PulseNet 2.0

A Future-Proof Infrastructure for Genomic Data Management and Analytics

White Paper Version 1.0

MAY 2023

APHL ASSOCIATION OF PUBLIC HEALTH LABORATORIES®

PulseNet™
The National Molecular Subtyping Network for Foodborne Disease Surveillance
CDC

Contents

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.

8515 Georgia Avenue, Suite 700, Silver Spring, MD 20910 | 240.485.2745 | www.aphl.org

This project was 100% funded with federal funds from Cooperative Agreement #NU60OE000104, funded by the US Centers for Disease Control and Prevention (CDC) for $1.17 million. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC or the US Department of Health and Human Services.

Definitions

ANI ..... Average Nucleotide Identity

API..... Application Programming Interface

AR..... Antimicrobial Resistance

AWS ..... Amazon Web Services

BAH ..... Booz Allen Hamilton

BaseSpace . Illumina's cloud-based analysis and storage platform

bMx ..... bioMerieux

CDC ..... Centers for Disease Control and Prevention

CE ..... Calculation Engine

cgMLST ..... core genome Multi-Locus Sequence Typing

CIA..... Confidentiality, Integrity and Availability

CLIA..... Clinical Laboratory Improvement Amendments

COTS ..... Commercial off-the-shelf product/software

CSV..... Comma Separated Values

Data lake.... Storage repository for all data types (e.g., structured, semi-structured, unstructured) in its native format

Data lakehouse .....
Open data management architecture implementing similar data structures and management features of data warehouses with low-cost storage of data lakes

DevSecOps . Development, Security and Operations

eLIMS..... electronic Laboratory Information Management System

FASTA/Q ..... Text-based file format for storing biological sequences (FASTQ contains sequence quality information)

FedRAMP.... Federal Risk and Authorization Management Program

FIPS..... Federal Information Processing Standard

FISMA ..... Federal Information Security Management Act

FOC..... Full Operational Capability

FOIA ..... Freedom of Information Act

GUI ..... Graphical User Interface

HIPAA..... Health Insurance Portability and Accountability Act of 1996

HPC ..... High Performance Computing

IT..... Information Technology

LIMS..... Laboratory Information Management System

MOU ..... Memorandum of Understanding

MVP..... Minimum Viable Product

NCBI ..... National Center for Biotechnology Information

NGS ..... Next Generation Sequencing

NIST..... National Institute of Standards and Technology

OSDB..... Organism-Specific Database

PFGE ..... Pulsed-field Gel Electrophoresis

PHI ..... Protected Health Information

PHL..... Public Health Laboratory

PII ..... Personally Identifiable Information

refID ..... reference Identification database

SAMS ..... Secure Access Management Services

SC..... Security Controls

SEDRIC ..... System for Enteric Disease Response, Investigation, and Coordination

SQL..... Structured Query Language

SRA ..... Sequence Read Archive

TOR ..... Terms of Reference

UAT ..... User Acceptance Testing

Viz..... Visualization (of data)

wgMLST ..... whole genome Multi-Locus Sequence Typing

WGS ..... Whole Genome Sequencing

Workflow managers.....
Push bioinformatics pipelines to containers with appropriate software to complete a task/job (e.g., Nextflow, Snakemake, Galaxy)

Introduction

This document is intended to provide a high-level overview of the proposed data architecture for PulseNet 2.0. If the details come across as vague, that merely reflects where we are in the development phase. Further questions are welcome and can be directed to pulsenet@cdc.gov with "PulseNet 2.0" in the subject line.

Background

For 25 years, PulseNet USA, has brought together federal, state, local and international public health for the purpose of identifying, tracking and preventing foodborne illnesses and outbreaks. This network currently consists of over 80 laboratories who use standardized laboratory workflows and data analysis tools to detect local and multi-state outbreaks in near real-time. The bioinformatics and information technology (IT) infrastructure that supports this work is a customized version of an end-to-end commercial off-the-shelf (COTS) software solution called BioNumerics (bioMerieux [bMx]/Applied Maths). For most of its 25 years, PulseNet has relied on pulsed-field gel electrophoresis (PFGE) for the detection and tracking of foodborne clusters and outbreaks. While PFGE has a long history of utility in enteric disease surveillance, some limitations include low discriminatory power, throughput, and flexibility. Next generation sequencing (NGS) technologies address these limitations and now at price points accessible to public health, most public health surveillance networks, including PulseNet, have transitioned to NGS.

PulseNet began exploring the feasibility of using whole genome sequencing (WGS) in 2012. In 2013, PulseNet launched a project using WGS to detect *Listeria* outbreaks. This project, known as the *Listeria* Pilot Project, showed that WGS could lead to greater detection of clusters of illness, an ability to solve small outbreaks faster, linkages of ill patients to likely food sources and identification of new food sources. This pilot showed great promise for the utility of WGS in public health and earned CDC a lot of acclaim in the field. Thus, PulseNet went on to validate WGS workflows for other PulseNet organisms, including *Salmonella*, *Campylobacter*, Shiga toxin-producing *E. coli* (STEC) and others.

In July 2019, PulseNet fully transitioned all enteric bacterial surveillance from PFGE to WGS. This included transitioning the PulseNet software, BioNumerics, from solely analyzing PFGE data to including WGS data. Keeping pace with the ever-growing volume of WGS data uploaded to BioNumerics and the PulseNet national databases presented new challenges that required customized solutions in BioNumerics to adjust to the changing needs. Additionally, multiple workflows were combined in BioNumerics, including reference identification, antimicrobial resistance gene identification and whole genome MLST, which require frequent updates to databases and underlying genomics analysis tools. Coupled with bMx's decision to retire BioNumerics at the end of 2024, these limitations prompted the need to develop a new bioinformatics and data management infrastructure: PulseNet 2.0.

Current PulseNet System

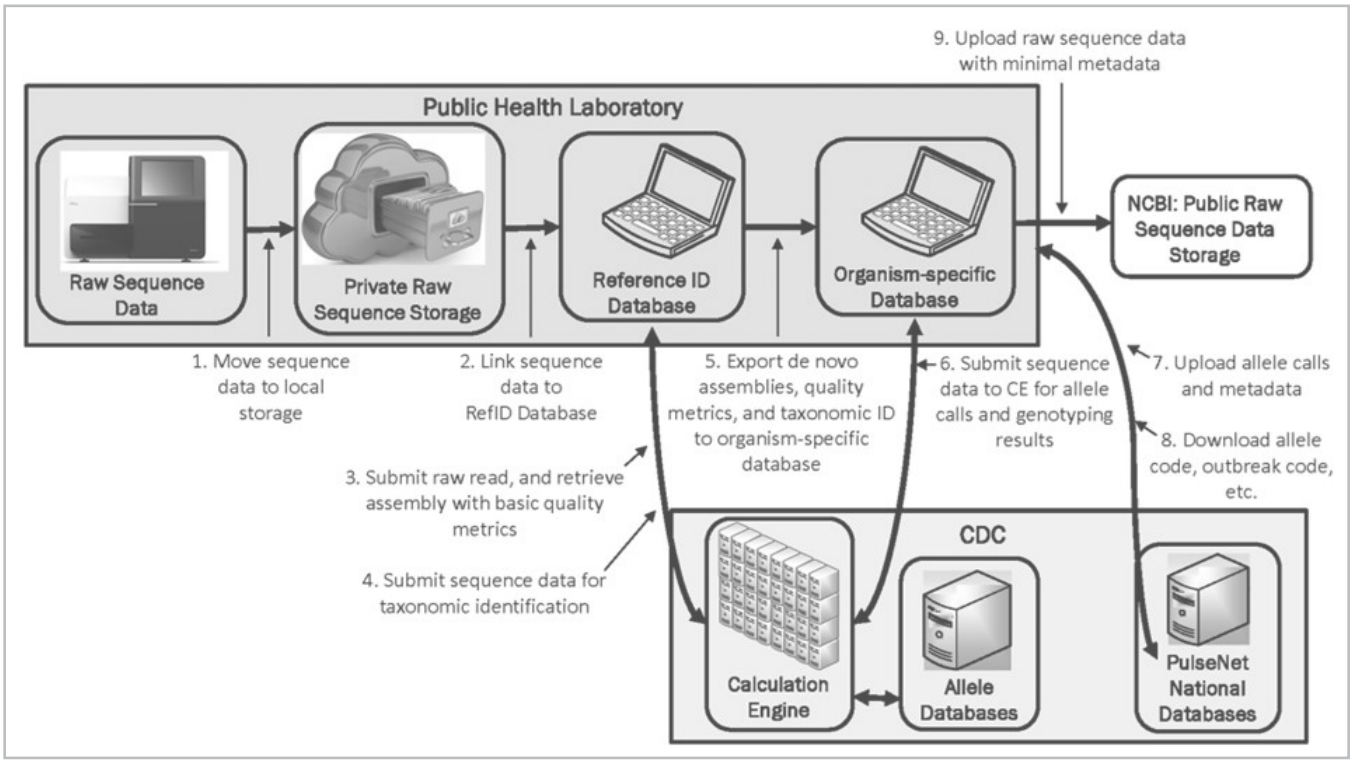
This section outlines the components of the current PulseNet system, used from its establishment in 1996 through BioNumerics' retirement in 2024.

