Contributing Laboratory Data to GenomeTrakr and PulseNet Networks:

A Focus Group Summary Report

JUNE 2022

APHL ASSOCIATION OF PUBLIC HEALTH LABORATORIES®

INTRODUCTION

In December 2021, APHL held two focus groups with laboratorians contributing data to both PulseNet and GenomeTrakr to discuss their experiences analyzing and submitting both required and requested data to both networks. To our knowledge, this was the first time a systematic approach to collecting this data had been generated to provide federal partners with user experiences for both PulseNet and GenomeTrakr networks. APHL hoped to identify advantages and challenges of contributing data to both networks, as well as opportunities for improving analysis or submission processes to make laboratory workflows more efficient. Eight people from seven laboratories participated in the focus groups, representing both low- and high-volume submission laboratories, as well as new and long-time contributors to both networks. Participants shared their experiences being a contributor to both PulseNet and GenomeTrakr, including workflows for quality control and data submission processes and tools used for real-time surveillance of enteric pathogens. For privacy reasons, the participants' answers have been deidentified.

WORKFLOW FOR DATA ANALYSIS AND SUBMISSION

Tools for Data Submission

APHL asked focus group participants which tools they use and their processes for submitting data to PulseNet or GenomeTrakr. Primarily, laboratories follow the PulseNet protocols until data submission. Six of the seven laboratories use BioNumerics to submit both PulseNet and GenomeTrakr data to the National Center for Biotechnology Information (NCBI). One laboratory uses BioNumerics to submit PulseNet data and either submits GenomeTrakr data using BioNumerics or directly uploads data using the NCBI submission portal; this is because they have separate staff that analyze PulseNet and GenomeTrakr isolates.

As PulseNet looks to sunset BioNumerics, the focus group participants noted their hope that BioNumerics' replacement can be harmonized with GenomeTrakr, allowing for a single, standardized "push button method" for both networks with a single NCBI template. There was also a clear desire for PulseNet 2.0 to be a more simplified system without a litany of tools or buttons that are not useful for PulseNet data, as well as simplified entry and upload processes.

Process for Checking Data Quality

APHL queried participants regarding the tools used and processes for quality checking (QC) data for PulseNet or GenomeTrakr. Six of the seven laboratories exclusively use the PulseNet protocols, using BioNumerics to QC their data. They noted that BioNumerics was already the established program in their laboratories prior to joining the GenomeTrakr network. Other notable findings include the following:

- One laboratory uses an internal taxonomic pipeline to identify genus/species for non-PulseNet organisms.
- One laboratory uses the CFSAN SNP pipeline to evaluate laboratory contamination by comparing new sequences with their laboratory positive control; this was developed prior to the implementation of BioNumerics.
- One laboratory surveyed has dedicated PulseNet and GenomeTrakr staff and differentiates sample workflows for each network, using GalaxyTrakr tools (including MicroRunQC) for GenomeTrakr samples. Those samples are mainly historical isolates or non-PulseNet organisms from environmental samples.

What Works Well?

One laboratorian noted that having one standardized "push button method" for both networks is greatly appreciated and saves precious time in public health laboratories where the analyst is both the bench scientist and the bioinformatician.

One laboratorian using BioNumerics noted that the initial QC metrics (e.g., concentration, purity) are easy to compile and give a quick sense of run quality. The laboratory that uses GalaxyTrakr tools noted that MicroRun QC is very easy to run and interpret, although it can be slow.

Participants noted that with a good connection, uploading the data to PulseNet via BioNumerics is smooth and straightforward. One laboratory also noted that BioNumerics is streamlined, familiar and integrates well with BaseSpace so that analysis is fast. One laboratory noted that while work is required upfront to create the individual templates for each project, using BioNumerics to submit data to NCBI is easy. Direct upload to NCBI using the submission portal also goes well. Participants claimed that submission is easier when laboratories are primarily running only one organism on a sequencing run since there is no need to switch between databases. One participant noted that it would be helpful if laboratories could submit all the run data into BioNumerics and have the program filter the sequences into their respective organism databases. These pipelines exist, as one participant noted that APHL's StaphB group uses a similar solution.

