Essential Elements of Genomic Epidemiology

Prepared by the Landscape Analysis Working Group* of the Pathogen Genomics Centers of Excellence Network

Preamble: Pathogen genomic sequencing is a powerful tool that bridges clinical care and public health by providing insights into pathogen transmission, identification, subtyping, resistance mechanisms, and outbreak management. This document represents a collation of concepts and capabilities that we, as the Pathogen Genomics Centers of Excellence (PGCoEs), believe are essential for effective genomic surveillance and actionable genomic epidemiology. Developing this document is meant to establish a baseline perspective so that gaps, needs, and opportunities can be meaningfully identified. Through different transformations, information here may contribute to a roadmap for pathogen genomics and genomic epidemiology.

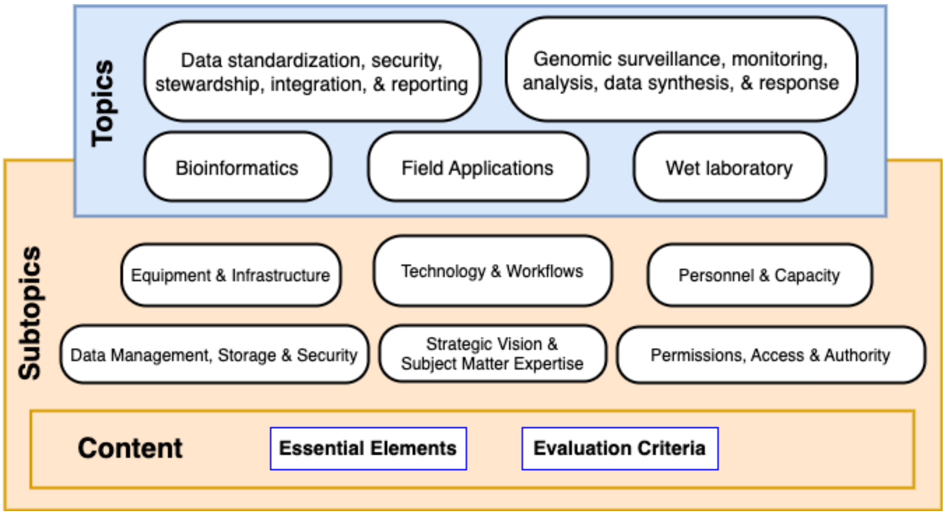
Intended audience: This document outlines suggested elements for integrating pathogen genomics into public health practice. It is intended to serve as a useful framework for laboratories and public health entities seeking to establish or enhance their genomic epidemiology capabilities, not as an exhaustive guide on the implementation of those elements. While primarily intended for laboratories and public health entities, this framework may also be valuable for a broader range of professionals in clinical and healthcare settings seeking to understand the role of pathogen genomics in public health and clinical investigations, and to explore avenues for utilization of genomic data.

Why Consult This Document?
This framework provides information for evaluating and building genomic epidemiology programs.

Laboratories and public health entities can use this information to:
Assess Current Capabilities: Identify strengths and weaknesses in existing infrastructure, technology, data management, personnel, and strategic vision.

Identify Gaps and Needs: Pinpoint areas requiring investment or improvement to achieve effective genomic surveillance and response.

Develop a Strategic Plan: Formulate a roadmap for developing or enhancing pathogen genomics programs, including resource allocation, training needs, and collaborative partnerships.


Topics
Data standardization, security, stewardship, integration, & reporting
Genomic surveillance, monitoring, analysis, data synthesis, & response
Bioinformatics
Field Applications
Wet laboratory
Subtopics
Equipment & Infrastructure
Technology & Workflows
Personnel & Capacity
Data Management, Storage & Security
Strategic Vision & Subject Matter Expertise
Permissions, Access & Authority
Content
Essential Elements
Evaluation Criteria

Framework composition

Essential Elements: Description of one aspect of what success looks like under each subtopic, e.g., "Equipment and Infrastructure." This description should be generalizable rather than specific to a particular manufacturer of tools or equipment, e.g., "sequencing platforms available for low, medium, and high throughput projects," *not* "3 Illumina MiSeq systems available."

Evaluation Criteria: Examples of evaluation questions that can be used to determine if institutions or agencies have the essential elements described. These criteria are not comprehensive (i.e., do not capture all of the essential elements identified in each category) but are meant to provide suggestions for how to evaluate or think about the essential elements above.

Category Definitions:

- **Equipment and Infrastructure:** *A key element of success in genomics (whether it be in the laboratory, on the computer, or with some other aspect of managing data or conducting analyses) is ensuring agencies have the instruments and space required to conduct the tasks at hand. This can include laboratory equipment, computer hardware, or other physical infrastructure such as dedicated laboratory space.*

- **Technology and Workflows:** *Some topic areas also require software and workflow infrastructure in addition to the hardware requirements listed above. Here, we aim to understand the pipelines, workflows, and technology needed to achieve success in different aspects of genomic epidemiology, whether that's the existence of specific software packages or capabilities, e.g. visualization and data anlaysis, or whether that refers to the ability to run complex workflows in the laboratory.,*

- **Data Management, Storage, and Security:** *All aspects of genomic epidemiology involve the generation, storage, and/or usage of specific types of data. In this section, we identify the systems important for facilitating the activities under each topic area. This can range from* *laboratory information management systems, to computational space and security concerns, to epidemiological data sharing mechanisms.*

- **Personnel:** *Integrating pathogen genomic analysis into public health practice requires many different types of personnel (e.g., microbiologists, bioinformaticians, epidemiologists, data scientists, and data engineers) and sufficient capacity across these roles to handle both routine and elevated data collection and sample processing (e.g., during pathogen emergencies). Here, we will try to get a sense of the different types of personnel relevant to each topic and the number of personnel that might be required.*