Workflow

The standardized workflow begins in network laboratories where isolates are cultured and sequenced by WGS (**Figure 1**). Once wet lab procedures are complete and data are properly stored, either locally, on an external hard drive or in the cloud (e.g., Illumina BaseSpace), raw sequence data are linked to a reference identification library in a local client database. Raw reads are submitted to a high-performance computing (HCP) environment, the CDC Calculation Engine (CE), for a number of different analyses or "jobs," including de novo assembly, species identification (ANI) and basic quality assessment (e.g., read quality, coverage, contamination (genomic analytics)). The CE not only provides access to a HPC environment, but it is also highly customizable with easy integration of both custom-made and open-source code.

High quality assemblies are then exported from the CE and laboratories manually import them into organism-specific databases (OSDB) on their local client. Select metadata are added and assemblies are manually submitted to the CDC CE again for additional analyses to include quality checks, assembly-free and assembly-based allele calling and genotyping (e.g., serotyping, AR markers, virulence). High quality sequences are uploaded (metadata and allele calls) to PulseNet national databases, while raw reads are deposited into the National Center Biotechnology Information (NCBI) for public long-term storage using user-created templates. Once sequences are submitted to the PulseNet national databases, CDC uses these sequences, together with epidemiological data, to detect clusters and prevent larger, more widespread outbreaks.

Figure 1. Standardized PulseNet workflow for original system.



Software and Hardware

At the time of PulseNet's inception, there were very few software suites available that could analyze 1-D gels. From 1996–1998, PulseNet relied on software called GelCompar or Molecular Analyst for analysis. In 1998, Applied Maths, currently a subsidiary of bMx, released its first iteration of BioNumerics, the software application that has since served as the end-to-end solution for storage, management and analysis of PulseNet data. With few IT requirements (e.g., a desktop or laptop computer with at least 8 GB RAM, a 64-bit processor, Windows 7 or later, network connection speeds between 100Mbps — 1Gbps, external storage options and IT support for SQL databases) and a user-friendly graphical user interface (GUI), even low-capacity laboratories quickly came up to speed and kept up with PulseNet's evolving needs.

Since 1998, PulseNet and the developers at Applied Maths have been trusted partners and this partnership has been vital, especially in the face of more complex NGS technologies. For WGS data, although BioNumerics only stored analyzed sequence results, these results were an order of magnitude greater than what had previously been stored in the BioNumerics SQL databases (Mbps versus Kbps), which impacted BioNumeric's efficiency and overall functionality. The changes needed to address these challenges would require updating some of the underlying technology components. Considering these challenges, as well as the need for greater flexibility, modularity, and rapid deployment and CDC Data modernization initiative principles that encourage the use of open-source tools and cloud-based approaches, BioNumerics no longer fully met the needs of the PulseNet network.

Data Transfer/Sharing

After generating WGS data, it is the PulseNet member's responsibility to transfer both isolate metadata and WGS data to the proper locations. With BioNumerics, this can mean manually linking an Excel spreadsheet containing metadata to WGS data stored locally or in the cloud for import into the reference identification (refID) database. This process can be time-consuming, but some laboratories have linked their LIMS system to BioNumerics, thus, automating the secure transfer of select metadata. Of note, for PFGE, laboratories were sharing up to ~500KB per isolate whereas for WGS, data submissions range from up to ~500 MB for the CE to up to ~90MB for national databases. Despite the use of protected health information (PHI) in PulseNet rather than personally identifiable information (PII), many states have IT policies in place preventing connecting their LIMS (which may contain PII) to BioNumerics. In the future of PulseNet section below, details about data security and privacy in the cloud are presented to help address these concerns.

Aside from getting isolate metadata into BioNumerics, PulseNet members also manually submit their WGS data to CDC's CE for analysis and retrieve the analyzed results. PulseNet members submit their WGS data to external data repositories including NCBI Sequence Read Archive (SRA) and submit analyzed sequencing results to the PulseNet national database. These manual, often multi-step processes make data transfer in BioNumerics a time-consuming and labor-intensive task requiring multiple return trips to a desktop/laptop computer to complete. Automation of the data sharing and transfer steps could be a secure and time-saving alternative to the current solution, streamlining the process for a faster and more efficient workflow.

Storage

Currently, the options for WGS data and isolate metadata storage are varied (e.g., cloud storage, local files, external hard drives, a variety of LIMS systems). PulseNet members will still retain the ability to store data in various sites, though cloud storage will be preferred and encouraged. Due to the size of sequencing files, we understand most laboratories do not/will not have the capability to store these files long-term, hence we encourage PulseNet members to upload raw sequencing data to NCBI for long-term storage. Data storage also depends on how the WGS data is used, since CLIA-validated workflows may require two years of on premises data storage depending on local CLIA requirements. While storage of sequence and limited metadata to NCBI is a requirement in PulseNet's Memorandum of Understanding (MOU) and Terms of Reference (TOR), it also increases transparency and encourages collaboration as it allows non-PulseNet participants to use the data for local surveillance or other research projects.

Genomics Analysis Capabilities

Much like the rest of the process, data analysis follows a standardized workflow, which is largely composed of open-source tools that perform different "jobs" or tasks (see Table 1). "Jobs" are selected by the PulseNet member who manually submits data to the CE in both refID and OSDB. The "jobs" submitted in the refID database include de novo assembly, speciation (ANI) and quality assessment. Those jobs submitted in OSDB are mainly for genotyping the isolate of interest (e.g., tools for serotyping and identifying AMR markers, virulence markers, plasmids, etc.) and allele calling. Of note, quality is also assessed for allele calls in OSDB.

Data Visualization and Reporting

Currently, data visualization capabilities are limited, consisting of bar graphs, basic charts and statistics, and phylogenetic trees (e.g., dendrograms and minimum spanning trees). Although more modern and sophisticated data visualization tools are available (and should be considered for PulseNet 2.0), the current tools in BioNumerics are sufficient to detect clusters both locally and nationally. Of note, the reporting features are also limited. PulseNet members often report the need to export data to Excel spreadsheets to create customized reports specific to the laboratory and epidemiologists' needs.

