MODERN DATA SYSTEMS TO SUPPORT PUBLIC HEALTH GENOMIC SURVEILLANCE

AMD Data Modernization Community of Practice

This document was developed by the AMD Data Modernization Community of Practice representing the collective expertise and insights of its members.

Special Acknowledgements: Florek, K., MacKellar, D., Bell, J., Baird, S., Jones, D., Boyd, L., Mitchell, M., Phung, T., Thayer, R., Haydel, D., Kampoowale, A., Azevedo, K., Ledin, K., Parker, J., Casiello, C., Doucette, M., & Tu, V.

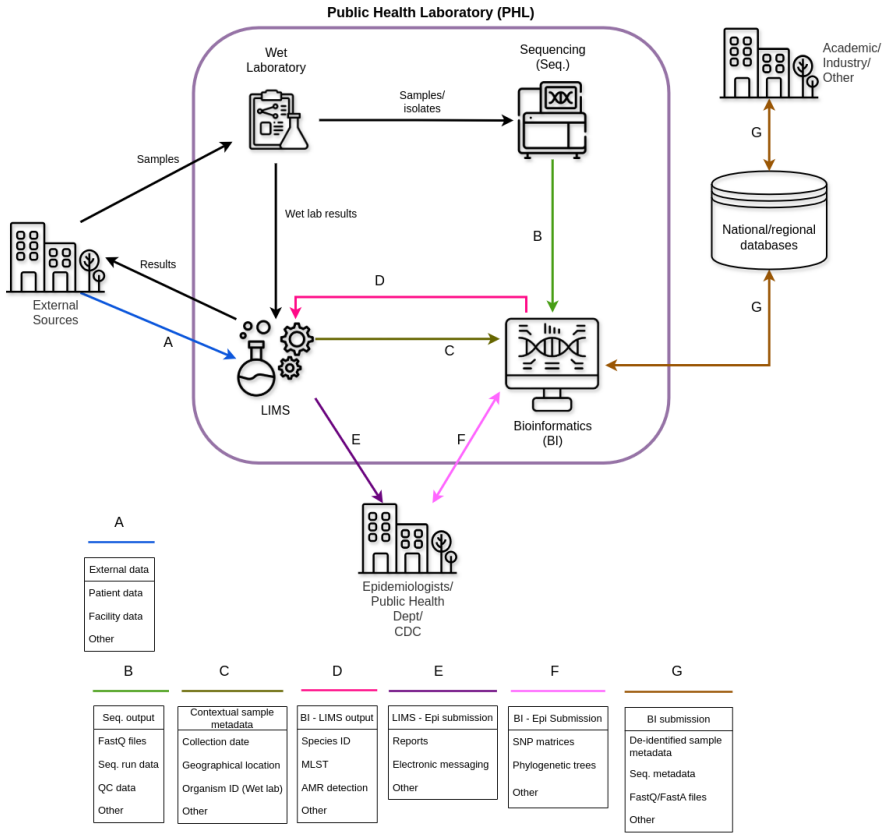
Introduction

The advent of sequencing and the rise of genomics has required the development of new tools and systems, many of them affected by more general developments in data science. Public health has its own particular requirements and history, both of which affect the successful incorporation of genomics as part of its work. In this white paper we examine legacy data systems, current usage, and modern data systems and document their challenges and benefits for public health genomic surveillance. We first discuss some of the major elements of modern data systems and trends in the computational needs of public health groups.

Modern data systems are foundational for working with massive and complex datasets, i.e., "big data". Big data originates from a variety of sources but shares a set of characteristics, commonly referred to as the 5 V's: large volume, high value, wide variety, high velocity, and high veracity. While genomic data can be considered a kind of big data, there are at least two unique challenges in working with it. First, esoteric formats are used to encode and store genomic data. Text-based file formats such as FASTA, FASTQ, SAM, and VCF are not easily incorporated into the standard databases or storage approaches commonly employed by organizations dealing with big data, especially given the size of genomic data. Second, the analytical processes used to convert raw genomic data into actionable or usable results involve computational approaches that range widely from classification problems to evolutionary modeling. The uniqueness of genomic data formats and the requirements of specialized analytics demand an architecture capable of supporting these characteristics.

