September 14, 2020

Re: CDC-2020-0051

The Association of Public Health Laboratories appreciates the opportunity to provide more detail on bioinformatics personnel, next generation sequencing data files and sequence analysis software to help inform discussions of the Clinical Laboratory Improvement Advisory Committee (CLIAC).

What are the roles and responsibilities for all personnel performing bioinformatics or pathology/laboratory informatics activities? What training is considered essential for each of the roles? What competencies are considered essential for each of the roles? What minimum educational requirements (degrees or courses) are required for each of the roles?

A foundational science or computing bachelor's degree is generally a minimum prerequisite for any laboratory bioinformatics personnel. Bioinformatics is a relatively new field, so experience and ability are commonly accepted over further formal education. Allowing a relevant degree *or equivalent*, and a competency-based system, potentially with a laboratory training component, would mean that personnel with diverse backgrounds, academically and experientially, could appropriately qualify for these roles. An understanding of rules relating to security and privacy, including data use agreements and limitations on use of data, are likely to become increasingly important. We have included some position descriptions below as examples; however, these may vary significantly across public health laboratories due to varying needs and capacity.

NGS Bench Scientist

Essential staff for all public health laboratories intending to perform sequencing of specimens/cultures onsite, or to institute analysis and reporting.

Roles and responsibilities: Performing routine NGS experiments with a fundamental understanding of bioinformatics principles. This individual should have an ability to interpret bioinformatics quality control (QC) reports, request and interpret analysis for cluster detection and/or genome annotation, and understand how wet-lab protocols relate to bioinformatics results, prepare reports for Technical Supervisor review for reporting to customers.

Essential Training: Required biological safety and chemical safety practices, conceptual understanding of sanger and next-generation sequencing theory and chemistries, proper laboratory techniques and molecular biology procedures, understanding of biosafety risk assessment process for new procedures, quality assurance procedures for documentation of testing equipment, testing performed, reagents and supportive equipment used, inventory management.

Competencies: Test sample quality assessment and label verification, techniques for preparing nucleic acid and assessing quality and quantity of nucleic acid, preparation of sequencing libraries and appropriate use of all associated instrumentation and kits, pooling libraries and preparing the sequencing run on the sequencing platform, interfacing with instrumentation software, assessing run quality metrics and per sample quality metrics, transferring data, performing preventive maintenance of instrumentation used in sequencing testing. Knowledgeable of bioinformatics principles for raw and assembled sequence quality review, interpretation of bioinformatics results obtained through a graphic-user interface (GUI) software or from bioinformatics personnel. Operation of GUI software for data analytics.

Educational requirements: Bachelor's degree in a scientific discipline, preferred training and work experience in molecular biology techniques including PCR-based methodologies

Bioinformatics Technician /Technologist

Essential staff for all public health laboratories performing bioinformatics analysis.

Roles and responsibilities: Bioinformatics scientist with a strong understanding of molecular biology as it relates to NGS technologies and bioinformatics applications and a proficiency in utilizing open-source and/or commercial software to perform bioinformatics tasks at the command-line interface of a Linux OS or through a developed (GUI) software, e.g. CLC Genomics Workbench, Bionumerics. Bioinformatics tasks include an ability to compile and formulate reports, draw public health conclusions from NGS and other sequencing data, troubleshoot issues with standard protocols, install and validate bioinformatics software, and provide interpretations of data to non-bioinformatics specialists.

Essential Training: Applied understanding of bioinformatics principles and approaches relevant to data QC, genome assembly and annotation, and cluster analysis, variant analysis (for e.g. newborn screening); navigation and utilization of the command-line interface; accessing and installing open-source software

Competencies: Execution of software installation and installation of associated dependencies, software testing/validation procedures, data transfer from NGS machinery or instrument/interface to workstation, data storage and retrieval, Cloud: spinning up a virtual machine, connecting to a virtual machine, retrieval of reports, knowledge of and ability to read/explain phylogenetic trees.

Educational requirements: Bachelors, MPH, MS preferred in biology, bioinformatics, or related life-science degree program.

Bioinformatics Developer

Recommended staff for regional bioinformatics support and development.

Roles and responsibilities: Bioinformatics scientist with a competency in scripting/programming languages and bioinformatics software development. This individual should have an ability to develop novel bioinformatics solutions (e.g. organism-specific analytical pipelines, data visualization schemes, and QC protocols) that promote reproducible, interoperable, and open-access principles; communicate with State IT personnel to describe needs for computational infrastructure development; and effectively collaborate with the greater public health bioinformatics community. Capable of developing, testing and implementing complex bioinformatics solutions with limited supervision. Good project management skills and capacity to manage multiple projects simultaneously. Provides consultation for work colleagues and customers regarding bioinformatics approaches, results and limitations. Facilitates the development of software and bioinformatic pipelines by planning or developing, performing and documenting detailed validations.

Essential Training: Expert understanding of bioinformatics principles and approaches relevant to data QC, cluster analysis, and genome annotation, variant interpretation (for e.g. newborn screening); scripting/programming languages for reproducible, interoperable, open-source software development; state compute-infrastructure development.

Competencies: Script generation, download and installation of new tools/dependencies, grant writing, use of version control software, use of containerization software, report building, software validation, spinning up cloud computational resources, transferring data using a cloud storage bucket, ability to run software from command line, knowledge of and ability to read/explain phylogenetic trees.

Educational requirements: Bachelors, MS, MPH or PhD preferred in biology, bioinformatics, computational biology and genomics or related life-science degree program or data science field.

What are the challenges for recruitment and retention of bioinformatics or pathology/laboratory informatics personnel?

