IT Security and Privacy Community of Practice

Cloud Computing for Public Health Bioinformatics

Understanding the Need and Communicating with IT

This document was developed by the IT Security and Privacy Community of Practice representing the collective expertise and insights of its members.

Special Acknowledgements: Polanco, D., Bankers, L., Fortes, E., Doucette, M., Matzinger, S., Manamon, M., Blader, T., Kampoowale, A., Wang, J., Creighton, M., Marcellus, S., Gallagher, G., Zeng, L., & Jones, D.

Funding: 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.

Cloud Computing for Public Health Bioinformatics: Understanding the Need and Communicating with IT

Advances in genomic sequencing have had profound positive effects on public health surveillance but implementing these innovations requires advanced computational infrastructure. Cloud services provide the required infrastructure; however, adoption remains challenging for many Public Health Laboratories (PHLs). This white paper aims to help PHLs advocate for investment in cloud services, assist Information Technology (IT) departments in understanding the complex needs of genomic analysis, and to facilitate communication between the two.

Next-Generation Sequencing (NGS) data and data analysis are complex and therefore require advanced solutions. There are several ways complex tasks manifest for PHLs including, but not limited to: (1) handling large and complex data, (2) installing and using non-commercial open-source software, (3) selection of an optimal pipeline, and (4) relying nearly exclusively on the Linux operating system.

Given the unique nature of NGS data and NGS data analysis, IT support needs also differ compared to other public health data needs. Again, while not limited to the following, we present three cases to illustrate that PHLs require unique IT support: (1) traditional office IT support specializes in Windows OS and often offers limited (if any) Linux OS support; (2) the nature of public health bioinformatics requires IT support nearly on-demand; and (3) a communication gap exists between PHLs and their IT departments.

CDC's Advanced Molecular Detection Platform

The Advanced Molecular Detection (AMD) Platform, developed by the Centers for Disease Control and Prevention (CDC), aims to create a unified, standardized cloud-based environment for genomic surveillance. By integrating collaborative bioinformatics and genomic epidemiology analysis across the nation, AMD Platform addresses a critical challenge in public health: how to bridge the gap between the abundance of sequence data being generated by PHLs and the ability to drive public health action through bioinformatics and genomic epidemiology analyses. To best support the necessary safeguards for PHLs and the CDC, AMD Platform will implement compliance, security, and monitoring processes in accordance with NIST and FedRAMP standards. In addition to compliance with security standards, AMD Platform will have robust access controls and user roles allowing jurisdictions independent control of their data.

Public health genomic surveillance requires scalability and flexibility. Public health laboratorians, bioinformaticians, and genomic epidemiologists must pivot quickly to develop and implement new workflows and pipelines in response to emerging public-health concerns and must ensure timely and accurate results. Cloud computing enables scalability and flexibility allowing these scientists to work fluidly and ensure robust timely results while meeting grant deliverables. Cloud infrastructure offers economies of scale, with relief from the burdens of budgeting, maintaining, and securing physical hardware for peak capacity. This is made possible by paying only for what you need and use, as well as dynamically scaling for demand.

CDC's Advanced Molecular Detection (AMD) Platform can assist with cloud migration including delineation of infrastructure roles and responsibilities. Managed cloud resources allow for meeting industry standard compliance goals for data at rest, in use, and in transit.

Security and Privacy Needs and Requirements

The National Institute of Standards and Technology's (NIST) guidelines provide recommended best practices for implementing industry standard security measures to protect both on-premises and cloud-based assets from cybersecurity threats. The NIST document with specific focus on securing these assets is titled "NIST 800-53 Security and Privacy Controls for Information Systems and Organizations [1]" and is currently at revision 5. In federal agencies, "800-

53" is often synonymous with the Federal Information Security Modernization Act (FISMA) and is the NIST recommendation (as stated in NIST 800-70r4) [2] for applied security controls to facilitate FISMA compliance.

The Health Insurance Portability and Accountability Act (HIPAA) is primarily focused on the security and privacy of protected health information (PHI). It sets standards and regulations to secure PHI to ensure its confidentiality, integrity, and availability while also guaranteeing the rights of individuals to access their health information.

Both NIST and HIPAA are essential for establishing a robust and adaptable security posture and meeting regulatory compliance. By implementing these controls effectively, organizations can protect sensitive data and reduce risk by using these industry standard guidelines.

