Genomic Epidemiology: An Emerging Tool for Public Health

Christopher Carter
Master of Public Health, Kansas State University

Kansas State University
College of Veterinary Medicine
Master of Public Health Program

Genomic Epidemiology

Genomic epidemiology is the use of pathogen genomic sequencing and epidemiological data to better understand outbreaks and disease transmission dynamics. The use of genomic sequencing data can help public health practitioners understand, at a fine resolution, how cases are linked (figure 1).

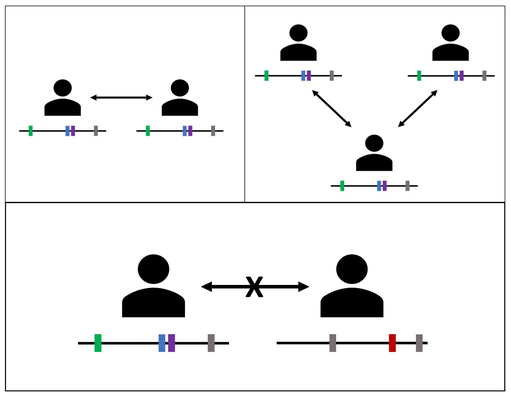

Figure 1; Cases with similar sequences are likely linked while those with distinct sequences are not. In the absence of other epidemiological data, however, it is difficult to infer a transmission chain.

Genomic epidemiology is also useful for guiding public health decisions. The resolution provided by genomic sequencing data can focus public health efforts and policy toward preventing the primary routes of introduction or transmission. For example, genomic epidemiology can be used to determine if increased case incidence is due to community transmission (similar sequences) or due to multiple introductions (distinct sequences). Knowing which form is predominant can inform intervention strategies.

Phylogenetic Trees

Phylogenetic trees offer a means to easily visualize your genomic sequencing data and infer links between samples. However, they are incapable of summarizing all relevant information. To maximize their utility to the public health practitioner, phylogenetic trees should be annotated with other epidemiological data (figure 2). Platforms that can create phylogenetic trees from whole genome sequence (WGS) data include the National Center for Biotechnology Information (NCBI) Pathogen Detection database, MicrobeTrace, and Nextstrain (see Additional Resources).