Security and Access

Security and access controls are a key component of data management in BioNumerics. While data are maintained and secure in local client databases, there are many instances where data are in-flight (data transfer/sharing). Any time PulseNet participants need to access CDC resources (e.g., CE and PulseNet national databases), they are required to authenticate to CDC's PulseNet using a SecurID key fob issued by CDC. This is also true when participants want to access data from BaseSpace—they are required to authenticate to Illumina's cloud. This added layer of security grants only certified participants with access to analysis tools on the CE, while protecting the privacy of these data during transit.

The Future of PulseNet: PulseNet 2.0

There are many lessons learned from the COVID-19 pandemic, specifically related to genomic surveillance which has been central to the federal, state and local response. Routine use of WGS has not only led to the rapid detection of viral variants, but also helped predict infection and hospital spikes within the community. As a coordinated response requires the support of a robust data architecture, modernization efforts at all levels of government have become a priority. State IT departments have had to reckon with the need for quick, high-quality data, while adhering to jurisdictional requirements and policies for data privacy and security (discussed below). The high-resolution and discriminatory power of WGS made it the method of choice for enteric diseases genomic surveillance, but the computing resources necessary to store, analyze and report these BIG datasets required state and local officials to shift their thinking from traditional methods and consider more modern approaches to data. Below is a model of how data could be handled in the minimum viable product (MVP) for PulseNet 2.0.

Of note, the full operational capability (FOC) will include additional functions and features, based on the feedback received from states, and will be available for full implementation by Fall 2024. Although this document will focus on the MVP, a future version of this document is anticipated by early 2024 and will include further details about the FOC. Additionally, there will be a transition period between the legacy on premise software, BioNumerics, and the cloud version of PulseNet 2.0. This will require coordination between the CDC, public health laboratory (PHL) directors, PulseNet staff and PHL IT departments to transition to the cloud. CDC and APHL have plans to hold informational sessions/webinars for laboratory staff and PHL IT staff to discuss technical specifications for PulseNet 2.0 later this year. As further details are available, they will be shared in future document updates.

The New Model: Cloud-based, Modular, Open-source

End-to-end operations are possible in the cloud—from storage to analysis to data visualization and reporting, the cloud represents a future-proof way forward. The preferred cloud platform for CDC is Microsoft Azure, thus the initial PulseNet 2.0 platform will be hosted in a Microsoft Azure environment. Additionally, modular functionality, such as NextFlow-based bioinformatics workflows open application programming interface (API) standards for data transfer, are planned to be cloud agnostic for future reuse across different platforms. The three main conceptual components of the PulseNet 2.0 platform are detailed below:

Data Storage

Currently, PulseNet data are stored in various locations. These include SQL or SQLite databases to store isolate metadata and analyzed sequencing results and either local servers or cloud-based resources to store the raw sequence data. In PulseNet 2.0, the WGS data and analyzed results will be stored together in the cloud and not require the push and pull of data from states to CDC to states and back again to CDC. This will cut down on network usage at PulseNet member laboratories and make submission of data to NCBI easier.

Data Compute and Analysis

The bioinformatics backend for PulseNet will continue to rely on open-source tools with stringent version control. Generally, these analytic tools are used to perform one step in a multi-step process and workflow managers, like Nextflow Tower, allow PulseNet members to package (or string together) singular tools into reproducible analytic pipelines. As it is likely that each step may need different software with different dependencies or resource requirements, workflow managers are ultimately responsible for pushing these pipelines to containers with the appropriate software for completing the task/"job." Each process will utilize containerized open-source bioinformatics tools and customized scripts optimized for PulseNet 2.0.

Container technology allows bioinformatics packages/applications to be developed, packaged with all necessary dependencies and configurations, and deployed reliably. With modularity and flexibility in mind, features like containers will allow PulseNet to expand or retract the infrastructure in real-time to meet the evolving needs of the network. PulseNet is currently exploring open-source container platforms and orchestration tools for the management, maintenance and orchestration of the containers. These solutions include modern tools like Docker/Singularity, Nextflow and Nextflow Tower. For the MVP, PulseNet 2.0 will make extensive use of containers for StaPH-B (The State Public Health Bioinformatics Group)-maintained open-source bioinformatics tools. Because each process has distinct dependencies and specifications, containers will be modified as needed. New containers will be created that did not exist in the StaPH-B, such as those for the contamination process (MIDAS). During FOC all containers will undergo version control and optimization.

The containerized bioinformatics workflows within the PulseNet 2.0 MVP and FOC will be initiated and orchestrated by an API call and executed in a High-Performance Compute (HPC) environment employing Azure Batch. Azure Batch creates and manages a pool of compute nodes (virtual machines), loads the containerized workflows and schedules jobs to run on the nodes in parallel. The execution progress is monitored and reported back to the client side via web service and the resulting analyses are then output and stored in the database.

Data Transfer and Sharing

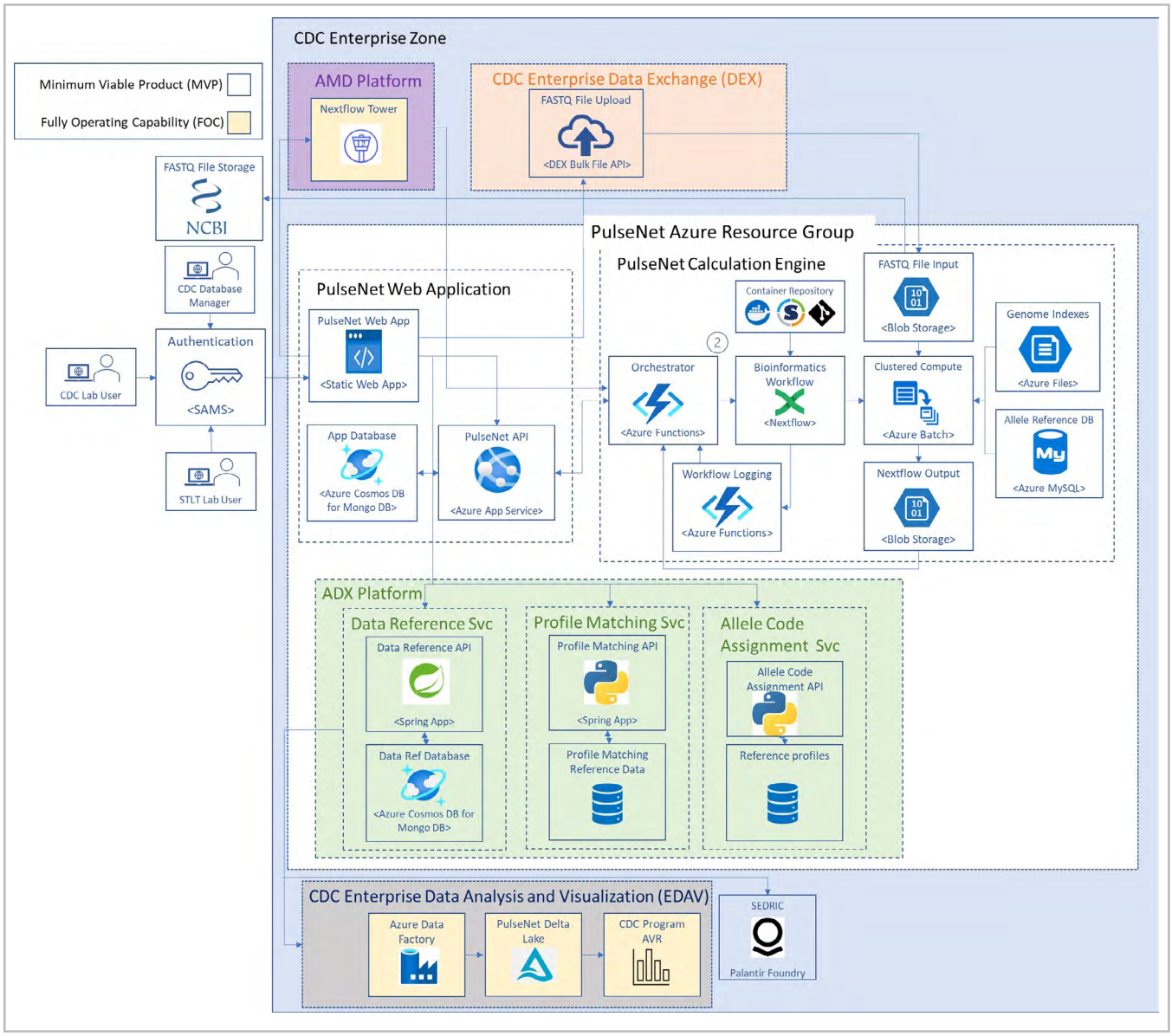
Centralizing the management of data in the cloud will help end PulseNet members make the most effective use of their data. A data lake structure offers the flexibility and functionality PulseNet members want for PulseNet 2.0. Further, a unified front-end with an API web services hub, for all users, will streamline upload and management of isolate sequencing data and metadata. In the future, this web services hub can be extended to allow for data connections to additional systems. Direct data lake connections to additional Azure services like Databricks can allow for supplemental downstream enterprise data processing and analytics.