- **Strategic Visioning and Subject Matter Expertise:** *Effectively incorporating pathogen genomic analysis into public health practice and sustaining that capacity over long periods of time requires the ability to change and adapt, bringing in new protocols, data tools, and methods of analysis. It can be easier to establish and foster the growth of genomic surveillance when agencies have access to subject matter expertise in pathogen genomics, either internally or through close external collaboration. Here, we enumerate the required subject matter expertise for success in each area, regardless of whether that expertise is available internally or through collaborations.*

- **Data and Workflow Permissions, Access, and Authority:** *Responding to pathogen threats requires some degree of authority to adapt to new situations and access sensitive data. Here, we explore the various regulatory and permissions aspects of genomic epidemiology. This can include regular or as-needed access to certain types of data (beyond just the ability to store it securely, as covered above), authority to implement and validate new protocols/workflows, etc.*

Topic Area: Wet Laboratory

Overall Vision: A successful and robust wet laboratory system for genomic epidemiology includes updated and well-maintained laboratory equipment, such as sequencers, operated by competent laboratorians. Protocols for equipment usage and testing procedures are governed by best practices and well-established quality standards, including seamless flow of sequence data among stakeholders from the point of origin. Laboratories may also benefit from a dedicated Research & Development team that performs or develops high quality sequencing assays and efficiently scale production around changes in volume, especially increases, in actionable turn-around-times.

I. Equipment and Infrastructure

Essential elements:

1. Equipment for extraction, amplification, quality assessment, and sequencing are available for low, medium, and high throughput projects.
2. Automation and other ancillary equipment that boost sequencing quality and throughput.
3. Sufficient laboratory space to accommodate routine sequencing as well as potential surge sequencing needs.

Evaluation criteria:

- What sequencing platforms are available and in use?
- What is the number of sequencers with each different sequencing platform?
- Does the laboratory have dedicated sequencing laboratory space?
- What protocols are in place to reduce contamination during the sequencing process?
- Are there maintenance service contracts in place for sequencers and other relevant laboratory equipment?
- What measures are in place to assess the compatibility of new equipment with laboratory computational infrastructure, budget, and chemistry?
- What equipment is available for nucleic acid extraction from a wide variety of samples?
- What automation systems are available for sequencing library preparation?
- Does the laboratory have needed intermediate equipment? (i.e., PCR, plate reader, automated electrophoresis system station, fridge/freezer, etc.)

II. Technology and Workflows

Essential elements:

1. Diversity of library preparation assays and sequencing approaches to meet different sequencing requirements (e.g., shotgun sequencing, amplicon-based sequencing, etc.).
2. Capacity to add new sequencing platforms, adapt associated assays, and/or develop new assays.
3. Systems for reducing cross-contamination of samples and effective experimental controls.

Evaluation criteria:

- What sequencing approaches and protocols are currently in use/validated for use?
- What are the infectious pathogens for which whole genome sequencing assays were previously performed and/or are currently being performed?
- What are the maximum and average monthly volumes for existing assays?

- What is the standing Quality Assurance / Quality Control system that ensures sequencing quality?
- What is the description of potential sample sources and types for each assay?
- What is the laboratory allocation plan to avoid contamination?
- What is the Standard Operating Procedure (SOP) for re-runs on failed samples?

III. Data Management, Storage, and Security

Essential elements:

1. Robust and regulation-compliant onsite and/or cloud-based genomic sequence storage for routine operational needs as well as surge needs (e.g., internal network servers, cloud storage, etc.).
2. Data sharing mechanisms and agreements that ensure smooth flow and exchange of raw sequence data among stakeholders and partners (e.g., secure file transfer systems, connected cloud storage, etc.).

Evaluation criteria:

- What is the number of sequencers online?
- What are the specifications of dedicated servers for raw sequence data storage?
- Is the data storage system HIPAA compliant?
- Is there a secured access system for data storage?
- What data sharing mechanisms exist, and what is the associated data transfer speed?
- What challenges does the lab face connecting databases containing matched sequence records and specimen records?
- What backup plans or redundancies are in place in the event of lost/interrupted sequence runs or specimen contamination?

IV. Personnel

Essential elements:

1. Adequate and competent genomic sequencing workforce dedicated to pathogen sequencing tasks.
2. Good staff retention rate to avoid operational gaps.
3. Ability to adjust existing staff work shifts and/or recruit additional staff to accommodate unexpected high-volume surges.

Evaluation criteria:

- What is the number of laboratory staff scientists, laboratory technicians, and research staff (including short-term contractors, interns, and fellows)?
- What are the associated staff education levels, years of experience, and expertise?
- What staff training system exists?
- What is the availability of research and assay development team?
- What recruiting/adding mechanisms exist if more staff are needed? What is its flexibility and timeliness?
- What is the staff retention strategy?
- What methods or protocols do you use to transfer knowledge for continuity of operations?

V. Strategic Visioning and Subject Matter Expertise

Essential elements:

1. Ability to rapidly develop and optimize new high quality sequencing assays that take advantage of new approaches and technologies to respond to new situations or emerging threats.
2. Internal and external stakeholders with expertise in microbial genomic sequencing that can guide the generation and interpretation of sequencing data as it pertains to public health.
3. Continuity of sequencing laboratory operations in the event of major disaster and other unforeseen emergencies (e.g., systems in place if bomb threat, laboratory pipe abruption, electrical outage, natural disasters, staff out due to sickness, etc.).
4. Consideration of testing modalities and methods that are CLIA approved where clinically relevant, ensuring results can be appropriately used in clinical settings when needed.

