Master of Public Health

Applied Practice Experience

by

Christopher Alan Carter

MPH Candidate

submitted in partial fulfillment of the requirements for the degree

MASTER OF PUBLIC HEALTH

Graduate Committee:
Dr. Justin Kastner
Dr. Ellyn Mulcahy
Dr. Thomas Platt

Applied Practical Experience Site:
Kansas Department of Health and Environment (KDHE)
Genomic Epidemiology Program
10 July 2023–11 August 2023

Applied Practical Experience Preceptor:
John Anderson, MSc, PhD

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Table of Contents

Chapter 1 - Portfolio Products

In October 2021, the Kansas Department of Health and Environment (KDHE) established its Genomic Epidemiology program with the mission of "analyzing, interpreting, and facilitating the use of genomic data that has public health importance to protect the health of Kansans."[1] The Genomic Epidemiology program achieves this mission by bridging the pathogen genomic sequencing of the Kansas Health and Environmental Laboratory (KHEL) and the disease investigations of KDHE's Infectious Disease Epidemiology and Response (IDER) section within the Bureau of Epidemiology and Public Health Informatics (BEPHI). KHEL regularly sequences the genomes of the following pathogens: SARS-CoV-2, *Shigella* spp., *Salmonella* spp., *Campylobacter* spp., *Escherichia coli*, Carbapenem-resistant *Acinetobacter baumannii* (CRAB), influenza, tuberculosis, and *Neisseria gonorrhoeae*. These pathogens are typically isolated and sequenced from patient samples (e.g., stool) but may also originate in environmental samples (e.g., wastewater). Once KHEL isolates, sequences, and assembles the genomic data, a bioinformatician uploads the genomic sequences to the National Center for Biotechnology Information (NCBI) for use in the Genomic Epidemiology Program.[2] The genomic epidemiologists then visualize the genomic sequences in phylogenetic trees annotated with other pertinent epidemiological data. Annotating the phylogenetic trees with additional data often requires collaboration with BEPHI epidemiologists and disease investigators. Through the collaboration between KHEL, the Genomic Epidemiology Program, and BEPHI epidemiologists, KDHE effectively generates a holistic and detailed understanding of disease cases and outbreaks.

Between July 10 and August 11, 2023, I conducted my Applied Practice Experience (APE) at the KDHE Genomic Epidemiology Program, precepted by program director Dr. John Anderson. The main goal of my APE was to explore the integration of genomic epidemiology with food-borne disease investigations by reviewing literature on the reasoning and methods of genomic epidemiology, creating a process map of the food-borne disease investigation process, and developing an annotated phylogenetic tree based on a food-borne disease outbreak. These three objectives correlate to my first three portfolio products. After reviewing genomic

[1] John Anderson, "Genomic Epidemiology" (presentation, IDER monthly staff meeting, Topeka, KS, 10 March 2023).

[2] While KHEL sequences all of the aforementioned pathogens, the Genomic Epidemiology Program typically only works with SARS-CoV-2, influenza, and CRAB sequences.

epidemiology literature, I wrote the *KDHE Genomic Epidemiology Field Guide*, a shareable, plain-language document that summarizes the uses and methods of genomic epidemiology for current public health practitioners. Once I was oriented to the field of genomic epidemiology, I began meeting with KDHE personnel responsible for key steps of the food-borne disease investigation process. Following these conversations, I developed a process map—a systems thinking tool for understanding complex processes—to better understand what steps could be aided by incorporating genomic epidemiology. Then, to demonstrate the effectiveness of genomic epidemiology in food-borne disease investigations, I created an example annotated phylogenetic tree using a recent food-borne disease outbreak.

The final three portfolio products communicate the information from the above products. On August 11, 2023—the final day of my APE—I presented the field guide, process map, and annotated tree to the monthly IDER staff meeting. On September 21, 2023, I presented a poster titled "Genomic Epidemiology: An Emerging Tool for Public Health" at the Kansas Public Health Association (KPHA) Annual Meeting. Finally, after my APE, Dr. Anderson and Mayra Trujillo, a genomic epidemiologist at KDHE, created an online training program based on the KDHE Genomic Epidemiology Field Guide. Dr. Anderson and Ms. Trujilo then authored an abstract to be considered for presentation at the 2024 Council of State and Tribal Epidemiologists (CSTE) Annual Conference.[3] The abstract is still pending approval. In addition to the above products, I participated in various activities throughout the APE including online training (e.g., HIPAA Awareness and EpiTrax) and daily meetings.

Table 1.1 Summary of Portfolio Products

Portfolio Product		**Description**
1	*KDHE Genomic Epidemiology Field Guide*	A 15-page document designed to introduce public health practitioners to the field of genomic epidemiology. The document guides the user through the following chapters: Genomic Epidemiology 101, Data Sources, Sampling, Making Phylogenetic Trees, Interpreting Phylogenetic Trees,

[3] While listed as an author on the abstract for my contribution to the project, I did not write the abstract text.

		Additional Resources, Glossary, and References.
2	Kansas food-borne disease investigation process map	A process map showing the order of steps that KDHE and the Kansas Department of Agriculture (KDA) take to investigate a food-borne disease complaint or outbreak.
3	Salmonellosis Annotated Phylogenetic Tree	A figure combining the phylogenetic tree of genomic data from patients of a recent food-borne disease outbreak with key results from KDHE's food-borne disease survey.
4	Presentation to IDER monthly staff meeting	A slide deck presenting portfolio products 1-3 along with biographical information.
5	Genomic Epidemiology: An Emerging Tool for Public Health	A poster summarizing the information presented in the *KDHE Genomic Epidemiology Field Guide*.
6	CSTE Annual Meeting Abstract	An abstract submitted for presentation at the 2024 CSTE Annual Meeting. This abstract shares the creation of an online training based on the *KDHE Genomic Epidemiology Field Guide.*

Table 1.2 Portfolio Products and Competency Addressed

Portfolio Product		**Number and Competency Addressed**	
1	*KDHE Genomic Epidemiology Field Guide*	18 19	Select communication strategies for different audiences and sectors. Communicate audience-appropriate public health content, both in writing and through oral presentation.
2	Kansas food-borne disease investigation process map	22	Apply systems thinking tools to a public health issue.
3	Salmonellosis Annotated Phylogenetic Tree	3	Analyze quantitative and qualitative data using biostatistics, informatics, computer-based programming and software, as appropriate.

		4	Interpret results of data analysis for public health research, policy or practice.
4	Presentation to IDER monthly staff meeting	18 19	Select communication strategies for different audiences and sectors. Communicate audience-appropriate public health content, both in writing and through oral presentation.
5	Genomic Epidemiology: An Emerging Tool for Public Health	18 19	Select communication strategies for different audiences and sectors. Communicate audience-appropriate public health content, both in writing and through oral presentation.
6	CSTE Annual Meeting Abstract	18 19	Select communication strategies for different audiences and sectors. Communicate audience-appropriate public health content, both in writing and through oral presentation.

Chapter 2 - Competencies

The portfolio products described in Chapter 1 allowed me to further develop several MPH Foundational Competencies, notably competencies #3, #4, #18, #19, and #22.

Table 2.1 Summary of MPH Foundational Competencies

Number and Competency		Description
3	Analyze quantitative and qualitative data using biostatistics, informatics, computer-based programming and software, as appropriate	Creation of the phylogenetic tree required analysis and visualization of genomic and survey data using R and RStudio
4	Interpret results of data analysis for public health research, policy or practice	After creating the phylogenetic tree, I inferred a likely cause of the outbreak.
18	Select communication strategies for different audiences and sectors	In creating the *Genomic Epidemiology Field Guide*, IDER presentation, genomic epidemiology poster, and the CSTE abstract, I chose different tones and approaches to presenting information based on assumptions of the intended audience's prior knowledge.
19	Communicate audience-appropriate public health content, both in writing and through oral presentation	The *Field Guide*, IDER presentation, poster, and abstract all communicate public health information in a variety of means and to a variety of audiences.
22	Apply systems thinking tools to a public health issue	By creating and using the food-borne disease process map, I could better understand the existing investigation process and decide where integration of genomic epidemiology would be the most effective.

