Augmenting Cluster-Tracker for Low Overhead, High Resolution Importations

Reid Farmer, Dr. Andrew Warren

Introduction

At UC Santa Cruz, researchers created a web application that can transform SARS-CoV-2 sample data into an interactive map of the United States that enables the user to visualize interstate pathogen cluster importations.

The functionality of this tool has great potential benefit to public health officials in regards to understanding the spread of disease. However, the tool would be of greater utility if users could more easily increase the resolution and visualize the spread at the county-level for a specified state.

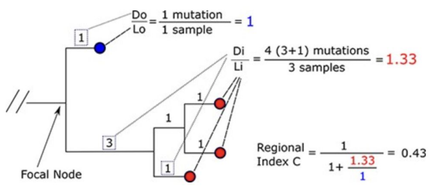
Background

The Cluster-Tracker utilizes phylogeny to transform millions of SARS-CoV-2 samples into a helpful bioinformatic tool. A cluster is a set of closely related samples from the same region and descended from a common ancestor with a regional introduction event. Introductions are identified in a three step process:

1. The samples are transformed into a **phylogenetic tree** via bioinformatic techniques.

2. A **post-order traversal dynamic programming** algorithm is conducted on the tree to compute the **regional index** for each node. The regional index is a heuristic defined by UCSC and is ultimately a weighted summary of a node's children within a phylogenetic tree.

3. An ancestry traversal is conducted for each sample in the tree. The most recent ancestor that possesses a regional index below the set threshold is inferred to be this lineage **introduction point**. Once an introduction point is identified for each sample, the samples are clustered by a shared introduction point.



$$Regional\ index\ (C) = \frac{1}{1 + \frac{\frac{D_i}{L_i}}{\frac{D_o}{L_o}}}$$

Figures 1 and 2. Images made by UCSC researchers to communicate 'Regional Index' heuristic

L_i: the number of downstream leaves that are in a given region

D_i: the minimum total branch length to a leaf descendent in the focal region

L_o and D_o are the same for out-of-region leaves

The Problem

For users wishing to visualize the pathogen spread throughout their state, there are two major limitations:

1. The software was designed to treat all admin geographic levels the same. This requires that users include their low-level data in the construction of the primary data structure, requiring that they understand the details of the workflow.
2. Increasing geographic resolution currently requires users to run expensive operations demanding high memory (500Gb) compute nodes.

The Solution

The strategy to solve this problem was **client-side recalculation**. Although the aforementioned computation is resource intensive and slow, the entire computation does not need to be rerun each time. Instead, in order to optimize accessibility, we can use existing national level results. Given user input that specifies the county source of samples, we can perform a client-side calculation that reassigns the introduction sample... without the expensive importation calculation.