In addition to the NIST guidelines and HIPAA regulations for reducing risk, additional NIST documentation is now available, which focuses on guidance for securing genomic data. The NIST document titled “NIST IR 8432 Cybersecurity Framework Profile for Genomic Data” [3] deals with challenges, current practices, and proposals involving genomic data, while a second document titled “NIST IR 8467, Cybersecurity Framework (CSF) Profile for Genomic Data” [4] provides actionable guidance to help manage and reduce risk to assets involved with processing genomic data.

Next Generation Sequencing data of pathogens can include small amounts of human sequencing data. As such, it is important to scrub data of unintended target organisms, including potential human data, prior to submission to public repositories such as NCBI SRA. NCBI HRRT (Human Read Removal Tool) and Hostile are open-source tools for removing human-related sequence data from sequencing output files (e.g. Fastq).

The Role of Open-Source Tools in Bioinformatics

Open-source tools are a critical component of bioinformatic and genomic analysis, with many being developed by academic institutions as well as other experts in the public health field. The use of open-source and community developed tools introduces unique security and privacy concerns. Below lists concerns and mitigation best practices that can be implemented to reduce security and privacy risks associated with open-source and community developed bioinformatics tools.

Concern	**Mitigation**
Lack of Vulnerability Scanning	• Implement third party automated vulnerability scanning tools • Establish a policy for regular patching and updates of software • Utilize AMD Platform with built in vulnerability scanning • Supported Linux distribution for threat management software (e.g. Cisco AMP, Qualys Cloud Agent, Workspace One, LUKS) • Use code reviews to identify potential vulnerabilities • Regularly update dependencies and use tools that check third-party libraries for vulnerabilities • Foster a security culture where individuals are empowered to help mitigate risk • Customized user and environment permissions
Abandonware	• Prioritize the use of actively maintained software by reviewing code commit history, open issues, and active discussion • Follow updates and stay connected with the community • Regularly audit for signs of declining maintenance or stale development • Develop a contingency plan for critical software (e.g. identify alternative projects) • Design systems to be modular so replacing software can be done with minimal disruption • Consider adopting the project, licensing it, or otherwise contribute to its development

Concern	Mitigation
Root Level Access	• Implement strict user access controls and limit user permissions to minimum necessary for task • Use role-based access controls to limit resource use based on job requirements • Ensure multi-factor authentication is required and use identity providers • Monitor and audit activity logs and use intrusion detection systems • Use infrastructure as code tools to deploy resources with secure configuration settings • Regularly update software and maintain backups with one copy off-site • Educate users on security best practices and the risks of root level access • Have an incident response plan in case the worst happens
Installation of dependency software	• Containerization is a way to package programs along with their necessary dependencies for ease of installation and use (e.g. docker, singularity). These containers also provide a stable or repeatable way to use the software
Version control	• The most common tool for version tracking is Git and there are services such as GitHub for accessing open-source software, displaying the version history, and tracking releases of software as the mitigation

Table 1: Cloud Computing as a Solution, addressing current system limitations and needs

	Limitations of Current Systems	Need	Cloud Computing/AMD Platform Solutions
Security Standardization	• Lack of security standardization and associated documentation across bioinformatic software. • Differences in governance, policies, and management between different PHL environments (IT, Lab, Epidemiology).	• Clear compliance requirements, security goals, and associated responsibilities	• Comply with security standards including FedRAMP, NIST 800.53, ISO 27017, and FISMA. • Shared responsibility model. • Third party security and compliance audits.
Accessibility and Data Management	• Limitations and unreliability of onsite storage (USB, share drive) combined with undefined retentional policies for NGS data. • Access restrictions can require local network connection and prevent remote access. • Network access restrictions to approved public data repositories	• Disaster recovery • Access control • Security controls for data classification • Expansive & efficient storage solutions to manage large sequencing datasets • Data governance and stewardship	• Application programming interface (APIs) for moving data to and from the cloud, encryption in movement and at rest. • Secure, cost-effective data storage with backup and disaster recovery options available. • Uptime for large cloud vendors (AWS, GCP, Azure) demonstrates commitment to reliability. • Various cloud vendors (AWS, GCP, Azure) offer solutions that allow organizations control over their data's storage, its access, its usage and disposal while maintaining regulatory compliance [5,6]. • These solutions enable data discoverability, metadata management, and data class-level controls that allow for the separating of sensitive data from other data [7].