Table 2.2 MPH Foundational Competencies Course Mapping

22 Public Health Foundational Competencies Course Mapping	MPH 701	MPH 720	MPH 754	MPH 802	MPH 818
Evidence-based Approaches to Public Health					
1. Apply epidemiological methods to the breadth of settings and situations in public health practice	x		x		
2. Select quantitative and qualitative data collection methods appropriate for a given public health context	x	x	x		
3. Analyze quantitative and qualitative data using biostatistics, informatics, computer-based programming and software, as appropriate	x	x	x		
4. Interpret results of data analysis for public health research, policy or practice	x		x		

22 Public Health Foundational Competencies Course Mapping	MPH 701	MPH 720	MPH 754	MPH 802	MPH 818
Public Health and Health Care Systems					
5. Compare the organization, structure and function of health care, public health and regulatory systems across national and international settings		x			
6. Discuss the means by which structural bias, social inequities and racism undermine health and create challenges to achieving health equity at organizational, community and societal levels					x
Planning and Management to Promote Health					
7. Assess population needs, assets and capacities that affect communities' health		x		x	
8. Apply awareness of cultural values and practices to the design or implementation of public health policies or programs					x
9. Design a population-based policy, program, project or intervention			x		
10. Explain basic principles and tools of budget and resource management		x	x		
11. Select methods to evaluate public health programs	x	x	x		
Policy in Public Health					
12. Discuss multiple dimensions of the policy-making process, including the roles of ethics and evidence		x	x	x	
13. Propose strategies to identify stakeholders and build coalitions and partnerships for influencing public health outcomes		x		x	x
14. Advocate for political, social or economic policies and programs that will improve health in diverse populations		x			x
15. Evaluate policies for their impact on public health and health equity		x		x	
Leadership					
16. Apply principles of leadership, governance and management, which include creating a vision, empowering others, fostering collaboration and guiding decision making		x			x
17. Apply negotiation and mediation skills to address organizational or community challenges		x			
Communication					
18. Select communication strategies for different audiences and sectors	DMP 815, FNDH 880 or KIN 796				
19. Communicate audience-appropriate public health content, both in writing and through oral presentation	DMP 815, FNDH 880 or KIN 796				
20. Describe the importance of cultural competence in communicating public health content		x			x
Intreprofessional Practice					
21. Perform effectively on interprofessional teams		x			x
Systems Thinking					
22. Apply systems thinking tools to a public health issue			x	x	

Chapter 3 - Student attainment of MPH Emphasis Area Competencies

Table 3.1 Summary of MPH Emphasis Area Competencies

MPH Emphasis Area:		
Number and Competency		**Description**
IDZ 1	Evaluate modes of disease causation of infectious agents.	This study examines drought as a risk factor for a variety of diseases and health-related topics.
IDZ 2	Investigate the host immune response to infection.	In considering various health-related consequences of drought, the host immune response influences disease causation, especially for issues related to air quality.
IDZ 3	Apply epidemiological methods to the breadth of settings and situations in public health practice	This study calculates and interprets such epidemiological measures as incidence using ecological methods.
IDZ 4	Examine the influence of environmental and ecological forces on infectious disease.	This study examines drought as a risk factor for a variety of health-related topics, including infectious diseases.
IDZ 5	Analyze disease risk factors and select appropriate surveillance.	In the conclusion of this study, I argue for the most salient health-related consequence of drought and propose several methods of response and mitigation.

Appendices

Appendix I – Portfolio Product 1

Appendix II – Portfolio Product 2

Appendix III – Portfolio Product 3

Appendix IV – Portfolio Product 4

Appendix V – Portfolio Product 5

Appendix VI – Portfolio Product 6

Genomic Epidemiology Field Guide

Contents

Introduction to this Resource

This guide introduces the field of genomic epidemiology and is intended for public health practitioners at any jurisdictional level in Kansas and beyond. First, the "Genomic Epidemiology 101" section introduces you to the biological basis of genomic epidemiology and the benefits of incorporating genomic analysis into existing disease investigation procedures. The remaining sections guide you through the workflow of analyzing genomic sequence data starting with "Data Sources" and "Sampling". Next, the "Making Phylogenetic Trees" section presents a variety of methods for creating phylogenetic trees. "Annotating Trees with Epidemiological Data" then discusses the value of supplementing your phylogenetic trees with other data from your investigation and how to display epidemiological relationships on your tree. The "Interpreting Phylogenetic Trees" section provides an overview of how to interpret the phylogenetic and epidemiological relationships displayed in phylogenetic trees. Finally, the "Additional Resources" section highlights key sources of information related to genomic epidemiology and the methods presented here.

Genomic Epidemiology 101

Genomic epidemiology is the use of pathogen genomic sequencing and epidemiological data to better understand outbreaks and disease transmission dynamics. This section will explore how pathogens acquire mutations and how you can use sequence data to aid epidemiologic investigations and guide public health practice.

Acquiring Mutations

Pathogens acquire new genetic material through recombination, transcriptional errors, conjugation, and damage. These genetic changes can be deleterious, neutral, or beneficial. Deleterious changes are those that reduce pathogen survival and replication, neutral changes are neither harmful nor beneficial, and beneficial changes facilitate pathogen survival and replication. Throughout the course of an infection, populations of a pathogen accrue neutral and beneficial changes while deleterious changes are out-competed. A sample of this in-host genetic diversity is transferred from one host to the next during a transmission event. If the changes continue to benefit the pathogens in the new host, the changes are conserved and transmitted to the next host. In this way, genetic changes in the pathogen link hosts of the same transmission chain.

Mutation Notation

Genetic mutations are notated according to one of the following standardized nomenclatures:

- Protein: original amino acid, the numeric location of the amino acid, new amino acid
- Gene: original nucleotide, the numeric location of the nucleotide, new nucleotide.

For example, a mutation shared by XBB.1.5 variants of SARS-CoV-2 is notated as S:S486P. This mutation occurs in the spike (S) protein in which a serine encoded (in part) by nucleotide 486 was changed to a proline. Similarly, a hypothetical mutation in the *stx1* gene of *E. coli* at

nucleotide 34 in which adenosine was erroneously substituted for guanine would be notated as stx1:G34A.

Using Sequencing Data in a Disease Investigation

The use of genomic sequencing data can help you to understand, at a fine resolution, how cases are linked. This resolution is helpful for both disease surveillance and outbreak investigation. For example, imagine that there are several new cases of COVID-19 in your community and the epidemiologic curve suggests a point-source outbreak with some sustained transmission. Upon the addition of genomic sequencing data, you conclude that what appeared to be a single introduction event was multiple introductions with little sustained transmission. As another example, consider three separate outbreaks at local restaurants. You would like to know if the three outbreaks are related. After analyzing the sequence data obtained from the cases, you find that the sequences obtained from restaurants 1 and 3 are very similar to each other and the sequences from restaurant 2 are distinct. You then conclude that the outbreaks in restaurants 1 and 3 are likely connected while the outbreak in restaurant 2 is unrelated. Genomic sequencing data can thus be powerful tools for understanding outbreaks and disease transmission, especially when paired with traditional epidemiological data.

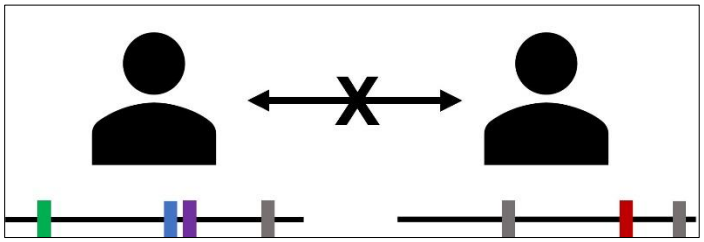


Figure 1; Cases with distinct sequences are unlikely to be linked to each other.

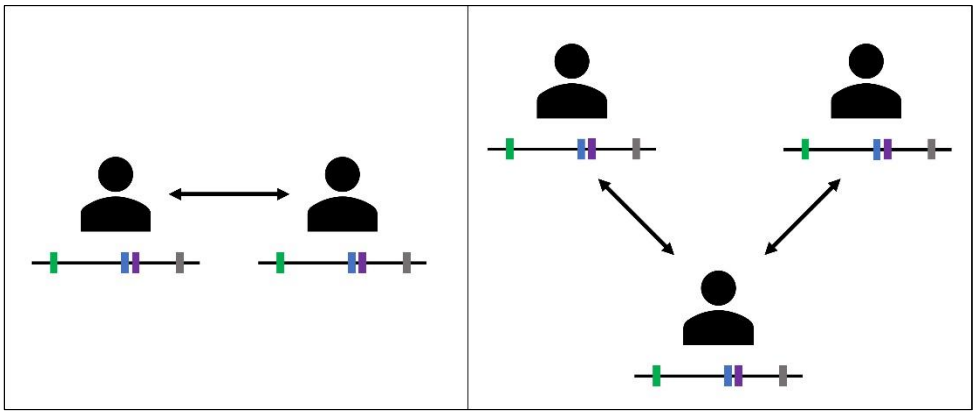


Figure 2; Cases with similar sequences are likely linked. In the absence of other epidemiological data, however, it is difficult to infer a transmission chain.

There are limitations to what genomic data can tell you. Genomic data cannot be used to definitively rule-in links between cases. This is due largely because transmission events do not

always occur simultaneously with genetic changes and because all cases are rarely sequenced. You can, however, use genomic data to rule-out case linkages. To illustrate this, imagine two sick roommates, one of whom became ill 3 days before the other. You might assume that the first roommate transmitted their infection to the other. With genomic sequencing data, you find that the sequences are distinct from one another and more similar to other cases in the community. You then conclude each roommate acquired their infection elsewhere and one did not infect the other (figure 1). Similarly, the two sequences being similar or identical is not enough evidence to conclude sequential transmission (figure 2). This is because sequential transmission events are rarely distinguishable from events in which two people acquire an infection at similar times from the same source. These limitations may be overcome with more sophisticated analyses in the future. Despite the current limitation to definitively rule-in linkages, genomic epidemiology provides critical information to disease investigators and public health officials.

