Northwest Pathogen Genomics Centers of Excellence

Sequencing Landscape Analysis

TOOLKIT V3.0

Table of Contents

Background

The Molecular Epidemiology program at the Washington State Department of Health (WA DOH) developed a comprehensive assessment tool with the goal of understanding the status of pathogen genomics data generation and data use. We sought this information to improve our ability to support programs across WA DOH in leveraging pathogen genomics data in their public health investigations, and to support our own strategic planning and program development in building out this support.

This tool focuses on four main areas: 1) how teams have historically generated and used sequence data, by program area and by pathogen, 2) teams' existing genomic epidemiology capacity and their training and education needs, 3) bottlenecks and limitations impacting teams' ability to use pathogen genomic data, and 4) teams' goals for future use of sequencing data. There are four components to the tool: a questionnaire to guide the long form interviews, a template profile to organize the information collected through the interviews, two scales that facilitate standardized scoring, and a worksheet to facilitate data organization and analysis. The data analysis component facilitates identification of overarching themes across teams, categorization of teams' current state of sequencing data-use, as well as highlighting gaps and opportunities.

Ideally the baseline results will provide information about:

- the availability of sequence data for genomic epidemiologic analysis by condition,
- existing data workflows and platforms that host genomic data, and
- system interoperability which can be used to frame program planning.

Follow up activities may include growing technical support and subject matter expertise within an agency, and the development of roadmaps to increase and sustain the use of pathogen genomics in public health.

This tool can be implemented multiple times to collect information at baseline, midline, and endline. This information can serve as an evaluation metric to ensure the successful implementation of a molecular epidemiology program or grant.

Intended Audience

This guidance is intended for use by teams or agencies seeking to assess their agency's current capacity in pathogen genomics.

Tools

There are four components that make up the landscape analysis tool:

1) A questionnaire that was designed for data collection including:

- Current landscape:
    - General questions about the team
    - Sequencing of samples
    - Handling of genomic data results
    - Data analysis
- Areas of opportunity:
    - Roadblocks and limitations
- Current training resources:
    - Existing trainings
- Future directions:
    - Developing genomic surveillance for pathogens that are not currently sequenced
    - Increasing accessibility and utility of sequencing data for guiding public health action

2) A sample template profile to organize the data collected during the interviews.

3) A data analysis tool which includes two scales to help quantify the data collected:

- Sequencing data generation and data management.
- Capacity using and interpreting sequencing data.

4) A worksheet to enter the scores and visualize the results to identify overarching themes and gaps.

Questionnaire Tool

Before starting the interviews, it is important to explain the main goal of the landscape analysis, which is to understand communicable disease epidemiology teams' current ability to access pathogen genomic data and to analyze and interpret those data to support their investigations. Understanding this landscape improves the agency's ability to support advanced molecular detection development through effective strategic planning and program development. The use of pathogen genomics in public health is a nascent field, and therefore many teams may not have extensive experience using pathogen genomic data at baseline. Therefore, the goal is NOT to score and evaluate the team itself, but rather understand their needs in terms of sequencing data use and interpretation as they build their capacity to integrate pathogen genomic data into public health practice.

There are 50 questions in the questionnaire. The interviews can be conducted in a single session or in multiple. We recommend conducting the interview over two, hour-long sessions, starting with Part I (Assessing the current landscape, identifying areas of opportunity, and training resources) and following with Part II (Assessing future direction).

Questionnaire by Area of Interest

PART I

Assessing the Current Landscape

1. General Questions

1.1	Describe the staffing composition of your team (I.e., how many epidemiologists, disease and research investigators, public health nurses, etc. work on these pathogens together?).
1.1.1	What are their positions?
1.1.2	How long have they been part of the work? (Range of time is fine)

2. Sequencing of Samples

2.1	Has your team used sequencing data in the past?
2.1.2	For which pathogens?
2.2	How were samples identified for sequencing? *Ask this question for all pathogens listed in question 2.1.2* *(If there is an ask for clarification, can elaborate - Was this process used for routine surveillance or in specific outbreak situations? Or only samples that meet certain criteria are selected for sequencing?)*
2.3	Where was the sequencing done? *Ask this question for all pathogens listed in question 2.1.2*
2.4	How long did it take between the time specimens were identified and getting the data? *Ask this question for all pathogens listed in question 2.1.2* • About 2 weeks or less • About 4 weeks • About 3 months • About 6 months • About 1 year