Considerations and Requirements

Front-end Features

Current plans have PulseNet members logging into PulseNet 2.0 through a web application. Users will no longer need to maintain their own database software or proprietary licenses to interface with the PulseNet program. All data and software components of PulseNet will now be maintained on the CDC hosted Azure cloud. External users will authenticate into the PulseNet 2.0 application via CDC's Secure Access Management Services (SAMS), a standard authentication protocol for CDC partners. For a user to authenticate via SAMS, a SAMS profile and permissions to access the application are required. SAMS accounts will be approved by PulseNet and a participant's access will be granted when they get analysis certified. Upon authentication, authorized members may then come to a fully customizable landing page or client GUI. From the GUI, members may have the ability to upload data, interact with recent isolate information, set up alerts, view refID and OSDB (state and national), submit and view running jobs, etc. The front end may also be connected to different data visualization applications and the PulseNet API may enable members to submit data to the calculation engine for analysis and publish data to the national database and NCBI (**Figure 2**).

Figure 2. Proposed system architecture for the minimum viable product.



Data Transfer/Sharing

Jurisdictions may be able to upload sequence data files (FASTQ) and isolate metadata to the PulseNet 2.0 infrastructure through a variety of HTTPS transport mechanisms (e.g., BaseSpace API, AWS S3 buckets, MinION services, CSV upload, eLIMS API, etc.). PulseNet 2.0 may allow users to upload data to the SRA NCBI through the API and provide SRA and isolate metadata through APIs to other CDC systems. Most jurisdictions have requirements for the protection of jurisdictional data. In accordance with the Privacy Act of 1974, PulseNet 2.0 will only maintain data about an individual as is relevant and necessary to accomplish the purpose of PulseNet. Regarding the Health Insurance Portability and Accountability Act of 1996 (HIPAA), CDC is considered a public health authority and therefore is not required to comply with HIPAA. CDC protects individually identifiable health information pursuant to other federal laws (e.g., FOIA and the Privacy Act), but NOT pursuant to HIPAA.

From a data security and encryption standpoint, all data ingested will be secured using SAMS authorization and roles. All APIs will require Oauth authenticated tokens to be present in the request and will only allow information to be accessed based on the role and location of the user. Please see Security & Access section below for more details.

HTTPS and TLS will be used at data transport levels to move encrypted data. Data at rest stored in Azure Storage Services will be encrypted using Advanced Encryption Standard (AES), additionally data at rest is encrypted by default using Azure CosmosDB.

Metadata

The current system only ingests, stores and manages protected health information (PHI). For PulseNet 2.0, PHI such as sample metadata and experiment character data submitted to PulseNet national databases will be stored permanently. Data not submitted will be retained indefinitely at first but may have a retention policy added in the future to manage cloud storage costs. We will continue to capture and use PHI for PulseNet 2.0 and are currently exploring the feasibility of including more health-equity related variables, namely race and ethnicity.

Sequence Data

Sequence read and assembly files will be stored only as needed for bioinformatic analysis and will have retention policies in place (not yet determined). Although sequence data generated from an Illumina instrument is the only data validated for current PulseNet workflows, CDC is currently exploring the use of other NGS technology, including long read sequencers like Oxford Nanopore Technology instruments for PulseNet 2.0.

Storage

Data entered, uploaded to, and captured by the application will be stored in a hybrid of relational database (SQL) and document database (NoSQL, such as CosmosDB for MongoDB) to provide data queried as fast as possible for users. With a need to support the processing of 40+ TB of data per year, plus the increase due to the inclusion of metagenomics data, it may be necessary to implement a data lakehouse (e.g., Databricks, Cloudera, etc.) to improve upon current PulseNet data storage practices—a proposed feature of the FOC. Data lakehouses merge the data structure and management benefits of a data warehouse with the open ability to store structured and unstructured data like a data lake. Flexibility is afforded when using a data lakehouse because they can be implemented on any cloud environment due to consistency in management, security and governance. Data lakehouses support atomicity, consistency, isolation and durability (ACID) transactions allowing multiple users to read and write data concurrently. Data governance and auditing capability is afforded by data lakehouse platforms. Due to the ability to decouple storage and compute, data lakehouses can support more simultaneous users and larger volumes of data. The largest drawback of data lakehouses is that some jurisdictions may not be able to interact with an environment that centralizes data from multiple jurisdictions.

In terms of data ownership, from a technical perspective, the CDC will own the data infrastructure and stewardship of all data captured by the application and stored in the database. However, from a usage/accessibility perspective, the data will be maintained by PHLs and only data shared by states will be made available in the national databases. Unshared data can be updated/deleted as necessary by the individual user/laboratory.

Genomics Analysis Capabilities

Containerized analytical pipelines will replace the current calculation engine (see Data Compute and Analysis above). Unlike BioNumerics, where end users manually select "jobs" for data analysis, the common jobs will automatically be selected, and users may deselect them manually if needed. On the backend, a workflow manager will automatically orchestrate the appropriate containers into analytical pipelines to complete a task/job. Use of a workflow manager allows for version control of containerized resources and parallel processing with multiple analytical pipelines establishing ensemble-based computation.

As users have become familiar with GUIs over time, we intend to maintain a user-friendly interface to help members navigate to the standardized PulseNet analysis workflow, but we also envision users may develop and validate pipelines that will be accessed through the new platform—capabilities that may be a part of the FOC. In addition, PulseNet 2.0 will be part of the future OAMD Genomic Epidemiology platform which may contain a sandbox of popular bioinformatics and data visualization tools/pipelines for ad-hoc/laboratory-specific analyses.

Data Visualization and Reporting

As data visualizations and reporting are quite limited in BioNumerics, PulseNet plans to access these capabilities through API calls to proprietary software like Power BI, Palantir Foundry (SEDRIC), Tableau, etc., and/or custom open-source web applications like GrapeTree, Nextstrain and MicrobeTrace. For reporting, end users may be able to select different data visualization and reporting platforms as well as navigate to/around customizable state and national dashboards.