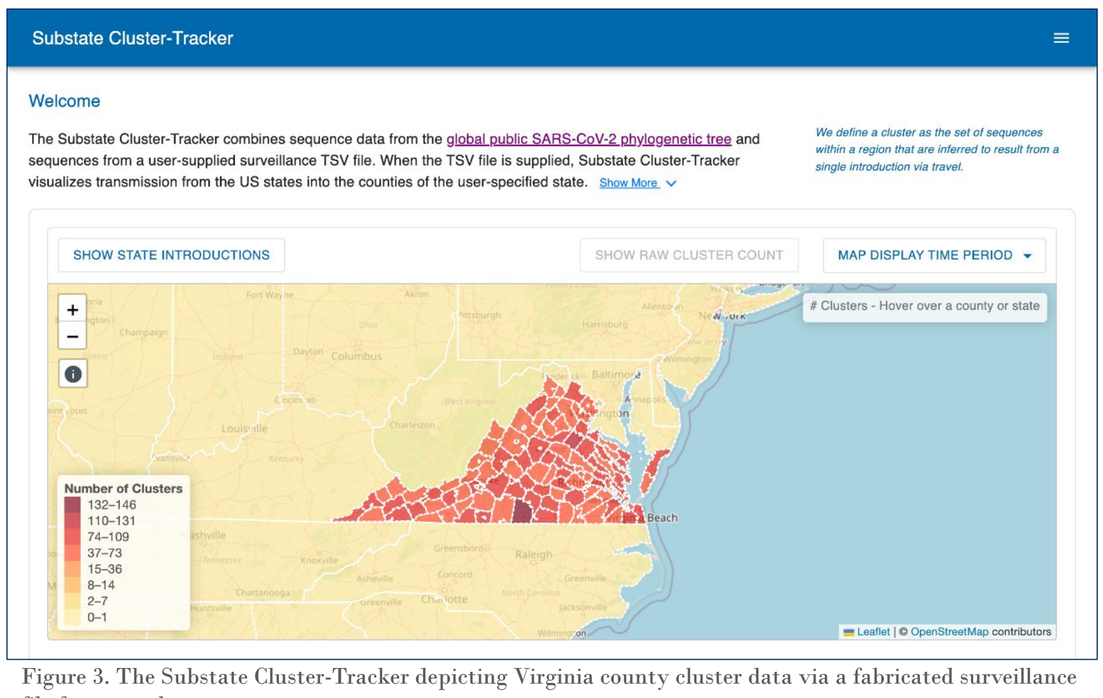


Figure 3. The Substate Cluster-Tracker depicting Virginia county cluster data via a fabricated surveillance file for example

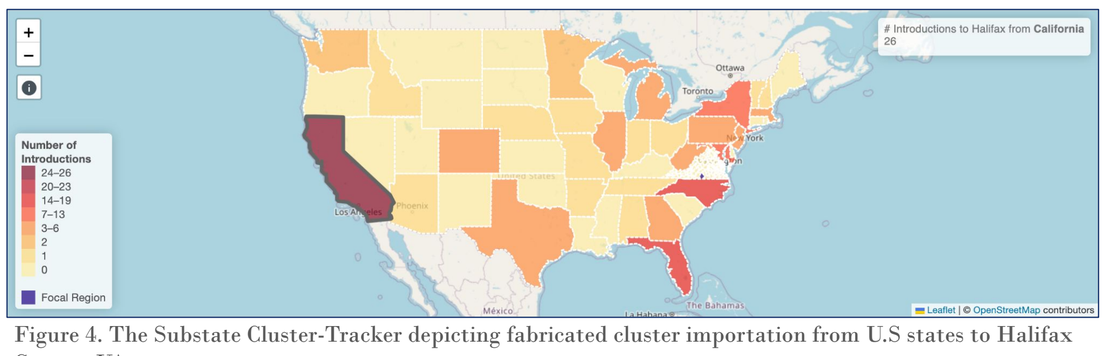


Figure 4. The Substate Cluster-Tracker depicting fabricated cluster importation from U.S states to Halifax County, VA

The Work

My contributions to the project include the following:

- Augmented Python scripts to properly generate data structure responsible for relevant geographical boundaries
- Strategically pinpointed optimal workflow stages for client-side recalculations
- Optimized both synchronous and asynchronous communication between client and server
- Designed and implemented JavaScript solutions to seamlessly recalculate and reassign importations on the client
- Overhauled BASH scripts to properly configure both new and previously existing workflows
- Containerized entire workflow with Apptainer and Docker to ensure portability and reproducibility

Future Work

Revising the original computation to enable origin specification, enabling the following two modes:

1. Intrastate transmission
2. County to outside state transmission

References

Jakob McBroome, *introduction-website*, 2021, https://github.com/pathogen-genomics/introduction-website

Jakob McBroome, Jennifer Martin, Adriano de Bernardi Schneider, Yatish Turakhia, Russell Corbett-Detig, Identifying SARS-CoV-2 regional introductions and transmission clusters in real time, *Virus Evolution*, Volume 8, Issue 1, 2022, veac048, https://doi.org/10.1093/ve/veac048

Special Thanks

Thank you to my mentor Dr. Andrew Warren, the Cluster-Tracker team, especially Dustin Machi and Dr. Bryan Lewis, and the G4GC program, such as Lucius Lichte, Erin Raymond, and Savanna Galambos

UNIVERSITY of VIRGINIA
BIOCOMPLEXITY INSTITUTE