What is Challenging?

Delays

Traffic on the Calculation Engine can cause delays in processing. The Calculation Engine processing time is highly variable depending on the time of day, internet bandwidth, the number of simultaneous logins to CDC and the organism database with which one is working.

Troubleshooting and resolving failed sequences are also time-consuming, ranging from hours to days. Using BioNumerics for submission works well when the data clear QC metrics, but it is very cumbersome to track issues when they fail at various stages of the process. Uploading to NCBI does not always occur in real time, which can affect that laboratory's turnaround time. Failed submissions to NCBI do not always prompt a feedback notification as to why they are failing through the Calculation Engine job tracking window. Additional information is available in a location on the NCBI website that is difficult for most users to find. Participants noted it would be helpful if the upload process gave consistent feedback when errors occur, or data fails.

Both focus groups discussed issues with communication during the QC analysis process. Participants find it difficult to know who to contact with issues—PulseNet, GenomeTrakr or NCBI. Responses from NCBI can be delayed anywhere from a few hours to several days. Several laboratories noted issues reading the error code and determining the problem in the calculation engine. Without a staff bioinformation, laboratories must reach out to CDC, FDA or NCBI contacts for assistance.

Time is an issue in the data submission process. Differences in genome sizes (e.g., Listeria v. Salmonella) and the size of the database can cause differences in how quickly jobs are processed in BioNumerics. As previously noted, NCBI uploads can take days to show in the database. The PulseNet firewall expires every 2–3 hours and can disconnect if there are lots of data to analyze and upload. This can cause an interruption with the server, creating an error. Participants noted that the workflow is disjointed and using multiple organism databases is timely. Searching PulseNet for matching sequences is a slow process that could be improved. The recommendation is to load 60 days' worth of data on the server and search for query words, but it does not work as well as a fast match against the entire server.

Protocols

Some participants find it difficult to transition from one network's SOPs to another, as the structure and organization of the SOPs are different. The focus groups noted that management of GenomeTrakr SOPs on protocols.io can be confusing, specifically trying to determine if comments on the SOPs are simply feedback or actual changes to the protocol. It was noted that the "Published Protocols" designation did help alleviate this confusion. Additionally, it was noted that sections of the GenomeTrakr protocol in a specific version did not match the PulseNet protocol (specifically the fields for the NCBI templates). This created confusion as to which protocol to follow. The focus groups suggested that PulseNet either directly insert the GenomeTrakr protocol language or state "Refer to GenomeTrakr protocol" in its instructions to reduce further misunderstanding.

Having different requirements for data quality between programs (e.g., clinical v. environmental) can be troublesome. For example, Salmonella needs 30X coverage for clinical isolates but only 20X coverage for environmental isolates.

Participants noted that standardization between PulseNet and GenomeTrakr on this issue would ease the analysis process (e.g., calculating how to load a cartridge would be less complicated).

No participants found places where duplicate data entry was required. However, it was noted that entries of the same information can be changed independently without concomitant update of the submitted data. For example, information updated in the PulseNet database are not automatically updated in NCBI. Additionally, updating data in NCBI can be an arduous task as it cannot be done directly by the laboratory, rather it is an emailed request to NCBI to make changes. NCBI staff do not always notify the laboratory that the changes have been completed, so the laboratory must monitor the data for these changes.

One laboratory spoke about their process for GalaxyTrakr dry lab submission, which is used to share data to an individual external laboratory for their projects. If the GenomeTrakr isolates cannot go through BioNumerics, it goes through their internal taxonomic pipeline and the sequences are uploaded into GalaxyTrakr. Because GalaxyTrakr does not allow sequence import from BaseSpace, the laboratory must download the sequences from BaseSpace to a flash drive before uploading to GalaxyTrakr; this adds quite a bit of time to the workflow.