	Limitations of Current Systems	Need	Cloud Computing/AMD Platform Solutions
Computational Requirements, Procurement and Data Modernization	• Outdated computational infrastructure. • Knowledge retention concerns related to configuration of local server (i.e. personnel changes and knowledge continuity). • Timely implementation of new software or systems as bioinformatic workflows can require new infrastructure to be invented, purchased, provisioned, and supported.	• Modern and efficient computational resources • Straightforward procurement flow/expansion of service • High RAM	• On-demand, scalable infrastructure. • Implementation support and infrastructure management provided by cloud vendor.
Cost and Budgets	• Difficult to determine the exact need to budget (e.g. surges) and requires a large upfront cost for implementation. • Systems managed centrally in IT or lab require time and expertise in-house.	• Scalable, transparent, and cost-effective computing solutions	• Pay as you go models mean no need to provision for 100% capacity from the start. • Infrastructure features and price benefits from economies of scale.

The Basics: Where to Start

Q: Our laboratory is interested in integrating cloud computing for bioinformatics. Who should we start conversations with?

A: When considering changes to IT infrastructure it's important to build a team with broad interests. Start at the top with laboratory leadership, they will be able to direct down through the chain of command and to the appropriate IT staff for the project. In addition to IT Staff, consider including legal, purchasing, and staff familiar with departmental structure and policies in early conversations.

Q: How can I secure IT buy-in and advocate for dedicated IT staff for implementing the AMD platform or cloud computing in my jurisdiction?

A: First, identify if your jurisdiction has dedicated personnel with cloud and/or Linux experience.

- If yes, meet with this person to discuss the AMD platform and identify any issues or consideration they may have and begin devising a plan for implementation.
- If no, determine whether you want/need to hire a person with this content knowledge and how that would be funded.
- Alternatively, determine if education/training could be provided for existing personnel.

Strategies to support buy-in:

- Come prepared with specific needs: Reach out to APHL, your local Bioinformatics Regional Resource (BRR), and other labs to get an idea of how they worked with their IT for their set-ups.
- In addition to talking with other labs about their set-ups, research some of the different resources they utilize as proof of meeting some IT security standards.
- Set up lab tours and visual schematics to help highlight workflows.

Q: How can I prepare for conversations with IT staff around cloud computing and the AMD platform?

A: To effectively prepare for AMD platform conversations, consider compiling the following to share with IT staff:

1. Document any challenges and specific needs related to data storage and computational infrastructure. Providing documentation of data types and security classifications for data uploaded to and downloaded from the cloud assists IT departments in understanding the specific requirements of your project.
2. Identify existing or planned cloud resources within your jurisdiction as well as the agency's acceptable use policies for the planned or proposed resources.
3. Quantify the volume of data generated from sequencing runs and the amount of storage needed to meet data retention policies. Include examples of sequencing output files to demonstrate the data type and volume (see Appendix A).
4. Review and document data retention policies and develop a jurisdiction-specific plan for long-term data storage, considering factors like file types/sizes, cloud storage vs. local storage, and data retention periods. Conduct a cost analysis to determine the most effective storage strategy.
5. Identify department-specific barriers and bureaucratic processes by obtaining a list of written policies specific to your state/jurisdiction/university.