The Role of Genomic Epidemiology in Public Health

Genomic epidemiology is also useful for guiding public health decisions. You can use the resolution provided by genomic sequencing data to 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. Furthermore, genomic epidemiology can be paired with existing surveillance methods such as wastewater surveillance. This pairing can allow you to better understand the pathogens circulating in your community and what interventions are most important. Genomic epidemiology thus aids public health decision-making by providing highly resolved data on pathogen attributes and transmission dynamics.

Data Sources

The Kansas Health and Environmental Laboratories (KHEL) routinely sequence SARS-CoV-2 and enteric pathogens. These sequences are then uploaded to the National Center for Biotechnology Information (NCBI) sequence database (https://www.ncbi.nlm.nih.gov/).

NCBI also houses over 6 million publicly available sequences and is a great resource for contextual sequence data for a wide variety of pathogens.

Another source of sequence data is the GISAID database (https://gisaid.org/). This database houses over 15 million SARS-CoV-2, influenza, and Mpox sequences.

Sampling

As with any epidemiological endeavor, good genomic epidemiology begins with an appropriate sampling of both cases and controls. Establishing a sampling scheme before data collection is time-consuming and may limit what questions can be answered. However, having an

established scheme is generally more cost-effective and allows for testing specific hypotheses. This section explores opportunistic, random, location- and time-informed, and points of entry sampling and discusses the strengths and weaknesses of each.

Opportunistic (Convenience) Sampling

Opportunistic sampling is the most common sampling scheme used in public health. In opportunistic sampling, samples are not chosen for sequencing systematically. This method exposes your analysis to bias and may reduce external validity. However, opportunistic sampling is quick and easy to implement and may be used to assess pathogens and transmission chains roughly. You can mitigate potential bias in opportunistic sampling by incorporating detailed meta-data such as epidemiological or demographic data.

Random Sampling

In random sampling, a proportion of samples are selected for sequencing independent of any associated data and every sample has an equal chance of selection. Random sampling encompasses methods such as stratified random sampling (SRS) or multistage sampling. In these versions, samples are divided into attribute-based strata (e.g., sex) before random selection. This ensures that all strata are well represented by the sample population, reducing the potential for bias.

Location and Time-informed Sampling

In location- and time-informed sampling, a proportion of samples from diverse locations and times are chosen for sequencing. This method can detect regional and temporal variations in the observed sequences, catching greater genetic diversity. However, location- and time-informed sampling requires a robust system of sample collection and a high degree of centralization.

Points of Entry Sampling

In points of entry sampling, samples acquired from incoming travelers are sequenced. Sampling from points of entry allows for understanding a greater diversity of sequences from a wider geographic area. This may be especially useful for highly interconnected areas. However, this sampling scheme will be ineffective at understanding local pathogens and transmission dynamics.

Contextual Data

In addition to analyzing genomic data that are relevant to the desired population or outbreak, it is also important to analyze genomic data from unrelated samples, called contextual data. Contextual genomic data are analogous to controls in a traditional epidemiological study. As with controls, contextual data are necessary to understand the relationships between cases. Contextual data also provide the “backbone” of the phylogenetic tree, without which, phylogenetic trees provide little information.

Making Phylogenetic Trees

Phylogenetic trees offer a means to easily visualize your genomic sequencing data and infer links between samples. This section will briefly overview how phylogenetic trees are created and discusses some common software and web-based tools you can use to create phylogenetic trees.

Multiple Sequence Alignment (MSA)

Assuming sequencing reads have already been assembled into a consensus genome, aligning sequences is the first step in creating a phylogenetic tree. This process, called multiple sequence alignment (MSA), is typically done in the background of software used to create phylogenetic trees. In an MSA, each row corresponds to a consensus sequence for a particular sample. These sequences are aligned such that nucleotide 1 of sample A is directly above nucleotide 1 of sample B (Figure 3A). The columns of an MSA thus correspond to a specific nucleotide position. After the sample sequences have been aligned, the software will infer phylogenetic relationships based on the number of shared mutations (Figure 3B).

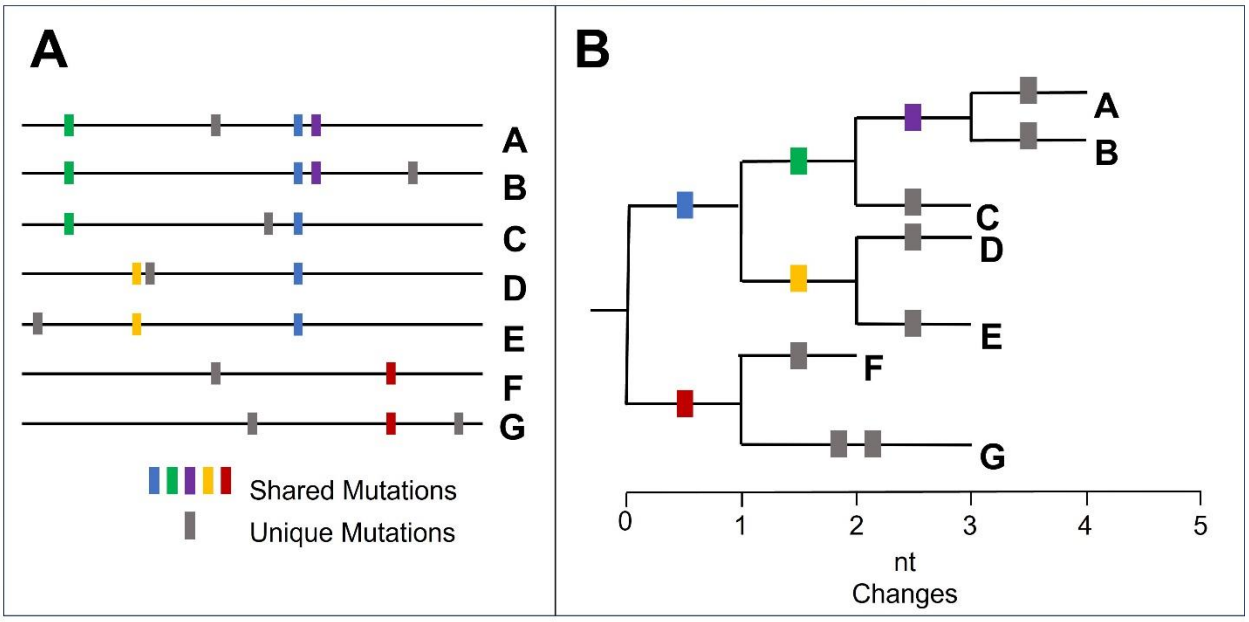


Figure 3; A: An example multiple sequence alignment (MSA) of sequences A-G. Each sequence has some shared mutations (colored bars) and some unique mutations (grey bars). B: The phylogenetic tree based on the MSA. Each tip corresponds to a sequence. Branch lengths are weighted to the number of nucleotide changes from the inferred common ancestor.

Distance Matrix vs Maximum Likelihood (ML) Methods

There are two common methods in which phylogenetic trees are constructed, distance matrix and maximum likelihood (ML). Distance matrix methods measure the genetic differences between samples and display a phylogenetic tree representing the similarities between samples. While distance matrix methods are faster than ML, they do not approximate

evolutionary relationships. ML methods, contrastingly, provide better estimations of evolutionary relationships. ML-based algorithms assess the probability of each sample being descended from an inferred common ancestor. The sum of all these probabilities is then used to assess the quality of a particular tree and all possible trees are then compared to each other. The algorithm selects the tree(s) with the highest probability. While slower than distance matrix methods, ML algorithms are more robust, better estimate evolutionary relationships, and are more widely used and developed.

UShER

Ultrafast Sample placement on Existing tRees (UShER) is a program that places sequences on an existing phylogenetic tree made from all available sequences. The main benefit of UShER is the speed at which phylogenetic placements are made. UShER is optimized for SARS-CoV-2 phylogenetic placements but can also be used for respiratory syncytial virus (RSV) and human Mpox sequences. The University of California Santa Cruz hosts the online UShER upload portal on their website (https://genome.ucsc.edu/cgi-bin/hgPhyloPlace). To use this resource, choose which pathogen you are working with from the drop-down menu, upload your sequence data in FASTA format to the webpage, and select a phylogenetic tree. You will then be directed to the output page. From there, you can view your phylogenetic tree in the browser, on Nextstrain, or download the tree as a Newick file.

UShER does not store data after it has produced the phylogenetic placements. However, if your data contains protected information, you can use an alternate program called ShUShER.

More resources for UShER are in the "Additional Resources" section at the end of this guide.