Metadata

Participants discussed challenges around metadata templates. Creating metadata templates and importing the metadata is a cumbersome process. It is also difficult to maintain separate templates for each organism. Given that each organism has its own database in BioNumerics, the metadata templates must be closely curated and monitored to assure consistency of the data uploaded. Additionally, information needed on the environmental sample templates seems to change frequently, and the information is not always relayed clearly to public health laboratories. If a change is needed on a template, there are likely 10–15 templates that must be changed.

Participants also noted that manual metadata entry into BioNumerics is very time consuming, especially having to switch between databases. Multiple organism-specific databases require curation and maintenance of each database separately. One participant from an agriculture laboratory mentioned they've created metadata templates for each of the databases; they partner with international submitters and need to clean the metadata (e.g., remove dashes, colons) and format it correctly, which is time consuming.

TOOLS USED FOR REAL-TIME SURVEILLANCE OF ENTERIC PATHOGENS

Participants were asked about which tools they use for real-time surveillance of enteric pathogens. All seven laboratorians surveyed use BioNumerics to conduct real-time, national and local surveillance of enteric pathogens. One laboratory uses GalaxyTrakr tools for research studies. One laboratory uses NCBI Pathogen Detection to look for matches within clusters; this laboratory also compares this information to what they find in PulseNet. Two laboratories also use SEDRIC to investigate outbreaks and find sources of contamination. Additionally, not all historical isolates are in the PulseNet national database—many laboratories are sequencing historical food and environmental isolates for GenomeTrakr that pre-date when PulseNet began accepting WGS data directly from the laboratories. Laboratories that are only using BioNumerics for surveillance are missing the isolate data that is only in NCBI. This creates a gap and can hinder outbreak investigations.

Participants noted that BioNumerics' feedback mechanisms are lacking. There is no feedback given to the submitter when outbreak or allele codes are assigned—epidemiologists may be notified, but this is not helpful to laboratories that lie outside of health departments. An automated response when a PulseNet ID is assigned an outbreak code would be much preferable to having to manually check for outbreak or allele code assignments.

One participant noted difficulties using the Vibrio database. The laboratory had to use their own tools to investigate an outbreak of Vibrio in shellfish, and there was a several week delay in PulseNet's ability to provide a SNP matrix for the outbreak.

APHL asked participants about why they prefer to use BioNumerics over NCBI's Pathogen Detection to conduct surveillance. In general, they found that the lack of metadata in the Pathogen Detection was a huge limitation for surveillance investigations. When compared with BioNumerics, participants found that Pathogen Detection was overly complicated.

DIFFERENCES IN ANALYSIS AND INTERPRETATION OF PULSENET AND GENOMETRAKR DATA

The focus group participants discussed the differences in analysis and/or interpretation of PulseNet and GenomeTrakr data and how these differences affect the processes in their laboratories. Participants mainly use BioNumerics to analyze both PulseNet and GenomeTrakr data. The main difference in analysis and interpretation between the networks is the use of allele versus SNP differences.

APHL asked the laboratorians how PulseNet's use of allele differences and GenomeTrakr's use of SNP differences affects the analysis of data across networks. One participant uses SNP distances for PulseNet/GenomeTrakr comparisons. One laboratory looks at both sets of data (alleles and SNPs) and noted that interpretation is highly dependent on the organism and subtype. This participant has seen instances where a sample looks like the allele code should be similar to a cluster but is actually 60–70 SNPs different.

One laboratory mentioned that it is very time-consuming to match a food isolate against the PulseNet national database, especially for a Salmonella isolate; the participant noted that the search could take up to 30 minutes. This laboratory is looking to use NCBI Pathogen Detection in these instances but have not yet begun this implementation.

Participants expressed their hopes for a system with complementary analysis tools. For instance, NCBI now has an MLST scheme on its back end, but it is not the same as CDC's MLST scheme. Participants also indicated that, in an ideal scenario, data could be entered into a database and a report would be generated that included both allele and SNP difference analyses. It can be difficult to navigate between the two databases, so having the analyses for both networks in one place would be ideal. One participant said they foresee laboratories gravitating more towards SEDRIC to perform analyses and investigations. Participants noted that they appreciate the ability to get a link to their query in NCBI's Pathogen Detection for their inspectors that shows the relationship between isolates of interest.