References and Resources

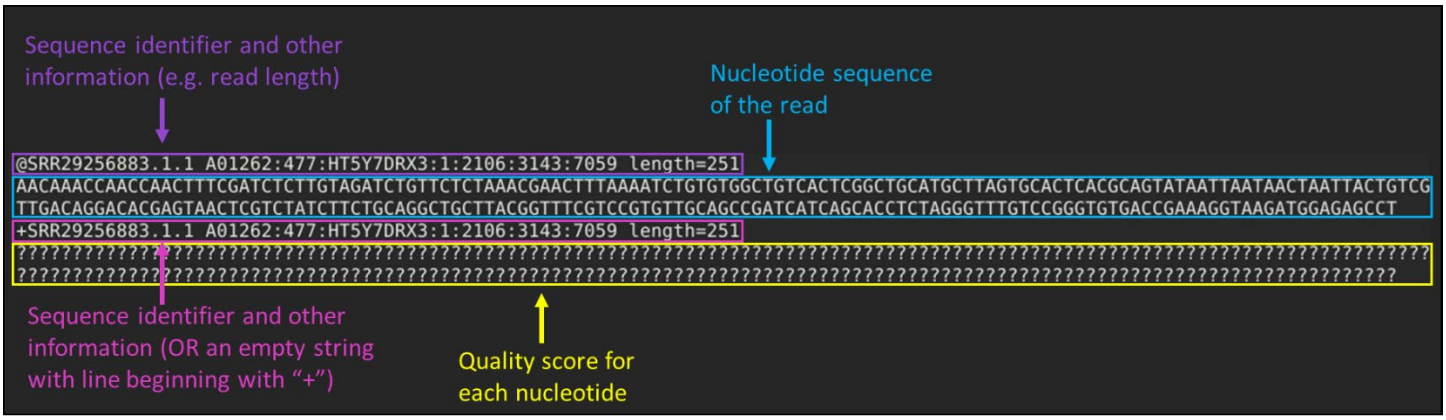
[1] https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final
[2] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-70r4.pdf
[3] https://www.nccoe.nist.gov/news-insights/nccoe-publishes-final-nist-ir-8432-cybersecurity-genomic-data
[4] https://doi.org/10.6028/NIST.IR.8467.ipd
[5] https://aws.amazon.com/what-is/data-governance/
[6] https://www.microsoft.com/en-us/security/business/security-101/what-is-data-governance-for-enterprise
[7] https://cloud.google.com/learn/what-is-data-governance?hl=en

Appendix A: Example Raw Data Sequencing Output File Size by Platform

Instrument	File Type	Raw Data
Illumina iSeq	.fastq	< 3 GB per run
Illumina MiniSeq	.fastq	< 8 GB per run
Illumina MiSeq	.fastq	< 26 GB per run
NextSeq 500/550	.fastq	39 – 170 GB per run
NextSeq 1000/2000	.fastq	35 – 170 GB per run
NovaSeq 6000	.fastq	20 – 2190 GB per run
Oxford Nanopore Technology	.pod5 (formerly .fast5) .fastq	9.9 Gbytes per 1 Gbase of sequence data Standard sequencing run 64 – 72 hours ~ 135 – 162 GB per run
Oxford Nanopore Technology	.fastq	1.1 Gbytes per 1 Gbase of sequence data Standard sequencing run 64 – 72 hours ~ 15 – 18 GB per run
PacBio	unaligned .bam	30 Gbases of data for a 5-10 kb library up to 90 Gbases of data for a 20-25 kb library (specs based on PacBio's Revio system)
Element Aviti	.fastq	75-300 Gbases depending on kit and flowcell output

Appendix B: Example Sequencing Data File

The most common format for raw sequencing data is a fastq file. Here is a link to download and explore a publicly available fastq file from a SARS-CoV-2 sample as an example of the file format and data size (https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&page_size=10&acc=SRR29256883&display=download). File size will vary across organisms, sequencing platforms, sequencing depth, etc. More information on fastq and other raw sequencing data file type can be found at this link (https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/). Below is a schematic of the main information found in a fastq file using the first read from the SARS-CoV-2 example found at the link above.



Appendix C: Important Definitions and Expanded Overview of Security Regulations

Abandonware: Software that is no longer maintained by the developer.

Bioinformatics: the multidisciplinary field that combines biology, computer science, and information theory to store, transfer, classify, and analyze complex biological information, including NGS sequencing data. Public health bioinformatics involves the development and use of software tools and algorithms to analyze biological data, especially at the molecular level. It is crucial in genomics for sequencing, assembling, and annotating genomes. Additionally, it plays a role in proteomics, transcriptomics, and metabolomics to identify patterns and relationships between biological molecules.

Cloud computing: the delivery of servers, storage, databases and software applications over the internet. Unlike on-premise computing services where physical infrastructure, such as high-performance computing clusters, is maintained by users themselves, cloud computing providers take responsibility for its security and automate implementation details.

FedRAMP: Federal Risk and Authorization Management Program. A set of federal security standards for cloud services.

