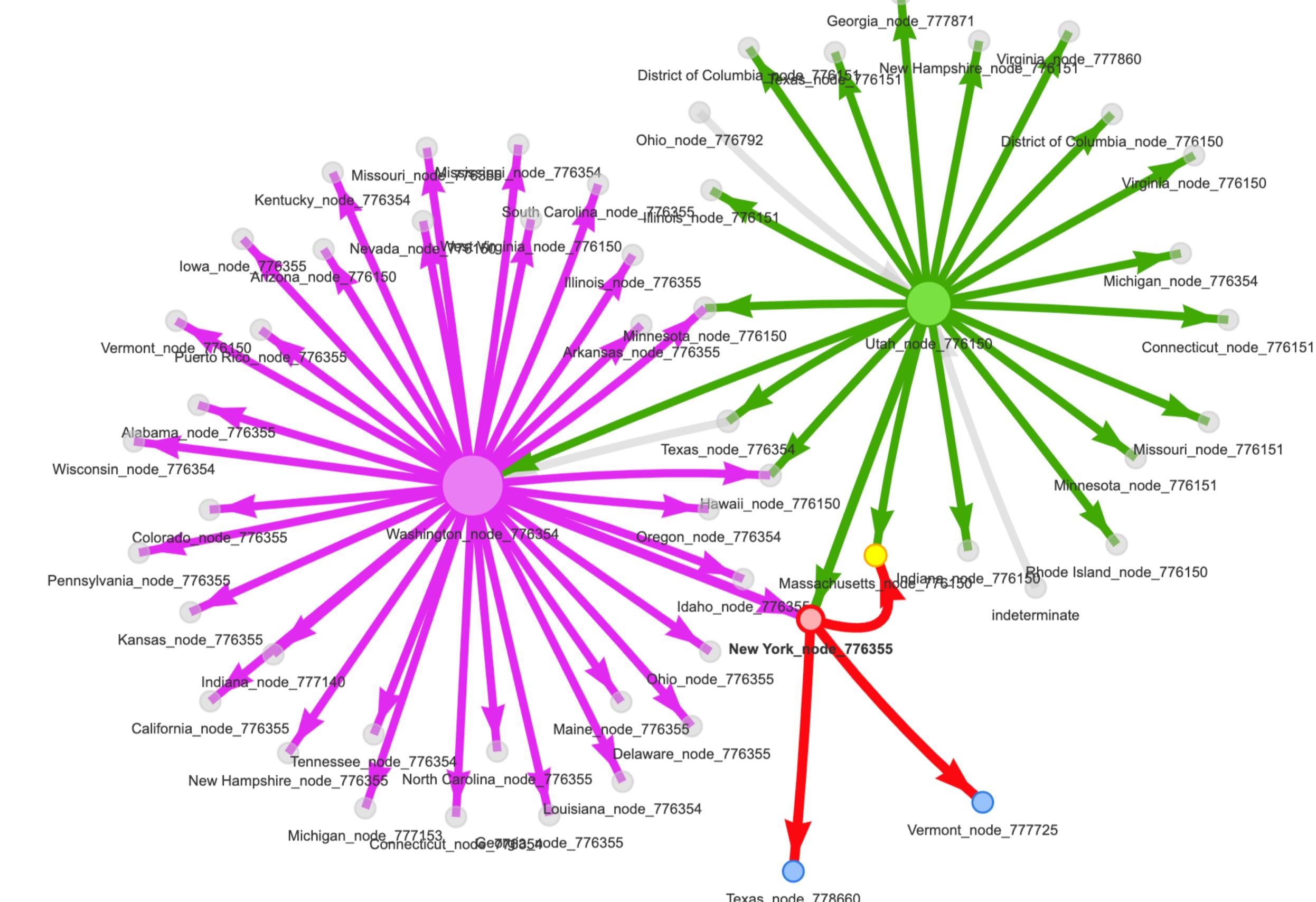
4. Cluster networks (continued)