CHALLENGES OF SUBMITTING REQUIRED OR REQUESTED METADATA TO BOTH NETWORKS

The focus group discussions identified two major challenges for laboratories submitting metadata to both networks. It can be difficult for laboratories to interpret what metadata is needed for both networks. Parameters frequently change, and communication on these changes does not come from both PulseNet and GenomeTrakr in a coordinated fashion. Additionally, laboratories find it difficult when GenomeTrakr requests additional metadata which are not included in the NCBI template for PulseNet.

The participants had several suggestions for how to address these challenges and improve their experience submitting metadata to the networks.

- First, it would be helpful to create version numbers for the NCBI templates for BioNumerics to help laboratories easily identify the most current version.
- There was also a request that GenomeTrakr delay requiring new metadata fields until PulseNet can fix the scripts in BioNumerics—states could then avoid the extra metadata upload (sometimes months after submission) and save valuable time and effort.

- One laboratory noted that they plan to pay for and process environmental samples linked to an outbreak through PulseNet funding rather than Laboratory Flexible Funding Model (LFFM) funding, which would require using additional resources to upload metadata at a later date.
- As CDC looks ahead to PulseNet 2.0, participants suggested looking at ways to update these data fields without needing to rewrite code.

EXPERIENCES BEING A CONTRIBUTOR TO BOTH NETWORKS

Enhancements

APHL asked focus group participants to identify how PulseNet and GenomeTrakr have complemented one another and assisted with outbreak investigations. Participants largely agreed that querying results to include both PulseNet and GenomeTrakr isolates has helped with investigations. Several participants explained that when they search for an individual sample in the PulseNet server, it is very helpful to have the GenomeTrakr samples included in that search and can help epidemiologists identify sources of contamination.

One participant mentioned a long-term (~5 years) Listeria monocytogenes outbreak associated with a food production facility. The clinical isolates matched by PFGE and WGS, and environmental samples from the facility matched the clinical samples. The environmental samples were sequenced with GenomeTrakr funds and run through BioNumerics for analysis.

Another participant spoke of an interesting case where they identified positive samples of butter from a production facility and found clinical isolates in the database with similar WGS results. Several months later, there was another clinical isolate that appeared very similar to the original outbreak strain. While the case was out of the time frame of the typical cluster definition, it matched within 0–1 alleles.

Challenges

Once again, communication between the networks and to the states was identified as a major issue for laboratories contributing data to both PulseNet and GenomeTrakr. While at the federal level there are usually many folks involved in sequencing (e.g., food testing, foodborne illness testing, rabies), there may be just one person at the state laboratory responsible for all sequencing. Identifying contacts within PulseNet, GenomeTrakr and NCBI can be challenging. For example, there are different email addresses and network contacts for PulseNet and GenomeTrakr at NCBI. It was mentioned that the GenomeTrakr protocol for data curation at NCBI lists email contacts, and PulseNet has a training module for how to log into the NCBI troubleshooting page. Participants suggested highlighting where these resources are located and emphasizing the correct contact for each network to help states know where to turn for assistance.

States also suggested that better coordination and communication with proficiency test instructions would be helpful. One state was unaware that PulseNet would not share its PT results with GenomeTrakr. Although the PT instructions do state that states must share their sequences with both networks, it may be helpful to emphasize this during the PT rounds. States also mentioned that the PT reports differ slightly in structure between the networks, mostly because the types of information they are looking for differ. One state did note that getting different information about the PT results was helpful and interesting.

Participants noted that the evolving nature of required metadata is challenging, especially when the requested metadata can be difficult to collect. States suggested that epidemiologists and investigators be brought into the conversation about what metadata they can reasonably collect.

For grant requirements, it can be difficult to meet LFFM deliverable numbers given the way the funding year is structured. One participant noted that they are always busy testing Salmonella and Campylobacter in the summertime. GenomeTrakr funding cycles begin on July 1, so it can be difficult to meet grant-required testing numbers when the deadline falls mid-summer. One participant noted that they are a low-throughput laboratory and are working to adjust how they fill sequencing runs to improve turnaround time.

