The Journal of Molecular Diagnostics, Vol. 20, No. 1, January 2018

ELSEVIER

the Journal of Molecular Diagnostics

jmd.amjpathol.org

SPECIAL ARTICLE

Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines

A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists

Somak Roy,[*†] Christopher Coldren,[*‡] Arivarasan Karunamurthy,[*†] Nefize S. Kip,[*§¶] Eric W. Klee,[*‖] Stephen E. Lincoln,[*,**] Annette Leon,[*††] Mrudula Pullambhatla,[‡‡] Robyn L. Temple-Smolkin,[‡‡] Karl V. Voelkerding,[*§§] Chen Wang,[*‖] and Alexis B. Carter[*¶¶]

From the Next Generation Sequencing Bioinformatics Pipeline Validation Working Group of the Clinical Practice Committee, The Association for Molecular Pathology,[‡‡] Bethesda, Maryland; the Division of Molecular and Genomic Pathology,[†] University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania; PathGroup,[‡] Nashville, Tennessee; Genetics and Genomics Center,[§] Icahn School of Medicine, Mount Sinai Hospital, New York, New York; Sema4 Laboratory,[¶] Branford, Connecticut; the Division of Biomedical Statistics and Informatics,[‖] Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota; Invitae,** San Francisco, California; Color Genomics, Inc.,[††] Burlingame, California; ARUP Laboratories,[§§] Salt Lake City, Utah; and the Department of Pathology and Laboratory Medicine,[¶¶] Children's Healthcare of Atlanta, Atlanta, Georgia*

Accepted for publication November 6, 2017.

Address correspondence to Somak Roy, M.D., Division of Molecular and Genomic Pathology, University of Pittsburgh Medical Center, 3477 Euler Way, Room 7028, Pittsburgh, PA 15213. E-mail: roys@upmc.edu.

Bioinformatics pipelines are an integral component of next-generation sequencing (NGS). Processing raw sequence data to detect genomic alterations has significant impact on disease management and patient care. Because of the lack of published guidance, there is currently a high degree of variability in how members of the global molecular genetics and pathology community establish and validate bioinformatics pipelines. Improperly developed, validated, and/or monitored pipelines may generate inaccurate results that may have negative consequences for patient care. To address this unmet need, the Association of Molecular Pathology, with organizational representation from the College of American Pathologists and the American Medical Informatics Association, has developed a set of 17 best practice consensus recommendations for the validation of clinical NGS bioinformatics pipelines. Recommendations include practical guidance for laboratories regarding NGS bioinformatics pipeline design, development, and operation, with additional emphasis on the role of a properly trained and qualified molecular professional to achieve optimal NGS testing quality. *(J Mol Diagn 2018, 20: 4—27; https://doi.org/10.1016/j.jmoldx.2017.11.003)*

Supported by the Association for Molecular Pathology.

Disclosures: C.C. is an employee of PathGroup; S.E.L. is employed by and holds stock in Invitae, and holds stock in Illumina; A.L. is employed by and has stock options with Color Genomics; N.S.K. has received speaker fees/honoraria from Nanostring; K.V.V. is an advisor to PierianDx and is a committee chair/member of College of American Pathologists; A.B.C. is a committee chair/member of College of American Pathologists and the Clinical and Laboratory Standards Institute and has received speaker fees/honoraria from the American Medical Informatics Association.

The Next Generation Sequencing Bioinformatics Pipeline Validation Working Group of the Clinical Practice Committee, Association for Molecular Pathology (AMP), with organizational representation from the College of American Pathologists (K.V.) and the American Medical Informatics Association (C.W.). The AMP 2016 and 2017 Clinical Practice Committee consisted of Marina N. Nikiforova (2016 Chair), Antonia R. Sepulveda (2017 Chair), Monica J. Basehore, Mark Boguski, Susan Butler-Wu, Jennifer Crow, Linda Cook, Birgit Funke, Meera R. Hameed, Lawrence J. Jennings, Arivarasan Karunamurthy, Keyur Patel, Jess F. Peterson, Benjamin Pinsky, Somak Roy, Mark J. Routbort, Kandelaria Rumilla, Ryan Schmidt, and David S. Viswanatha. The AMP 2016 and 2017 Informatics Subdivision Leadership consisted of Alexis Carter (Chair), Mark Boguski, Christopher Coldren, Nefize S. Kip, Eric Klee, Roy E. Lee, Annette Meredith, Mark Routbort, Jeremy P. Segal, Jorge L. Sepulveda, Brian Shirts, and Somak Roy.