MicrobeTrace

MicrobeTrace is a web-based tool developed by the CDC that visualizes epidemiological and genomic relationships between cases. MicrobeTrace creates network diagrams with lines representing genetic, person-to-person, and person-to-place links (Figure 4). To use this tool, upload your genomic data as either sequences in FASTA format, distance matrices in CSV format, or a phylogenetic tree as a Newick file to the webpage (https://microbetrace.cdc.gov/MicrobeTrace/). You can also upload spreadsheets containing

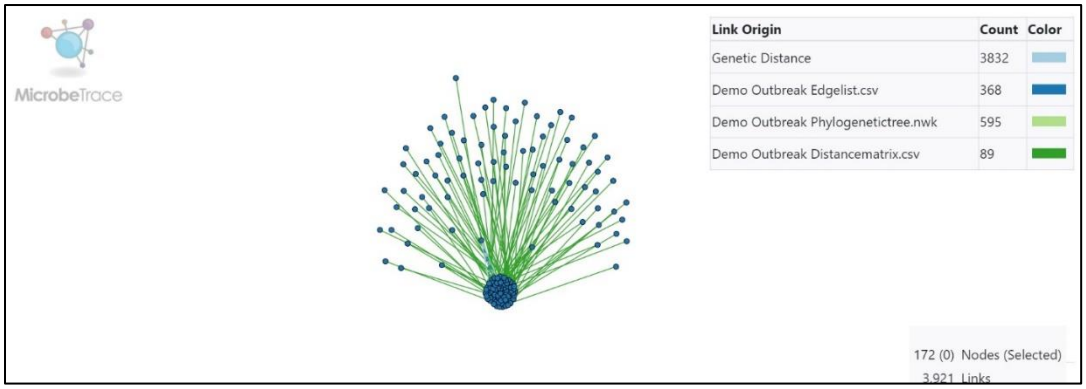


Figure 4; Example of a MicrobeTrace visualization.

epidemiological or other meta-data. MicrobeTrace offers a unique way to visualize both epidemiologic and genomic data.

More resources for MicrobeTrace are in the "Additional Resources" section at the end of this guide.

Nextstrain

Nextstrain is an online repository of genomic sequencing data, a toolkit for creating your phylogenetic trees, and a data-sharing platform. The online repository is accessible via nexstrain.org and contains phylogenetic trees and other related data for several pathogens including SARS-CoV-2, influenza, and measles. These phylogenetic trees are composed of publicly available sequence data from NCBI, GISAID, and similar sources.

Nextstrain's main data analysis tools are Augur and Auspice. Nextstrain's Augur tool consists of several modules that filter and align sequencing data inputs and then construct, refine, and export phylogenetic trees. The Auspice tool then creates interactive visualizations of the phylogenetic tree created by Augur. To install these tools, follow the installation instructions provided here: https://docs.nextstrain.org/en/latest/install.html.

Finally, Nextstrain provides a platform to share your genomic data and phylogenetic trees locally or with the wider scientific community. These Nextstrain Groups are housed on the main Nextstrain website and can be private or public. Public NextStrain Groups can be found here: https://nextstrain.org/groups/. To create your own Group, email the NextStrain administration at hello@nextstrain.org.

More resources for NextStrain are in the "Additional Resources" section at the end of this guide.

BEAST

Bayesian Evolutionary Analysis Sampling Trees (BEAST) is a software for creating time-rooted trees using Bayesian methods. BEAST allows for modeling how populations change over time, called phylodynamics. You can also use BEAST to model how interventions may affect pathogen phylodynamics. Analysis using BEAST can be quite time-consuming and is not the best option for quickly informing disease investigators or public health officials. The BEAST website (https://beast.community/index.html#downloading-beast) provides instructions for installing the software as well as several tutorials.

More resources for BEAST are in the "Additional Resources" section at the end of this guide.

