Use of Bioinformatics in Genomic Epidemiology, and Pathogen Surveillance

Julien A. Nguinkal

1. Theoretical Background

1.1 Definitions and Principles of Genomic Epidemiology

Epidemiology and Genomic Epidemiology:
Epidemiology is the study of the distribution and determinants of health-related states or events (like disease) in specific populations, and the application of this study to control health problems. Genomic Epidemiology applies this concept at the genomic level, using pathogen genomic data to determine the distribution and spread of an infectious disease in a population, and applying that knowledge to control the disease. In essence, genomic epidemiology marries classical epidemiology with genomics, allowing scientists to track *pathogens* by their genetic fingerprints.

Pathogen Genomics in Public Health:
Pathogen genomics involves sequencing the entire genome of an infectious agent (bacteria, virus, or parasite) to glean detailed information about its characteristics and evolution. By comparing genomes from multiple isolates, researchers can infer how pathogens are related. This approach has emerged as an *indispensable mechanism for sharing sequence and metadata*, facilitating global health security. Public health agencies now use genomic data to detect outbreaks, monitor pathogen evolution, and design interventions.

Evolutionary and Transmission Insights:
Genomic epidemiology provides unparalleled resolution for distinguishing even closely related strains. For example, Whole Genome Sequencing (WGS) has replaced older techniques like pulsed-field gel electrophoresis (PFGE) as the “gold standard” for subtyping many pathogens. By examining single-nucleotide polymorphisms (SNPs) or sequence differences across genomes, we can discern transmission chains and outbreak clusters with high precision. During the 2014 Ebola outbreak, sequencing 99 Ebola virus genomes helped trace the outbreak to a single introduction and track its human-to-human spread, illustrating the power of genomic data in understanding disease dynamics.

One Health and Genomic Surveillance:
Genomic epidemiology aligns with the “One Health” concept, which recognizes the interconnected health of humans, animals, and the environment. Many emerging infectious diseases are zoonotic (e.g., Ebola, avian influenza), jumping from animals to humans. Genomic surveillance enables us to monitor these cross-species transmissions by analyzing pathogen genomes from various sources. For instance, sequencing of *H5N1 avian influenza* from a human patient and comparison with bird and cattle isolates identified a specific mutation (PB2 E67K) associated with adaptation to mammals. This underscored the need for vigilant genomic surveillance to catch early signs of increased virulence or transmissibility in zoonotic viruses.

Precision Epidemiology:
Precision epidemiology is an emerging approach that leverages detailed genomic data alongside clinical and ecological information to target public health interventions more effectively. Instead of one-size-fits-all measures, interventions can be tailored to specific strains or transmission networks (e.g., focusing vaccination or containment on identified clusters). As Ladner et al. (2019) describe, *precision epidemiology for infectious disease control* emphasizes integrating genomic sequencing into routine surveillance for quick, data-driven responses (as referenced in CDC's AMD Module).

Genomics in Outbreak Response:
Genomic epidemiology helps answer critical questions during outbreaks: How are cases connected? Is this a single outbreak or multiple unrelated ones? What is the source of infection? By building phylogenetic trees of pathogen genomes, we can visualize transmission links and estimate when the outbreak started (the *most recent common ancestor* of the outbreak strains). This was key in the early COVID-19 pandemic; for example, genomic analyses in early 2020 revealed cryptic transmission of SARS-CoV-2 in Washington State weeks before widespread testing was available, indicating the virus had been spreading undetected. Such insights inform public health action by highlighting the need for broader testing, travel restrictions, or targeted interventions.

1.2 Role of Bioinformatics in Pathogen Surveillance and Outbreak Detection

Bioinformatics Integration:
Modern genomic epidemiology would be impossible without bioinformatics – the suite of computational tools and methods for analyzing biological data. Advanced Molecular Detection (AMD) programs, such as those at CDC, *integrate next-generation genomic sequencing technologies with bioinformatics and epidemiology expertise to find, track, and stop spread of infectious pathogens*. AMD integrates traditional epidemiology with next-generation sequencing and bioinformatics to provide detailed information on disease-causing organisms. Bioinformatics pipelines process raw sequencing data (which may be billions of DNA bases per sample) into meaningful information: high-quality genome sequences, lists of genetic mutations, phylogenetic trees, and more.

Data Processing and QC:
A typical surveillance bioinformatics workflow starts with quality control (QC) of raw reads from next-generation sequencing (NGS). Tools like FastQC (for assessing read quality) and fastp/Trimmomatic (for trimming low-quality sequences) ensure that only high-quality data enter analysis. For example, in a *whole-genome sequencing* run of *Salmonella* from a foodborne outbreak, bioinformaticians would remove adapter sequences and low-quality bases before attempting genome assembly, variant calling or antimicrobial resistance genes (ARGs). This QC step is crucial: downstream analyses depend on accurate data, and poor-quality reads can lead to false variants or misassemblies (Garbage in ⇒ garbage out !).

Genome Assembly and Alignment:
Bioinformatics enables assembly of genomes (putting short reads together into longer contigs or complete chromosomes) or alignment of reads to a reference genome. In outbreak detection, *read mapping* is common – sequencing reads from an isolate are aligned to a reference genome of the species to identify differences (mutations). For instance, with *Mycobacterium tuberculosis*, mapping

reads to a reference genome can reveal specific mutations linked to drug resistance or transmission clusters. Bioinformatic tools like BWA or Bowtie2 (for alignment), Flye/Spades (for assembly) and SAMtools (for processing alignments) are standard in this step.

Variant Calling and Interpretation:
After alignment, *variant callers* (e.g., FreeBayes, GATK, or Snippy) identify SNPs and small insertions/deletions (indels). In a surveillance context, variant calling allows comparison of isolates. A cluster of clinical isolates with zero or few SNP differences suggests a tightly linked transmission chain (potential outbreak), whereas tens of SNP differences likely indicate unrelated cases. Public health labs often set SNP thresholds (e.g., ≤5 SNP differences within an outbreak cluster for *Listeria monocytogenes*), based on empirical studies, to decide if cases should be investigated as related. Bioinformatics makes such analysis routine and rapid, often with automated pipelines flagging new clusters in real-time as sequences are uploaded to databases.