Security and Access

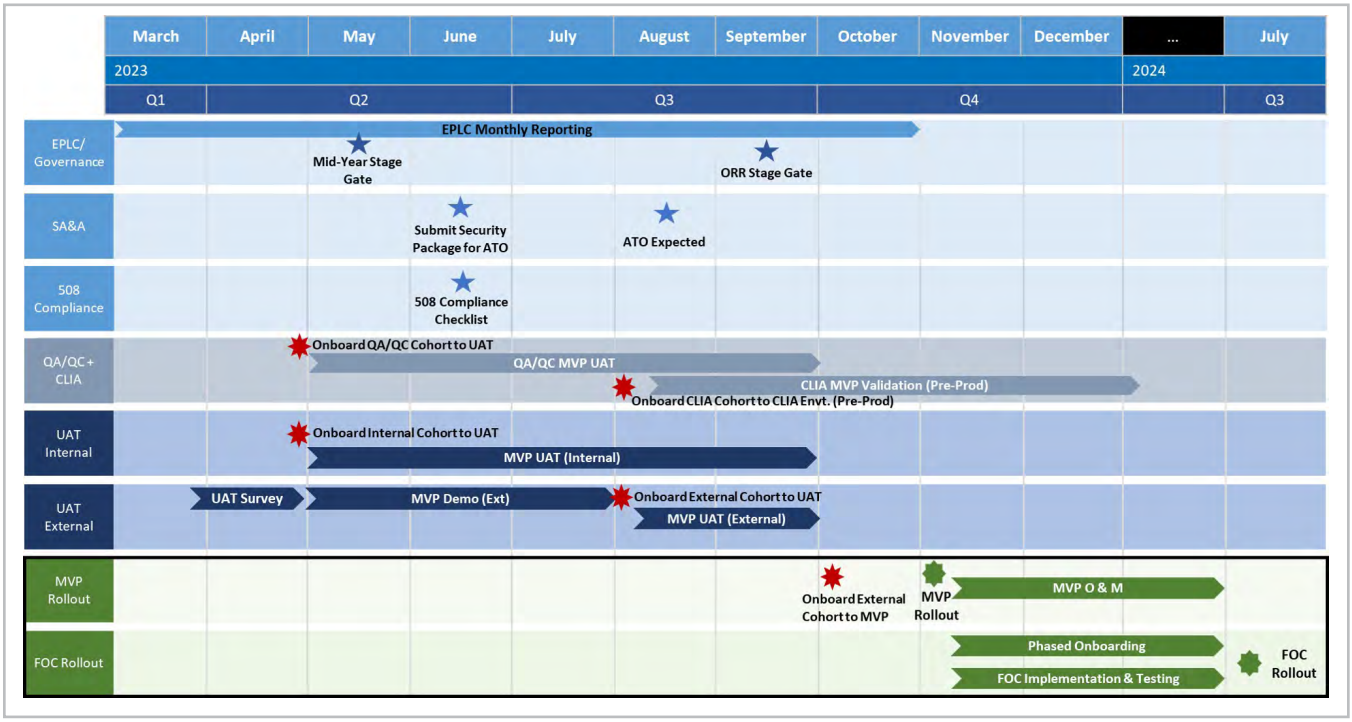
PulseNet 2.0 will be hosted on CDC Cloud Enterprise in Azure. As a cloud service, PulseNet 2.0 is subject to full security assessment, authorization and continuous monitoring under the Federal Risk and Authorization Management Program (FedRAMP). Currently, PulseNet is secured for confidentiality and integrity at a low level, and availability at a moderate level based on the requirements of the Federal Information Security Management Act (FISMA). As such, data is protected from unauthorized access, use, disclosure, duplication, modification, diversion or destruction—whether accidental or intentional—to maintain confidentiality, integrity and availability (CIA Triad). The security and privacy controls that provide this protection meet minimum federal requirements with additional risk-based and business-driven control implementation achieved through a defense-in-depth security structure. Data will be encrypted in transit and at rest following the National Institute of Standards and Technology's (NIST) Federal Information Processing Standard (FIPS 140-2) for Security Requirements for Cryptographic Modules. FIPS 140-2 specifies the security requirements that will be satisfied by a cryptographic module 2 are accepted by the federal agencies for the protection of sensitive information. Per FedRAMP requirements, PulseNet 2.0 employs more than 300 security controls (SCs). More information on the required FedRAMP SCs for systems deemed low can be found here. In addition to the extensive SCs in place, there are access control measures (role-based access) that jurisdictions can implement to further protect the data.

Future Planning for PulseNet 2.0

Timeline

In September 2022, contracts were awarded to Booz Allen Hamilton (BAH) for the development and implementation of the PulseNet 2.0 platform and bioMerieux (bMx) to help transition some proprietary functions of BioNumerics to open-source tools. Although the timeline is ambitious, we remain on track and anticipate that the MVP will be available by September 2023. This initial product will be tested internally before piloting in select laboratories. Once the MVP is implemented, BAH will work with those piloting to resolve performance issues and prepare for the transition and rollout of the FOC, which is projected by May 2024. **Figure 3** provides a graphic of this timeline.

Figure 3. Development and Operations cycle for PulseNet 2.0



Due to the short timeline, BAH is following an Agile DevSecOps methodology using 10-week increment planning, with each increment consisting of five 2-week sprints. During the first 2-month phase of the project, Planning and Investment Decisions, BAH gathered, documented, analyzed and validated business and system requirements, recommending potential improvements to existing processes. The Booz Allen team also performed system, conceptual design studies and integration requirements analysis. This increment was completed in November 2022 and BAH provided a clear way forward to the MVP and the transition to FOC.

The second phase, System/Application/Database Development, began in December 2022 and will end in September 2023 with the MVP as its main deliverable. The third phase, Agile Deployment Implementation: MVP & Rollout to FOC, is set to start in September 2023 with the implementation of the MVP and finish with the rollout of the full platform. Throughout this phase, the platform will be continuously developed and maintained as PulseNet members onboard and issues arise. These issues will be addressed with updates and enhancements. In the fourth phase, Operations and Maintenance, beginning in September 2024, BAH will provide "lights on" performance, availability, reliability and security standards for PulseNet. They will help maintain the platform and provide technical support as they correct defects and make small/minor changes and enhancements to the system. Along with these four phases, a constant fifth phase, Management Reporting, will run the length of project implementation (September 26, 2022 – March 25, 2027). The purpose is to provide overall direction, coordination, implementation, execution, control and completion of the project.

BAH has been meeting with the PulseNet 2.0 team to provide progress updates which include roadmaps, key decisions, meeting minutes, assumptions, technical and functional specs, etc. We debuted the most recent wireframes for the GUI PulseNet members at 2023 regional InFORM meetings and will continue to update as we get more information.

User Group On-boarding Plan

The initial MVP product will be tested internally via User Acceptance Testing (UAT) before piloting to a group of select laboratories. The below section describes both the UAT and MVP pilot processes.

User Acceptance Testing

User acceptance testing (UAT) is a phase of software development in which the software is tested in the real world by its intended audience. The goal of UAT is to ensure the application can handle real-world tasks and perform up to development specifications.

During UAT, participants will be given the opportunity to interact with the MVP application before its official release to production to test features and identify any potential bugs. The results from UAT will be recorded and forwarded to the developers, who will then make final changes before releasing the application.

A full User Acceptance Plan with clear instructions and testing checklist will be provided to participants prior to UAT kickoff. A general outline of the UAT process is outlined below:

1. **UAT Plan:** The business requirements, time frame and strategies for UAT are outlined.
2. **UAT Test Scenarios:** These test scenarios should cover as many functional cases as possible that end users may face.
3. **Test and Document:** The end users begin testing the software, logging any potential bugs or other issues. All bugs should be recorded in a bug tracker with notes on how to reproduce the errors.
4. **Update code, retest and sign off:** The BAH development team adjusts the code based on test results—resolving any bugs or making suggested changes—and then retests. Once the software meets the users' criteria, the tester signs off on the changes.