Standard of practice is not defined by this article, and there may be alternatives. See *Disclaimer* for further details.

https://doi.org/10.1016/j.jmoldx.2017.11.003

Bioinformatics pipelines are an integral component of next-generation sequencing (NGS). Processing raw sequence data to detect genomic alterations has significant impact on disease management and patient care. Because of the lack of published guidance, there is currently a high degree of variability in how members of the global molecular genetics and pathology community establish and validate bioinformatics pipelines. Improperly developed, validated, and/or monitored pipelines may generate hidden, inaccurate, and/or inscrutable results, which may have negative consequences for patient care. To address this unmet need, the Association of Molecular Pathology (AMP), with organizational representation from the College of American Pathologists and the American Medical Informatics Association, has developed best practice consensus standards and guidelines for the validation of clinical NGS bioinformatics pipelines. The AMP believes it is the responsibility of professional organizations to establish guidelines for professional practice; as such, we routinely engage with other professional associations to publish evidence-based practice guidelines. Our members are among the early adopters and users of NGS technology in a clinical setting, and have accumulated substantial knowledge and expertise as it relates to this novel and powerful technology.

The Need for NGS Bioinformatics Guidance

The democratization of NGS technologies has contributed to their rapid adoption in clinical practice, but constant technology evolution and the absence of clear recommendations for analytical validation of NGS bioinformatics pipelines have contributed to inconsistencies in clinical laboratory practice. Examples of analytical validation of NGS tests have been published in the medical literature (vide infra); however, existing documents lack clarity on requirements for analytical validation of NGS assays for both germline and somatic variants. These deficiencies are particularly evident with relation to NGS bioinformatics pipelines. This is further complicated by the proprietary nature of bioinformatics pipelines supplied by NGS instrument manufacturers. An understanding of the process required to validate fully a set of pipelines in which the full algorithmic details are unknown is critical for providing safe patient care. Furthermore, bioinformatics methods and principles for NGS data analysis are constantly evolving and may be customized for specific platforms and assays types, and individuals unfamiliar with the validation processes necessary to perform clinical patient care may make changes to existing pipelines without any or adequate revalidation. As a result, this consensus recommendation guideline was developed as a set of broad principles, which are applicable to the validation of any clinical NGS bioinformatics pipeline.

Description of NGS Technology

NGS is a generic term used to describe several different massively parallel and high-throughput sequencing technologies. Compared with dideoxy sequencing (Sanger), NGS is faster and cheaper by orders of magnitude but is also dependent on a highly complex computational data analysis infrastructure. As a result, Sanger sequencing and other less complex and less computationally dependent techniques continue to be widely used for validating NGS results.

Defining the Bioinformatics Pipeline

NGS generates massive amounts of data that require multiple computationally intensive steps for appropriate analysis to be performed.[1,2] Bioinformatics is the discipline that conceptualizes biology in terms of macromolecules and then applies informatics techniques (applied math, computer science, and statistics) to understand and organize the information associated with these molecules, on a large scale.[3] Bioinformatics algorithms executed in a predefined sequence to process NGS data are collectively referred to as the NGS bioinformatics pipeline (Figure 1). A glossary of NGS bioinformatics pipeline-related terminologies is provided in Supplemental Table S1.[4,5] A bioinformatics pipeline progressively shepherds and processes massive sequence data and their associated metadata through a series of transformations using multiple software components, databases, and operation environments (hardware and operating system). A typical clinical implementation of a bioinformatics pipeline is automated, necessitating appropriate quality control (QC) to ensure the generated data are robust, accurate, reproducible, and traceable. As with all hardware and software used for clinical patient care, each step of a clinical NGS pipeline emits several data points that can be used as metrics for bioinformatics pipeline QC. This is critical not only for good patient care but also for troubleshooting and compliance with regulatory requirements.

Bioinformatics Analysis of NGS Data

NGS bioinformatics pipelines are frequently platform specific and may be customizable on the basis of laboratory needs. A bioinformatics pipeline consists of the following major steps.

Sequence Generation

Sequence generation (signal processing and base calling) is the process that converts sensor (optical and nonoptical) data from the sequencing platform and identifies the sequence of nucleotides for each of the short fragments of DNA in the sample prepared for analysis. For each nucleotide sequenced in these short fragments (ie, raw reads), a corresponding Phred-like quality score is assigned, which is sequencing platform specific. The read sequences along with the Phred-like quality scores are stored in a FASTQ file, which is a de facto standard for representing biological sequence information.[4]