> "The extreme size, novelty, and potential uses of sequencing data have created a need for PHLs to rapidly modernize infrastructure..."

To effectively support public health genomic surveillance, a modern data system must directly address the needs of bioinformaticians, data scientists, and epidemiologists. Bioinformaticians and Data Scientists are new roles in public health with unique responsibilities and needs that health departments and public health laboratories (PHLs) may not be used to supporting. Historically, PHLs have not employed scientists who combine responsibilities in software development, IT system administration,



research project management, and data analysis and management. As such, PHLs have faced barriers adjusting traditional laboratory practices and policies to meet the demands of this new and growing workforce.

Next-generation sequencing (NGS) is clearly creating a technology disruption in PHLs. The extreme size, novelty, and potential uses of sequencing data have created a need for PHLs to rapidly modernize infrastructure and laboratory systems to support this new technology and its accompanying workforce. The history of public health is rich with well-documented technological revolutions. Most recently, nucleic acid amplification and its associated molecular techniques have significantly changed the testing approaches used by PHLs and clinical laboratories. Given the breadth of application of molecular-based approaches used today and the added value sequencing provides, it is likely sequencing-based tests will continue to expand, placing new demands on laboratory data systems.

New cloud-based data systems have been a critical advancement for data-centric organizations over the last decade. Cloud-based systems leverage cutting-edge technologies to provide a scalable architecture that supports analytical data workflows and data governance while maintaining cost effectiveness and data security. These cloud systems have facilitated new developments in analytical platform design, including data warehouses that focus on structured data for analytics, data lakes that combine unstructured data and structured data, and data lakehouses that combine the two, offering flexibility across diverse data. All these platform designs collocate data and analytics and provide various approaches to support complex data needs. At their core, these systems have been born out of the need of organizations to manage a changing and complex data landscape. While the IT departments of many public health groups have resisted moving some of their organization's computational work to the cloud, the need for flexibility and scalability in both storage size and computational power suggest that access to the cloud will become only more necessary for public health departments.

Figure 1. Public Health Laboratory Flow

The flow of materials and data involved in genomic sequencing and bioinformatics within a typical public health laboratory (PHL) is complex and involves multiple systems and connections before data is reported to external partners.

Legacy data systems

With the acquisition of high throughput next-generation sequencing technologies, Public Health Laboratories employed a wide variety of approaches for the analysis and storage of genomic data. PHLs often adopted existing systems, storing data in Laboratory Information Management Systems (LIMS), Microsoft Access/Excel, or on-premises storage solutions such as Network Attached Storage (NAS) and/or Storage Area Networks (SAN). Genomic data analysis often included off-the-shelf solutions such as Bionumerics, Galaxy, and CLC Genomics Workbench. A small number of PHLs had access to university-based High Performance Computing clusters or deployed Linux workstations / servers / virtual machines running open-source community driven software. While these solutions met a critical need in PHLs, they present challenges in unifying and integrating data systems and the development of standards across public health. Additionally, despite the wide variety of solutions available, many laboratories still face challenges that include lack of capacity, lack of standardization, lack of IT or administrative support, lack of scalable solutions, lack of interoperability and interfacing, and restrictive policies and procedures. These challenges have resulted in many laboratories delaying the adoption of NGS or relying heavily on external partners or contractors to support genomic data analytics, further limiting the application of sequence-based testing in laboratories.

