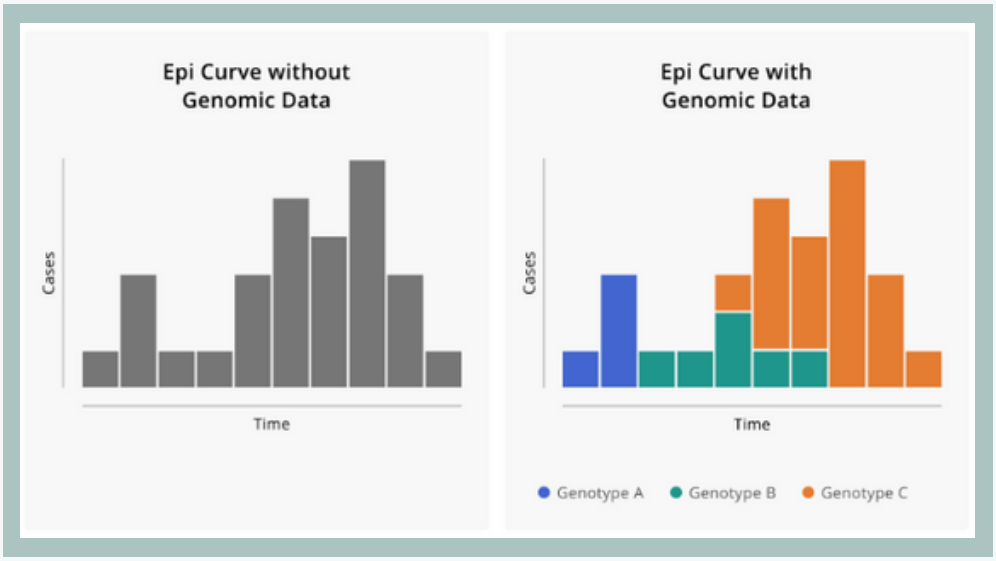
Genomic Epidemiology Fundamentals

I. Genomic Information Enhances Epidemiology

According to the CDC, genomic epidemiology is defined as “use of pathogen genomic data to determine the distribution and spread of an infectious disease in a specified population and the application of this information to control health problems.” In short, genomic data offers an extra layer of information on top of classic epidemiological techniques which can be used to extrapolate the relationships between cases, creating distinct transmission chains. This extra layer of information increases both the scale and resolution of outbreak investigation, allowing a more targeted approach to infectious disease control at both an individual and population level



While classic epidemiological techniques can gather valuable outbreak information, sometimes these epidemiological questions can only be answered through sequencing:

- Is the pathogen novel, or does it represent a re-emergence of a known pathogen?
- What is its mode of transmission?
- Where does the pathogen come from (i.e. reservoir or geographic source)?

- What ecological factors surround its emergence?
- How many introductions into humans have there been, and what is the timing of these introductions?
- Was there undetected transmission before the first reported case?
- What is the nature of pathogen evolution?

During the detection of an outbreak, the primary advantage of sequencing is the ability to detect novel pathogens. This answers the question of whether there have been any human introductions and whether there are treatments available. Genetic diversity, measured as the average number of nucleotide differences between samples in the population, increases as an outbreak progresses. In this way, pathogen mutations are used as markers of transmission events. This lends itself to phylogenetic analysis, which can be used to determine the spatial and temporal scales of transmission, allowing for effective public health intervention.

By monitoring genotype-specific incidence rates, we can get an idea of the fitness of a given lineage and how differing symptoms can be associated with it. Repeated sequencing from the same individual can also provide longitudinal information to identify genetic determinants of disease progression. As sequencing has become cheaper over time, researchers can characterize entire pathogen genomes of infected individuals in real time. As these capabilities increase, rapid genomic surveillance can start to become a standard part of outbreak investigation.

Real World Examples of Genomic Epidemiology in Practice

Pathogen	Location	Findings
MRSA	Cambridge, UK	Whole-genome bacterial sequencing was used to help reconstruct transmission chains and identify a likely source for a sustained outbreak of MRSA within a hospital ward. This investigation led to targeted decolonization.
Ebola Virus	West Africa	West Africa & Whole-genome virus sequencing was used to help reconstruct transmission chains and confirm the first documented case of sexual transmission of Ebola virus. This investigation led to immediate changes to guidance for male survivors that included a recommendation to have semen tested for presence of viral RNA.