Cluster #2



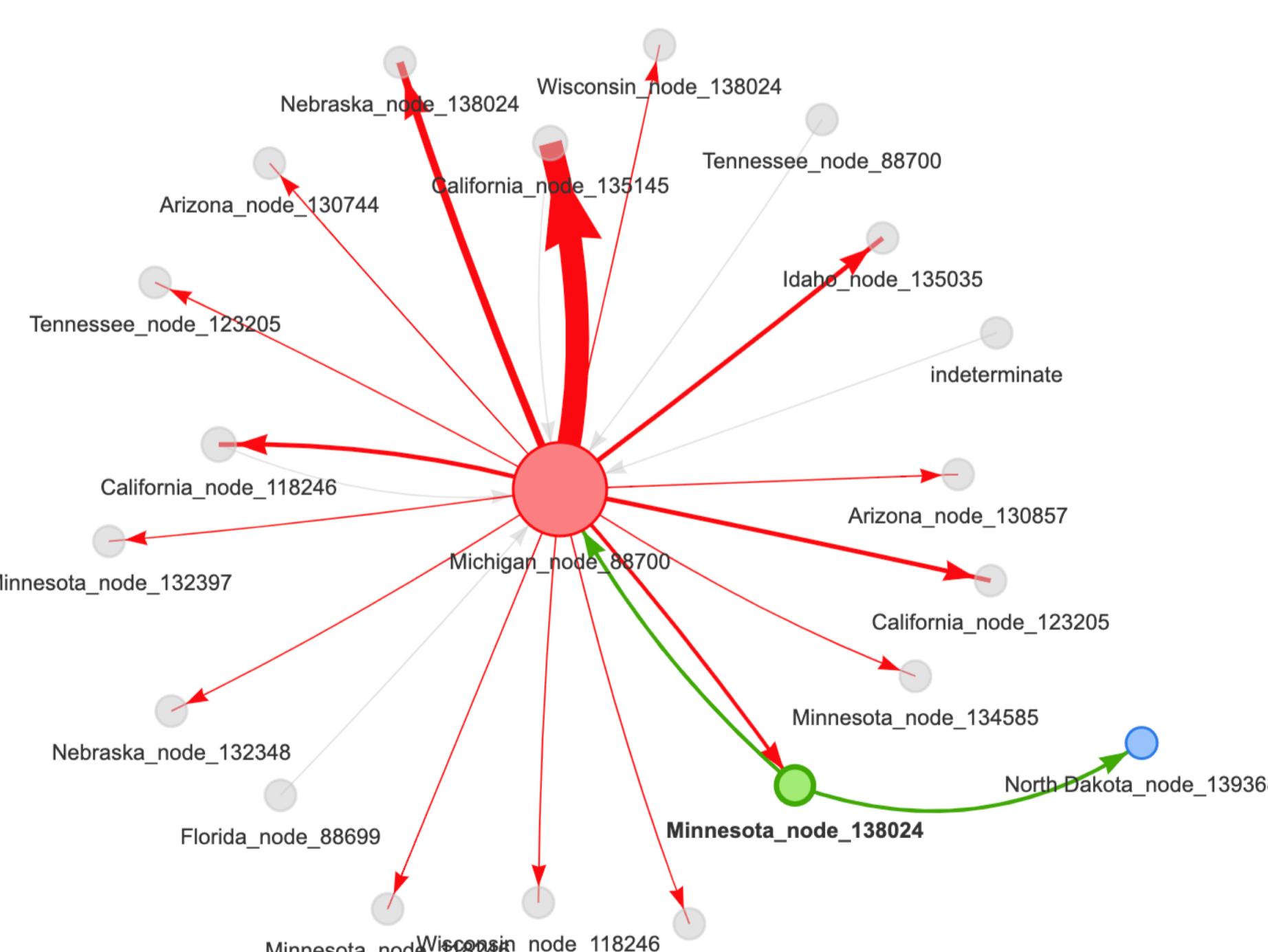
Network of Utah Node 31820. Network shows that Utah received transmissions from the neighboring states of Washington and Colorado, while it seems to have originated six new clusters in distinct states. Nodes are scaled by Degree centrality (larger nodes = larger the number of transmissions).

Cluster #3



Network of New York Node 776355. Network shows that New York received transmissions from clusters from Washington and Utah states, while it seems to have originated two new clusters in distinct states and contributed to the spread in a Massachusetts cluster. Nodes are scaled by Degree centrality (larger nodes = larger the number of transmissions).

Cluster #4



Network of Minnesota Node 138024. Network shows that Minnesota received transmissions from the state of Michigan and spread to North Dakota as well as back to Michigan. This network revealed a Michigan Node as central to multiple other clusters. Nodes are scaled by Degree centrality (larger nodes = larger the number of transmissions).

5. Implementation and data availability

UShER is available at <https://github.com/yatisht/usher/>. Cluster-Tracker is available at <https://clustertracker.gi.ucsc.edu/>. StrainHub is available as a standalone tool at strainhub.io, additionally it can be used through Docker or as an R package (see here: <https://docs.strainhub.io/>). The snakemake workflow encompassing the three tools can be downloaded from: <https://github.com/yatisht/usher/tree/master/workflows>.

¹ Biomolecular Engineering and Genomics Institute, University of California, Santa Cruz, United States of America; ² Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC, United States of America; ³ School of Data Science, University of North Carolina at Charlotte, Charlotte, NC, United States of America; ⁴ Electrical and Computer Engineering, University of California, San Diego, United States of America.