The broad and diverse genomic solutions employed by PHLs present a wide variety of challenges. As previously noted, bioinformatics and data science are new domains of expertise in public health; as such, there are a limited number of personnel with these skills in the public health community. Along with a limited workforce, laboratories have faced difficulties integrating these novel roles into the existing laboratory and epidemiology framework. Additionally, laboratory support systems often face challenges providing resources for new bioinformaticians and data scientists. IT departments are often ill-equipped to support the technological demands of these new roles and struggle with the unique data types and workflows that are common with genomic data. Supervisors and directors can also face challenges around the guidance of bioinformaticians and applicability of genomics in the lab. PHLs that are unprepared for these new roles may face incredible challenges in recruiting and retaining their staff.

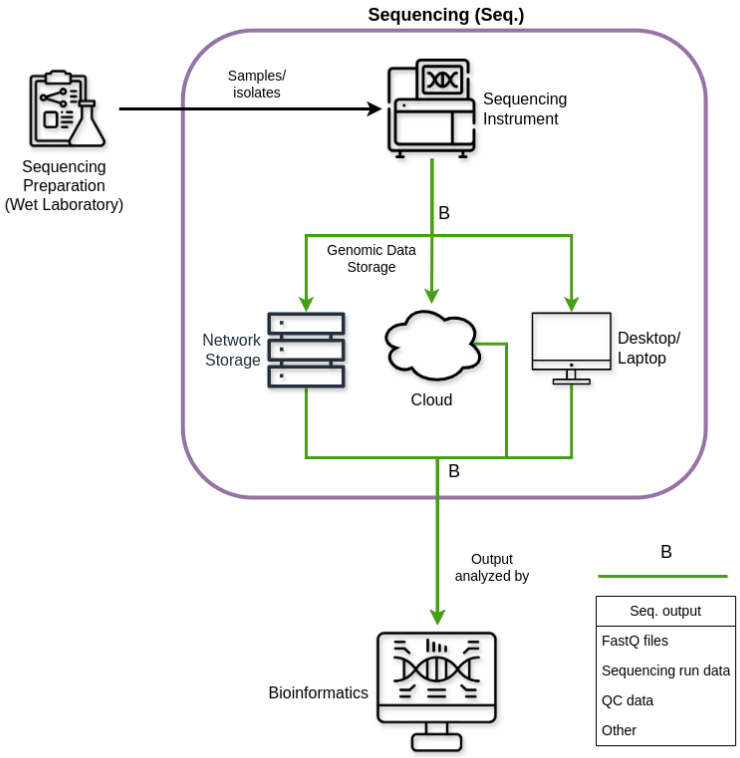


Figure 2. Sequencing Flow

The sequencing process begins with a sequencing library preparation, which is then loaded on to a sequencing instrument. Once sequencing has started output files containing sequencing data may be stored on a shared drive connected to a high-performance computing (HPC) cluster, cloud storage, or on a desktop or laptop computer usually maintained by the PHL. The files from one or more of these sources act as input to the bioinformatics process.

Across public health, reproducibility and reliability are critically important; it is no different in bioinformatics. However, due to the variety of systems being employed by PHLs, it has been difficult to establish standardized infrastructure, workflows, and processes for many pathogens. Public health agency administrations, PHL administrations, and associated IT teams have often chosen solutions that were the easiest to deploy at the time given existing policies and capabilities and the urgency to enact genomic pathogen surveillance. Thus, these solutions are unlikely to meet the current needs and requirements of PHLs to standardize and scale computational infrastructure to support genomic surveillance across a variety of pathogens and diseases. Many legacy data systems cannot scale to meet the demand of high-volume computation and lack the required logging and tracing to support "data provenance" requirements. PHLs also often adopt highly manual processes that result in higher operating costs for limited throughput, have limited data security and suffer from a lack of reproducibility and quality.

Owing to the exigencies and limitations described above, PHLs have taken a widely varied set of approaches to solving genomic surveillance problems in public health. Each PHL's strategic plan for genomic sequencing and surveillance had to address the immediate needs of sequencing while navigating complex procurement processes, policies, and an inexperienced workforce, all while dealing with the impact of the SARS-CoV-2 pandemic. The variety of creative solutions laboratories developed while navigating these issues has contributed to an inconsistent approach to public health genomic surveillance, which has raised concerns around the standardization of methods and data. In order to meet the field's needs for high quality and integrated data systems, a modern data system to support public health genomic surveillance must be developed that can meet the needs of PHLs while democratizing access and supporting changes in the public health workforce.