R and ggtree

ggtree is an R package that displays phylogenetic trees from a variety of sources including Newick and BEAST files using the ggplot2 system. ggtree is highly customizable and supports the annotation of phylogenetic trees with other epidemiological data. To install ggtree, start R and enter the following:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("ggtree")
```

More resources for ggtree are in the “Additional Resources” section at the end of this guide.

Annotating Trees with Epidemiological Data

While phylogenetic trees are helpful to understand genetic relationships, 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.

Simple annotations such as highlighting and labeling specific tips or clades are useful in directing a user’s attention to areas of note. This is particularly useful for large, complex trees. Tips or clades can also be labeled with epidemiological relationships (figure 5).

Numerous types of data may be useful to display along with your phylogenetic tree such as maps, symptoms, related exposures, notable dates, and the laboratory that performed sequencing. Maps and heat plots can be added to the side of a phylogenetic tree to display further data from an epidemiological investigation. For example, a map can be added displaying the county of residence for each case or a heat map shows what restaurants are common to each case. The most helpful data depends on the pathogen, the outbreak, and the objective of the visual.

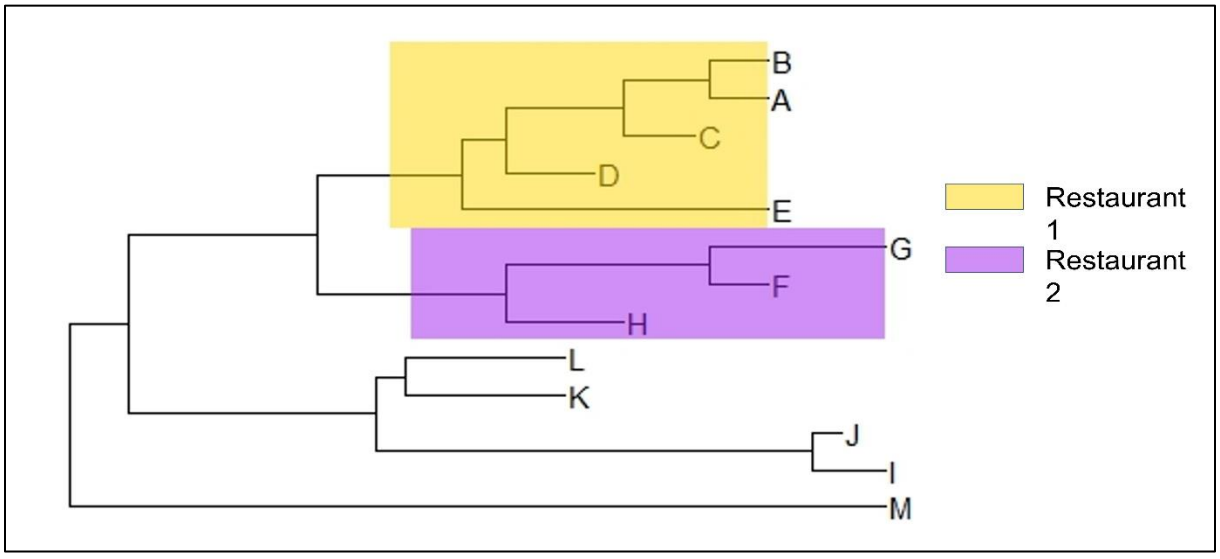


Figure 5; Phylogenetic tree with a simple annotation showing how you can display relationships between phylogenetic and epidemiological data.

Interpreting Phylogenetic Trees

Phylogenetic trees are composed of four parts, the root, internal nodes, branches, and tips (figure 6). Tips (also called leaves) 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.

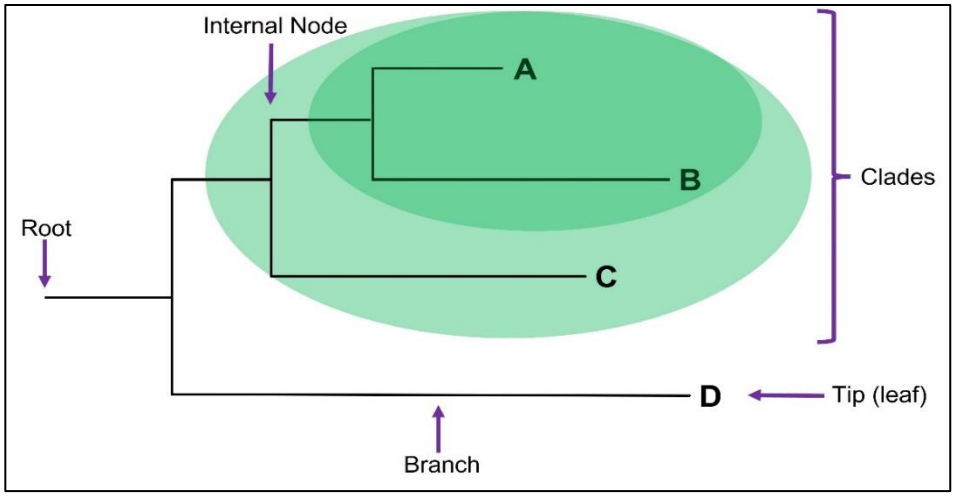


Figure 6; Anatomy of a rooted phylogenetic tree. Tips represent observed samples, internal nodes represent hypothetical common ancestors, branches connect nodes to tips, and the root represents the common ancestor for all sequences present. Clades are sequences that share a common mutation. As shown, clades are hierarchical with multiple levels of grouping.

Phylogenetic trees organize sequences into groups called clades (figure 6). Sequences within a clade all share a common ancestor. A single sequence may belong to many clades such as sequence A in Figure 6. Sequence A shares its most recent common ancestor with sequence B, but there is a more distant ancestor that is shared by sequences A, B, and C. Furthermore, the section shown in Figure 6 may be a part of a larger tree in which sequences A, B, C, and D belong to the same clade. 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.

Not all phylogenetic trees are rooted. Unrooted trees display the same phylogenetic relationships but do not infer the direction of evolution and are less useful for genomic epidemiology.

Phylogenetic trees can also be displayed as spirals or circles. The information displayed in these phylogenetic trees is the same as in horizontal trees. In circular trees, however, the radius becomes the representation of genetic divergence or time rather than an x-axis. Spiral phylogenetic trees are best used for displaying a high number of sequences.

Additional Resources

***An applied genomic epidemiology handbook* by Allison Black and Gytis Dudas** — This handbook provides a more in-depth overview of the biological and molecular basis for genomic epidemiology and its applications for public health. The handbook is open-source and accessible via GitHub here: https://alliblk.github.io/genepi-book/index.html.

BEAST — BEAST is software used to create time-rooted trees and run phylodynamic analyses. The BEAST website contains installation instructions and tutorials here: https://beast.community/index.html#downloading-beast.

- For further reading consult the following journal article:
  Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*(1), vey016. https://doi.org/10.1093/ve/vey016

CDC COVID-19 Genomic Epidemiology Toolkit — Consisting of three parts, Introduction, Case Studies, and Implementation, this toolkit guides the user through the basics of using genomic epidemiology methods for COVID-19 investigations. You can access this toolkit via the CDC's website here: https://www.cdc.gov/amd/training/covid-19-gen-epi-toolkit.html.

ggtree Tutorial — This tutorial guides the user through the importation, creation, and annotation of phylogenetic trees using the ggtree package in R. You can access the tutorial on GitHub here: https://4va.github.io/biodatasci/r-ggtree.html.

- For further reading, consult the following journal article:
  Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

GISAID sequence database — This publicly available database houses over 15 million SARS-CoV-2, influenza, and Mpox sequences. The database can be accessed through the GISAID website here: www.gisaid.org.

MicrobeTrace — MicrobeTrace is a web-based tool that visualizes both epidemiological and genomic data. You can access the tool here: https://microbetrace.cdc.gov/MicrobeTrace/. More information and further resources, including a full user manual, are available on GitHub here: https://github.com/CDCgov/MicrobeTrace. MicrobeTrace is also discussed in detail in the CDC Genomic Epidemiology Toolkit described above.

- For further reading, consult the following journal article:
  Campbell, E. M., Boyles, A., Shankar, A., Kim, J., Knyazev, S., Cintron, R., & Switzer, W. M. (2021). MicrobeTrace: Retooling molecular epidemiology for rapid public health response. *PLOS Computational Biology*, *17*(9), e1009300. https://doi.org/10.1371/journal.pcbi.1009300

NCBI sequence database — This publicly available database houses over 6 million sequences from more than 1000 species, including a variety of pathogens. KHEL uploads SARS-CoV-2 and enteric pathogen sequences here as well. The database can be accessed through the NCBI website here: https://www.ncbi.nlm.nih.gov/.

Nextstrain Documentation — These documents from Nextstrain provide an overview of how to use the interactive dashboards hosted on the website and Nextstrain's tools for your phylogenetic analyses. Also included are tutorials and how-to guides on topics ranging from creating workflows to sharing your analyses with the wider scientific community. You can access these resources via Nextstrain's website here: https://docs.nextstrain.org/en/latest/.