MVP Pilot

The BAH team will create a survey to be distributed to PulseNet laboratories to gauge interest in participating in the MVP pilot. The survey was sent to laboratories in March 2023. The intended purpose of the survey is to determine interest amongst laboratories in participating in the pilot as well as gain a better understanding of different skillsets, user needs, and current use of BioNumerics software (e.g., BaseSpace, Command Line Interface (CLI) tools, local file uploads, etc.) to ensure the pilot group captures a wide range of skills for more robust use of the MVP application.

Based on results and responses from the pilot interest survey a group of 5–10 laboratories will be selected to participate in the MVP pilot. The MVP pilot is projected to begin following the release of the MVP application in November 2023 and will span approximately four to six months. The purpose of the MVP pilot is to test features of the MVP solution in a production environment with production data, resolve performance issues and prepare for the transition and rollout of the FOC. Official state user onboarding will take place incrementally following the MVP pilot prior to the release of the FOC; a detailed plan and timeline for when laboratories will onboard to the new application is expected in the coming months.

Cloud Use

Although ~80% of PulseNet laboratories report using the cloud, we recognize some jurisdictions may have policies restricting them from sharing data within a centralized system like PulseNet 2.0. In these use cases, cross-cloud capability may be leveraged via the future OAMD Genomics Epidemiology platform, wherein a mesh for data portability could allow jurisdictions to maintain control of data within their local environments and use the computational capabilities of PulseNet 2.0. Similarly, in the future, the PulseNet 2.0 web services hub could offer direct access-approved API based compute triggers for the underlying PulseNet 2.0 calculation engine. For further information regarding data and system security, refer to the Security and Access section above. We are also exploring how the former laboratories can gain access to resources provided by PulseNet 2.0.

Sequence Data

While only data generated from Illumina instruments is validated in the PulseNet workflow, many members are interested in using other instruments to generate sequence data. Ingestion of these different sequence file types into the cloud may be virtually seamless, but we will need to confirm that isolates sequenced from different technologies generate comparable analyzed results in our bioinformatics workflows. This work is ongoing.

Metadata

While PulseNet is interested in including additional data elements in PulseNet 2.0, particularly those related to health equity, we want to understand the challenges that members might face with obtaining and sharing this type of data. We are currently consulting with data security and privacy experts at the agency to understand how the addition of health-equity variables, particularly race and ethnicity, could trigger changes in current security and privacy controls before modifying these elements. Data security and privacy remain central to modernization efforts and will continue to be key in the PulseNet 2.0 development process.

What Laboratories Can Do Now

Create Database Archives

With PulseNet 2.0, there will no longer be an on-premises SQL database for PulseNet members to maintain and participants will need to consider archiving those databases. Additionally, there will be less flexibility for PulseNet member-specific fields and if data is stored in BioNumerics that is not submitted to PulseNet national databases, members will need to come up with alternative solutions for that data storage. PulseNet 2.0 will only contain data and fields that are in PulseNet national databases, the historical data in PulseNet currently will be used to populate the new PulseNet 2.0 databases.

Communicate with IT Departments

Use this white paper as a resource to initiate and/or facilitate conversations with PHL IT departments about PulseNet 2.0 infrastructure, timelines and future plans/needs.

Appendix A: Roadmap to PulseNet 2.0 MVP

Figure 4a-c provides a formal project schedule with a high-level overview of the MVP roadmap and planned activities for the five 10-week planning increments that span the life of this project from kickoff to rollout. Each increment is further broken down to provide a more granular overview of the activities with prioritized features and user stories as well as a completion timelines, with anticipated dates for future activities.

Figure 4a: Integrated Master Schedule/Roadmap to MVP—Kickoff, Management, Planning and Investment Decisions & Increment 1

IMS Version: 2.0		PROGRESS	START	END	DEPEND-ENCIES	DAYS	Milestone Completed Date
ID	**Kickoff, Management, Planning & Invesment Decisions**		**9/26/22**	**11/28/22**		64	
3	Project Kickoff Meeting	100%	9/26/22	10/4/22		9	10/4/2022
4	Completed Project Initiation Form & Business Needs Documentation	100%	10/3/22	10/24/22		22	10/24/2022
5	Conduct EPLC Project Initiation Meeting & Submit EPLC Documentation ◇1	100%	9/30/22	11/15/22		47	11/15/2022
6	Request and review Legacy BioNumerics user profile demos	100%	10/11/22	10/27/22		17	10/27/2022
7	Review requirements, system designs, architectural challenges & current processes	100%	10/3/22	10/31/22		29	
8	Obtain access to applicable legacy systems, tools, documentation, repositories	100%	10/4/22	10/31/22		28	
9	Initial application wireframing and review of wireframe mockups	100%	10/28/22	1/25/23	4, 5, 6	90	1/25/2023
10	Discussions with AMD, OCIO, DMI, CSELS on coordinating efforts with NF Tower & DEX	90%	10/21/22	5/2/23		194	
11	Initiation ITDG Investment Review / Governance Engagement	100%	10/1/22	10/26/22		26	10/26/2022
12	Completed Baseline/Preliminary Design Review Stage Gage Documentation	100%	11/18/22	11/30/22	7	13	11/18/2022
13	Conduct Baseline/Preliminary Design Review Stage Gage ◇1 ◇2	100%	12/9/22	12/9/22	14	1	12/9/2022
14	Completed Task 1 Documentation Deliverables & Established MVP Roadmap	100%	10/3/22	11/30/22	5-13	59	12/2/2022
15	Increment Planning Sessions, TFS Product Backlog Buildout/User Story Mapping- **Ongoing**	45%	11/7/22	9/25/23		323	
16	ITDG Architecture Review Team (ART) Meeting	100%	11/7/22	12/14/22	11	38	12/14/2022
17	ITDG Architecture Review Team (ART) Update Meeting	30%	1/5/23	5/4/23	16	120	
18	Obtain ATD for PulseNet 2.0	100%	2/15/23	4/13/23		58	
19	Complete Mid-Year Stage Gate Documentation	0%	3/8/23	4/25/23	13	49	
20	Mid-Year Stage Gate Meeting ◇1 ◇2 ◇3	0%	4/25/23	5/11/23	19	17	
21	Submit and Obtain ATO for PulseNet 2.0	0%	6/15/23	8/15/23		62	
22	508 Compliance - Checklist	0%	5/1/23	9/15/23			
23	Operational Readiness Stage Gate Meeting ◇1 ◇2 ◇3 ◇4	0%	8/1/23	9/15/23	20	46	
24	508 Testing	0%	9/11/23	9/25/23			
	Task 2: Increment 1		**12/5/22**	**2/10/23**		68	
26	QC Workflow for Salmonella, Ecoli, Campy, Listeria, Vibrio	98%	12/5/22	4/24/23		141	
27	Configure Local Development Environment	100%	12/5/22	2/10/23		68	
28	Users should be able to select/deselect all items	100%	12/5/22	2/10/23		68	
29	Users should be able to customize their query	100%	12/5/22	2/10/23		68	
30	Search database based on database fields	100%	12/5/22	2/10/23		68	
31	User can see number of results in the selected View	100%	12/5/22	2/10/23		68	
32	User can select entries and initiate an action on them from the toolbar.	100%	12/5/22	2/10/23		68	
33	Users can create custom database views, export current database views, and open database viewer with an existing view loaded	65%	12/5/22	4/24/23		141	
34	Users can perform inline edits of data fields.	72%	12/5/22	4/24/23		141	
35	Users can query their current view	100%	1/16/23	1/27/23		12	
36	User can open, edit, and create isolate record in form view	100%	12/5/22	12/16/22		12	
37	Users can access all database fields	60%	12/19/22	12/30/22		12	
38	Users can customize their metadata views on landing page	100%	1/30/23	2/10/23		12	
39	Users can editing a current view	60%	1/2/23	1/13/23		12	
40	Users can navigate to all sections of the app using the top and side navigation bars	100%	1/30/23	2/10/23		12	
41	Users can view their user profile type on top bar	100%	1/30/23	2/10/23		12	