Evaluation criteria:

- What is the number of internal personnel who have job duties dedicated to research and development?
- What is the number of internal personnel with comprehensive genomic epidemiology expertise?
- What rank of skill level of genomic epidemiology for each internal personnel involved in relevant activities?
- What sequencing-based collaborations exist with academic institutions and clinical laboratories?
- Is a genomic sequencing-related COOP plan in place?
- Are CLIA-approved testing modalities considered or utilized when clinical applications are anticipated for genomic data?

VI. Data and Workflow Permissions, Access, and Authority

Essential elements:

1. Existing regulations and internal institutional policies are in place to facilitate genomics work.
2. Data usage agreements among stakeholders are in place, ensuring that contextual data (e.g., epidemiolocal, clinical, and environmental) can be shared with raw sequencing data for maximum impact.

Evaluation criteria:

- What are the institutional policies and procedures for access, exchange, and retention of sequence data (existence of a Data Management Plan, DMP)?
- What are the existing data usage agreement templates and approval processes for raw sequencing data?
- What are the existing data usage agreement template and approval process for Public Health Information/Personal Identifiable Information which accompany the raw sequencing data?
- What are the standing contracts or MOUs in place with partner sequencing laboratories or third-party industry vendors for sequencing outsourcing?

Topic Area: Bioinformatics

Overall Vision: Success in bioinformatics means the ability to run containerized workflows from the broader bioinformatic pipeline ecosystem, including the ability to find or develop new workflows and apply them to local public health needs and pathogens of interest.

I. Equipment and Infrastructure

Essential elements:

1. Availability of computer hardware in service of the goal of running containerized workflows, which can include hardware to connect to cloud-based infrastructure or the ability to run processes locally.
2. Availability of hardware that can store genomic data (including bioinformatic intermediates) to meet the demands—including possible surge demands—of the workflow systems in place (e.g., more local storage may be needed if analyses are performed locally instead of in the cloud).
3. Server and network hardware that allows for efficient handling and transfer of genomic data (including bioinformatic intermediates) between local systems and cloud systems.

Evaluation criteria:

- Is the current computing infrastructure for processing, storing, and analyzing genomic data sufficient for both current needs and potential surge capacity?
- What containerized workflow systems are currently in use? What workflow management systems are currently in use?
- Is the amount of on-site and/or cloud storage used for genomic data sufficient to supplement clinical and epidemiological work?
- Does network connectivity ever inhibit or present a bottleneck to uploading and/or downloading necessary bioinformatic data (e.g., sequences, cluster analysis reports)?
- Where are bioinformatic and analytical steps performed (e.g., cloud versus local)? Are there seamless pathways for connecting outputs across steps?

II. Technology and Workflows

Essential elements:

1. Availability of an adaptable bioinformatics ecosystem for genomics workflows (ideally open-source, extensible and well supported).
2. Bioinformatics workflows are optimized for local pathogens and local throughput.
3. Bioinformatics workflows that are portable and consistent (e.g., open source, containerized, etc.) and take advantage of existing genomic databases where relevant.
4. Capacity to upgrade software and pipelines.
5. Robust workflows for transferring raw data from sequencing machines to servers/cloud for internal use as well as for uploading raw data and assemblies to public repositories.
6. Interoperable data exchange standards and formats.

Evaluation criteria:

- How are raw genomic data transferred from sequencers to your bioinformatics processing engine?

- How are data uploaded to public repositories?
- What public repositories are currently used for genomic uploads?
- What genomic workflows are currently in use?
- How many samples can be processed through a chosen bioinformatics pipeline per day, and what is the current number of samples being processed per day?
- What compute environments are supported by your bioinformatics team/pipelines?

III. Data Management, Storage, and Security

Essential elements:

1. Availability of hardware or cloud infrastructure that can store genomic data (including bioinformatic intermediates) to meet the demands of the workflow systems in place (e.g., more local storage may be needed if analyses are performed locally instead of in the cloud).
2. Safeguards in place that control access to stored and in-process genomic data.

Evaluation criteria:

- Who has access to raw sequencing data, and what access controls are in place, if any?
- Who has access to bioinformatics workflows and output intermediates, and are there clear protocols in place for making genomic data accessible to outside partners and/or public repositories?
- What version control mechanisms are in use for bioinformatics pipelines?

IV. Personnel

Essential elements:

1. Dedicated personnel for managing and organizing genomic data.
2. Personnel with expertise to run bioinformatics workflows (laboratorians, bioinformaticians, or analysts).
3. Personnel with the ability to troubleshoot errors or issues with workflows in use.
4. Personnel dedicated to performing routine interpretation of data and report generation (including integration of bioinformatics analyses with sensitive metadata).
5. Engagement of IT Personnel to help troubleshoot hardware and network issues, including local, remote, or cloud based experts.

Evaluation criteria:

- Do you have sufficient data management personnel for managing and organizing genomic data in the context of epidemiological and laboratory data?
- Do you have enough bioinformaticians/analysts/microbiologists who can run bioinformatics workflows?
- What is the background of people who run bioinformatics workflows?
- Do you have enough personnel and redundancy to perform basic genomic analysis and genomic surveillance report generation?
- Do you have sufficient access to IT personnel to support the genomic surveillance (system administrators, database managers, etc.)?

V. Strategic Visioning and Subject Matter Expertise

Essential elements:

1. Ability to bring in and validate bioinformatics workflows for new pathogens or scenarios.
2. Ability to customize bioinformatics workflows to meet local needs or limitations.
3. Ability to interpret genomic data, including integrating multiple data types and bringing in or performing new analyses and methodologies.