The field of clinical and epidemiological bioinformatics is relatively new and its application to public health is even newer. There are relatively few degree programs available in bioinformatics and many of these programs and fellowships focus on either prokaryotic or eukaryotic genomics. It does require additional training for the skills to be transferrable to both clinical diagnosis and/or outbreak investigations of infectious diseases in public health. Furthermore, there is yet to be standardization across bioinformatics programs or technical certifications.

The limited number of suitable applicants for public health labs is further compounded by lack of awareness of public health as a career path. There are also few entry-level bioinformatics positions available in public health labs, most require PhD or MS or IT and server maintenance experience and outdated technology within laboratories may be unappealing. Our members are not in a position to recruit as aggressively as industry or academia and compensation in public health is significantly lower than in other positions that may be available to bioinformaticians. Even within the local government system, the salary is low, typically closer to bench scientists than IT professionals. If a candidate remains interested, some states use processes that make it difficult for bioinformaticians to qualify, e.g. civil service exams or have no applicable or specific, tiered position set for bioinformatics personnel. Bioinformaticians may also need to have skills unrelated to their position to qualify as CLIA supervisors.

Retention of bioinformatics staff can be a challenge for PHLs. The pace of work within a PHL may be slow compared to academia or industry, and limited policies for open source software can cause frustratingly slow implementation of solutions. They may have to do other work, outside of their primary duties, if there is not enough throughput. There may be limited opportunities for career advancement for bioinformaticians within this skillset and work benefits such as remote working may be less than with other career opportunities.

What are examples of how NGS data files are used in addition to generating a clinical test result?

While the clinical test result is often the primary goal of a sequencing data file, public health laboratories also use this information in a broad and expanding variety of important ways. Some examples include genomic epidemiology to support outbreak and surveillance investigations such as used in various food safety outbreaks. NGS data is also useful for identifying new or re-emerging viral variants and understanding transmission networks, detecting virulent genes and monitoring antibiotic resistance. Aside from the tremendous uses in pathogen detection, surveillance, and diagnosis, sequencing data has the potential to aid in assay development and algorithm modification. Its data can be used for building models and predicting disease spread, and monitoring for mutations which influences vaccine and drug development as well as identifying newborn screening disease causing variants.

What NGS data files should be retained for quality assurance, repeat analyses, or subsequent analyses? How long should these NGS data files be retained?

A number of different data files are generated during sequencing. From the initial image files to the eventual FASTQ (or FAST5 as generated by Oxford Nanopore Sequencers) files, different laboratories have different protocols for data storage. In general, raw data should be retained. Information regarding how the analytical process was performed should also be captured (e.g. which bioinformatics pipeline and if non-default parameters were invoked, which version of a software). Storage of data regarding whole

genome sequencing (WGS) library prep should also be retained in some manner. These fundamental files should be retained as long as there may be a need to explain the particular result. Storage solutions vary and range from on-site hard-drives, servers, private cloud accounts, or dependence on public repositories such as NCBI. If this need is possible but unlikely, the files can be put in deep storage, where the files may not be immediately accessible. Data retention policies will vary by sample type and may be influenced by state rules. For example, newborn screening programs have varying data retention policies, ranging from 2 years or less to 20 or more. Alternatively, sequencing files with potential legal implications may have specific storage requirements.

One example of data storage protocol that a laboratory follows is that raw sequence data files, Fastq files, i.e. basecall files with an associated quality score generated by the sequencing platform, should be retained for quality assurance, repeat analysis, or subsequent analysis. These should be maintained for two years plus current if on a public repository, and potentially indefinitely if the laboratory is the only repository for the data (i.e. data is not shared in a public repository). If MinION sequencing is performed, Fast5 and raw files direct from the sequencer should be maintained. As long as storage permits, compressed Fastq files should be kept indefinitely.

Greater guidance around data retention standards which take into account the different purposes of sequencing would be helpful as the quantity of data from sequencing makes the cost and space for storage cumbersome.

What are the challenges and approaches for laboratories to maintain and utilize previous versions of sequence analysis software?

One of the biggest challenges around sequence analysis software is that new versions are constantly being developed. Public health laboratories require validation and at times certification of software, which prevents them from switching to each new version of software. However, due to a number of reasons, many times companies do not support the older versions and the newer versions are not compatible with the older ones. Companies often do not give timely notice when versions are being phased out or information about how compatible newer versions will be. Another challenge with software is that the documentation is rather limited. Most commercial software companies understandably keep much of their analysis proprietary, which can make it difficult for laboratories to troubleshoot. Additionally, even in-house created pipelines are reliant on public repositories, which may become unavailable or deleted, thus making the pipeline redundant.

Laboratories have found ways to combat some of the challenges above. One key solution for command-line pipelines is containerization. Containerization environments of bioinformatics software, as demonstrated by the StaPH-B Docker-Hub Initiative, ensure a static compute environment that can capture multiple versions of the most commonly utilized public health bioinformatics programs and allow for easy sharing and deployment across laboratories. Furthermore, laboratories are moving to standardize tagging and nomenclature to easily query and employ previous versions of software. Documenting the software version at time of data analysis is a key step in order for laboratories to compare results with historic samples. There is also now package management software that enables compatibility between older versions of software. All of these solutions are being deployed in some public health laboratories, however, they do require a level of bioinformatics expertise which is currently not available in all laboratories.

Please contact Kuki Hansen, manager Regulatory and Public Policy (kuki.hansen@aphl.org) with any questions.

Sincerely,

Scott Becker, M.S.
Chief Executive Officer

Bill Whitmar, M.S.
President

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.