HIV	USA	Next-generation sequencing was used to identify low frequency drug resistance variants (≥1–3%) within individual patients. Baseline presence of a resistance variant, even at low frequency, increased probability of virologic failure.
HIV	British Columbia, Canada	An automated phylogenetic system was established for monitoring HIV outbreaks using routinely collected virus genetic data. This system was used to identify case clusters in near real time, thus directing public health interventions.
Candida auris	Oxford, UK	Whole-genome fungal sequencing of patient and environmental isolates was used to help identify contaminated equipment as the source of many infections acquired within a hospital intensive care unit.
Yellow Fever	Brazil	Whole-genome virus sequencing was used to show that the recent Yellow fever outbreak in Brazil was caused by repeated sylvatic ('jungle') spillover and not urban transmission. As sylvatic transmission involves different mosquito species than urban, this finding informs vector control strategies.
Zika Virus	Florida, USA	Sequencing of virus genomes from cases and mosquitoes infected with Zika virus in Florida showed that multiple introductions of the virus from the Caribbean (perhaps hundreds) were required to sustain the outbreak, suggesting that traveler education and surveillance could reduce future outbreaks.
Lujo virus	Zambia and South Africa	One of the earliest studies to use metagenomic sequencing of human samples to discover a novel virus responsible for a cluster of fatal hemorrhagic fever.

Listeria monocytogenes	USA	By using whole-genome sequence data, investigators were able to substantially improve their ability to identify the source and cause of Listeria monocytogenes outbreaks.
Influenza virus	Worldwide	This paper shows that serological changes of influenza virus can be captured by studying virus genomic sequences. Such findings can be used to direct selection and design of seasonal influenza vaccines.
E. coli	Germany and France	Whole-genome sequencing of E. coli isolates was used to dissect a European outbreak of bloody diarrhea and hemolytic uremic syndrome caused by Shiga-toxin-producing E. coli

II. Pathogen Evolution and Transmission

Pathogen evolution is shaped by a variety of factors, including host movements, transmission dynamics, environments, and other selective pressures.

Pathogen diversity accumulates not only over the course of an epidemic, but throughout an individual infection. RNA viruses use RNA-dependent RNA polymerases during replication, which lack proof-reading abilities. DNA polymerases can also leave small mistakes in the genome after replication. As a result, as the pathogen's cells divide continuously throughout an infection mutations accumulate in the genome creating wider diversity. This is referred to as **within-host diversity**. The changes that are carried through each replication cycle have no specific purpose, they simply make the pathogen progeny more or less fit to the environment.

- **Deleterious mutations** make the pathogen less fit
- **Neutral mutations** do not have any effect on pathogen fitness
- **Beneficial mutations** make the pathogen more fit

The extent to which mutations impact pathogen fitness exists on a spectrum from lethal to highly beneficial.

The **transmission bottleneck** refers to how many virions/bacteria are sampled from the infector and how many are required to cause an infection. This varies from pathogen to pathogen. Since infections occur at different time points over the course of an infection, the pathogen diversity differs between infections.

When looking at within-host diversity, we often look at the **consensus genome**. This represents the most frequently observed nucleotide at each site in the genome at the time of sample collection. There are some sites that have greater diversity than others, this can be represented using the nucleotide ambiguity code.

Code	Represents	Complement
A	Adenine	T
G	Guanine	C
C	Cytosine	G
T	Thymine	A
Y	Pyrimidine (C or T)	R
R	Purine (A or G)	Y
W	weak (A or T)	W
S	strong (G or C)	S
K	keto (T or G)	M
M	amino (C or A)	K
D	A, G, T (not C)	H
V	A, C, G (not T)	B
H	A, C, T (not G)	D
B	C, G, T (not A)	V
X/N	any base	X/N
-	Gap	-

Below is some general terminology to describe changes in gene sequences:

- **Single Nucleotide Polymorphism (SNP)**: Also referred to as a mutation, is a change in the genetic sequence at a single site.
- **Allele**: Can also refer to mutation.
- **Genotype**: The pattern of mutations observed across a gene sequence.
- **Substitution**: Used when a mutation has become completely dominant.