Current Data Systems

At the time of this white paper's preparation, the US is emerging from the COVID-19 pandemic, an experience that considerably altered and expedited the use of pathogen genomics in public health. As discussed in the previous section, prior to 2020 a variety of software platforms and hardware architectures had been adopted for the analysis and hosting of genomic data, for a modest number of pathogens, and which were often idiosyncratic to a given lab or state. The reasons for these differences are largely historic, dependent on the policies and personnel in place in each state's laboratory and health department framework and generally reflective of the relatively nascent and decentralized nature of pathogen genomics in public health at the time. The emergence of new SARS-CoV-2 variants of interest in late 2020, however, drove many labs around the world to rapidly stand-up SARS-CoV-2 sequencing. This focus on a single pathogen, the availability of cutting edge open-source solutions, and the scale of sequencing required for successful surveillance led many labs to converge on a small set of workflows and approaches. These solutions shared characteristics derived from their relatively recent design and are distinct from the previously described legacy systems. This section will summarize and contrast the characteristics of these systems and the road they pave towards a modern genomic data system.

The first relevant trend to emerge is the adoption of dedicated workflow languages to organize analytical pipelines. These workflow languages enable running pipelines in a parallel fashion on local or cloud environments and include detailed logging and fault tolerance, among a number of other features. The analytical pipelines typically used for genomics are composed of a series of bioinformatic tools, each of which performs a different function: cleaning, processing, or characterizing the raw genomic data. Bioinformatic tools may require a wide variety of libraries, dependencies, and computational power. These requirements can generate conflicts within a workflow that can cause considerable difficulties in deploying it in different environments. To mitigate these difficulties, public health bioinformaticians began adopting containerization in the years leading up to the pandemic in order to abstract away details of software dependencies. By wrapping each tool in its own Docker or Singularity container, software could be packaged with all of its required dependencies and run in an easily reproducible environment, separate from other tools in the workflow. Organizing the tools into the steps of the workflow was initially accomplished through general purpose scripting languages such as Shell, Python, or Perl. However, this often led to a new set of challenges around operating systems and hardware requirements and often lacked fault tolerance and detailed logging capacity and scalability.

Two groups have developed pipeline orchestration languages or workflow languages specifically for bioinformatics that address many of these challenges: the Workflow Description Language (WDL) developed by the Broad Institute, and Nextflow developed by Seqera Labs. At the time of the SARS-CoV-2 pandemic these languages were in a sufficiently mature state that their application to SARS-CoV-2 sequencing was a relatively obvious step. In addition to other features, both were designed to be compatible with a variety of execution

environments and to support cloud resource provisioning, which simplified the scaling of computational resources, increasing efficiency and turnaround time. Furthermore, both languages are utilized through bespoke browser-based platforms that support initiating pipeline runs, summarizing run metrics, and accessing run outputs. Nextflow workflows are accessed in the Seqera Platform (https://seqera.io/platform/), while WDL can be run on Terra (https://terra.bio/). These browser-based resources for bioinformatics address the ongoing staffing challenge for PHLs by allowing non-bioinformaticians, such as the laboratorians responsible for generating the raw sequence data, to access workflows and data. These workflow languages improved the scalability of sequencing, and many PHLs have since gained experience with one or both platforms. Issues with the application of these languages in PHLs, including standardization, required infrastructure, and data management, persist. Still, containerization paired with a specialized workflow language are critical technological steps towards a modern genomic data system going forward.