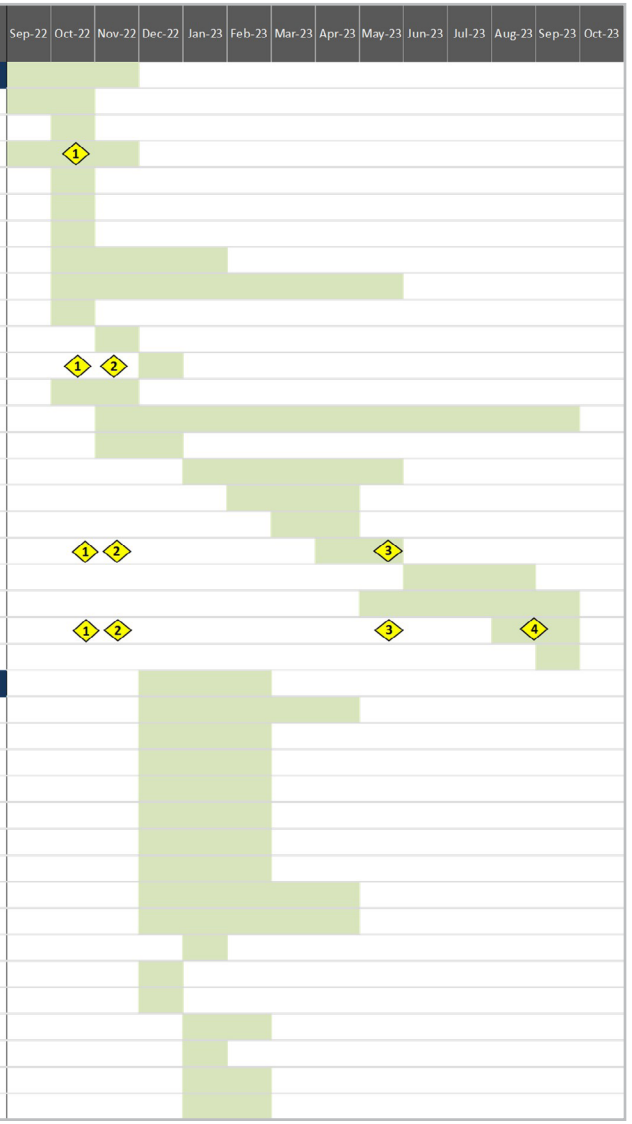


Figure 4b: Integrated Master Schedule/Roadmap to MVP—Increment 2 & 3

	IMS Version: 2.0	PROGRESS	START	END	DEPEND-ENCIES	DAYS	Milestone Completed Date
	Task 2: Increment 2		**2/13/23**	**4/21/23**		68	
43	Dev Azure environments established (1)	86%	3/27/23	4/7/23	18	12	
44	IRDG, ITDG, ART approvals for ADX integration	25%	4/6/23	5/11/23		36	
45	Submit AHB CRs for ADX Components	5%	4/11/23	5/11/23		31	
46	Allele code assignment for Salmonella	63%	3/3/23	4/7/23		36	
47	Comparisons feature complete- users has ability to compare, filter, save, edit, and copy experiments comparisons	54%	3/13/23	4/21/23		40	
48	Allele Calling for Salmonella	75%	2/13/23	3/21/23		37	
49	ANI Workflow for Salmonella, Ecoli, Listeria, Campy, Vibrio, Cronobacter (RGDv3 provided by program)	67%	3/27/23	4/21/23		26	
50	Implement Python Wrapper and Nextflow Workflow for Salmonella isPCR and Serotyping	95%	2/27/23	4/21/23		54	
51	Deploy Nextflow Runner Azure Functions Framework	100%	3/24/23	4/7/23		15	
52	Deploy MVP to Dev (1) (2)	80%	4/10/23	4/21/23	43	12	
53	Make bioinformatics code available for scicomp team to start testing/optimizing	100%	3/13/23	3/24/23		12	
54	Establish SAMS Profile for PulseNet	90%	3/27/23	4/21/23		26	
55	Users can import metadata from a csv file into the database	75%	3/13/23	3/24/23		12	
56	Users can create nwk files for MS tree	80%	3/27/23	4/21/23		26	
57	Users can import raw reads through local file server	90%	4/10/23	4/21/23		12	
58	API Logging in Azure with Application Insights	100%	4/10/23	4/21/23		12	
51	OAuth Flow	95%	4/10/23	4/21/23		12	
	Task 2: Increment 3		**4/24/23**	**6/30/23**		68	
61	UAT Environment established	10%	4/24/23	6/30/23		68	
62	Users can see the results of QA/QC for raw sequences	10%	3/27/23	4/20/23		25	
63	Add an option to 'Select All' fields instead of manually checking each one when creating a new view	0%	4/24/23	6/30/23		68	
64	Add identifiers to the metadata form to indicate which fields are required	0%	4/24/23	6/30/23		68	
65	Allow for entries to be entered manually, typed, into the source state field	0%	4/24/23	6/30/23		68	
66	Create new views based on already existing custom view	0%	4/24/23	6/30/23		68	
67	Data format in grid view needs to be in YYYY-MM-DD format	0%	4/24/23	6/30/23		68	
68	User can select PulseNet application link from SAMS landing page (sams.cdc.gov)	0%	4/24/23	6/30/23	54	68	
69	User can access PulseNet application via application URL	0%	4/24/23	6/30/23		68	
70	Users should have the ability to automatically submit CE jobs after data import	0%	4/24/23	6/30/23		68	
71	Users can import raw reads through local file server	0%	4/24/23	6/30/23		68	
72	Users can run an end-to-end bioinformatics NGS pipeline (QC, assembly, allele calling, genotyping)	0%	4/24/23	6/30/23		68	
73	Users can view job status window pop-up after CE submission	0%	4/24/23	6/30/23		68	
74	Users should not have to manually kick off the next job in queue; jobs dependent on a previous job output will run automatically	0%	4/24/23	6/30/23		68	
75	Users can view rejected data on the left-hand navigation panel	0%	4/24/23	6/30/23		68	
76	Users can view samples that need further review from landing page	0%	4/24/23	6/30/23		68	
77	Users can see the assembly/Genotyping results	0%	4/24/23	6/30/23		68	
78	Perform vulnerability scans	0%	4/24/23	6/30/23	21	68	
79	Schedule Fortify Scan	0%	4/24/23	6/30/23	21	68	
80	ADX Integration into PN 2.0 OCIO Azure Environment	0%	4/7/23	6/30/23	44, 45	85	
81	Internal UAT	0%	4/24/23	6/30/23	61, 79	68	
82	Users can view run status (Not Started, In-Progress, Complete, Failed/Stalled) of selected isolates and allele calls from navigation pane	0%	4/24/23	6/30/23		68	
83	Deploy and Test Data Management Service	0%	4/4/23	5/20/23		47	
84	Deploy MVP to UAT (1) (2) (3)	0%	4/24/23	6/30/23	52, 80	68	
85	National Database: Users can select samples to submit and view the National Database(s) they are certified for	0%	5/5/23	5/30/23		26	
86	Implement Nextflow Workflow for PlasmidFinder and AMRDetection wrapper provided by CDC program	0%	4/24/23	6/30/23		68	