The **mutation rate** is the actual rate that the DNA or RNA polymerase makes errors during replication. This is not easy to measure without specialized experimental techniques so the mutation rates of most pathogens are not known.

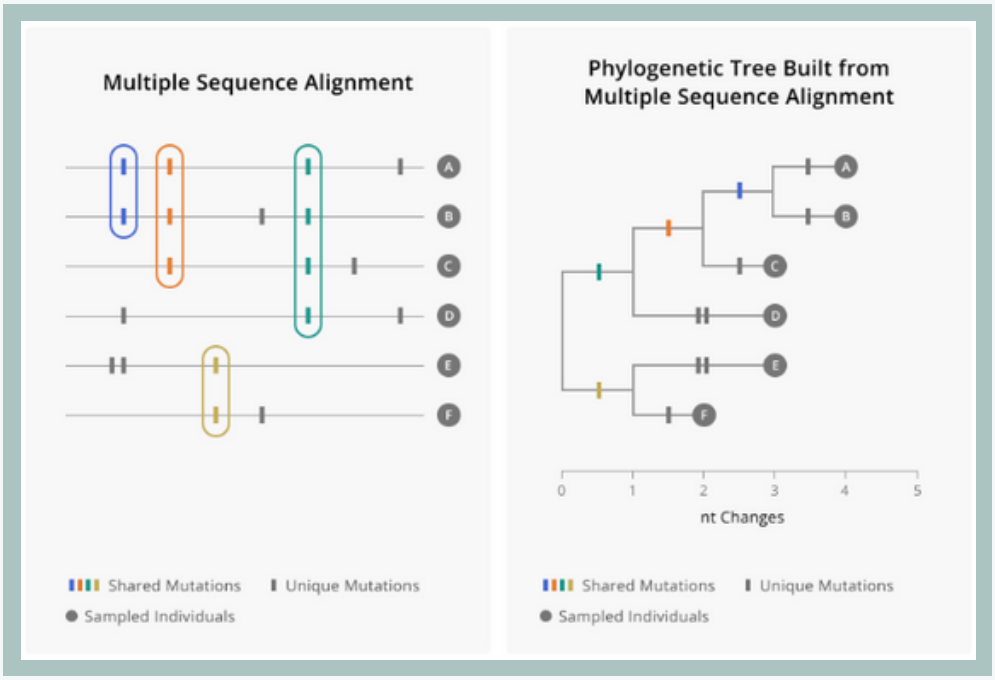
Since most deleterious mutations are lethal, they are often filtered out of the population quickly. The **evolutionary rate** is the rate that mutations accumulate after selection has filtered out deleterious variation. The number of mutations that have accrued can give us an idea of how much time has passed. This temporal signal of evolution is referred to as the **molecular clock**.