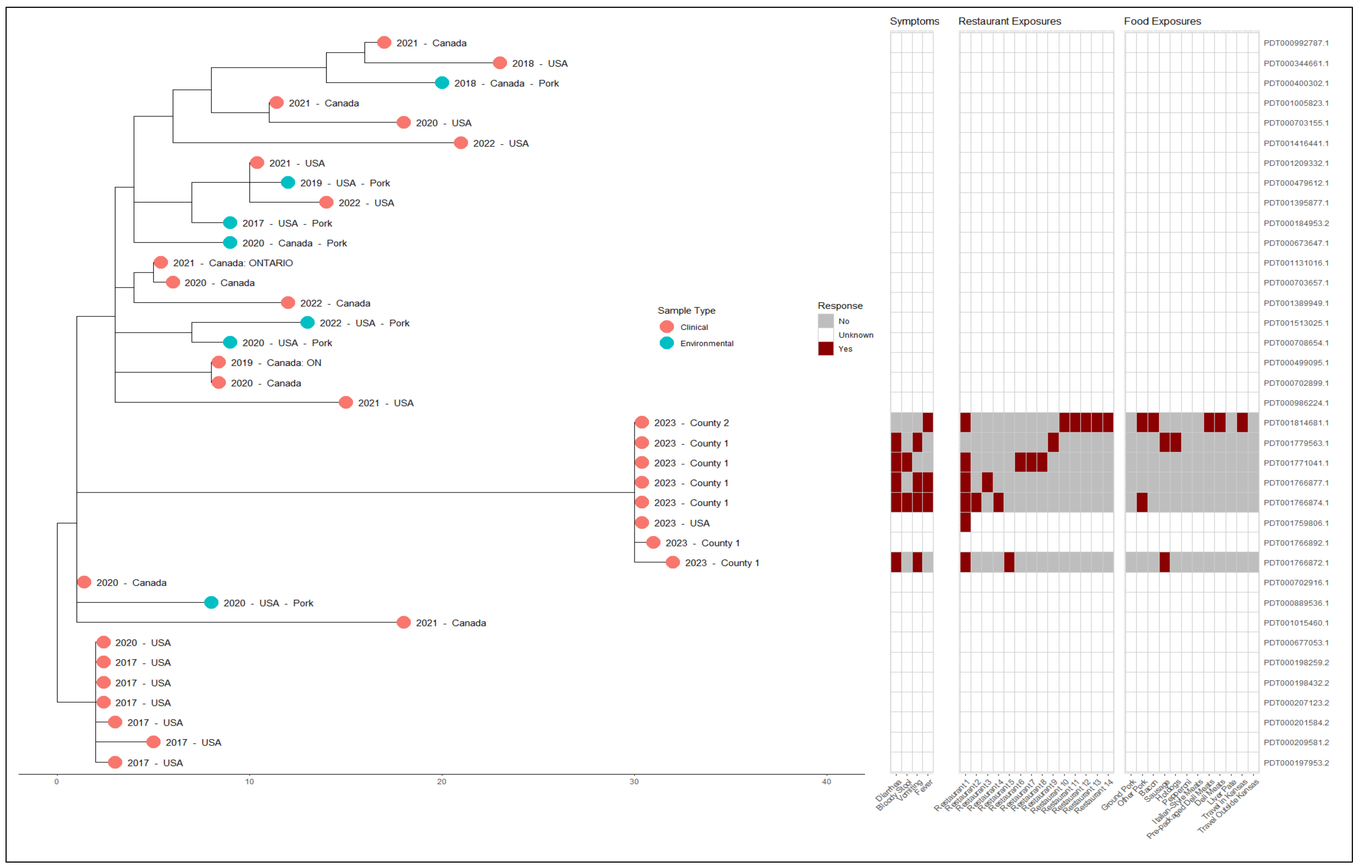

Figure 2; Example of an annotated tree. Tree tips are colored according to sample type and labeled with epidemiological information. The heat maps to the right display symptom and exposure data for each sample.

Interpreting Phylogenetic Trees

Phylogenetic trees are composed of four parts, the root, internal nodes, branches, and tips (figure 3). Tips represent the observed consensus sequences. Tips are connected by internal nodes that represent hypothetical common ancestors. Branches connect tips to the nearest node. The root of the phylogenetic tree represents the inferred common ancestor for all displayed samples. The x-axis and branches of a phylogenetic tree can be weighted to either the number of nucleotide diversions from the root (root-to-tip distance) or time. The y-axis has no meaning and exists only to provide space between the tips of a tree for readability.

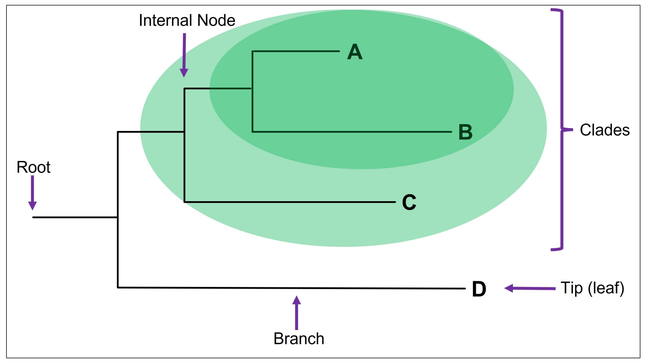

Figure 3; Anatomy of a rooted phylogenetic tree.

Phylogenetic trees organize sequences into groups called clades (figure 3). Sequences within a clade all share a common ancestor. A single sequence may belong to many clades. Sequences that belong to the same clade share genetic similarities and are more likely to be epidemiologically linked to each other than to sequences in other clades.

Additional Resources

An applied genomic epidemiology handbook **by Allison Black and Gytis Dudas — https://alliblk.github.io/genepi-book/index.html**

CDC COVID-19 Genomic Epidemiology Toolkit — https://www.cdc.gov/amd/training/covid-19-gen-epi-toolkit.html

MicrobeTrace — https://microbetrace.cdc.gov/MicrobeTrace/

NCBI Pathogen Detection Database — https://www.ncbi.nlm.nih.gov/pathogens/

Nextstrain — https://docs.nextstrain.org/en/latest/

Acknowledgments

The author thanks John Anderson for his invaluable and continued mentoring, support, and guidance.