3. Handling Results *Ask these question for all pathogens listed in question 2.1.2*

3.1	How were the data/results returned to you?

Questionnaire by Area of Interest

3.1.2	What format were the results in?
3.1.3	Was it raw sequencing data or a summary? *(If there is an ask for clarification, can give examples like the full-length sequence in fasta files, or an accession identifier for finding the sequence in a public repository, or summary information about the sequences such as a lineage call, presence/absence of an AMR allele, or identification as part of a cluster or a match with another specimen)*
3.2	Is there a processing step to get the data ready for storage in a database? If so what?
3.3	How was the data stored? What databases were sequences uploaded to?
3.3.1	If the sequencing data are stored in an external database – where does the team store the IDs used to access the sequence data record in the external database (i.e. CDDB, a tracking spreadsheet on Sharepoint)? *(If there is an ask for clarification, can give examples like, where do you store the sequence data accession IDs that point to a record in NCBI GenBank, GISAID,or other database like SEDRIC)*
3.3.2	Who can access the sequencing data? Is a specific account needed?
3.3.3	If the sequencing data is stored in a database, do you have training resources for this database?
4.	**Data Analysis** *Ask these questions for all pathogens listed in question 2.1.2*
4.1	What types of analyses were done with the sequencing data?
4.1.1	How frequently are sequencing analyses conducted?
4.1.2	Who performed the analysis? (ie. Was it someone on your team or an outside group?)
4.1.3	How was this used? Was this analysis useful? Why or why not?
4.1.4	Have you been able to directly use the sequencing data to explore different types of questions such as epidemiological linkage, strain frequency, disease phenotype, evolutionary motifs?
4.2	How have you used epi data (person, place, time) with sequencing data?

Questionnaire by Area of Interest	
4.2.1	To what extent are sequencing and epidemiological data integrated and analyzed jointly?
4.2.2	What was the source of the epi data?
4.2.3	How were the sequencing results linked back to the patient?
4.2.4	How comfortable are you using multiple data visualization tools to analyze sequencing and epidemiological data together?
4.2.5	Do you have specific types of data you would like to see integrated with the sequencing data? (ie. Healthcare encounters, vax, geospatial, etc.)
4.2.6	How comfortable are you using sequencing data during investigations and interpreting findings for all pathogens?
4.3	Have you worked with any external partners in your sequencing work in the past? (CDC, LHJs, Tribal Partners, ect.)
4.4	Have you used sequencing data in any analyses in the last year? What were they?
Identifying Areas of Opportunity	
5. Roadblocks	
5.1	What have been the "pain points" so far? Have there been any limitations that have prevented you from doing more with this data?
Training Resources	
6. Existing trainings	
6.1	What resources has your team used to gain their current level of comfort with sequencing?
PART II	
Assessing Future Direction	
7. Pathogens Not Currently Sequenced	

Questionnaire by Area of Interest

7.1	What pathogens do you want to sequence and are currently unable to?
7.1.1	If there are multiple pathogens you want to expand sequencing to, how would you order them in terms of priority?
7.2	Are there pathogens you currently send for sequencing at a lab other than your routine Public Health Lab, but that you would prefer to have your routine Public Health Lab sequence?
8.	**Pathogens Sequenced at the Public Health Lab**
8.1	Which pathogens are already being sequenced at the Public Health Lab?
9.	**Increasing Accessibility of Sequencing**
9.1	What types of sequencing results would be most useful to you? (ie. WGS, sequencing of a single gene, sequencing of key phenotypic loci such as antimicrobial resistance loci etc.)
9.2	Have you already developed a process for how specimens should be selected? If so, what is that process?
9.2.1	From whom do specimens need to be requested?
9.2.2	Where will the funding for sample acquisition come from?
9.3	How would you like to receive the data?
9.3.1	What format would you like the results in?
9.3.2	Would you rather have raw sequencing data or a summary? *(If there is an ask for clarification, can elaborate this means the full written-out sequence versus identification as part of a cluster or a match with another specimen)*
9.4	Are you aware of any databases where the data should be uploaded? Will the specimen come with a clinical accession number?
9.5	Are you aware of work being done by any other agencies?
9.5.1	Is it of value to be able to compare results with these agencies (ie. other states, CDC)?

Questionnaire by Area of Interest	
9.5.2	Are there any agencies you know you would like to be able to share data with (ie. LHJs, tribal partners, CDC, ect.)
9.6	What types of end deliverables are you hoping for?
9.7	Do you have specific types of data you would like to see integrated with the sequencing data? (ie. Healthcare encounters, vax, geospatial, ect.)
9.8	What capacity would your team ideally have for analysis of paired epidemiologic and sequencing data?
9.9	What training is needed to address the gaps?