Evaluation criteria:

- How many people can identify new workflows, bring them in, and troubleshoot as necessary?
- What programming languages are your bioinformatic staff familiar and comfortable with?
- What background and training do bioinformatics staff generally have (e.g., formal education, practical background via wet laboratory work)?

VI. Data and Workflow Permissions, Access, and Authority

Essential elements:

1. Clear regulatory processes in place for adapting and integrating new bioinformatics workflows.
2. Ability to publicly share genomic data after generation (including human read removal).
3. Clear processes in place for linking and de-linking sequencing data from sensitive epidemiological information and PII.

Evaluation criteria:

- What is your process for bioinformatic workflow validation?
- What is the approval process for submitting genomic data to public repositories?
- Are there clear pathogen-specific processes for linking sequencing and laboratory data?

Topic Area: Data standardization, security, stewardship, integration, and reporting

Overall Vision: Relevant data in genomic epidemiology includes sequencing, epidemiological, and laboratory information, each with different inherent levels of sensitivity depending on context, collection, aggregation, and communication. Success in this area involves the following components: Data Standardization, which enables data to move from one platform to the other to allow sharing and analysis; Data Security, which ensures data integrity, confidentiality, and availability of data and services by providing appropriate access and protection mechanisms; Data Stewardship, which involves active management and the use of mechanisms for acquiring, storing, safeguarding, and using data following practical rules that reflect legal requirements and assigning personnel to ensure compliance; Data Integration, which enhances the real-time picture of an outbreak by combining different types of data together, using tools that enable combined views for understanding the situation; Data Linkage, which improves data discoverability by making data available in a way that the existence of relevant parts can be identified, such as through project pages for investigations.

I. Equipment and Infrastructure

Essential elements:

1. Equipment and infrastructure that can support common formats, protocols, and vocabularies for genomic data exchange, such as cloud-based platforms, web services, application programming interfaces (APIs), or Linux environments.
2. Infrastructure that can implement encryption, authentication, authorization, backup, and recovery mechanisms for genomic data and associated epidemiological and laboratory metadata, such as firewalls, antivirus software, secure sockets layer (SSL) certificates, and two-factor authentication.
3. Equipment and infrastructure that enables data discovery and access across different platforms and networks, such as data catalogs, registries, indexes, and portals.

Evaluation criteria:

- Do your current epidemiological, laboratory, and bioinformatic infrastructure allow for the efficient linkage, transmission, and review of data, analyses, and reports between these groups?
- What encryption methods are used for sequencing, laboratory, and epidemiological data?
- What are the maintenance requirements of your physical storage (e.g., dedicated personnel, periodic upgrades, etc.) for sequencing, laboratory, and epidemiological data?

II. Technology and Workflows

Essential elements:

1. Availability of updateable operating systems with latest bug fixes in systems dedicated to genomic, laboratory, and epidemiological data.
2. Software that can encrypt data and provide access control mechanisms for the viewing and sharing of sequencing, laboratory, and epidemiological data across groups and divisions.
3. Availability of software that can combine different types of data together and provide a unified view of the data across epidemiological, sequencing, and other laboratory sources.
4. Availability of functioning software, systems, or programs for data transfer (e.g., multisite data managers, integration engines, etc.).

5. Software that can manage report workflows, both internally and externally (e.g., submitting institutions).
6. Standardized protocols for metadata collection alongside each assay, capturing information such as sample source, collection date, clinical or epidemiological context, and relevant laboratory processing details.

Evaluation criteria:

- What operating systems and versions are being used to store and process sequencing, laboratory, and epidemiological data?
- What tools are used to enable visualization of multiple data types for the purpose of genomic surveillance?
- What encryption software is used for sequencing, laboratory, and epidemiological data?
- What system, software, and/or protocols are in place for genomic surveillance reporting workflows (e.g., sequencing to epidemiologists, epidemiologists to leadership)?
- Are reporting workflows configured to provide timely notification and situational awareness to public health leadership?
- Are reporting workflows configured to provide appropriately aggregated information to the medical or health community?
- What pathogens are data joins and queries between epidemiological, sequencing, and laboratory data possible with?

III. Data Management, Storage, and Security

Essential elements:

1. In place access controls for sequencing, laboratory, and epidemiological data with ability to grant / change access, e.g. edit permissions at the end of emergency use authorization
2. SOP for agreements that need to be in place for internal and external access to sequencing, laboratory, and epidemiological data.
3. SOP for transfer mechanisms of laboratory, sequencing, and epidemiological data within and between participating teams.
4. Regularly scheduled data element and data systems backups.
5. Ability to restore data in the event of a disaster or other infrastructure failure.
6. Ability to transform, aggregate, and visualize across different sources (epidemiological, laboratory, and sequencing) and types of data, such as data warehouses, data lakes, extract-transform-load (ETL) tools, and dashboards (e.g., access to an expandable internal data storage resource).
7. Metadata associated with raw sequence data is stored in standardized formats (e.g., using JSON or XML schema) and adhere to established vocabularies or ontologies (internal or external) to ensure interoperability and facilitate data sharing and analysis across different platforms.

Evaluation criteria:

- For laboratory data, what Laboratory Information Management System (LIMS), type(s) and version(s), and/or other storage systems are used?
- What data storage system is used for epidemiological data?