Phylogenetics and Phylogeography:
Bioinformatics tools like IQ-TREE, RAxML, or FastTree allow construction of phylogenetic trees from genome sequences, which are then visualized with tools such as FigTree or interactive platforms like Nextstrain. Phylogenetic analysis reveals evolutionary relationships – in an outbreak, the tree might show all patient isolates forming a tight cluster distinct from other strains, confirming a common source. Phylogeography goes a step further, integrating the phylogeny with metadata (like sample location or date) to infer how a pathogen spreads in time and space. A prime example is Nextstrain's maps for COVID-19: as new sequences come in from around the world, Nextstrain's pipeline (Augur for analysis, Auspice for visualization) updates global trees, showing how different lineages (Alpha, Delta, Omicron, etc.) emerged and spread across regions.

Epidemiological Modeling with Genomic Data:
Beyond simply mapping out relationships, bioinformatics outputs feed into epidemiological models. For example, the number of SNP differences between viral genomes can be plugged into a molecular clock model to estimate how long an outbreak has been going on. During COVID-19, genomic data were used to estimate the reproductive number of emerging variants and the timing of introduction events in various countries. Likewise, phylogenetic clustering combined with case data can improve estimates of transmission rates or the effectiveness of control measures, a field known as *phylodynamics*. Tools like BEAST (Bayesian Evolutionary Analysis Sampling Trees) integrate genetic data with models of infection spread to infer epidemiological parameters (e.g., the date of the most recent common ancestor of an outbreak, which approximates when that outbreak started spreading).

Automation and High-Throughput Analysis:
The sheer volume of data in genomic surveillance is enormous – for instance, NCBI's Pathogen Detection program has analyzed over **2 million** pathogen genomes for outbreak detection and AMR tracking. Bioinformatics pipelines are automated to handle these loads. Pipeline managers or workflow languages (Snakemake, Nextflow) are used to orchestrate multi-step analyses from raw data to final reports. This automation allows real-time surveillance: as soon as a lab sequences a pathogen and uploads data (e.g., to NCBI or GISAID), pipelines can automatically incorporate it, re-run clustering analyses, and alert if a new cluster of concern is detected. For example, NCBI's

system clusters genomes that are genetically close (using SNP distances) to signal potential transmission chains.

Advantages in Outbreak Detection:
By using bioinformatics in surveillance, outbreaks can be caught and investigated *faster and with greater confidence*. Traditional methods might have missed that a set of cases across states were linked, but WGS analysis can reveal a tight genetic cluster, prompting an investigation. A case in point: a 2018 multistate *Salmonella* outbreak in the US had a common PFGE pattern (traditional DNA fingerprinting couldn't distinguish them), but WGS showed not all cases were related – some were a separate strain coincidentally sharing the PFGE pattern. Simultaneously, WGS linked other cases with different PFGE patterns to the outbreak. This refined the case definition, leading to a more accurate recall of contaminated food. Without bioinformatics analysis of the WGS data, investigators might have been misled, illustrating how crucial computational analysis is for interpreting complex genomic information.

1.3 Core Methodologies in Genomic Epidemiology

Next-Generation Sequencing (NGS):
At the heart of genomic epidemiology is Next-Generation Sequencing. NGS technologies (Illumina, Oxford Nanopore, PacBio, etc.) allow rapid sequencing of entire genomes. In public health labs, Illumina short-read sequencing is widely used for its accuracy and throughput. Sequencing can be done on cultured isolates (common for bacteria and some viruses) or directly from clinical samples (often for viruses like SARS-CoV-2, or for metagenomics). The output of NGS is millions of short sequences ("reads") that represent fragments of the pathogen's genome.

- **Whole-Genome Sequencing (WGS):** Sequencing the complete genome of a pathogen isolate. WGS yields the highest resolution data for comparing strains. It's now used routinely for surveillance of foodborne bacteria (e.g., PulseNet labs sequence *Salmonella, E. coli, Listeria* from patients and food). WGS data can be analyzed by SNP-based methods or gene-by-gene methods like core genome MLST (cgMLST). SNP analysis compares genomes base-by-base to find even single nucleotide differences. cgMLST, on the other hand, compares a set of thousands of conserved genes and assigns sequences types, which is useful for standardizing across labs. Both approaches are valuable; SNP analysis can offer finer resolution (more precise phylogeny), while cgMLST provides easier comparison across databases and is commonly used in cross-border investigations.
- **Targeted NGS (amplicon sequencing):** Sometimes instead of whole genomes, specific genomic regions are sequenced from many samples – for instance, 16S rRNA gene sequencing in microbiome studies, or amplicon deep sequencing of a particular viral gene to detect minority variants. In genomic epidemiology, targeted sequencing might be used for *genotyping* (e.g., sequencing the TB spoligotyping regions) but increasingly WGS is preferred due to cost reductions and greater information.
- **Metagenomic Sequencing:** Sequencing all DNA/RNA in a sample without prior culturing. This is crucial for identifying pathogens in complex samples (like stool, environmental swabs, or even wastewater). Metagenomics can reveal *entire communities of microorganisms* and detect pathogens that standard methods might miss. An example is using metagenomics for *AMR surveillance*: instead of isolating specific bacteria, sequencing a sewage sample can

show the “resistome” (the collection of AMR genes present) in a community. Metagenomics offers breadth (many organisms at once) but requires powerful bioinformatics to filter host DNA and analyze mixed data. For outbreak scenarios, it’s been used in cases where the causative agent is unknown or difficult to culture (e.g., investigating mysterious pneumonia cases by sequencing bronchoalveolar lavage fluid to identify a novel coronavirus).

Bioinformatics Pipelines:
Core methodologies involve a series of steps (often automated) to turn sequencing data into actionable results:

1. **Quality Control (QC):** As mentioned, tools like FastQC and adapter trimming tools ensure sequence data is of high quality. Any pipeline begins with QC to avoid garbage-in, garbage-out issues.
2. **Genome Assembly or Mapping:** Depending on the question, either de novo assembly (using assemblers like SPAdes, SKESA for bacterial genomes) or mapping to a reference (using aligners like BWA-MEM for variant calling) is performed. For bacteria, assembling a draft genome can help with gene finding (for virulence or resistance genes via Prokka or similar annotation tools). For viruses, mapping to a known reference genome (like mapping SARS-CoV-2 reads to the Wuhan reference) is common, followed by calling variants relative to that reference.
3. **Variant Calling:** Identifying SNPs and small indels relative to a reference. Tools such as *Snippy* provide an all-in-one pipeline for bacterial variant calling and even core genome alignment generation (Snippy will align reads to a reference and produce a core SNP alignment for phylogenetics). *GATK* and *FreeBayes* are common for variant calling, especially in viral genomics to identify intra-host variants or consensus changes.
4. **Phylogenetics:** Using the SNP alignment or whole genome alignment, build phylogenetic trees to explore relationships. *IQ-TREE*, *RAxML-NG*, *FastTree* are widely used, each balancing speed and accuracy. *Nextstrain’s Augur* pipeline wraps many of these steps specifically for real-time virus tracking, including phylogenetic inference and temporal dating of trees.
5. **Cluster Analysis and Typing:** For bacteria, classical *molecular typing* methods are also applied in silico. For example, Multi-Locus Sequence Typing (MLST) sequence types can be derived from WGS data (tools like mlst or sequencescape do this). *cgMLST* schemes cover thousands of loci and assign allelic profiles to compare strains; many public health labs have cutoffs (number of allele differences) to define clusters in cgMLST space. Separately, SNP clustering can define genetic clusters by threshold (like “up to 5 SNP differences” cluster definition). NCBI’s Pathogen Detection uses a SNP clustering approach: it automatically clusters isolates that are genetically similar into putative outbreak clusters, visible via their Pathogen Detection browser. This method has helped identify multistate outbreaks and even link cases internationally that would otherwise appear unrelated.
6. **Epidemiological Data Integration:** The power of genomic epidemiology is fully realized when genomic data are combined with epidemiological metadata (who, where, when, exposures). Core methodology includes linking sequence data with patient data. This can be as simple as using a spreadsheet or as complex as building a database or using tools like MicrobeTrace or CIDRAP dashboards. Some specialized tools and formats (like Nextstrain’s JSONs or GISAID’s metadata submission system) encourage standardized data linkage (e.g., including date of

sample collection, location, patient age/gender, etc., with sequences). Downstream, analytic methods like *transmission network inference* (e.g., using phybreak or outbreaker2 R packages) can use genomic distances plus epidemiological dates to infer who likely infected whom in an outbreak, bridging the fields of genomics and traditional shoe-leather epidemiology.

Phylogenetic Tree Interpretation:
Understanding phylogenetic trees is fundamental. Nodes represent sequences (viruses/bacteria isolates), and branches represent genetic change. A *rooted phylogeny* can suggest the direction of evolution; for example, the root might be the earliest case in an outbreak. Short branches indicate closely related samples (perhaps a quick transmission), while longer branches suggest more time or undetected cases. *Molecular clock* models assume a roughly constant rate of mutations over time, which allows dating of ancestral nodes (e.g., estimating the date a particular variant of SARS-CoV-2 first emerged). A key principle is that the fewer the genetic differences between two isolates, the more likely they are epidemiologically linked, but one must consider sampling and random mutation – not every transmission will accumulate the exact same number of mutations.

Epidemiological Modeling:
Core methodologies also span into modeling like SIR (Susceptible-Infected-Recovered) models or more complex agent-based models that incorporate genomic data. For example, genomic data might inform a parameter like introduction frequency in a region (if phylogenetics indicates multiple separate introductions vs. one spreader event). Techniques like *coalescent-based phylodynamics* (using BEAST, or newer tools like Nextstrain's tree time) merge phylogenetics with models of population growth. In public health practice, simpler approaches are often used – like looking at doubling times of genomic lineages to see if an intervention slowed spread, or mapping movement of strains between regions to adjust travel advisories.

Summary: Core methodologies in genomic epidemiology revolve around sequencing pathogens and analyzing their genomes. Next-Generation Sequencing provides the raw data; bioinformatics pipelines turn that into variants and phylogenies; epidemiology principles guide how to interpret and act on those results. The synergy of these methodologies yields powerful insights: from pinpointing a food source in a national outbreak to tracking a virus's mutations as it circles the globe, ultimately informing interventions and saving lives.

2. Practical Workflows : Real world use cases

2.1 Genomic Surveillance Protocols by Pathogen Type

Overview: Each pathogen type – bacterial, viral, and metagenomic – requires a tailored genomic surveillance workflow due to differences in genome structure, mutation rates, and typical data volumes. However, the overarching steps (sample collection → sequencing → analysis → interpretation) are similar. Below, we outline step-by-step protocols for each category, highlighting both commonalities and unique considerations.

2.1.1 Bacterial Pathogens (e.g., *E. coli*, *Salmonella*, *Mycobacterium tuberculosis*)

1. **Sample Collection & Culture:**
    - **Clinical sampling:** Collect specimens from infected individuals (blood, stool, sputum, etc.) following standard clinical protocols. For food/environmental surveillance, obtain food samples, water, surfaces swabs etc.
    - **Isolation:** Culture the bacterium on appropriate media (for *Salmonella*, use selective agar; for *M. tuberculosis*, use Löwenstein-Jensen or liquid culture).
    - **Confirmatory tests:** Confirm identity via biochemical tests or MALDI-TOF MS.
2. **DNA Extraction:**
    - Extract high-quality genomic DNA from bacterial cultures. Use kits (e.g., Qiagen DNeasy) or automated systems; ensure RNA removal for a pure DNA prep.
    - For Gram-positive bacteria or mycobacteria, include a bead-beating or enzymatic lysis step to break tough cell walls.
3. **Library Preparation:**
    - If using Illumina sequencing, prepare a short-insert library (~300-600 bp inserts typically). This could be via Nextera XT (transposase-based) or ligation-based kits. Normalize DNA input (e.g., 1 ng for Nextera XT, ~100 ng for ligation protocols).
    - Optionally, for *long-read* sequencing (Oxford Nanopore/PacBio) to resolve plasmids or repeats, prepare libraries per those protocols (often requiring more DNA and different cleanup).
4. **Sequencing (NGS):**
    - Sequence using the platform available. For surveillance, many public health labs multiplex dozens of bacterial genomes per Illumina run (e.g., using unique dual indices). Aim for >30x coverage for robust variant calling.
    - Check run quality metrics (e.g., % bases above Q30, total yield per sample).
5. **Data Transfer & Storage:**
    - Demultiplex sequences and generate FASTQ files for each isolate. Securely transfer data to analysis servers or cloud storage, maintaining sample metadata (lab ID, patient ID, date, location).
    - Adhere to naming conventions linking sequence files to sample records to avoid mix-ups.
6. **Bioinformatics Pipeline:**
    - **Quality Control:** Use FastQC to get per-sample quality reports; trim adapters and low-quality tails using tools like Trimmomatic or fastp.
    - **Genome Assembly (if needed):** For gene detection (AMR genes, virulence factors), assemble the genome using SPAdes or Unicycler (the latter especially if combining short and long reads to close bacterial genomes). Check assembly metrics (N50, total length, number of contigs).
    - **Reference Alignment & Variant Calling:** Alternatively, map reads to a reference genome of the same species using BWA-MEM. Use Snippy for a quick all-in-one SNP calling (particularly popular for bacterial surveillance – Snippy yields a core SNP alignment and variant list).
    - **Phylogenetic Analysis:** If comparing multiple isolates, take the core genome alignment from Snippy (or construct one from assembled genomes using Roary or Parsnp) and build a phylogenetic tree (FastTree for speed, or RAxML/IQ-TREE for more accuracy). This tree will show relatedness among isolates.