Innovation and advancement of NGS has increased throughput and decreased costs resulting in an upward trend data beyond what many PHLs have prepared for. NGS platforms now generate gigabytes to terabytes of data and are expected to grow as sequencing technology advances and PHLs expand the development and application of NGS-based tests. Although typical NGS tests may only return binary (yes/no) results, such as the presence or absence of a drug resistant gene or mutation, NGS generates many more results that may not be immediately applicable to testing but nevertheless useful. These results are often driven by the analysis workflow used and may vary widely from workflow to workflow or from disease to disease. PHLs are likely to focus on the needs of reporting key components of a NGS test, and there is currently no consistent methodology for saving and tracking genomic data or pipeline metadata over time in support of public health genomic surveillance. Ideally, NGS testing workflows would output key results in a standardized format and store genomic data and metadata in a queryable data structure for future use by the PHL. While most data systems currently used by PHLs lack the ability to store and easily query

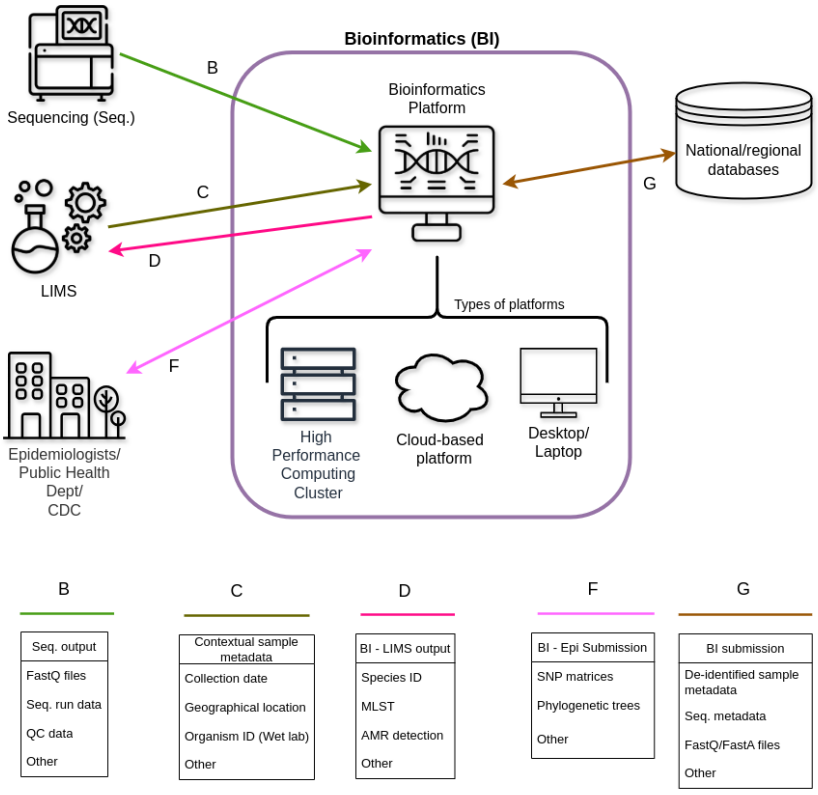


Figure 3. Bioinformatics Flow

The outputs from the sequencing procress are analyzed by specialized pipelines and tools on a bioinformatics platform. This could be an HPC cluster, a cloud-based platform such as the Easy Genomics or AMD platform, a web-based platform or a local desktop or laptop computer.

genomic data or metadata, transition to cloud-based data systems may allow these types of efforts in the future. Google's BigQuery, Amazon Web Service Redshift or Amazon Web Service Glue/Athena, and Microsoft Azure Synapse Analytics are all examples of cloud technologies built for querying large amounts of stored data and could be adapted to genomic data.

Beyond querying and managing data, many PHLs also currently face challenges in connecting genomic data to data essential for epidemiological case evaluation that may include Personal Identifiable Information (PII) and Protected Health Information (PHI). The choice to summarize major outputs in universal file formats such as CSV, while limiting the accessibility of some contextual details of how a pipeline ran, does at least guarantee compatibility with LIMS and other database systems for importing and disseminating critical results. More ideal, however, would be the direct integration of genomic data systems with a LIMS or disease surveillance system combined with a data governance layer ensuring role-based least privileged access. Integrated systems expedite data reporting, reduce both manual data tasks and unnecessary access to PHI/PII, and allow rapid, data-informed public health action. Advances in the integration of genomic and epidemiological data are a crucial element of a modern genomic data system.