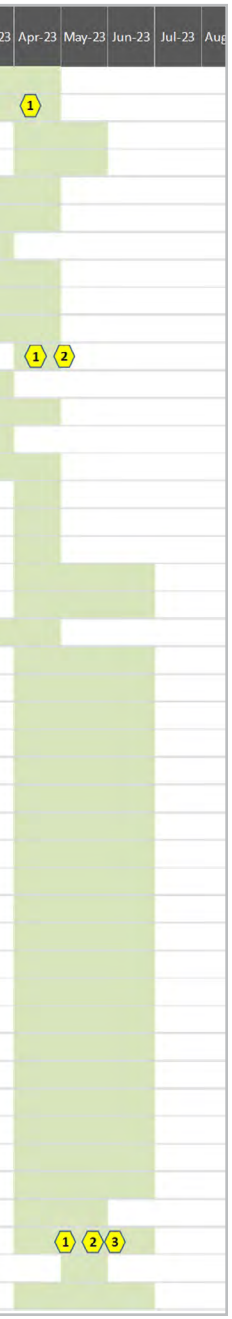
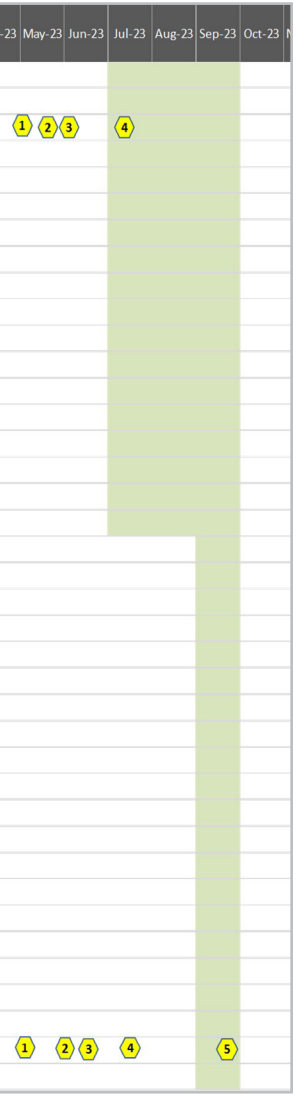


Figure 4c: Integrated Master Schedule/Roadmap to MVP—Increment 4 & 5

	IMS Version: 2.0	PROGRESS	START	END	DEPEND-ENCIES	DAYS	Milestone Completed Date
	Task 2: Increment 4		**7/3/23**	**9/8/23**			
88	Configure Production Environments: Pre-Prod and Production	0%	7/3/23	9/8/23	83	68	
89	Deploy MVP to Pre-Prod Environment (1) (2) (3) (4)	0%	7/3/23	9/8/23	84, 88	68	
90	Users can see and export allele difference matrices	0%	7/3/23	9/8/23		68	
91	Users can view test completion data at the isolate level in a pop-up window alongside the entry	0%	7/3/23	9/8/23		68	
92	Users can use cgMLST as a primary cluster detection method	0%	7/3/23	9/8/23		68	
93	Users can filter to only view core loci, all loci, or MLST for wgMLST	0%	7/3/23	9/8/23		68	
94	Users can view a static phylogenetic tree for the clusters they chose in the comparision window	0%	7/3/23	9/8/23		68	
95	Users should see that the CE submission window automatically populate with jobs that have not been run	0%	7/3/23	9/8/23		68	
96	Users can access their most recently viewed database entries from the landing page	0%	7/3/23	9/8/23		68	
97	Users have access to 'Import Data', 'Submit Jobs', 'View Recently Submitted Jobs' quick access tiles in landing page	0%	7/3/23	9/8/23		68	
98	Users can see their recent saved comparisons from landing page	0%	7/3/23	9/8/23		68	
99	Users can view notifications from top bar	0%	7/3/23	9/8/23		68	
100	Users can see raw error/log of analysis for each entry	0%	7/3/23	9/8/23		68	
101	Database Admins can add new species for new entry and customize quality setting for each pathogen	0%	7/3/23	9/8/23		68	
102	Database Admins can set default for quality setting per pathogen	0%	7/3/23	9/8/23		68	
103	External UAT	0%	7/3/23	9/8/23	61, 68, 21	68	
104	Users can export the results of test	0%	7/3/23	9/8/23		68	
	Task 2: Planning Increment 5		**9/11/23**	**9/25/23**		15	
106	Unit Tests - MVP API, MVP Web App	0%	9/11/23	9/25/23		15	
107	System Admins can create/edit shareable views	0%	9/11/23	9/25/23		15	
108	System Admins can manage roles	0%	9/11/23	9/25/23		15	
109	System Admins can manage user groups	0%	9/11/23	9/25/23		15	
110	System Admins can manage users	0%	9/11/23	9/25/23		15	
111	Users can export differences from wgMLST comparison – distance matric & png	0%	9/11/23	9/25/23		15	
112	Users can oscillate between locus number and gene name for wgMLST results	0%	9/11/23	9/25/23		15	
113	Users can import raw reads from Amazon S3 bucket	0%	9/11/23	9/25/23		15	
114	Users can import raw reads through direct BaseSpace connection	0%	9/11/23	9/25/23		15	
115	Users can import raw reads through NCBI	0%	9/11/23	9/25/23		15	
116	Users can select or create metadata/raw read templates	0%	9/11/23	9/25/23		15	
117	Users can re-submit already processed data to the CE	0%	9/11/23	9/25/23		15	
118	Users can search the database fields	0%	9/11/23	9/25/23		15	
119	Users can set filters and sort from header rows of each view	0%	9/11/23	9/25/23		15	
120	Users cannot submit samples with a failed status to the National Database	0%	9/11/23	9/25/23		15	
121	Users can import metadata from an excel file into the database	0%	9/11/23	9/25/23		15	
122	Data will be validated upon bulk upload import to ensure it follows the data dictionary format	0%	9/11/23	9/25/23		15	
123	Users will be notified upon file bulk upload failure/warning	0%	9/11/23	9/25/23		15	
124	MVP build deployed to production (1) (2) (3) (4) (5)	0%	9/11/23	9/25/23	88, 89	15	
125	Migrate legacy production data to database	0%	9/11/23	9/25/23		15	



Appendix B: PulseNet 2.0 Bioinformatics Tooling

Table 1. Open-source tools currently in use in BioNumerics.

BioNumerics Environment	Function	BioNumerics Open-Source Tool
RefID	De Novo Assembly	SPAdes
	Contamination	MIDAS
	Average Nucleotide Identity (ANI)	MUMmer
Organism-specific Database	Serotyping	Serotype Finder
		SeqSero2
	AMR	ResFinder v 4.1
	Plasmids	PlasmidFinder 2.1
	Virulence	VirulenceFinder 2.0
	Lineage/Subspecies	MUMmer ANI
	Core Genome and Whole Genome MLST (cgMLST & wgMLST)	Strain-level surveillance/typing using cgMLST & wgMLST
	Tree Building (Genome-Based)	Cluster analysis based on pairwise similarities (cgMLST & wgMLST alleles)
	Distance Matrix (Genome-Based)	Pairwise differences based on cgMLST & wgMLST alleles (Categorical Values, UPGMA)

Appendix C: Data Migration Plan

The creation of a data migration plan is in progress, and will be added to this document as it is developed.