Template Profile

Name of the team: Current Sequencing Efforts

Sequencing Data Production

Pathogens Sequenced	Sequencing Frequency	Sequencing Labs	Type of Sequencing	Timing of Sequencing	Sequencing Results	Data Storage

Sequencing Data Analysis

Types of Analyses	Who Performed the Analyses	Analyses Use	Source/Use of Epi data	Other Useful Epi Data	External Partners previous analyses	Frequency of Seq. Data Use

Training Resources

In house Trainings	Other Trainings	Training needs

Limitations

Name of the team: Future Sequencing Directions

Pathogens Sequenced at PHL

Pathogens Sequenced at PHL	Sequencing Frequency	Data Format	Other Types of Deliverables

Pathogens Not Currently Sequenced at PHL

Pathogens NOT Sequenced at PHL	Work Done at other Agencies	Other Types of Deliverables	Other Data Specifications

Data Analysis

The data analysis component facilitates organization of the information collected during the interviews. This enables the analyst(s) to identify overarching themes across teams as well as gaps and opportunities for improvement. The baseline results will provide information about the availability of sequence data for genomic epidemiology analysis by condition, existing data workflows and platforms that host genomic data, and system interoperability. This information can then frame the agency's future activities, including through technical support and/or the development of roadmaps that guide how the agency will increase and sustain the use of pathogen genomics in public health.

There are two scales that we designed to help quantify the data collected. One scale aims to score the areas of sequencing data production and data management, while the other focuses on the current capacity building needs within the team and the support they have received so far with pathogen genomic data analysis and interpretation. The data analysis tool also includes a worksheet to enter the scores and visualize the results to identify overarching themes.

Data management assessment scale

The data management assessment was developed to understand the status of sequencing data generation, timeliness, accessibility, and linkage for each pathogen currently supported by each team being interviewed.

The data management scoring scale ranges from 1 to 5, where 1 represents that the sequencing data generation, timeliness, accessibility, and linkage of genomic and epidemiologic data for that specific pathogen is not adequate to support the team's needs, and 5 represents the best-case scenario where the sequencing data generation, timeliness, accessibility, and linkage of genomic and epidemiologic data are supportive of all surveillance and outbreak response needs for that specific pathogen.

Data Management Scale by Pathogen	
Sequencing of samples	
5	Representative surveillance *and* intensive outbreak are both conducted
4	Representative surveillance *or* intensive outbreak sequencing are conducted
3	All samples that meet specific criteria are sequenced
2	Samples are sequenced ad-hoc not following specific criteria
1	Samples are not being sequenced using WGS
Sequencing labs	
5	All sequencing is conducted at the agency's routine Public Health Lab (PHL)
4	Most of the sequencing is conducted at the PHL but samples are sometimes submitted to other labs for sequencing and bioinformatic assembly
3	Most of the sequencing is conducted at CDC but the PHL is developing internal capacity to sequence this pathogen
2	All sequencing is conducted at CDC
1	No sequencing is being conducted for this pathogen currently
Timeliness from collection date to receipt of sequencing results	
5	Less than 2 weeks
4	2-4 weeks
3	3 months
2	6 months
1	1 year
Data Storage and accessibility	
5	Sequencing data are uploaded to public repositories and the accession number is stored in a database or document at the public health agency for easy access
4	Sequencing data are uploaded to public repositories, but the accession numbers are not recorded
3	Sequencing data are stored somewhere locally, but the team is unsure if data is uploaded to public repositories
2	Sequencing data are stored somewhere locally but will not be made available in public repositories
1	Raw sequencing data is not stored in public repositories or locally or the status of raw sequencing data storage is unknown

Data Management Scale by Pathogen	
Data Interoperability	
5	It is very easy to access and join sequencing and epidemiological data together
4	Sequencing and epidemiological data can be accessed and pulled together, but data interoperability could be improved
3	Sequencing and epidemiological data exist in separate databases, and the data can be linked, but it is somewhat challenging and manual
2	Sequencing and epidemiological data exist in separate databases and even though some data fields can be linked, not all datapoints, such as sequencing accessions, are collected/stored with the epidemiological data
1	Sequencing and epidemiological data may exist in separate databases but there is no known crosswalk to link the data sources

Capacity assessment scale

The capacity assessment contains a scale to categorize the current support that each team has received from agencies such as CDC, the agency's routine Public Health Lab (PHL), the State Department of Health (DOH), or other subject matter experts, in analyzing, interpreting, and communicating pathogen genomics results. The capacity assessment also aims to assess how comfortable each team is conducting sequencing analysis on their own at a decentralized level, meaning without support from CDC. The main goal of collecting this information is to plan training and tailor strategies for improved pathogen genomic data technical support within the agency.

The capacity assessment scoring scale ranges from1 to 5, where 1 represents that the interviewed team has been provided limited support in using, conducting, and interpreting sequencing data, and where 5 represents the best-case scenario where the team being interviewed has become highly proficient with genomic data analysis and interpretation, with both extensive experience and advanced technical skills.