"A modern genomic data system must also integrate with downstream laboratory and epidemiology tools and resources that support public health genomic surveillance."

Modern Data Systems

The fundamental reasons to use a modern data system are to reduce data siloing and to increase the usefulness of an organization's data by enabling data-driven actions and insights. For genomic data where contextual information is critical to public health action, a modern data system is a necessity. Scalable and robust cloud-based approaches are well-suited to manage the rapid advance of sequencing technology and applications. While a cloud-based infrastructure is not an absolute requirement for a modern genomic data system, on-prem solutions of the same capability and magnitude are extremely difficult to develop and maintain and are also more costly. Cloud-based services emphasize scalability, interoperability, resiliency and availability, all of which are important considerations for an organization dealing with genomic data.

The "Ten recommendations for supporting open pathogen genomic analysis in public health" by Black et al. align strongly with the desired characteristics of a modern data system. Supporting data interoperability, application programming interfaces, improving scalability and reproducibility of bioinformatic workflows, and improving data stewardship and governance are all critical pieces of a modern genomic data system. Cloud-based resources allow for the combination of a centralized system that can support genomic surveillance across a broader number of labs and organizations, with decentralized resources that empower organizations to take control of their data and support innovation. An example of this centralized approach was realized in the CLIMB-COVID SARS-CoV-2 project described by Nicholls et al. that centralized genomic analysis among a network of universities, academic institutes, regional centers, and public health agencies. Regarding interoperability, cloud-based approaches commonly employ a microservice style architecture that relies on application programming interfaces (APIs). These APIs allow greater integration with instruments and systems (e.g., sequencers, LIMS, etc.) in the PHLs as well as services and software applications in the cloud, thereby promoting interoperability. APIs are also tolerant of changes in connecting systems over time and allow new automation tasks to be built, which provides adaptability in a data system. Furthermore, SaaS solutions such as Seqera Platform, Terra, and Easy Genomics (https://www.easygenomics.org/) have provided an on-ramp to genomics and bioinformatic analysis in the cloud, increasing accessibility and function. Cloud-based systems allow GUI systems like these to enable laboratory scientists and technicians to handle the day-to-day analysis of genomics data, thereby freeing up time for bioinformaticians, data scientists, and genomic epidemiologists to focus on research and development, pipeline optimization, complex cases, and downstream analyses of genomic data.

A modern genomic data system must also integrate with downstream laboratory and epidemiology tools and resources that support public health genomic surveillance. Many PHLs have started using tools such as Microreact (https://microreact.org/), Nextstrain (https://nextstrain.org/), and Microbetrace (https://microbetrace.cdc.gov/MicrobeTrace/), for visualizing genomic data and exploring clustering (trees), geographic (map), and temporal (timeline) data. These secondary analyses have been critical to making actionable use of genomic data. Some labs have begun to leverage artificial intelligence and machine learning libraries in Python and R to develop custom approaches. These can be used to support

advanced analytics capable of classifying the impact of variants or genes or predicting phenotypic characteristics from the genomic data. Any modern data system must allow PHLs to store and query historical data in a variety of ways and support connecting genomic data to tools and resources that exist outside of the managed services that are traditionally provided by cloud systems.

In addition to integrating with secondary analytics, a modern data system must also be capable of connecting to laboratory and epidemiology data systems. The traditional LIMSs used by PHLs are not designed for the complexity and scale of genomic data and often require extensive customization to support any kind of genomic data. Currently available genomics LIMS support several features specific to genomics including wet lab workflows for DNA/RNA extraction with the capture of quality control data throughout the process; inventory management; integration with robotics such as automated liquid handlers; integration with sequencing instruments; integration with bioinformatics analysis; and handling of complex data files generated throughout the process, such as FASTQ and VCF. While a LIMS should not be expected to manage highly complex and resource intensive bioinformatic workflows, management of the data before and after analysis is a feature many PHLs need.