During this discussion on challenges, several participants also mentioned that they are seeing a similar situation with wastewater surveillance through the LFFM. CDC and FDA are both involved in WWS testing, with several states testing for each agency. CDC and FDA are asking the states for different things for collection, testing and sequencing. Disconnects in communication leave states unsure of how to proceed. Many states are taking it upon themselves to innovate and create methods and hope they fall in line with what FDA and CDC are requesting.

IMPACT OF COVID

APHL asked the focus groups how their workflows changed during the COVID-19 pandemic. Many participants noted that their laboratories fill sequencing runs with multiple types of samples, including SARS-CoV-2. At the time of the focus groups, many still had staff pulled into COVID testing, carving out time to perform both their daily work and COVID samples. Some laboratories noted they have had to put some testing on the backburner, shifting enterics testing from once weekly to once monthly; luckily, foodborne illness (or at least clinical samples) seemed to decrease during the pandemic.

Supply chain issues have also caused delays in testing, leaving laboratories without access to needed reagents and consumables. Several participants noted that they automated their library prep process during the pandemic, which has greatly improved their testing capacity and turnaround times, and this automation will stay in the laboratories beyond the pandemic.

One participant noted that their laboratory has allowed them to analyze sequences while remote working through a secure VPN; having the ability to process a run on a Saturday morning bypasses the typical traffic on the Calculation Engine on a Monday morning.

There was also an observation during one of the focus groups that COVID has diminished the sharing of lessons learned between colleagues. Participants noted that they miss small conversations where folks share their challenges and various solutions.

SUMMARY

The focus group participants were very open and honest about their experiences as contributors to both GenomeTrakr and PulseNet. While being a contributor to both networks is not necessarily challenging or duplicative once you understand the workflows, it does require a lot of time and attention to detail. While there isn't much duplication of effort, there are many repeated steps throughout the process. Any opportunities to automate workflow steps should be explored. Due to their PulseNet contributions, much of the feedback received concerned the BioNumerics workflow. With PulseNet 2.0 on the horizon, this feedback can be vital for understanding states' current challenges and hopes for an improved system that can be harmonized with GenomeTrakr workflows. During these early stages, federal partners should talk to state laboratories—large and small, agricultural and strictly clinical—and learn about their workflows and build something that works for every participating laboratory.

One notable comment was on the lack of state input on the Gen-FS working group. Participants felt that while it is important for the federal partners to work together on harmonization efforts between the agencies and between the networks, it is also critical to have states involved in the work groups, particularly the Metadata working group. States can voice what is feasible and provide input on the direction of the two networks.

A common theme that emerged during both focus groups was the importance of communication: among network contributors, between networks and contributors and between the networks themselves. Coordinated and consistent messaging will be greatly appreciated by network contributors.

Collected data came from state respondents who primarily work through PulseNet workflows. It would be valuable to solicit feedback from federal and non-federal laboratories that use GenomeTrakr protocols as their primary workflows, identifying challenges encountered in those laboratories when they pull data from PulseNet, thus revealing additional opportunities for improvement. APHL recommends holding an additional focus group discussion with federal partners

from FDA and USDA to provide a more well-rounded picture of the data submission experience.

APHL thanks the focus group participants for their time and contributions to these efforts and asks our federal partners to consider the feedback collected when making improvements to the GenomeTrakr and PulseNet networks.

Association of Public Health Laboratories

The Association of Public Health Laboratories (APHL) works to strengthen laboratory systems serving the public's health in the US and globally. APHL's member laboratories protect the public's health by monitoring and detecting infectious and foodborne diseases, environmental contaminants, terrorist agents, genetic disorders in newborns and other diverse health threats.

This project was 100% funded with federal funds from a federal program of $0.5 million. This publication was supported by Cooperative Agreement #U18 FD 006937 from the U.S Food & Drug Administration (FDA). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of FDA.

8515 Georgia Avenue, Suite 700
Silver Spring, MD 20910

Phone: 240.485.2745

Fax: 240.485.2700

www.aphl.org