- **Cluster Detection:** Analyze SNP distances between isolates. Many labs use a simple matrix of pairwise SNP differences and apply a threshold (e.g., ≤5 SNPs difference as a cluster threshold for *Listeria* or *E. coli* O157). Tools like snp-dists can compute this matrix from an alignment.
- **Annotation and Typing:** Run tools like:
  - **MLST:** to get traditional sequence type (ST) from assemblies or reads (e.g., using mlst command with PubMLST schemas).
  - **AMR Gene Detection:** use ResFinder or Abricate (with ResFinder and CARD databases) to list acquired resistance genes. For *M. tuberculosis*, call known resistance-associated SNPs (in genes like rpoB, katG; TB-Profiler is a specialized tool).
  - **Virulence/Plasmid:** use tools like VirulenceFinder or PlasmidFinder for additional context.
- **Metadata Integration:** Prepare a table that includes for each isolate: sample ID, collection date, location, patient metadata, ST, key resistance genes, etc., plus any epidemiological links known (e.g., "Isolate A and B from same restaurant on same date").

7. **Interpretation and Surveillance Action:**
   - **Cluster investigation:** If a cluster of genetically related isolates is identified (and particularly if they share an unusual resistance gene or come from the same time frame), epidemiologists are alerted. Investigate common exposures, sources, or contacts. For example, a cluster of *Salmonella* with the same rare serotype and nearly identical genomes might all trace back to a specific food distributor.
   - **Database submission:** Share data via public databases. For routine surveillance, labs upload raw reads to NCBI's Sequence Read Archive (SRA) and/or assembled genomes to NCBI's Pathogen Detection system, which will further cluster and compare with international data. Also consider uploading to GenomeTrakr (FDA's network for food isolates) or PulseNet (which now uses WGS data).
   - **Report generation:** Create a surveillance report or update. This typically includes a dendrogram (phylogenetic tree) highlighting clusters of interest, a map if relevant, and tables of key genomic findings (like resistance profiles). This textbook-style content might feature a figure of a phylogenetic tree of bacterial isolates from an outbreak investigation.

2.1.2 Viral Pathogens (e.g., SARS-CoV-2, Influenza, Ebola)

1. **Sample Collection:**
   - **Clinical specimens:** For respiratory viruses like SARS-CoV-2 or influenza, collect nasal or throat swabs (often in viral transport medium). For blood-borne or systemic viruses (like Ebola), blood samples or oral swabs in certain cases. Ensure proper biosafety since some viruses (Ebola) require BSL-4 conditions for handling live virus.
   - **Inactivation or RNA Extraction:** Many protocols inactivate viruses using lysis buffers and then extract nucleic acid. For RNA viruses, use kits that preserve RNA or use TRIzol/column extraction. Include DNase treatment if the library prep is RNA-specific.
2. **Reverse Transcription (for RNA viruses):**

- Convert RNA to cDNA using reverse transcriptase. Some sequencing library kits (like Illumina DNA prep) can work off cDNA directly; others use amplicon approaches (like the ARTIC protocol for SARS-CoV-2, which uses ~400 bp tiled amplicons).
- *Amplicon vs. Metagenomic:* For targeted surveillance (like SARS-CoV-2 during the pandemic), a common approach was ARTIC amplicon sequencing: design primers to amplify the whole viral genome in ~100 overlapping pieces, sequence those. Alternatively, metagenomic (random-primed) sequencing can capture viral genomes but may have more host contamination.

3. **Library Preparation:**
    - **Illumina:** If using amplicons, you can barcode and pool multiple samples (up to 384 or more with combinatorial dual indices). Use a DNA library prep kit on the cDNA or amplicon pool.
    - **Oxford Nanopore:** For rapid results (e.g., during an outbreak), nanopore sequencing is often used. E.g., the MinION device was used in the 2014 Ebola outbreak; protocols like ARTIC have a nanopore version (with different primers and an Oxford Nanopore-specific library prep using PCR barcoding kits).
    - **Unique Molecular Identifiers (UMIs):** For some virus sequencing, especially for intrahost variant analysis, methods add UMIs to avoid PCR bias and error inflation.
4. **Sequencing:**
    - Generate sequence data. For surveillance, a moderate depth (e.g., 1000x coverage for a viral genome) is often enough to generate a high-quality consensus sequence. If tracking minor variants within a host (like minor quasispecies in an HIV sample), higher depth might be used.
    - Monitor run performance: balanced coverage across genome (amplicon dropouts are common, e.g., some SARS-CoV-2 primer sets had dropouts in presence of certain variants). If using Illumina, check percentage of viral reads (metagenomic runs might yield a lot of host reads, lowering efficiency).
5. **Data Processing:**
    - **Demultiplexing and QC:** For multiplexed runs, separate reads by barcode. Perform quality trimming (especially important for nanopore, which often has higher error rates; for Illumina, quality is usually high but remove any adapter or primer sequences).
    - **Reference Alignment:** Map reads to reference genome of the virus. For SARS-CoV-2, use the Wuhan-Hu-1 reference; for Ebola, the appropriate species/Zaire ebolavirus reference. Tools: Bowtie2 or BWA for Illumina; Minimap2 for nanopore.
    - **Consensus Calling:** Derive the consensus sequence for each sample (the most common nucleotide at each position, perhaps with IUPAC ambiguities for tie/low coverage). Tools like iVar (commonly used with ARTIC) or samtools/bcftools can call a consensus.
    - **Variant Calling:** Identify differences from the reference. In outbreak surveillance, these differences help assign lineage or clade. E.g., call SNPs for each SARS-CoV-2 sample to feed into Pangolin (for lineage assignment) or Nextstrain (for clade assignment).

- **Quality Checks:** Remove samples that fail criteria (like <90% genome recovered at acceptable coverage, or too many Ns in the sequence). Many surveillance efforts set thresholds to ensure only good sequences are analyzed/shared.