UShER — UShER is a program for phylogenetic placements of SARS-CoV-2, RSV, and Mpox sequences. The online genome browser is hosted by the UCSC website here: https://genome.ucsc.edu/cgi-bin/hgPhyloPlace. More information on UShER and associated programs can be found on the GitHub repository here: https://github.com/yatisht/usher.

- For further reading, consult the following journal article:
  Turakhia, Y., Thornlow, B., Hinrichs, A. S., De Maio, N., Gozashti, L., Lanfear, R., Haussler, D., & Corbett-Detig, R. (2021). Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics*, *53*(6), Article 6. https://doi.org/10.1038/s41588-021-00862-7

Glossary of terms associated with genomic epidemiology

Clade — Groups or subgroups of sequences that share a common ancestor.

Consensus genome — An imperfect summary of within-host pathogen genetic diversity that represents the most frequent nucleotides at each site in the genome.

Contextual data — Sequence data from pathogens outside of the investigation that are used to build the "backbone" of a phylogenetic tree.

Evolutionary rate — The rate at which mutations accumulate after undergoing selection. At the pathogen level, this rate is subject to replication time, population size, and selection pressures. Sampling schemes may also influence the apparent evolutionary rate. Infrequent sampling allows more time for the population to select against deleterious variations, and intensive sampling may capture a higher proportion of (potentially) deleterious mutations. Infrequent or intensive sampling thus inaccurately estimates the evolutionary rate.

Mutation rate — The rate at which a pathogen's DNA or RNA polymerase makes replication errors.

Phylogenetic placement — Similar in appearance to a phylogenetic tree, a phylogenetic placement adds observed sequencing data to an existing phylogenetic tree (as used in UShER).

Phylogenetic trees — A diagram used to describe the genetic relationships between pathogens. Tips (leaves) represent observed samples, nodes represent inferred common ancestors, and branches connect nodes to tips, the length of which represents either number of mutations or time.

Root (phylogenetic tree) — The representation of the inferred ancestor for all sequences in the dataset.

Root-to-tip distance — A measure of genetic divergence calculated from the number of different mutations between the root (ancestor) and the tip (relevant sequence).

Substitution — A mutation that has become completely dominant or "fixed" in a population.

Transmission bottleneck — A phenomenon in which only a portion of the within-host pathogen genetic diversity is passed from an infected individual to a recipient.

References

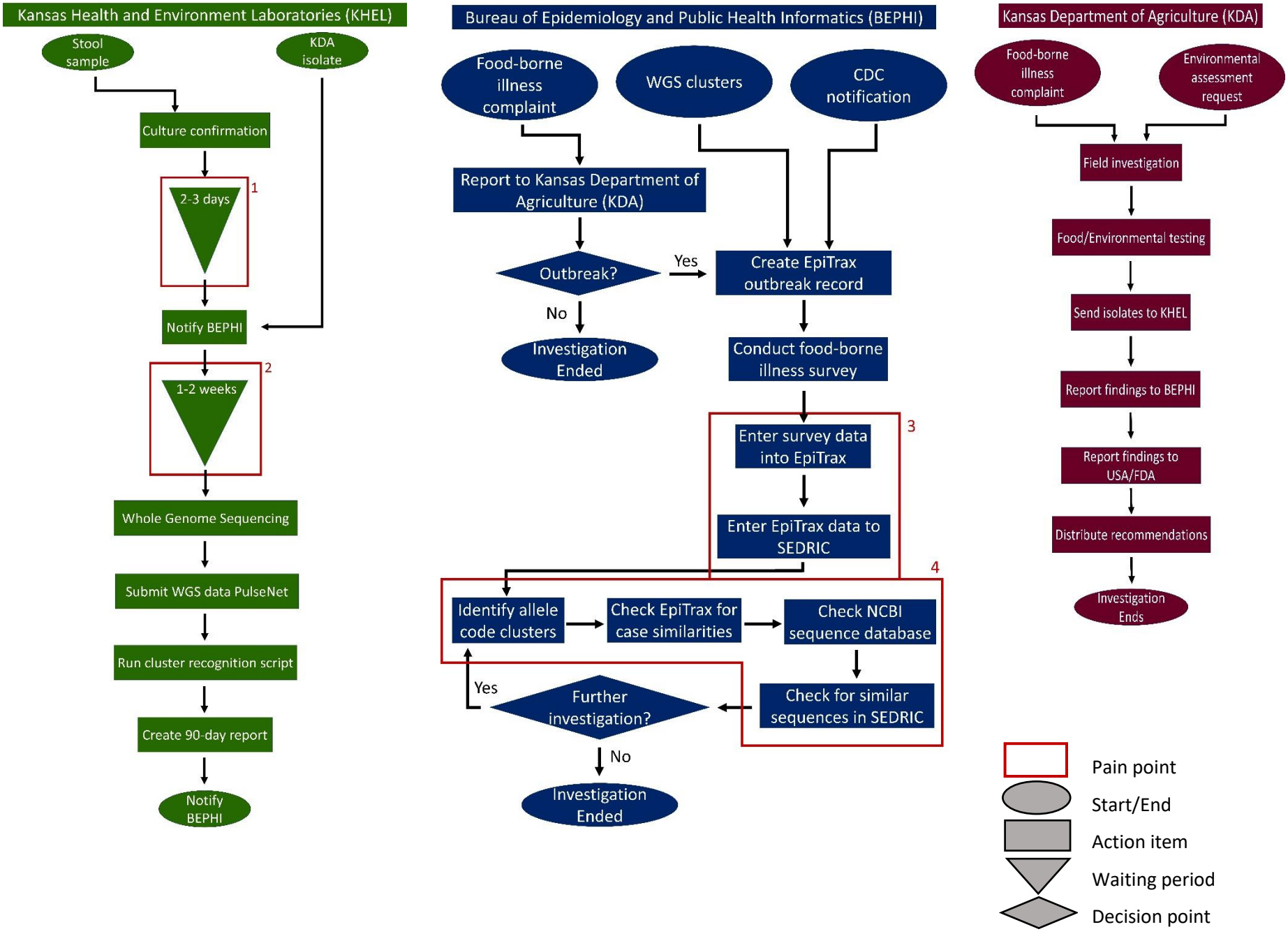
Campbell, E. M., Boyles, A., Shankar, A., Kim, J., Knyazev, S., Cintron, R., & Switzer, W. M. (2021). MicrobeTrace: Retooling molecular epidemiology for rapid public health response. *PLOS Computational Biology*, *17*(9), e1009300. https://doi.org/10.1371/journal.pcbi.1009300

Hill, V., Ruis, C., Bajaj, S., Pybus, O. G., & Kraemer, M. U. G. (2021). Progress and challenges in virus genomic epidemiology. *Trends in Parasitology*, *37*(12), 1038–1049. https://doi.org/10.1016/j.pt.2021.08.007

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*(1), vey016. https://doi.org/10.1093/ve/vey016

Turakhia, Y., Thornlow, B., Hinrichs, A. S., De Maio, N., Gozashti, L., Lanfear, R., Haussler, D., & Corbett-Detig, R. (2021). Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics*, *53*(6), Article 6. https://doi.org/10.1038/s41588-021-00862-7

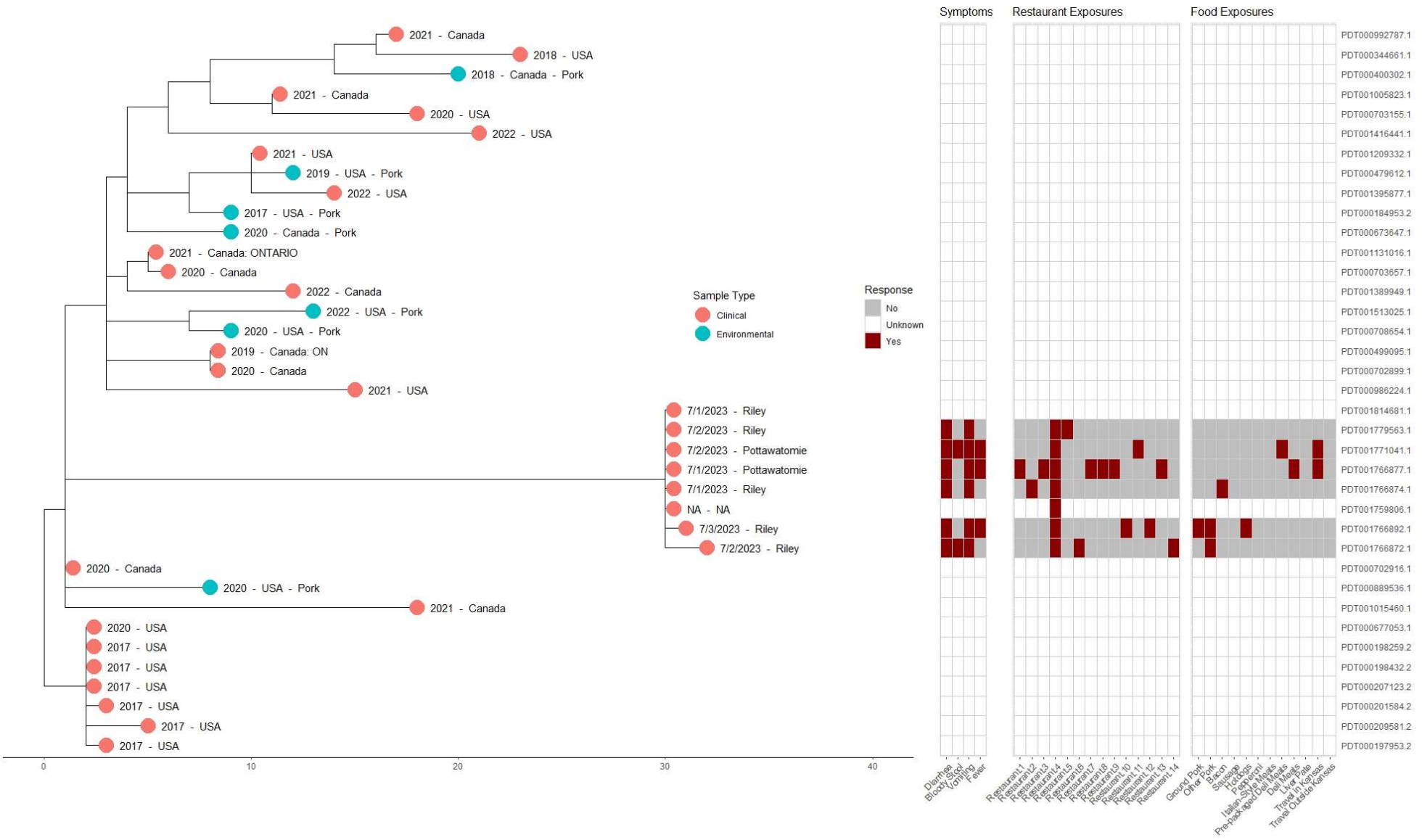
Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

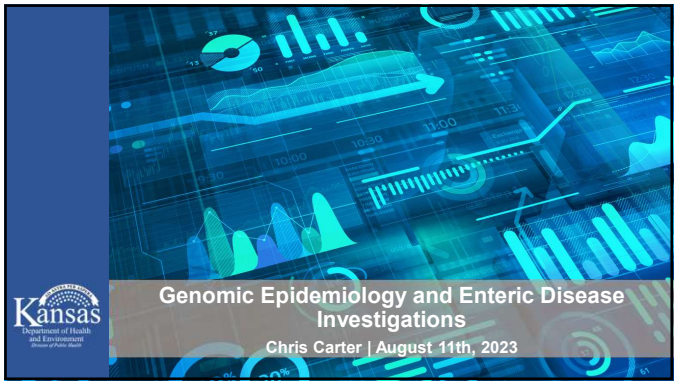

Kansas Health and Environment Laboratories (KHEL)
Stool sample
KDA isolate
Culture confirmation
2-3 days
1
Notify BEPHI
1-2 weeks
2
Whole Genome Sequencing
Submit WGS data PulseNet
Run cluster recognition script
Create 90-day report
Notify BEPHI
Bureau of Epidemiology and Public Health Informatics (BEPHI)
Food-borne illness complaint
WGS clusters
CDC notification
Report to Kansas Department of Agriculture (KDA)
Outbreak?
Yes
No
Investigation Ended
Create EpiTrax outbreak record
Conduct food-borne illness survey
3
Enter survey data into EpiTrax
Enter EpiTrax data to SEDRIC
4
Identify allele code clusters
Check EpiTrax for case similarities
Check NCBI sequence database
Check for similar sequences in SEDRIC
Yes
Further investigation?
No
Investigation Ended
Kansas Department of Agriculture (KDA)
Food-borne illness complaint
Environmental assessment request
Field investigation
Food/Environmental testing
Send isolates to KHEL
Report findings to BEPHI
Report findings to USA/FDA
Distribute recommendations
Investigation Ends
Pain point
Start/End
Action item
Waiting period
Decision point

Pain Point	Description	Items
1	Culture confirmation can take up to several days to identify a pathogen. This waiting period may delay epidemiological investigations. Culture independent identification of the pathogen may be faster but does not provide a colony that is needed for sequencing.	2-3 day waiting period
2	It is more cost-effective to run many samples on a large capacity sequencing cartridge. During times of low throughput, it may take up to several weeks until there are enough samples to fill a sequencing run. This waiting period may delay investigations.	1-3 week waiting period
3	It is time-consuming to enter the same information into two databases.	Enter survey data into Epitrax; Enter EpiTrax data to SEDRIC
4	This process of investigating multiple data sources is difficult, clunky, and time consuming.	Identify allele code clusters; check EpiTrax for case similarities; Check NCBI sequence database; Check for similar sequences in SEDRIC

Section	Item	Description	Pain Point
KHEL	Sample Received	KHEL receives stool samples and isolates from Kansas hospitals. KDA will also send isolates to KHEL.	
	Culture confirmation	KHEL uses culture-based diagnostic methods to identify the pathogen. If an outbreak is already identified, outbreak-associated samples are first tested using multiplex PCR.	