Ultimately, a modern genomic data system reduces barriers to accessing and integrating data and leveraging cutting-edge solutions and tools for bioinformatics and genomic epidemiology analysis. As mentioned above, cloud computing is a way to remove some of these barriers. The Enhanced Detection System for Healthcare-Associated Transmission (EDS-HAT) approach described by Sundermann et al. is an example of how enabling near real-time genomic data connected with health care data can result in actionable infection prevention. A data system that can simultaneously manage complex genomic workflows and data and connect data across systems will enable PHLs to gain additional value from infectious disease genomic data.

No public health data system can function without strict management of data. Operationally, “data governance”, the management of the usability, integrity, and security of the data, is a primary consideration in the feasibility of cloud usage. Many cloud architectures rely heavily on managed cloud services that make data governance workable and compliant with regulatory agencies. Public health data can include PHI or PII, which require more measures to comply with HIPAA. Additionally, any organization working with cloud-based data systems needs to have control over which data is kept internal vs shared publicly. A modern genomic data system must support all data management standards and protocols and allow the adoption of strategies or frameworks to support data governance.

In summary, a modern genomic data system must meet a variety of requirements to support public health genomics this includes: an ability to increase in size both in terms of computational scaling and data storage, allow modular processes to be run on the data, produce data usable by genomic epidemiology tools and allow more general inquiries, manage the status of data as it is processed, integrate genomic data and metadata to allow public health action, and to enable good data governance, including possible HIPAA compliance needs. While some of these requirements are common to most modern organizations, some are unique to genomic data. As described above, there are many domain-specific requirements that must be met. Many of these requirements are highlighted by the unique and specialized data science role that bioinformaticians play in supporting the development and management of tools, resources, and data interpretation. A modern genomic data system must accommodate the needs of bioinformaticians and allow flexibility in the design and implementation of genomic computing workflows and processes. Thus, effective data systems for public health genomics should draw on recent developments in the industry and advance them to best suit public health.

This project was supported by the Centers for Disease Control and Prevention (CDC) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award totaling $75.2M with 100 percent funded by CDC/HHS. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by CDC/HHS, or the U.S. Government.

This project was 100% funded with federal funds from a federal program $1,681,122 by Cooperative Agreement number #NU600E00104, funded by the US Centers for Disease Control and Prevention (CDC). 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.

Technical Term Glossary

1. **Application programming interfaces (APIs)**: A set of rules and protocols that allow two or more computer programs or components to communicate with each other.

2. **Big data**: Extremely large data sets that are too large to be used with traditional data processing software.

3. **Bioinformaticians**: Scientists who use computational methods to analyze and interpret biological data.

4. **Bionumerics, Galaxy, CLC Genomics Workbench**: Software tools with a graphical interface that are used for analyzing genomic data.

5. **Centralized systems**: Systems where all data is stored and accessed from a single point.

6. **Cloud-based data systems**: Systems that use remote servers hosted on the internet to store, manage, and process data, rather than a local server or a personal computer.

7. **Containerization**: A lightweight alternative to full machine virtualization that involves encapsulating an application in a container with its own operating environment.

8. **Data governance**: The overall management of the availability, usability, integrity, and security of data used in an organization.

9. **Data governance layer**: A set of policies or procedures that manage and maintain data in an organization.

10. **Data interoperability**: The ability of different systems and technologies to communicate and exchange data.

11. **Data lakes**: a central repository or system that integrates data from one or more disparate sources for reporting and data analysis, typically supporting structured and unstructured data.

12. **Data Lakehouse**: A new, open data management paradigm that combines the best elements of both data lakes and data warehouses.

13. **Data provenance or data lineage**: The record of the origins and whereabout of data, which can help provide a historical context about the data.