FISMA (Federal Information Security Management): a federal law that establishes a framework for protecting government information systems by requiring agencies to develop and implement comprehensive security programs. It aims to ensure the confidentiality, integrity, and availability of federal information and systems.

Genomics: the multidisciplinary field that studies the complete genetics code (genome)of an organism or group of organisms that ranges from mapping the sequence and structure of the genome to evolution.

HIPAA (Health Insurance Portability and Accountability Act): a law focused on the security and privacy of protected health information (PHI). It sets standards and regulations to secure PHI to ensure its confidentiality, integrity, and availability while also guaranteeing the rights of individuals to access their health information.

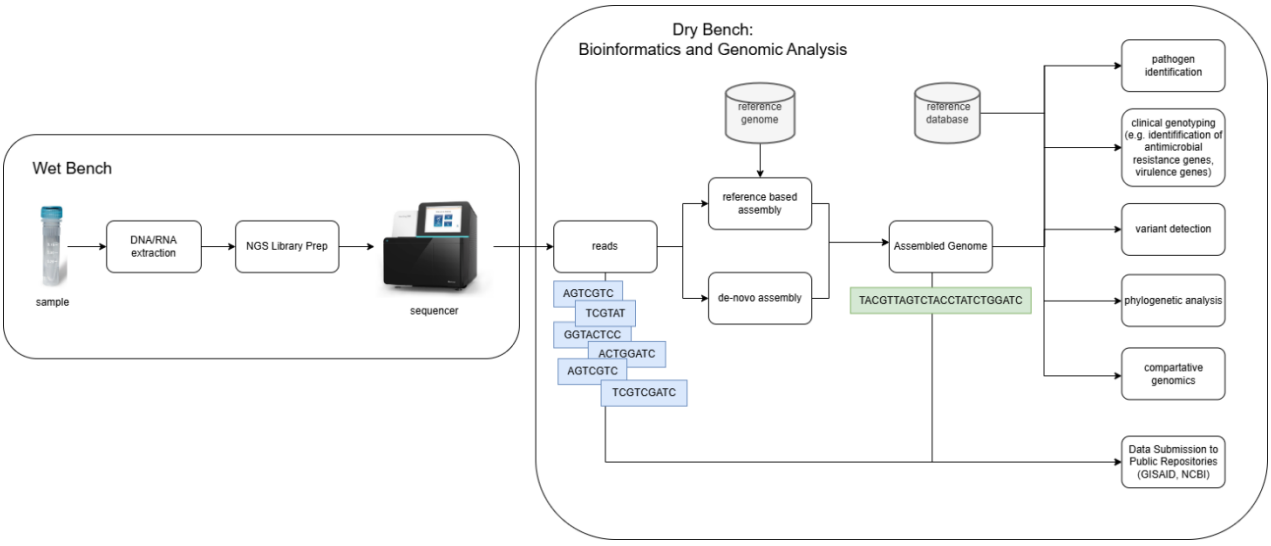
Next Generation Sequencing (NGS): high throughput technology used to sequence the DNA or RNA of an organism.

NIST (National Institute of Standards and Technology's): organization that recommends guidelines and best practices for implementing industry standard security measures to protect both on-premises and cloud-based assets from cybersecurity threats.

Pipeline: a series of steps chained together that are involved in the assembly and analysis of genomic data, where the outputs of one step serve as the inputs for the next step. These steps typically include checking raw data quality, assembling raw data into complete genomic sequences, and annotating and performing statistics on the complete genomic sequence. Pipelines are typically written in workflow languages such as Nextflow, WDL or Snakemake which help organize inputs and outputs as well as track runtime details.

Workflow: In the context of bioinformatics and genomics, a workflow is a series of computational steps to transform raw data into processed data and results.

Appendix D: AMD-Platform or Dataflow/Workflow Schematic



General NGS workflow. The workflow is divided into a wet bench phase performed by laboratorians in a physical laboratory and a dry bench phase performed by bioinformaticians in a computational environment. During the wet bench phase, the genetic material is extracted from the sample and prepared for sequencing. Sequencing generates millions of reads, which are then assembled during the dry bench phase. Additional genomic analyses are performed using the assembled sequences depending on the epidemiological questions. Reads and assembled sequences are submitted to public repositories so other PHLs and researchers have access to the data.