6. **Downstream Analysis:**
    - **Lineage/Clade Assignment:** Use Pangolin for SARS-CoV-2 to assign a Pango lineage (e.g., B.1.1.529 for Omicron) which is epidemiologically informative. For influenza, use Nextstrain clade naming or WHO clades (e.g., H3N2 3C.2a1 etc.); for Ebola, often each outbreak gets a clade designation.
    - **Phylogenetic Tree:** If analyzing a batch, include global context sequences. For example, when analyzing a new Ebola sequence, one might download recent Ebola sequences from GenBank to see where the new case fits in (is it part of a known chain or a new introduction?). Build a tree (using Nextstrain's augur toolkit or standard phylogenetics software).
    - **Phylogeographic Mapping:** If numerous sequences over time, tools like Nextstrain offer a temporal analysis (tree with time on x-axis showing progression of mutations over the months of an outbreak) and geographic mapping (coloring branches by region, animating spread). This can highlight, for instance, multiple introductions of a virus vs. local spread.
    - **Mutation Surveillance:** Track mutations of interest. For SARS-CoV-2, labs monitored specific spike protein mutations (E484K, N501Y, etc.) that affect transmissibility or immunity. A simple script or tool can scan each consensus for presence/absence of key mutations and output a table.
7. **Integration & Action:**
    - **Reporting to Health Authorities:** Provide results to epidemiologists and decision-makers. For a viral outbreak, this could mean informing that "all cases are genetically similar, suggesting one introduction" or "we have multiple distinct lineages, so multiple introductions occurred." E.g., in COVID-19, genomic analysis helped determine whether a surge in cases was due to a new variant introduction or local spread of existing lineages.
    - **Database Submission:** Upload sequences to databases like GISAID (for influenza and SARS-CoV-2, which require data sharing agreements), or GenBank for open access. These platforms allow global scientists to see and use the data, and often require accompanying metadata (collection date, location, etc.).
    - **Public Sharing and Dashboards:** Many jurisdictions have dashboards (like Nextstrain builds specific to regions or a website showing the latest local sequences). Keeping these updated helps maintain transparency and public awareness of variant spread. For instance, Nextstrain Community builds let local groups upload data and share interactive phylogenies easily.
    - **Contact Tracing Augmentation:** In some cases, use sequences to aid contact tracing. If two patients claim no contact but have virtually identical virus genomes, investigators might dig deeper for a hidden link or overlapping location (e.g., same grocery store). Conversely, if a suspected transmission pair have very different sequences, maybe the disease was acquired elsewhere, adjusting the investigation.

2.1.3 Metagenomic Surveillance (Pathogen Discovery & Microbiome-based)

1. **Sample Collection:**
    - **Environmental Surveillance:** Samples like wastewater, sewage, or river water for broad pathogen monitoring. For example, many cities started **wastewater surveillance** for SARS-CoV-2, where weekly sewage samples are analyzed for viral RNA as an early indicator of community infection levels.
    - **Clinical/One Health Samples:** Stool samples for gastrointestinal outbreaks where multiple pathogens are possible, nasal swabs for respiratory microbiome analysis, or samples from animal reservoirs (soil, farm effluents). The One Health approach uses metagenomics to look at AMR genes across human, animal, and environmental interfaces.
    - **Pre-processing:** Often required due to sample complexity. E.g., filter water to concentrate microbes, or use centrifugation to pellet bacteria/viruses. For stool, a homogenization and heavy centrifugation or filtration step might separate bacteria from particulate matter.
2. **Nucleic Acid Extraction:**
    - Because metagenomic samples contain mixtures (bacteria, viruses, human DNA/RNA), extraction methods should recover all types. Protocols might include a bead-beating step to break tough cells (releasing bacterial DNA) and column capture or magnetic beads to collect total nucleic acids.
    - For RNA viruses in metagenomes, include an RNA extraction and reverse transcription. Some workflows do separate DNA and RNA sequencing to capture both DNA genomes (bacteria, DNA viruses) and RNA viruses.
3. **Library Prep – Shotgun Sequencing:**
    - Typically random amplification or shotgun library prep is used (no specific target). This yields a metagenomic library of all DNA (and cDNA from RNA) in the sample.
    - To increase pathogen yield, methods like **host DNA depletion** are common: e.g., in blood samples, target and remove human DNA (using DNase or enrichment kits) so that more sequencing reads come from microbial DNA.
    - Unique to metagenomics: sometimes ultra-low DNA inputs require amplification (like using whole genome amplification kits, though those can introduce bias).
    - If high-throughput, can multiplex multiple samples, but metagenomic runs often use a lot of reads per sample (to dig deep into diversity), so sometimes only a few samples per sequencing run.
4. **High-Throughput Sequencing:**
    - Illumina is common for its accuracy (useful when assembling unknown pathogen genomes). NovaSeq or HiSeq lanes might sequence dozens of millions of reads per sample. Nanopore is also used in field situations (like sequencing directly from human samples to find an unknown pathogen, e.g., during the early COVID outbreak in Wuhan had a metagenomic component).
    - The output is a big jumble of reads from potentially thousands of organisms.
5. **Bioinformatics Processing:**
    - **Quality Control:** As usual, remove low-quality reads. Also, **remove host reads**: For human clinical samples, map reads to the human genome and discard matches, to focus on non-human sequences (this is both to enrich for pathogens and for privacy

considerations since human reads could cover patient's genome). Similar removal can be done for animal samples (map to cow genome if analyzing cow feces, etc.).

- **Taxonomic Classification:** A key step in metagenomics is identifying what organisms are present. Tools like **Kraken2** (and its database of k-mer signatures) classify reads into taxa very quickly. Others include Centrifuge or Kaiju. They provide outputs like: X% reads are *E. coli*, Y% *Clostridium*, Z% *Norovirus*, etc. This gives a snapshot of microbiological composition.
- **Pathogen Detection:** Scan the taxonomy report for known pathogens. E.g., a stool metagenome might reveal a high proportion of *Shigella* reads, indicating likely *Shigella* infection, even if routine tests hadn't picked it up. Or wastewater might suddenly show poliovirus reads, triggering a public health response.
- **Genome Assembly:** If a particular pathogen is of interest (especially novel ones), attempt to assemble its genome from the metagenomic data. For bacteria, one can often assemble draft genomes if coverage is sufficient. For viruses, target assembly via specialized tools or extract the reads classified as that virus and assemble just those (which reduces interference from other data).
- **Binning (for complex samples):** In environmental samples, you might assemble many contigs that derive from different bacteria. Binning algorithms (Metabat, MaxBin) group contigs that likely come from the same organism based on coverage patterns and GC content, creating **Metagenome-Assembled Genomes (MAGs)**. This way, you can reconstruct new species or strains from environmental data.
- **AMR Gene Profiling:** Use tools like **AMR++** or directly run ResFinder or CARD's RGI on metagenomic assemblies or even raw reads to identify resistance genes in the sample. Metagenomics can uncover a wide range of AMR genes present at low abundance that targeted surveillance might miss. For instance, checking sewage from an international airport might show the diversity of AMR genes from travelers around the world (as has been done in some studies).
- **Variant Analysis:** If doing longitudinal metagenomics (e.g., weekly sewage samples), you might track the relative abundance of certain pathogen strains over time. Tools like Freyja (for SARS-CoV-2 in wastewater) deconvolute mixed lineage signals to estimate proportions of variants.