- What file formats are being used for storage of sequencing data, and how are they stored (e.g., physical, cloud)?
- Is there an established data access control layer infrastructure that allows users from epidemiology, laboratory, and sequencing groups to access the data required for genomic surveillance?
- What is your storage access plan, and how is data access layered? Who controls access, and what is the process for granting data access?
- Do you have the ability to monitor data access of sensitive information, including ingress and egress of data?
- What agreements are required to share data with external partners (DUA, etc.)?
- What data management practices are being followed (FISMA, FedRAMP, internal) for epidemiological, sequencing, and laboratory data?
- Are sequencing, laboratory, and epidemiological data systems linked automatically, or is communication between sequencing and laboratory data sources done ad hoc?
- What is the backup schedule of systems used for data storage (epidemiological, sequencing, and laboratory)?
- How difficult is it to restore previous versions of data (epidemiological, sequencing, and laboratory)?
- What types of data are stored long-term (inputs, intermediate analysis files, outputs, etc.) for epidemiological, sequencing, and laboratory data?
- How long do different data types get stored?
- What SOP exists for promoting standardized and consistent metadata, allowing for the unambiguous identification and interpretation of data elements across different systems and organizations?

IV. Personnel

Essential elements:

1. There is at least one data engineer or data scientist assigned to work on sequencing data-related queries or analysis, which may involve integration of multiple data types.
2. Availability of database administrators or managers assigned to evaluate, maintain, and access sequencing, laboratory, and/or epidemiological data.
3. Availability of IT personnel familiar with genomic epidemiological workflows, including on-prem network managers, permissions administrators, and a security head.
4. There is at least one person with expertise working in the cloud.
5. Staff members have access and are aware of appropriate continued education/training in the realm of genomic epidemiology and/or associated data types.
6. Availability of personnel, either internal or external, that can assume the "data steward" functional role who will be responsible for ensuring data collection, processing, and sharing complies with the law.

Evaluation criteria:

- How many personnel are employed with the following titles/responsibilities:
    - Data scientist
    - Database administrator/manager
    - Systems engineer

    - Bioinformatician
- Are there dedicated IT personnel to support data linkage queries across sequencing, laboratory, and epidemiological data systems?
- Are there personnel with cloud expertise?
- What trainings and/or certifications in data security and management are provided by the home institution?
- What trainings and/or certifications in data security and management are provided outside of your institution?

V. Strategic Visioning and Subject Matter Expertise

Essential elements:

1. Expertise on privacy concerns and combinations of metadata/sequence for the purpose of collaboration with external partners.
2. Expertise on data joining and derived views that enable situational awareness.
3. Plan for infrastructure deployment, management, upgrades, and adaption/surge.

Evaluation criteria:

- Is there expertise in privacy and data sharing regulations that allow for ease of collaboration, both currently and for future projects?
- Is there expertise that can be relied on for data joining and reporting for situational awareness as the pathogen landscape changes and evolves?
- Are SOPs available for sharing deidentified data from laboratory, sequencing, and epidemiological sources with internal and external collaborators?
- Are there standardized pathogen-specific reports generated that can be used as a foundation to build actional public health interventions?

VI. Data and Workflow Permissions, Access, and Authority

Essential elements:

1. Training compliance with standards (e.g., IRB, NIH InfoSec).
2. Agreements in place that facilitate routine access and use of data.
3. Ability to develop and efficiently create agreements (e.g., MOAs, NDAs, DURSA, DSA, MOU) as needed for new pathogens or unexpected situations.
4. Access and proficiency to use software systems for data analysis (i.e., dataframe centric and bioinformatics centric analysis environments).
5. Responsible authority identified for data governance, i.e., what organization is responsible for the data to comply with the law.

Evaluation criteria:

- Is there awareness and management of data security and stewardship compliance trainings?
- How many personnel are trained in data compliance standards?
- Do you have protocols in place describing how personnel can gain access to the necessary systems?
- Is there proficiency in the creation of research agreements (e.g., MOAs, NDAs, DURSA, DSA, MOU)?

- Are workflows hindered due to access bottlenecks to the different data management systems (e.g., LIMS)?
- Who are the people responsible for the laboratory, sequencing, and epidemiological data?

Topic Area: Genomic surveillance, monitoring, analysis, data synthesis, and response

Overall Vision: Successful genomic surveillance synthesizes the standardized data from laboratory and bioinformatic workflows in order to produce robust monitoring of pathogens of concern. Genomic surveillance encompasses the monitoring, analysis, and data synthesis of genomic information to advise in responses and inform public health practice. Effective genomic surveillance systems generate consistent and reproducible outputs, such as reports or dashboards, at intervals that can be used for field responses, and which are flexible enough to capture relevant information as response needs change. Systematic collection, curation, and evaluation of component relationships are vital for well-functioning surveillance. Well-functioning genomic surveillance systems for genomic data may be able to detect when it is time to scale to the needed response. They may have a combination of pathogen agnostic systems and more tailored systems that are organism specific.

I. Equipment and Infrastructure

Essential elements:

1. Technology and workspace space allow members of different parts of the system (e.g., laboratories and epidemiologists) to communicate clearly and securely in real-time.
2. Computers with internet connection and an effective amount of RAM readily available to any laboratory and epidemiological staff involved in the system.

Evaluation criteria:

- What is the typical total file size of data collected in a 90-day period?
- How many samples, by pathogen, expected to be sequenced in a 90-day period?
- What kind of conferencing technology is available to surveillance teams?
- How far away are laboratory, bioinformatic, and epidemiology members from one another (e.g., sharing a common workspace, all in the same building but spread across the floors, in completely different buildings)?

II. Technology and Workflows

Essential elements:

1. When a subset of a live dataset is used for analysis, it is versioned or dated and shared at intervals that can make investigations incrementally more effective.
2. Availability of computational pipelines or programs that can produce phylogenetic or comparative genomic analyses for inferring epidemiological linkage and transmission dynamics.
3. The system has flexible reporting media (e.g., dashboards or markdown documents) that can be shared between personnel.

Evaluation criteria:

- How often are sequences and their associated metadata added to surveillance databases?
- What software is primarily used to generate phylogenetic trees, distance matrices, or heatmaps? Are multiple software used for the same purposes within a single public health agency?

