# Generating Count and Tree Files for TreeScan

## Project File Inventory

### `BCD006\_GitHub\_TreeScan\_count\_and\_tree\_files.sas` - SAS program that reads in sample data, aggregates cases by lineage (or variant or allele code) and specimen collection date, and outputs count and tree files

### `BCD006\_GitHub\_sample\_data.sas7bdat` - Simulated sample data with specimen collection date and lineage, the minimum required fields to perform TreeScan analysis, along with simulated unique ID

### `BCD006\_GitHub\_sample\_count\_file.txt` - Tab-delimited count file with one row per unique lineage (column 1) and collection date (column 2) with aggregated case count (column 3)

### `BCD006\_GitHub\_sample\_count.sas7bdat` - Simulated source data for count file

### `BCD006\_GitHub\_sample\_tree\_file.txt` - Tab-delimited hierarchical tree file with all node (column 1) and parent (column 2) relationships necessary to construct tree with one root that contains all lineages in count file and their closest node one level up on the tree

### `BCD006\_GitHub\_sample\_tree.sas7bdat` - Simulated source data for tree file

### `BCD006\_GitHub\_sample\_parameters.prm` - TreeScan parameter file for use with sample count and tree files. Parameters are set to run a prospective analysis using the tree and time scan options, aggregated by day, and conditioning on node and time. To run an analysis using these parameter settings on the simulated count and tree files, please update the Tree File and Count File fields on the “Input” tab and Results File field on the “Output” tab to reflect the location of the downloaded sample input files.

## Summary

### [TreeScan](www.treescan.org) is freely downloadable software that implements the tree-based scan statistic, a data mining method primarily used by FDA and academic scientists for drug and vaccine safety surveillance. To identify unexpected adverse reactions, this method scans, for example, a hierarchical tree of ICD-10 codes, looking for excess risk over any time range after drug or vaccine exposure of any potential adverse event and groups of related adverse events, while adjusting for multiple testing.

### NYC DOHMH Bureau of Communicable Disease (BCD) research scientists engaged the TreeScan software developers and fundraised to support TreeScan public release v.2.0.0, which added a prospective analysis option to facilitate scanning for emerging clusters in whole-genome sequencing (WGS) data.

### BCD is currently using TreeScan prospectively to detect emerging SARS-CoV-2 variants and is preparing to launch weekly analyses of Salmonella WGS data.