6. **Results Interpretation:**
    - **Pathogen Discovery:** If the goal was to find an unknown pathogen causing illness, once assembly yields a full genome (or significant partial), check if it's novel. Align it against databases (BLAST search). The 2019 novel coronavirus was identified by such means: metagenomic sequencing of patient BAL fluid, assembly of a novel coronavirus genome, which had ~88% similarity to SARS-like bat viruses – clearly something new.
    - **Surveillance Alerts:** For routine metagenomic surveillance (like food or environment), define what triggers an alert. It could be detection of a certain pathogen where it shouldn't be (e.g., Listeria in a food plant's drain), or a sudden increase in a particular virus in wastewater. Because metagenomics can detect many things, having clear reporting criteria is important (perhaps set thresholds or look for known outbreak signatures).

- **One Health Insights:** Metagenomics often reveals interesting cross-connections: e.g., finding *antibiotic resistance genes* in farm soil that are identical to those in hospital pathogens implies a flow of resistance in the environment. Surveillance teams may incorporate such findings into risk assessments or policymaking (like restricting certain antibiotic use in agriculture).
- **Database Sharing:** Submit assembled genomes of novel findings to GenBank (ideally with the "Metagenome" keyword). For microbial communities, international repositories exist (like MG-RAST or the European Nucleotide Archive's metagenome section) where raw reads can be shared. If a public health threat is found (like a new virus), rapid sharing (e.g., posting the sequence on NCBI or GISAID if applicable) is essential to alert the global community.

7. **Follow-up Actions:**
    - If a known pathogen is found, confirm with targeted tests. For instance, if metagenomics flags *Salmonella* in a food sample, follow up by culturing *Salmonella* from that sample if possible (to have an isolate for further testing and regulatory action).
    - If a novel organism is discovered, initiate research or emergency measures: e.g., when a new virus is discovered, immediately start developing diagnostic PCR assays to detect it in other patients, notify WHO if it's of public health significance (potential PHEIC – Public Health Emergency of International Concern).
    - Integrate with *traditional surveillance*: For example, wastewater data may suggest COVID-19 cases are rising even before clinical testing does, prompting health departments to prepare resources or issue warnings.

Metagenomic Workflow Example – Sewage Surveillance:
A city collects weekly sewage samples from the inlet of the wastewater treatment plant. The protocol: concentrate viruses from sewage (via filtration and ultracentrifugation), extract RNA, do random-primed cDNA and Illumina sequencing. The bioinformatics pipeline removes human reads, then uses Kraken2 to identify viruses. Over weeks, the city sees a rise of SARS-CoV-2 reads corresponding to a specific lineage (detected via lineage-specific mutations). This rise precedes a clinical uptick by two weeks. The city also intermittently detects *norovirus* and *rotavirus* spikes, reflecting GI illness in the community. These data are shared with hospitals and the public. Additionally, AMR genes like NDM-1 (a carbapenemase) are tracked; a sudden increase might prompt investigation into a source (perhaps a healthcare facility contributing to sewage). Metagenomic surveillance thus acts as an early warning and a comprehensive monitoring tool that complements case-based surveillance.

3. Case Studies

In this section, we delve into concrete case studies that showcase how genomic epidemiology and bioinformatics are applied in public health scenarios. Each case emphasizes different pathogens and challenges, demonstrating the versatility of genomic tools in solving real-world epidemiological cases.

3.1 COVID-19 Genomic Surveillance: Variant Tracking, Lineage Assignment, and Outbreak Monitoring

Context: COVID-19 (caused by SARS-CoV-2) is the hallmark example of genomic epidemiology in action. Never before have so many genomes of a single pathogen been sequenced in such a short time – By 2024, over **16 million SARS-CoV-2 sequences** had been uploaded to GISAID from 219 countries. Genomic surveillance has been pivotal in tracking the emergence of variants, understanding transmission, and informing public health decisions.

Early Days – Detecting Introduction and Spread: In January 2020, the first SARS-CoV-2 genome from Wuhan was shared, kicking off global sequencing efforts. As COVID-19 cases appeared worldwide, genomic data helped identify how the virus spread. For example, in Washington State (USA), the first COVID-19 case was in January 2020 from a traveler. Weeks later, a *high-school student* with no travel history tested positive. Sequencing revealed that the student's virus was genetically very similar to the traveler's virus sequence, with only a couple of differences, indicating the virus had been silently spreading in the community for weeks (cryptic transmission). This genomic evidence underscored the need for broader testing and that travel restrictions alone were too late.

Lineage and Variant Nomenclature: To manage the burgeoning genetic diversity of SARS-CoV-2, scientists developed lineage naming systems. The **Pangolin** tool (Phylogenetic Assignment of Named Global Outbreak Lineages) assigns each sequence a lineage label based on phylogeny. For instance, the initial strains were in lineage A and B, the European wave was linea

ge B.1, and many subdivisions happened thereafter. *Variants of Concern (VOCs)* got Greek letter names via WHO (Alpha, Beta, Gamma, Delta, Omicron correspond to lineages B.1.1.7, B.1.351, P.1, B.1.617.2, and B.1.1.529 respectively). Genomic surveillance allowed quick identification of these VOCs:

- **Alpha (B.1.1.7):** First detected in the UK (late 2020). Genomics showed it had an unusual number of mutations and was rapidly growing in frequency. Within weeks, it was found via sequencing in numerous countries, often seeded by travelers from the UK.
- **Delta (B.1.617.2):** Emerged in India (early 2021) and was tracked as it outcompeted other variants. Countries with strong genomic surveillance (e.g., UK) identified Delta variant cases increasing and informed global peers.
- **Omicron (B.1.1.529):** Detected by scientists in South Africa and Botswana in late 2021, who swiftly uploaded sequences to GISAID and alerted the world. Nextstrain visualizations made it immediately clear that Omicron was genetically very distinct (long branch on phylogenetic tree) and likely had been evolving under the radar, possibly in an immunocompromised patient or unsampled region. Omicron's many spike mutations correlated with immune evasion, explaining why it reinfected people and reduced vaccine effectiveness to some extent.

