## Romblon transmission tree-related processing/analysis/plotting scripts

All in outbreak\_romblon\_PH/R

## Initial data processing

1. **process\_population\_size\_data.R**
   * Function: Extracts pop size data and centroid coordinates (UTM) for all 100m2 cells within Romblon province
   * Outputs: output/population\_sizes/Romblon\_pop\_by\_cell.csv
   * Run again in case: only if you want to change the underlying population size dataset
2. **process\_outbreak\_dat.R** (Ellie’s data cleaning script)
   * Function: Pulls outbreak linelist data from google sheet & does initial data cleaning
   * Outputs: data/raw/dat\_\*.csv; data/clean/dat\_human.csv; data/clean/dat\_animal.csv; data/clean/dat\_outbreak\_all.csv;
   * Run again in case: you need to pull new data from the Google sheet, or want to add an additional data column to the clean dataset.
3. **patristic\_dist.R** and **process\_genetic\_data.R** 
   * Function: Generation of patristic distance matrix from phylogenetic tree, cluster analysis, and lineage assignment
   * Outputs: output/phylogenetic\_distances/romblonSeq\_24\_patristicDist\_matrix.txt; output/figures/Hamming\_distance\_frequencies.jpeg; output/phylogenetic\_distances/0.002clust.csv
   * Run both scriptsagain in case of new genetic data (a new phylogenetic tree); run process\_genetic\_data.R again in case you just want to change the clustering cut-off.
4. **process\_outbreak\_dat\_for\_Romblon\_tm\_trees.R**
   * Function: Further data processing and filtering of dat\_animal.csv, to prepare data for transmission tree reconstruction
   * Outputs: data/clean/dat\_animal\_for\_Romblon\_tm\_trees.csv
   * Run again in case:
     + You want to include a different set of cases (lines 41-45 for date filtering)
     + You want to change included columns, dates, lineage assignment, any other data

## Transmission trees + related plots

1. **run\_treerabid.R**
   * Function: sets up a table with different scenarios (parameter combinations), and runs treerabid functions (boot\_trees, build\_consensus\_links, build\_consensus\_tree) over these in parallel, to reconstruct transmission trees
   * Requires **boot\_trees\_simulate\_location.R** for an adapted boot\_trees version, that simulates case locations within each bootstrap
   * Outputs: output/tm\_trees/trees\_all.gz; output/tm\_trees/links\_consensus\_raw.csv; output/tm\_trees/links\_consensus\_consistent.csv; output/tm\_trees/scenarios.csv
   * Run again in case: you’ve made changes to the dataset and want to rerun analyses
2. **plot\_tm\_trees\_comparisons.R**
   * Function: creates comparative plots with trees for different scenarios
   * Requires **plot\_lineage\_ts.R** for plotting
   * Outputs: output/figures/consensus\_tree\_check.jpeg, output/figures/consensus\_tree\_subset.jpeg
   * Run again in case: you’ve updated the underlying transmission trees and want plots to reflect latest datasets
3. **plot\_tm\_trees\_epi\_gen\_sim\_loc\_pruned.R**
   * As above but for epi + genetic data / simulated locations / pruned by distance and time 0.99
   * Uses simulated case locations from the MCC tree for mapping purposes
   * Outputs: various plots and animation in output/figures/epi\_gen\_simulated\_locations\_prunedDT99