- What software is used to generate reports? Are reports automated, or do they need to be generated manually?
- What kinds of procedures are in place to validate that genomic data are complete, valid, and have been correctly associated with epidemiological or environmental metadata?

III. Data Management, Storage, and Security

Essential elements:

1. LIMS workflows that promote linking of laboratory outputs to epidemiological case databases.
2. Clinical microbiologists or field microbiologists are able to securely share relevant genomic information with public health personnel.
3. Normalized record schema that avoids duplication.

Evaluation criteria:

- What challenges do laboratorians in your department face when interacting with the LIMS system? What challenges do data scientists/epidemiologists face when interacting with the LIMS system?
- What kinds of unique identifiers are used to link epidemiological "cases" to genomic sequences and laboratory specimen metadata?
- Are there limits on which personnel have access to test results and patient data? How are these limits set?
- How are relevant sequence, clinical, and epidemiological data joined together?

IV. Personnel

Essential elements:

1. Bioinformaticians are available for managing genomic tools and computational pipelines.
2. Epidemiologists are available for the interpretation and synthesis of case data with genomic data.
3. Supportive information technologists are available for permissions management, software installation, and for troubleshooting complications with digital communications.

Evaluation criteria:

- How many bioinformaticians that spend at least 50% of their time working on genomic surveillance of pathogens?
- How epidemiologists that spend at least 50% of their time working on genomic surveillance of pathogens?
- What is the availability of IT/Informatics support and willingness to build or incorporate new genomics-related software infrastructure into existing surveillance systems? As new surveillance systems?

V. Strategic Visioning and Subject Matter Expertise

Essential elements:

1. Epidemiologists can make informed decisions about genomic cut-offs for cluster analyses
2. Personnel can perform systematic evaluations of the systems for stability, usefulness, and scalability.

3. Bioinformaticians can assess genomic data quality to ensure its effective use for downstream analysis.
4. Personnel know how to accurately interpret specific genomic analyses, including understanding sampling limitations and evaluating the strength of evidence from phylogenetic analyses.
5. Laboratorians and analysts have open communication regarding analysis goals and concerns.

Evaluation criteria:

- For team members dedicated to data analysis, what are their levels of understanding of statistics and study design?
- When and how do data analysts consult laboratorians involved in sequencing about results?
- What kinds of sampling frames (e.g., time intervals and sample size) are used for different pathogens that are sequenced?
- Do you have at least one team member dedicated to surveillance system evaluation activities?
- Do you have at least one team member trained in bioinformatics or computational biology?
- How much training have bioinformaticians received to understand the needs and goals of surveillance epidemiologists?
- How confident are the surveillance epidemiologists regarding how pathogen genomic data can be used to support their public health investigations?

VI. Data and Workflow Permissions, Access, and Authority

Essential elements:

1. Ways to bring secure, identifiable epidemiological data together with less secure pathogen genomic sequence data to enable joint visualization and analysis while respecting security needs and user permissions levels.

Evaluation criteria:

- How are reporting outputs stored for each type of surveillance activity?
- How are reporting outputs shared for each type of surveillance activity?
- What software is available to teams to create visualizations of the surveillance reports for all genomically surveyed pathogens? Which ones are used the most?
- Are protocols in place to change or update a surveillance team member’s access to certain databases along the surveillance system?
- Are there separations in database access? What governs the separation? (e.g., type of pathogen, type of team member, affiliation to the public health department)?

Topic Area: Field Applications

Overall Vision: We define field application of genomic epidemiology to public health investigations as the ability of frontline public health agencies (e.g., state and local health departments) to take effective actions during outbreak responses, surveillance, and other public health investigations based on conclusions derived from genomic data and surveillance reports. Successful response efforts rely on effective information sharing among jurisdictional public health entities, clinical partners, community partners, and the public.

I. Equipment and Infrastructure

Essential elements:

1. Frontline public health agencies have clear and documented protocols for requesting or conducting sequencing.
2. Appropriate sample collection materials (e.g., sterile tubes, swabs, etc) and transport to the testing laboratory are available to field-based practitioners.
3. Frontline public health agencies have ready access to sequencing technologies through in-house procedures or strongly established partnerships.
4. Field partners have digital equipment that allows them to easily interact with sequence results or receive sequence data.

Evaluation criteria:

- What kinds of services or partnerships exist with the public health agency that allow for the transport of specimens to a sequencing lab?
- Are collection materials and associated packaging (e.g., swabs, tubes, specimen bags) in sufficient quantity for typical public health responses? Are there differences in availability based on pathogen or specimen type?
- How long does it take between initiating a request for sequencing and specimen collection?
- What is the relationship between pathogen volume/frequency and sequencing volume/frequency?
- In instances where a frontline public health agency wants to use pathogen genomic data as part of their investigation, but can’t generate those data, what barriers are associated with sequencing hardware?
- If the frontline public health agency does in-house sequencing, do they have sufficient equipment and materials to perform that sequencing?
- When the frontline public health agency requests sequencing or performs sequencing themselves, what is the turnaround time for them to receive pathogen genomic data?

II. Technology and Analytic Workflows

Essential elements:

1. The frontline public health agency has the capacity to easily integrate pathogen genomic data with highly detailed epidemiologic data from investigations with reproducible digital workflows that minimize individual handling.
2. Sampling frames and intensity of sampling for response activities are well documented for a variety of scales based on the needs of the response.

3. Reporting and visualization of sequencing and epidemiological data can be interpreted easily by the field investigators.
4. Epidemiological metadata collection is standardized based on biological understanding of the pathogens and the goals of the responses.