III. Using Phylogenetic Trees

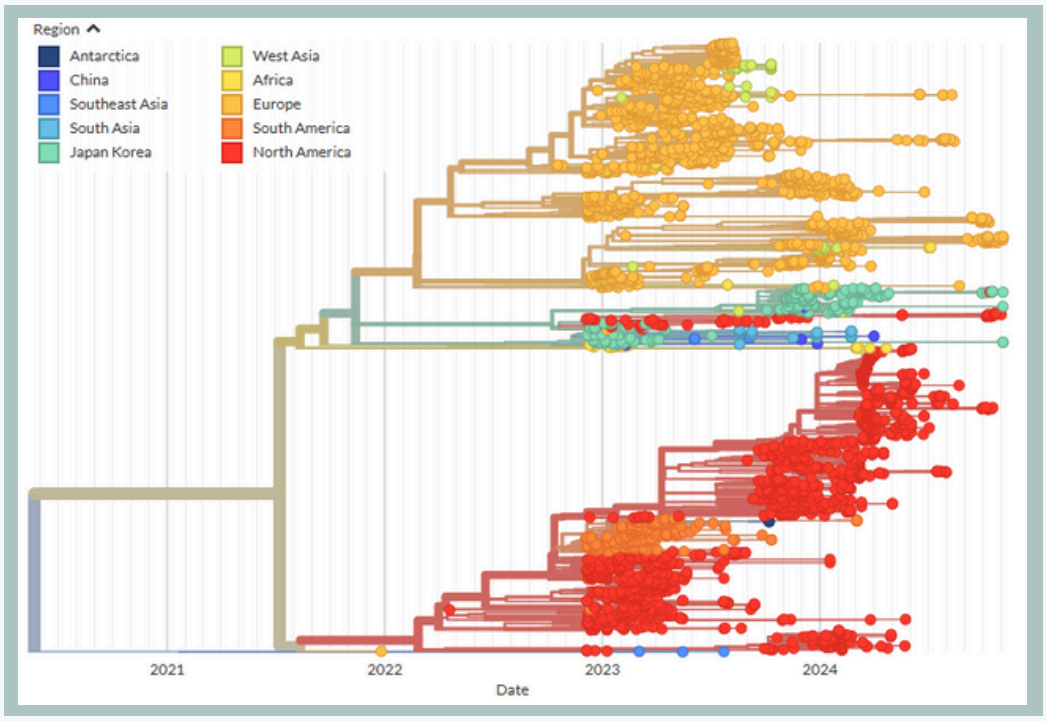
Phylogenetic trees are visualizations of the evolutionary relatedness between organisms. These trees are composed of tips, internal nodes, and branches.

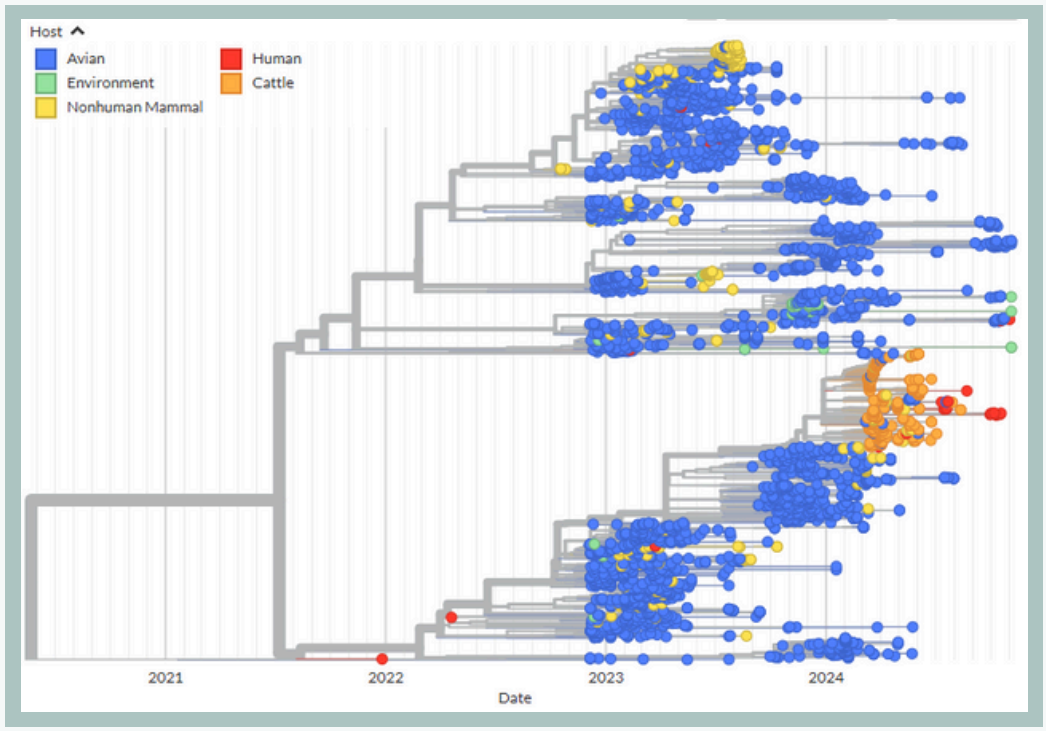
- **Tips**: directly observed samples
- **Internal nodes**: hypothetical common ancestors between samples that are not directly observed
- **Branches**: form connections between nodes and tips or nodes and other nodes