Capacity Assessment Scale by Team	
Cadence of sequencing analyses for pathogens	
5	Analyses of pathogen genomic sequence data are conducted regularly for all pathogens
4	Analyses of pathogen genomic sequence data are conducted regularly for some pathogens and ad-hoc for others

Capacity Assessment Scale by Team	
3	Analyses of pathogen genomic sequence data are conducted ad-hoc for all pathogens
2	Analyses of pathogen genomic sequence data are conducted ad-hoc for some pathogens
1	Analyses of pathogen genomic sequence data are not conducted for any of the pathogens for which genomic data are available
Sequencing analyses performed in-house	
5	All sequencing analyses are performed in-house (within the team or agency)
4	Most sequencing analyses are performed in-house, with a couple of exceptions
3	Some sequencing analyses are performed in-house, but most are conducted by CDC or by an external lab conducting the sequencing
2	None of the sequencing analyses are conducted in-house, but CDC provides strong support in this area. For example, CDC provides summary results and there is space to discuss how to interpret the results and the assumptions/caveats.
1	None of the sequencing analyses are conducted in-house and CDC is only able to provide the bare minimum support. For example, CDC provides summary results but there is not space for iterative discussion about what the analyses mean and the assumptions/caveats.
Types of sequencing analyses that are performed	
5	The team is able to directly use the sequencing data to explore different types of questions, such as investigating epidemiological linkage, estimating strain frequency, assessing disease phenotype, exploring evolutionary motifs of interest.
4	The team understands that sequencing data can be used to conduct different types of analyses that explore different aspects of a pathogen's epidemiology, and they have a couple of questions they would like to explore, but they are not able to conduct the analyses on their own.
3	The team can directly use the sequencing data, but they are confined to one type of analysis that is used regularly.
2	Only summaries of sequencing analyses (e.g. lineage call, sequence type, etc.) have been used, and the team does not conduct direct analysis of the sequence data.
1	Sequencing analyses are not conducted for any of the pathogens.
Use of sequencing and epidemiological data	
5	Sequencing and epidemiological data are always integrated, analyzed jointly, and there are no other epidemiological datapoints that need to be integrated

Capacity Assessment Scale by Team	
4	Sequencing and epidemiological data are often integrated but there are other epidemiological datapoints that would be useful to analyze that are not currently integrated with the genomic data analyses.
3	Sequencing and epidemiological data are sometimes integrated for joint analysis and visualization.
2	Sequencing and epidemiological data are rarely integrated for joint analysis and visualization.
1	Sequencing and epidemiological data are not integrated for joint analysis and visualization.
Data visualization of sequencing data	
5	The team is comfortable using multiple data visualization tools and/or approaches for exploring analyses of pathogen genomic sequence data
4	The team is comfortable with a couple of data visualization tools and/or approaches for exploring analyses of pathogen genomic sequence data
3	The team is familiar with a couple of data visualization tools for exploring analyses of pathogen genomic sequence data, but they don't fulfill all the needs of the team
2	The team is aware of data visualization tools for exploring analyses of pathogen genomic sequence data, but haven't used them
1	Unsure about which data visualization tools for exploring analyses of pathogen genomic sequence data are available and how to use them
Sequencing data interpretation	
5	The team is comfortable using sequencing data during investigations and interpreting findings, for all pathogens
4	The team is familiar using sequencing data during investigations but not fully comfortable interpreting findings for some pathogens independently
3	The team can use sequencing data during investigations, and interpreting findings for all pathogens, but with support from other teams/subject matter experts
2	The team relies heavily on other teams/agencies/subject matter experts to be able to use sequencing data during investigations and to interpret findings
1	The team is not comfortable using sequencing data during investigations, nor with interpreting findings
Capacity building and training needs	
5	There is no need for further capacity building or trainings

Capacity Assessment Scale by Team	
4	Minimal training and some capacity building would still be useful
3	There is need for some capacity building and trainings in specific areas of genomics for some of the pathogens
2	There is a large need for capacity building and trainings in specific areas of genomics, and for all pathogens
1	There is a large need for capacity building and trainings in all areas of genomics and for all pathogens

Data Collection and Data Analysis Worksheet

Use the Excel worksheet for data collection and data analysis.

Limitations

It is recommended that the group conducting the landscape analysis outlines the limitations as part of a final report. For example, some of the limitations we noted were the need to update the Program Profiles as advance molecular detection is a field that is evolving rapidly within our agency and globally. Thus, the Program Profiles needed to be treated as living documents that should be updated regularly for them to retain their utility. Another limitation we noted was the need to include other programs that handle and use genomic data that were not included in the baseline landscape analysis.

Recommendations

The recommendations are one of the most important products of conducting this landscape analysis. Based on the overarching themes identified discuss which actions are needed to address gaps and improve the generation and use of sequencing data in epidemiological work. One way of structuring the recommendations is classifying them in two categories: 1) those that are meant to facilitate collaboration within the agency and with external partners, and 2) those that are meant to improve systems and infrastructure. These recommendations will be useful when work-planning and even grant writing.