	2-3 day waiting period	Culture-based methods may take up to several days for results.	1
	Notify BEPHI (1)	Once the pathogen is identified, KHEL notifies epidemiologists at BEPHI	
	1-3 week waiting period	It is more cost-effective to run many samples in a large sequencing cartridge. It may take several weeks until there are enough samples to fill a sequencing run. This waiting time can vary depending on sample availability and time of year.	2
	Whole Genome Sequencing	KHEL sequences samples using Illumina or Oxford nanopore sequencers. KHEL then assesses the quality of the sequencing read and assembles a consensus genome.	
	Submit WGS data to CDC	KHEL submits consensus genomes to the CDC and to NCBI.	
	Run cluster recognition script	KHEL enters consensus genomes into a Python script that recognizes genetically similar samples.	
	Create 90-day report	KHEL summarizes the output from the cluster recognition script for samples from the past 90 days.	
	Notify BEPHI (2)	KHEL sends the 90-day report to BEPHI.	
BEPHI	Food-borne illness complaint	BEPHI receives food-borne illness complaints from the KDHE epidemiology hotline or the online KDHE food-borne illness report.	
	WGS clusters	WGS clusters identified by KHEL are stored in a spreadsheet. If there are further connections between cases, BEPHI opens an investigation.	
	CDC notification	If the CDC detects a cluster of cases, they will notify BEPHI.	

	Report to Kansas Department of Agriculture (KDA)	BEPHI forwards all food-borne illness complaints to KDA. KDA then investigates.	
	Outbreak?	To meet the outbreak definition, there has to be 2 or more cases from different households. If these criteria are not met, KDA is informed and the BEPHI investigation ends. If these criteria are met, BEPHI opens an investigation.	
	Create EpiTrax outbreak record	Once BEPHI opens an investigation, they create an outbreak record in EpiTrax.	
	Conduct food-borne illness survey	BEPHI interviews cases involved in outbreaks. The interview consists of several questions regarding demographics and potential exposures.	
	Enter survey data into EpiTrax	BEPHI then enters the information obtained during the interview into EpiTrax.	3
	Enter EpiTrax data into SEDRIC (as requested)	BEPHI enters the EpiTrax data into the System for Enteric Disease Response, Investigation, and Coordination (SEDRIC).	3
	Identify allele code clusters	BEPHI investigates the allele code clusters identified by the KHEL cluster recognition script.	4
	Check EpiTrax for case similarities	BEPHI uses EpiTrax to identify similar exposures between cases.	4
	Check NCBI sequence database	BEPHI searches the NCBI sequence database for sequences similar to the cases.	4
	Check for similar sequences in SEDRIC	SEDRIC houses a national database of enteric disease outbreak information. BEPHI uses SEDRIC to identify if outbreaks in Kansas are related to outbreaks in other states.	4
	Further investigation?	If more associated cases are identified, further investigation is needed. If the outbreak and its causes are well understood and no further cases are identified, the investigation is ended.	
	Investigation ended	BEPHI may reopen an investigation if additional cases are identified.	
KDA	Food-borne illness complaint	KDA receives food-borne illness complaints from BEPHI.	
	Environmental assessment request	KDHE may request environmental assessments for suspected facilities.	
	Field investigation	KDA conducts investigation of facilities following the Council to Improve Food-borne Outbreak Response (CIFOR) and National Environmental Assessment Reporting System (NEARS) procedures.	
	Food/environmental testing	KDA may perform testing procedures on samples obtained during the field investigation.	
	Send isolates to KHEL	After KDA tests the samples, it sends isolates to KHEL for further testing.	
	Report findings to CDC	KDA sends a report of its findings to the CDC.	
	Report findings to BEPHI	KDA sends a final outbreak report to BEPHI.	
	Distribute recommendations	KDA advises facilities and other stakeholders on how to prevent future outbreaks.	
	Investigation ends	KDA concludes their investigation. If problems continue, KDA may reopen an investigation.	


2021 - Canada
2018 - USA
2018 - Canada - Pork
2021 - Canada
2020 - USA
2022 - USA
2021 - USA
2019 - USA - Pork
2022 - USA
2017 - USA - Pork
2020 - Canada - Pork
2021 - Canada: ONTARIO
2020 - Canada
2022 - Canada
2022 - USA - Pork
2020 - USA - Pork
2019 - Canada: ON
2020 - Canada
2021 - USA
7/1/2023 - Riley
7/2/2023 - Riley
7/2/2023 - Pottawatomie
7/1/2023 - Pottawatomie
7/1/2023 - Riley
NA - NA
7/3/2023 - Riley
7/2/2023 - Riley
2020 - Canada
2020 - USA - Pork
2021 - Canada
2020 - USA
2017 - USA
2017 - USA
2017 - USA
2017 - USA
2017 - USA
2017 - USA
0
10
20
30
40
Sample Type
Clinical
Environmental
Response
No
Unknown
Yes
Symptoms
Restaurant Exposures
Food Exposures
PDT000992787.1
PDT000344661.1
PDT000400302.1
PDT001005823.1
PDT000703155.1
PDT001416441.1
PDT001209332.1
PDT000479612.1
PDT001395877.1
PDT000184953.2
PDT000673647.1
PDT001131016.1
PDT000703657.1
PDT001389949.1
PDT001513025.1
PDT000708654.1
PDT000499095.1
PDT000702899.1
PDT000986224.1
PDT001814681.1
PDT001779563.1
PDT001771041.1
PDT001766877.1
PDT001766874.1
PDT001759806.1
PDT001766892.1
PDT001766872.1
PDT000702916.1
PDT000889536.1
PDT001015460.1
PDT000677053.1
PDT000198259.2
PDT000198432.2
PDT000207123.2
PDT000201584.2
PDT000209581.2
PDT000197953.2
Diarrhea
Bloody Stool
Vomiting
Fever
Restaurant1
Restaurant2
Restaurant3
Restaurant4
Restaurant5
Restaurant6
Restaurant7
Restaurant8
Restaurant9
Restaurant10
Restaurant11
Restaurant12
Restaurant13
Restaurant14
Ground Pork
Other Pork
Bacon
Sausage
Hotdogs
Pepperoni
Italian-Style Meats
Pre-Packaged Deli Meats
Deli Meats
Liver Pate
Liver In Kansas
Travel Outside Kansas



1

A brief introduction

KANSAS STATE UNIVERSITY
Master of Public Health

- Education
  - MPH, Infectious Diseases and Zoonoses, Kansas State University, May 2024
  - BS, Medical Microbiology, Kansas State University, May 2022
- Research Experience
  - "Droughts and Diseases: A comparative analysis of the health challenges in the Po and Colorado River Basins", Present
  - "Shiga Toxin Production Reduces the Population Growth of *Shigella flexneri* BS937", 2021-2023
  - "Of Fevers and Agues: Disease During the American Revolution (1775-1783)", 2020-2023