Mutations occur along branches where the parent node does not have a mutation that is present in its descendants. This pattern of shared mutations enables hierarchical clustering which is visualized in the phylogenetic tree. We can obtain information about shared mutations through multiple sequence alignment as shown here:



Subgroups descending from a common ancestor are referred to as **clades**. These are defined by the mutations that are shared by the samples within the clade. Embellishments can be added to trees to add another layer of information. Common embellishments include geographical location or host, as shown below in trees from from Nextstrain:


Region
Antarctica
China
Southeast Asia
South Asia
Japan Korea
West Asia
Africa
Europe
South America
North America
2021
2022
2023
2024
Date


Host
Avian
Environment
Nonhuman Mammal
Human
Cattle
2021
2022
2023
2024
Date

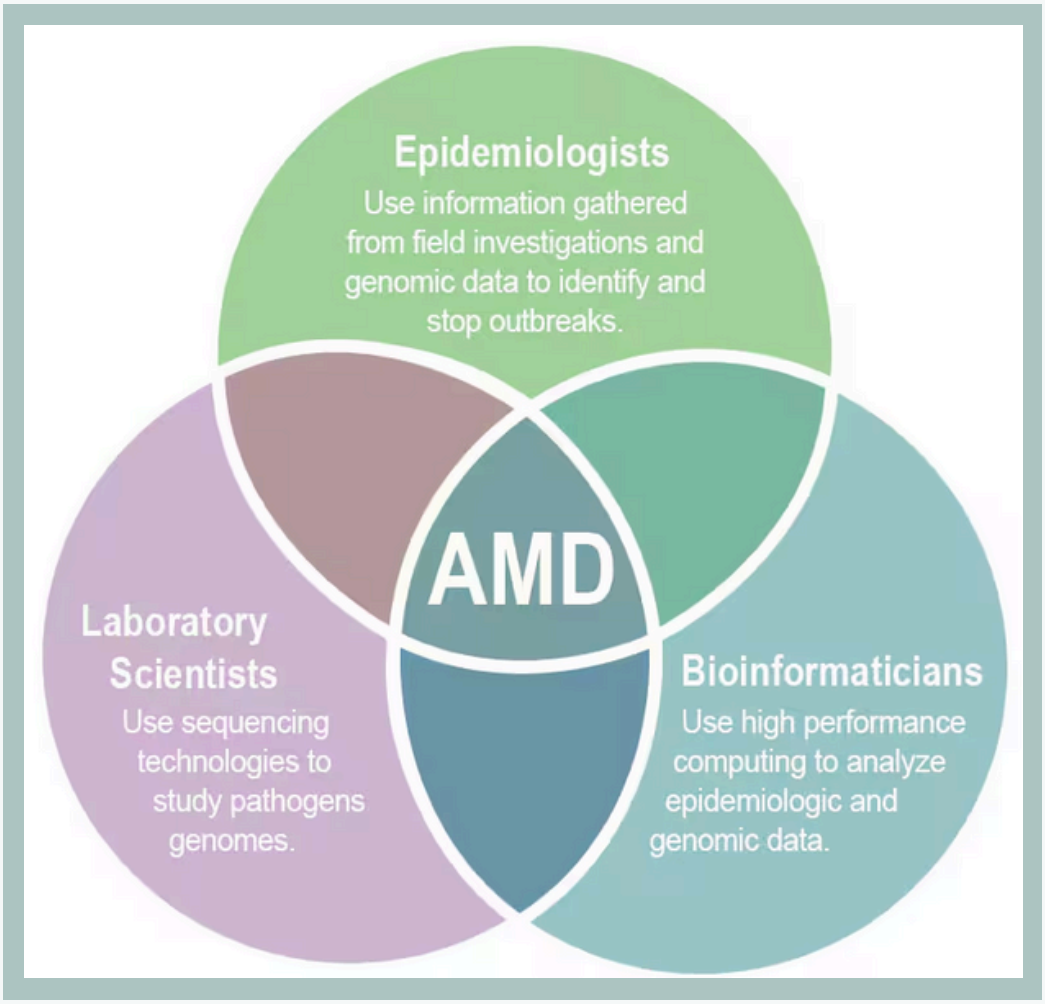
IV. Genomic Epidemiology in Practice

While genomic data is a powerful addition to traditional epidemiological practice, its utility is still dependent on epidemiological data. Primarily, timely access to clinical samples and data is most necessary. For a clinical sample to be sequenced, a set of metadata related to the case must also be obtained:

- The date of sample collection (and onset of symptoms if possible)
- The location of sampling

Other information including travel/contact history of the patient, suspected source of infection, and clinical outcomes can also further enhance the utility of sequencing. Assessment of rick factors including age, sex, and economic status can also be included.

There have been various programs and initiatives created to increase the use of genomic data in epidemiological investigations in public health laboratories. The foremost of these being the CDC's **Advanced Molecular Detection (AMD)** program, which is a $40 million per year initiative established by congress in 2014 to integrate laboratory,, bioinformatics, and epidemiology technologies across the nation.



V. Sources

Black, Allison and Dudas, Gytis 2023. An Applied Epidemiological Handbook. https://alliblk.github.io/genepi-book/

Ladner, J.T., Grubaugh, N.D., Pybus, O.G. et al. Precision epidemiology for infectious disease control. Nat Med 25, 206–211 (2019). https://doi.org/10.1038/s41591-019-0345-2

Grubaugh, N.D., Ladner, J.T., Lemey, P. et al. Tracking virus outbreaks in the twenty-first century. Nat Microbiol 4, 10–19 (2019). https://doi.org/10.1038/s41564-018-0296-2

Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. 2018 Jan;19(1):9-20. doi: 10.1038/nrg.2017.88. Epub 2017 Nov 13. PMID: 29129921; PMCID: PMC7097748.

Chow, N. (2024, April 11). Module 1.1 – what is genomic epidemiology?. Centers for Disease Control and Prevention. https://www.cdc.gov/advanced-molecular-detection/php/training/module-1-1.html

Sources for Examples of Genomic Epidemiology in Practice

Pathogen	Location	Findings
MRSA	Cambridge, UK	S. R. et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet. Infect. Dis. 13, 130–136 (2013).
Ebola Virus	West Africa	S. E. et al. Molecular evidence of sexual transmission of ebola virus. N. Engl. J. Med. 373, 2448–2454 (2015).; Butler, D. What first case of sexually transmitted Ebola means for public health. Nature News https://doi.org/10.1038/nature.2015.18584 (2015).; Christie, A. et al. Possible sexual transmission of Ebola virus - Liberia, 2015. MMWR. Morb. Mortal. Wkly. Rep. 64, 479–481 (2015)

HIV	USA	Simen, B. B. et al. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. J. Infect. Dis. 199, 693–701 (2009).
HIV	British Columbia, Canada	Poon, A. F. Y. et al. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. Lancet HIV 3, e231–e238 (2016).
Candida auris	Oxford, UK	Oxford, UK & Eyre, D. W. et al. A Candida Auris outbreak and its control in an intensive care setting. N. Engl. J. Med. 379, 1322–1331 (2018).
Yellow Fever	Brazil	Faria, N. R. et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. Science 361, 894–899 (2018).
Zika Virus	Florida, USA	Grubaugh, N. D. et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature 546, 401–405 (2017).
Lujo virus	Zambia and South Africa	Zambia and South Africa & Briese, T. et al. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. PLoS. Pathog. 5, e1000455 (2009).
Listeria monocytogenes	USA	Jackson, B. R. et al. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. Clin. Infect. Dis. 63, 380–386 (2016).

Influenza virus	Worldwide	Neher, R. A., Bedford, T., Daniels, R. S., Russell, C. A., Shraiman, B. I. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proc. Natl Acad. Sci. USA 113, E1701–E1709 (2016).
E. coli	Germany and France	Germany and France & Grad, Y. H. et al. Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011. Proc. Natl Acad. Sci. USA 109, 3065–3070 (2012).