14. **Data scientists**: Professionals who use scientific methods, processes, algorithms, and systems to extract knowledge and insights from data.

15. **Data warehouse**: a central repository or system that integrates data from one or more disparate sources for reporting and data analysis, typically supporting structured and semi-structured data.

16. **Decentralized resources**: Systems where data is distributed across multiple points or locations.

17. **Docker, Singularity**: Popular tools that enable software containerization.

18. **FASTA, FASTQ, SAM, VCF**: File formats used for storing genetic information.

19. **Genomic data**: Information with respect to an organism's complete set of genes, or genetic material.

20. **High Performance Computing cluster**: Groups of computers that work together to perform complex computations more quickly than a single computer could accomplish alone.

21. **HIPAA**: The Health Insurance Portability and Accountability Act, a US law designed to provide privacy standards to protect patients' medical records and other health information.

22. **Laboratory Information Management Systems (LIMS)**: Software used in laboratory and scientific settings to manage complex sample workflows and associated scientific data.

23. **Microreact, Nextstrain, and Microbetrace**: Tools used for visualizing genomic data and exploring clustering, geographic, and temporal data.

24. **Microservice style architecture**: An architectural style that structures an application as a collection of services that are highly maintainable and testable, loosely coupled, independently deployable, and organized around organizational capabilities.

25. **Next-generation sequencing (NGS)**: High-throughput methodology that enables rapid sequencing of the base pairs in DNA or RNA samples.

26. **Network Attached Storage (NAS)**: Dedicated file storage that enables multiple users and heterogeneous client devices to retrieve data from centralized disk capacity.

27. **Personal Identifiable Information (PII)**: any data that could potentially identify a specific individual.

28. **Protected Health Information (PHI)**: any information about health status, provision of health care, or payment for health care; that is created or collected and can be linked to a specific individual.

29. **Seqera Platform, Terra, Easy Genomics**: SaaS platforms that support the execution of scientific workflows, particularly in the field of genomics.

30. **Software as a Service (SaaS)**: A software model in which cloud-based software is centrally hosted and allows users to connect to and use the application over the Internet.

31. **Storage Area Networks (SAN)**: A network which provides access to consolidated, block-level data storage.

32. **Workflow Description Language (WDL), Nextflow**: Workflow languages used to specify data processing workflows, primarily in the field of bioinformatics.

References

1. Black A, MacCannell DR, Sibley TR, Bedford T. Ten recommendations for supporting open pathogen genomic analysis in public health. Nat Med. 2020 Jun;26(6):832-841. doi: 10.1038/s41591-020-0935-z. Epub 2020 Jun 11. PMID: 32528156; PMCID: PMC7363500.

2. Nicholls SM, Poplawski R, Bull MJ, Underwood A, Chapman M, Abu-Dahab K, Taylor B, Colquhoun RM, Rowe WPM, Jackson B, Hill V, O'Toole Á, Rey S, Southgate J, Amato R, Livett R, Gonçalves S, Harrison EM, Peacock SJ, Aanensen DM, Rambaut A, Connor TR, Loman NJ; COVID-19 Genomics UK (COG-UK) Consortium. CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. Genome Biol. 2021 Jul 1;22(1):196. doi: 10.1186/s13059-021-02395-y. PMID: 34210356; PMCID: PMC8247108.

3. Sundermann AJ, Chen J, Kumar P, Ayres AM, Cho ST, Ezeonwuka C, Griffith MP, Miller JK, Mustapha MM, Pasculle AW, Saul MI, Shutt KA, Srinivasa V, Waggle K, Snyder DJ, Cooper VS, Van Tyne D, Snyder GM, Marsh JW, Dubrawski A, Roberts MS, Harrison LH. Whole-Genome Sequencing Surveillance and Machine Learning of the Electronic Health Record for Enhanced Healthcare Outbreak Detection. Clin Infect Dis. 2022 Aug 31;75(3):476-482. doi: 10.1093/cid/ciab946. PMID: 34791136; PMCID: PMC9427134.