Nextstrain's Role: Throughout the pandemic, **Nextstrain's interactive phylogenetic trees** became a go-to resource for scientists and the public. Nextstrain showed how variants related to each other and their geographic spread. For example, a Nextstrain view might highlight clades like 20E (EU1) – a

variant that spread in Europe in summer 2020 – which was traced via genomes to holiday travel in Spain and then outward, demonstrating how travel policy impacted variant distribution.

Outbreak Investigation within COVID-19: Genomic data were also used on smaller scales:

- **Hospital Outbreaks:** If patients in a hospital COVID ward had sequences that were identical or nearly so, it suggested an in-hospital transmission (or common exposure). If they were very different, transmission likely happened elsewhere. This helped infection control teams to confirm or refute suspected hospital outbreaks, guiding isolation or ward closure decisions.
- **Superspreader Events:** For instance, a notable case – in Boston in 2020, a conference led to hundreds of secondary cases. Genomic analysis by Lemieux et al. showed that the genomes from conference-associated cases had distinct markers and ended up accounting for a significant proportion of cases in Boston area later (a single event seeding broader community spread).
- **Regional lineage dynamics:** In, say, California, genomic surveillance could show how lineages rose and fell with interventions. After lockdowns, diversity shrank (fewer introductions), but when travel resumed, new lineages appeared.

Public Health Impact: Genomic data directly influenced public health measures:

- Travel restrictions were placed on regions when variants of concern surfaced (e.g., many countries banned flights from the UK when Alpha was announced, or from Southern Africa after Omicron sequences emerged). There was controversy on effectiveness and ethics of these bans, but they underscore how genomics drove policy in real time.
- Vaccine updates: When Beta variant (B.1.351) with immune-escape mutations was found to reduce efficacy of some vaccines, companies began considering booster updates. Later, Omicron's emergence sped up development of Omicron-specific boosters. All based on genomic surveillance signaling antigenic drift.
- Fine-grained contact tracing: In Taiwan, an interesting small outbreak in 2021 was traced by genomics to a single hotel quarantine breach; despite multiple introductions, genomics helped pinpoint which introduction led to community cases, refining the search for the protocol failure.

Data Sharing and Global Collaboration: The COVID-19 pandemic showcased unprecedented data sharing via GISAID. However, it also revealed tensions. South Africa's quick sharing of Omicron data in 2021 was initially met with punitive travel bans, leading to fears that countries might hesitate to share data if it leads to economic harm. This sparked discussions on equitable data sharing: how to ensure credit and benefits to the data generators, and not just downsides. The global scientific community largely coalesced around open data for the common good, but with calls to avoid penalizing transparency.

Case Outcome: As of 2022-2023, genomic surveillance of SARS-CoV-2 started being scaled back in some areas (due to costs and a sense of pandemic fatigue), but many experts argue for its continuation as a model for future pandemics. The ability to watch a pathogen mutate and spread in

near-real-time was revolutionary. In summary, the COVID-19 genomic surveillance case study illustrates:

- Quick adaptation of genomic tools (some originally developed for Ebola/flu) to a new pathogen.
- The importance of standardizing analysis (everyone using Pangolin and similar schemes).
- Direct influence on public health (identifying variants that need targeted responses).
- Issues of data sharing ethics and equity.
- How an engaged global community (Nextstrain updates, outbreak.info dashboards, etc.) can keep both scientists and the public informed during a health crisis.

3.2 Antimicrobial Resistance (AMR) in Bacteria: WGS for Resistance Gene Detection

Context: Antimicrobial resistance is a growing public health crisis often described as a "silent pandemic." Genomic methods provide a powerful lens to detect and track AMR. This case study explores how WGS is used to identify resistance genes in pathogens and inform public health interventions, using examples such as multi-drug resistant *Tuberculosis* and *carbapenemase-producing Enterobacteriaceae*.

Case A – Extensively Drug-Resistant Tuberculosis (XDR-TB):
Tuberculosis can evolve resistance to multiple drugs, and genomic sequencing has transformed its surveillance. Traditional TB drug susceptibility testing takes weeks to months; WGS can predict resistance in days by identifying known genetic mutations:

- An outbreak of XDR-TB in South Africa was sequenced, revealing that the strains shared common mutations (in genes katG for isoniazid resistance, rpoB for rifampicin, gyrA for fluoroquinolones, etc.), confirming them as XDR. Genomics also showed that these cases were all closely related (few SNPs apart), indicating transmission of XDR strains (as opposed to separate acquisitions of resistance).
- The insight allowed public health officials to focus on containment (isolation of patients, tracing contacts) to stop the spread of that particular clone, and also to ensure patients get individualized therapy (sometimes new drugs like bedaquiline if traditional ones won't work due to resistance).
- On a routine basis, many countries now sequence TB isolates at diagnosis. One benefit: sometimes *identifying resistance genes before lab tests catch up*, enabling early regimen adjustment. Conversely, if WGS shows no resistance mutations, a patient can be kept on first-line therapy if lab capacity is limited.

Case B – Carbapenem-resistant *Klebsiella pneumoniae* in a Hospital:
Carbapenems are last-resort antibiotics. A hospital ICU sees a cluster of *K. pneumoniae* infections with high mortality. WGS of these isolates finds they all carry the gene *bla_KPC-2*, a carbapenemase, on a plasmid. The sequences are nearly identical (suggesting a single strain spreading in the ICU), and plasmid analysis shows the plasmid is also nearly identical in each.

- Epidemiologists use this info to implement stringent infection control: cohorting patients, deep cleaning, possibly closing the ICU to new admissions briefly.

- They also check past isolates (retrospective sequencing) and find that a few months earlier, a patient had a similar strain, indicating that might have been the index case, and the strain persisted in the ICU environment.
- The genomic data also triggers a broader alert: *bla_KPC* is a serious gene, often transferable. The hospital notifies public health authorities, and nearby hospitals start screening their *Klebsiella* isolates for that sequence type or plasmid. This preemptive search, thanks to genomics, can catch early spread to other facilities.

Use of ResFinder and Similar Tools:
In labs, tools like ResFinder are employed on each bacterial genome. For instance:

- A *Salmonella* from an outbreak is sequenced; ResFinder reports genes *aac(3)-IV, bla_TEM-1B, sul1, tetA*, which corresponds to resistance to aminoglycosides, penicillins, sulfonamides, tetracycline. The outbreak strain is thus multi-drug resistant. This may elevate concern and lead to advice on antibiotic choices for treating cases (perhaps avoid those drug classes).
- Another scenario: routine surveillance of commensals on farms finds *mcr-1* gene (colistin resistance) in *E. coli* from a pig farm, discovered through genomic sequencing. Colistin is last-resort in humans for some infections; finding *mcr-1* prompts a One Health response: perhaps restrict colistin use in veterinary practice in that region, and surveillance to see if *mcr-1* carrying bacteria have spread to humans locally (like through food chain).