Evaluation criteria:

- When pathogen genomic data are available to a frontline public health agency, does it always get integrated into the analysis and investigation? In instances where it does not, what are the reasons it is not used?
- What is the typical process to link genomic data to epidemiologic data? Does it vary by pathogen?
- What challenges are frontline public health agencies facing when trying to link their line list data or other epidemiologic data to whole genome sequence data from the relevant case?
- Who manages decisions about executing specimen sampling and associated sequencing for an individual public health response? For the overall response efforts of the public health agency?
- What kind of reports are shared between response personnel, sequencing personnel, and data science personnel? What visualizations are most relied upon for an effective response?
- Are there missing visualizations or aspects of the reporting that would improve responses overall?

III. Data Management, Storage, and Security

Essential elements:

1. The frontline public health agency can access pathogen genomic data at the level of analysis (raw, assembled, analyzed) that fits that agency's needs, interests, and workforce capacity.
2. The agency is able to receive genomic data (at the appropriate level of analysis from above) quickly enough that pathogen genomic data interpretations can guide prospective actions and decision-making during outbreaks.
3. Data storage capacity reflects the expected volume of surveillance for a sufficient response effort and is flexible in times of surges.
4. Clear documentation of specimen flow from clinical care, research, or surveillance settings to genomic testing laboratories.
5. Defined processes for relaying results to public health agencies and returning results to specimen sources.

Evaluation criteria:

- In what formats do public health agencies interact with pathogen genomic data (e.g., laboratory summaries, raw sequence files, or processed genomes)? Are any formats more useful or preferred by the agency than others?
- Do frontline public health agencies have the data storage and computing resources to receive and store their desired genomic data at typical capacity and at surge capacity?
- How much time is between sequencing of a specimen and a frontline agency receiving the sequencing data/sequencing report? Is the time feasible for use of that data for different responses?
- Is the specimen flow clearly mapped and understood across clinical and public health entities?
- How efficiently are results relayed back to healthcare providers or public health agencies?

IV. Personnel

Essential elements:

1. A workforce that is comfortable making interpretations from pathogen genomic data and understands the caveats and uncertainty associated with those analyses.
2. A workforce that is comfortable communicating pathogen genomic interpretations, as well as caveats and uncertainty, to clinical partners, other epidemiologists, policy makers, and the general public.
3. Members of the workforce that have a defined role to liaise and share information between laboratory, epidemiology, and clinical partners.

Evaluation criteria:

- What duties or roles do molecular or genomic epidemiologists have at the frontline public health agency? What skills and competencies are required at introductory, mid-level, and senior positions?
- In general, how many members of the workforce are designated to have a focus on genomic surveillance or molecular epidemiology?
- What funding sources support pathogen genomics personnel at the public health agency and for the public health agency if personnel are not centrally located within the frontline agency?

V. Strategic Visioning and Subject Matter Expertise

Essential elements:

1. Frontline public health agencies (e.g., county or major metro-area health departments, hospital infection control) know how and when pathogen genomic data can support their investigations or infection prevention and control efforts.
2. Frontline public health agencies know what types of samples to collect for sequencing during different types of investigations and for different pathogens.
3. Information relevant to clinicians and clinical microbiologists is readily shared and correctly interpreted to the appropriate point of contact for patient-level care.
4. Guidelines on how and when clinical providers should order pathogen genomic testing.
5. Defining the role of infection preventionists in prioritizing interventions based on genomic data.
6. Frameworks for communicating genomic data, indications, and limitations to clinicians and infection prevention teams.

Evaluation criteria:

- For what kind of responses and community health goals does the public health agency use pathogen sequencing (e.g., for outbreak management, environmental/wildlife monitoring, prophylaxis or partner contact, for clinical diagnostic partnerships, etc.)?
- What trainings do personnel engage in regarding analyzing and communicating about genomic epidemiology, bioinformatics, or similar supportive topics? Is the availability and diversity of topics sufficient for improving everyday practice?
- Who will provide the knowledge and training to frontline staff? How will understanding be assessed, and quality assurance be monitored?
- Are clinical providers aware of how to order pathogen genomic testing and understand its applications?
- Is there clear guidance for infection preventionists on applying genomic data to interventions?

- How effectively is genomic information communicated to clinical teams for decision-making?

VI. Data and Workflow Permissions, Access, and Authority

Essential elements:

1. Frontline public health agencies have mechanisms to share and mask patient identifying metadata so that relevant information can be shared for cluster detection in multiple jurisdictions.
2. Frontline public health agencies have cross-jurisdiction partnerships and data sharing plans in place ahead of outbreaks.

Evaluation criteria:

- Are mechanisms/permissions in place that allow frontline public health agencies to see phylogenetic clustering of their samples with samples from other jurisdictions?
- Do frontline public health agencies have a protocol, mechanism, or platform to interact with each other if they observe genomic relationships between cases in their respective jurisdictions?
- What are the regulations for HIPAA and privacy related to pathogen genomic data in the public health agency's jurisdiction?
- What measures are put into place to provide de-identified genomic sequences and associated metadata in public repositories?

Concluding Remarks

The “essential elements of genomic epidemiology” describe foundational resources and capabilities for use of pathogen genomic technologies in public health. While not comprehensive, the criteria for these elements described here allow assessment of current strengths and gaps in application of genomic epidemiology to understand, prevent, and respond to infectious disease threats. For the rapidly evolving field of pathogen genomics, this tool may also help identify opportunities for growth, for example in application of new genomic and related technologies as they become available, and in partnerships (e.g., public health organizations, academia, industry, healthcare sector) that may expand capability, utility and impact of the use of pathogen genomic epidemiology to improve public health.