  - "Competition and Conjugation between Agrobacterial Cooperators and Cheaters", 2018-2020

2

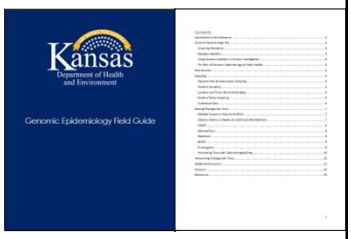
Internship Objectives

1. Conduct a review of genomic epidemiology literature and summarize the field in a shareable document.
2. Create a process map of enteric disease investigations to identify steps where genomic data can be useful.
3. Create and annotate a phylogenetic tree to demonstrate integration of genomic data in enteric disease workflows.

3

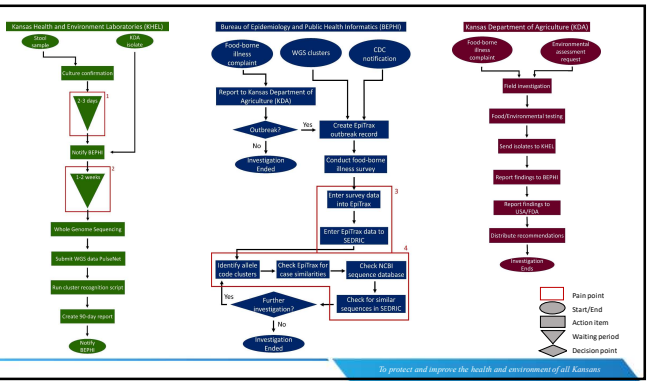
Genomic Epidemiology Field Guide

- Objectives
  - Provide background on the field
  - Demonstrate the usefulness of genomic epidemiology to disease investigations and public health
  - Present a workflow from obtaining data to creating and interpreting phylogenetic trees
  - Connect users to further resources
- Major Sections
  - "Genomic Epidemiology 101"
  - "Making Phylogenetic Trees"
  - "Interpreting Phylogenetic Trees"
  - "Additional Resources"



To protect and improve the health and environment of all Kansans

4

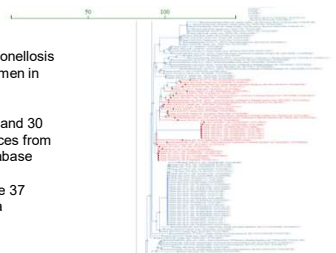


To protect and improve the health and environment of all Kansans

5

Creating a Phylogenetic Tree

- Analyzed data from the 2023 Salmonellosis outbreak associated with Jinya Ramen in Johnson County
- Selected the 7 Kansas sequences and 30 of the most closely related sequences from the NCBI Pathogen Detection Database
- Exported the subtree containing the 37 sequences (highlighted in red) as a NEWICK file



https://www.ncbi.nlm.nih.gov/pathogens/tree/#Salmonella/PDG000000002.2708/PDS000149998.327accessions=PDT001759806.1

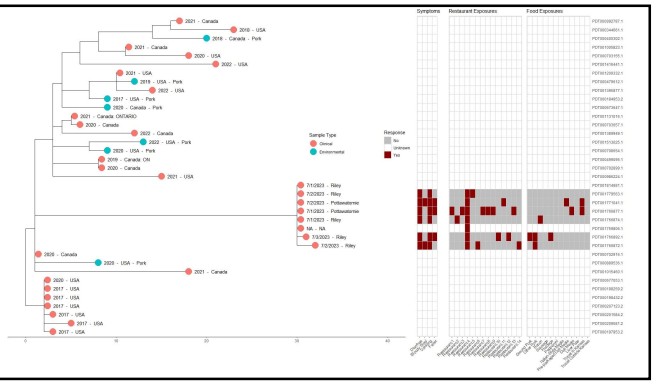
To protect and improve the health and environment of all Kansans

6

EpiTrax Exports

- Exported responses from the Salmonella Survey including symptoms, restaurant exposures, and food product
- Also exported county at diagnosis, symptom onset date, and CMR

7



8

Next Steps

1. Add PulseNet, NCBI, and KHEL accession numbers as an exportable feature in EpiTrax
2. Upload KS consensus genomes to NCBI
3. Add the Outbreak Surveys as an exportable feature in EpiTrax
4. **Automate workflow**

9

Summary

- The Genomic Epidemiology Field Guide provides a quick overview of the uses and methods of genomic epidemiology
- There are several pain points in the enteric disease workflow, one of which can be improved by combining genomic analysis and existing investigation data
- We present a helpful figure for displaying phylogenetic and epidemiological relationships
- When automated, this figure will aid in enteric disease investigations

10

Thank You/Questions

11

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).

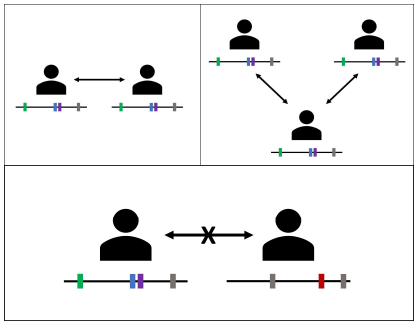


Figure 1; Cases with similar sequences are likely linked while those with distinct sequences are not. In the absence of other epidemiological data, however, it is difficult to infer a transmission chain.

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).

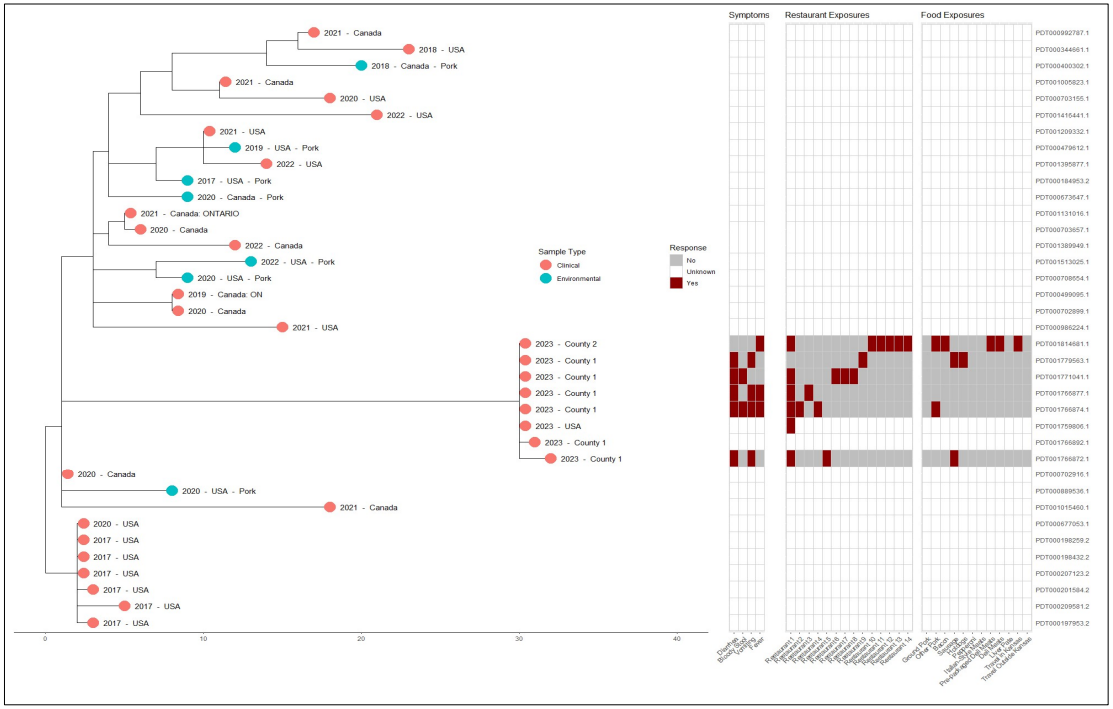


Figure 2; Example of an annotated tree. Tree tips are colored according to sample type and labeled with epidemiological information. The heat maps to the right display symptom and exposure data for each sample.

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.

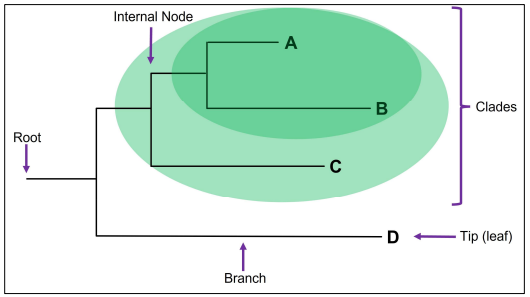


Figure 3; Anatomy of a rooted phylogenetic tree.

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.

Application of the Learnr Package and R Shiny to Train Staff to Apply Genomic Epidemiology in Public Health

Mayra Trujillo, MS, Chris Carter, and Dr. John Anderson, MSC PhD
Kansas Department of Health and Environment

BACKGROUND: Genomic epidemiology is a rapidly advancing field in public health. The application of genomic data in public health has the potential to facilitate the detection of emerging diseases, determine the distribution of diseases, and evaluate therapeutic interventions. Due to the complex nature of genomic analyses, the public health workforce needs additional training that is tailored to different disciplines to facilitate the use of genomic data. A self-paced, wide-reaching, and low-cost tutorial and accompanying exam were created using R to train public health staff on the use of genomics in public health.

METHODS: The genomic epidemiology tutorial was designed using the learnr package from R to create an interactive learning experience. The tutorial introduces the biological basis of genomic epidemiology, then reviews data sources and sampling strategies, and finally presents methods used to create/visualize SNP matrices, phylogenetic trees, and annotations of phylogenetic trees with epidemiological data. After users complete the tutorial, they are guided to a separate survey that employs questions formatted as a 5-point Likert scale to measure their interest in, confidence in conducting, and perception of genomic epidemiology before and after the tutorial. Finally, an exam is given to test comprehension using questions based on the “remember”, “understand”, and “apply” levels of Bloom’s Taxonomy. Users can take the exam as many times as they want until they reach a score of 80%. Survey and exam submissions are scored and populated into a Google Sheets spreadsheet to track performance.

RESULTS: Over two weeks, 21 participants took the training and exam. After taking the training, the median Likert score for participants’ confidence in conducting genomic epidemiology significantly increased by 1 point on a 5-point scale ($p<0.05$) when compared to participants’ confidence in conducting genomic epidemiology before the tutorial. All participants passed the exam and 76% percent of participants passed the exam on their first attempt.

CONCLUSIONS: The application of the learnr package and R Shiny ecosystem to provide training on genomic epidemiology resulted in improved confidence in conducting genomic epidemiology among participants. By using tools in the R ecosystem, the training was simple and inexpensive to distribute and update and provided the capability to automate the evaluation of survey and exam data in R to track tutorial performance. Overall, these tools facilitate the quick development of trainings and the measurement of performance in the rapidly developing field of genomic epidemiology.