Whole-Genome Sequencing replacing Phenotypic Tests:
It's noteworthy that regulatory frameworks are adjusting. The EU's EFSA (food safety authority) considered allowing WGS + ResFinder to replace routine phenotypic antibiotic susceptibility testing for surveillance by 2021 and making it mandatory by 2025 for certain bacteria. This is a significant policy shift acknowledging that genomic predictions are robust enough for surveillance purposes:

- *Accuracy:* Studies have shown >95% concordance between genotype and phenotype for many drug-bacteria combinations.
- *Comprehensiveness:* WGS finds *all* known resistance genes at once, and even shows novel mutations or genes. Traditional tests only check a set of antibiotics; WGS might reveal an unnoticed resistance (e.g., a disinfectant resistance gene or heavy metal resistance gene co-carried, hinting at selective pressures).
- *Outbreak clusters with mixed phenotypes:* WGS has solved puzzles where phenotypic tests confused things. For example, in a *Salmonella Heidelberg* outbreak in 2011, some isolates were pan-susceptible and others resistant, so they seemed unrelated by phenotype. WGS showed they were highly related genetically – it was one outbreak with varying plasmids in different cases. It turned out to be a contamination event involving multiple strains and plasmids, which WGS clarified.

Data Sharing in AMR Genomics:
Platforms like NCBI's Pathogen Detection and FDA's GenomeTrakr, while initially focused on foodborne pathogens, also gather data on resistance elements. Pathogen Detection's AMRFinder screens every genome for AMR genes and includes that in its database. This enables cross-analysis:

- If a certain resistance gene appears in multiple species from the same region, it could indicate horizontal gene transfer hot-spot (maybe a plasmid spreading in a hospital).
- If similar resistance plasmids are found in a human isolate and a farm isolate, that suggests a link (direct or via environment) – prompting perhaps investigations into farm hygiene or waste management.

Public Health Decision-Making:

- Genomic data on AMR informs treatment guidelines. E.g., if an outbreak strain of *Shigella* is sequencing and found to carry *macrolide resistance* genes, health departments may advise against using azithromycin even before full lab AST is back.
- Policy changes: The detection of *mcr-1* globally by genomics led WHO and CDC to push for monitoring colistin use and ban on certain animal uses. Similarly, discovery of NDM-1 (another plasmidic gene) in environmental samples in New Delhi was heavily facilitated by sequencing and spurred efforts to track AMR in the environment.
- AMR genomic surveillance also ties into predicting *trends*: e.g., by sequencing *Salmonella* from many sources each year, one might see resistance gene trends (a rise in ESBL genes year over year). If a certain class is rising, maybe that antibiotic should be used more cautiously or alternatives sought.

Conclusion of AMR Case: WGS in AMR surveillance is both a detective and a sentinel. It can retrospectively solve why an outbreak didn't respond to a certain antibiotic (detecting a gene after the fact), and prospectively alert to emerging resistance (e.g., a new variant of a resistance gene or a new combination in a pathogen). Public health benefits by anticipating shifts in resistance patterns and tailoring both clinical guidelines and preventive measures (like antimicrobial stewardship programs, infection control, and awareness campaigns). Genomic epidemiology thus adds a proactive layer to the fight against drug-resistant infections, which is crucial as we strive to preserve the effectiveness of existing drugs.

3.3 Zoonotic Disease Surveillance: Genomic Tracking of Ebola and Avian Influenza

Context: Zoonotic diseases are those that can jump from animals to humans. Genomic epidemiology is crucial in these contexts to differentiate between multiple spillover events versus sustained human-to-human transmission, and to monitor changes that might increase pandemic risk. We examine Ebola (a high-fatality virus with sporadic outbreaks) in Africa.

Case Study – Ebola Virus Outbreak in West Africa (2014-2016):
This was the first Ebola outbreak to be tackled with large-scale genomics.

- Early in the outbreak (mid-2014), Gire et al. sequenced 99 Ebola genomes from patients in Sierra Leone. The data revealed:
    - The outbreak was caused by a single spillover introduction (likely from a bat to a human) and then human-to-human spread, as evidenced by the genetic similarity of the viruses. Previous outbreaks in other countries were from separate introductions.

    - Two genetically distinct viral lineages in Sierra Leone, both tracing back to a funeral where infected individuals from Guinea spread it to attendees, seeding two chains in SL.
    - A high mutation rate: the virus was accumulating mutations as it spread. Several were non-synonymous (amino acid changing), raising concerns about potential changes in virus behavior (though fortunately none fundamentally changed transmission or virulence in that outbreak).
- The genomic data helped confirm that containing human-to-human transmission was key (as opposed to worrying about repeated new jumps from animals during that outbreak). It also guided vaccine and therapy developers by providing many viral genome sequences to target.
- Later in the outbreak, genomics helped trace how the virus spread between countries. For example, a case in Mali was quickly sequenced and matched to the Guinea strain, confirming its origin. Similarly, when cases flared up in Nigeria, sequencing confirmed they stemmed from a traveler from Liberia.
- Near the end of the outbreak, a surprising use of genomics: a case re-emerged after Ebola was thought gone in Liberia. Sequencing showed the virus from that case closely matched earlier Liberian strains, suggesting it wasn't a new introduction, but rather persistence (likely in a survivor who relapsed or sexually transmitted it). This highlighted how Ebola can hide in immune-privileged sites (like eyes, testes) and led to updated policies on survivor monitoring and safe practices (like advising condom use for a time even after recovery).
- Genomic surveillance was also used in *ring vaccination* trials. By sequencing viruses from vaccinated vs. unvaccinated clusters, scientists could ensure that vaccine failures (cases in vaccinated individuals) were not due to a new divergent strain escaping vaccine protection (none were, vaccine worked well).

Conclusion on Zoonotic Diseases Surveillance: The zoonotic case studies underscore that genomic epidemiology is a critical tool not just after spillover into humans, but in the animal realm to **pre-empt** or better understand these jumps. Surveillance at the human-animal interface, and rapid sequencing when an outbreak happens, can literally be the difference between a contained event or a global crisis. For Ebola, that meant stopping the West African outbreak and quickly controlling later ones in DRC with real-time sequencing in the field.