Funding Acknowledgement

This document was supported by the Office of Advanced Molecular Detection, Centers for Disease Control and Prevention through Cooperative Agreement Numbers CDC-RFA-CK22-2204, CDC-RFA-OE20-2001. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention.

To cite this document, please use:

Pathogen Genomics Centers of Excellence Landscape Analysis Workgroup. (2024). Essential Elements of Genomic Epidemiology. https://github.com/PGCoE/resources.

*Working Group

Leads

Brooke Talbot (Georgia PGCOE; Emory University)
Andrew Warren (Virginia PGCOE; University of Virginia Biocomplexity Institute)

Members

Itika Arora (Georgia PGCOE; Emory University)
Alli Black (Northwest PGCOE; Washington Department of Health)
Justin Bahl (Georgia PGCOE; University of Georgia)
Kenneth Beckman (Minnesota PGCOE; University of Minnesota)
Eleanor Click (Centers for Disease Control and Prevention)
Kathleen Conery (Northwest PGCOE; Washington Department of Health)
Matt Doucette (New England PGCOE; Massachusetts Department of Health)
Philip Dykema (Northwest PGCOE; Washington Department of Health)
Logan Fink (Virginia PGCOE; Division of Consolidated Laboratory Services)
Esther Fortes (New England PGCOE; Massachusetts Department of Health)
Travis Glen (Georgia PGCOE: University of Georgia)
Christin Hanigan (Association of Public Health Laboratories)
John Houghton (Georgia PGCOE; Georgia State University)
Rebecca Hutchins (Georgia PGCOE; Georgia Tech Research Institute)
Samantha S. Katz (Centers for Disease Control and Prevention)
Tatyana Kiryutina (Georgia PGCOE; Georgia Department of Public Health)
Erin K. Lipp (Georgia PGCOE; University of Georgia)
Ruth Lynfield (Minnesota PGCOE; Minnesota Department of Health)
Hayleigh McDavid (Association of Public Health Laboratories)
Amy Mathers (Virginia PGCOE; University of Virginia)
Larry Madoff (New England PGCOE; Massachusetts Department of Public Health)

Melissa A. McDonald (Centers for Disease Control and Prevention)
Tonia Parrott (Georgia PGCOE; Georgia Department of Public Health)
Arunachalam Ramaiah (Georgia PGCOE; Georgia Department of Public Health)
Timothy Read (Georgia PGCOE; Emory University)
Emily Snavely (Virginia PGCOE; University of Virginia)
Shruti Subramaniam (Centers for Disease Control and Prevention)
Zachary Thompson (New England PGCOE; Massachusetts Department of Health)
Sean Wang (Minnesota PGCOE; Minnesota Department of Health)
Shirlee Wohl (New England PGCOE; Massachusetts Department of Health)
Joseph Yao (Minnesota PGCOE; Mayo Clinic)
Harrison Yu (Georgia PGCOE; Georgia Department of Health)

Meeting facilitation and coordination provided by Association for Public Health Laboratories (APHL)
Christin Hanigan (Association of Public Health Laboratories)
Hayleigh McDavid (Association of Public Health Laboratories)

Additional support provided by Council of State and Territorial Epidemiologists (CSTE)
Abby Hoffman (Council of State and Territorial Epidemiologists)

Additional resources

CDC Quality management system tools for assessing proficiency https://www.cdc.gov/lab-quality/php/ngs-quality-initiative/qms-tools-resources.html

CLIA (Clinical Laboratory Improvement Amendments): This federal regulatory program establishes quality standards for clinical laboratory testing, including genomic sequencing. https://www.cms.gov/medicare/quality/clinical-laboratory-improvement-amendments

The College of American Pathologists (CAP) provides an accreditation along with a set of guidelines which helps meet or exceed CLIA standards. https://www.cap.org/laboratory-improvement/accreditation

CDC provides competency guidelines for Public Laboratory Professionals with many domains and subdomains that touch on elements highlighted in this document https://www.cdc.gov/MMWR/pdf/other/su6401.pdf

CSTE (Council of State and Territorial Epidemiologists) provides an Applied Epidemiology Competencies Evaluation tool to identify areas for professional improvement https://aecs.cste.org/

Limited Glossary of Terms and Acronyms

Containerized Workflows: A method of packaging software and its dependencies into a standardized unit (a container) for easy deployment and execution across different computing environments. This ensures reproducibility and portability of bioinformatics pipelines.
Workflow Management System: Software that helps automate and manage the execution of complex bioinformatics workflows, often involving multiple steps and dependencies.
Human Read Removal: The process of filtering out sequencing reads that originate from human DNA from a sample containing a mixture of human and pathogen DNA, typically done during analysis of clinical specimens.
API (Application Programming Interface): A set of definitions and protocols that allows different software systems to communicate and exchange data with each other.
FISMA (Federal Information Security Management Act): A US federal law that requires federal agencies to implement information security programs to protect sensitive data.
FedRAMP (Federal Risk and Authorization Management Program): A US government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services.
Sampling Frame: The specific population or subset of a population from which a sample is drawn for analysis. In genomic epidemiology, this might be the set of cases, locations, or time periods from which samples are selected for sequencing.
InfoSec: Information security

COOP: Continuity of Operations Program

DUA: Data use agreement
DURSA: Data use and reciprocal support agreement
DSA: Data sharing agreement
HIPAA: Health Insurance Portability and Accountability Act of 1996
IRB: Institutional review board
IT: Information technology
LIMS: Laboratory information management system
MOA: Memorandum of agreement
MOU: Memorandum of understanding
NDA: Non-disclosure agreement
NIH: National Institutes of Health
PII: Personally identifiable information
R&D: Research and development
SOP: Standard operating procedure