The Journal of Molecular Diagnostics, Vol. ■, No. ■, ■ 2017

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

Q32 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 Q1 Alexis B. Carter[*§§]

Q2 Q3 *From the Next Generation Sequencing Bioinformatics Pipeline Validation Working Group of the Clinical Practice Committee,[*] The Association for Molecular Pathology,[††] Bethesda, Maryland; the University of Pittsburgh Medical Center,[†] Pittsburgh, Pennsylvania; PathGroup,[‡] Nashville, Tennessee; the Mount Sinai Hospital,[§] New York, New York; 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., Molecular Informatics, 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 advisement 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 2017, ■: 1–24; https://doi.org/10.1016/j.jmoldx.2017.11.003)*

Q4 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

Q5 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.

Q6 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 [F1] 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]

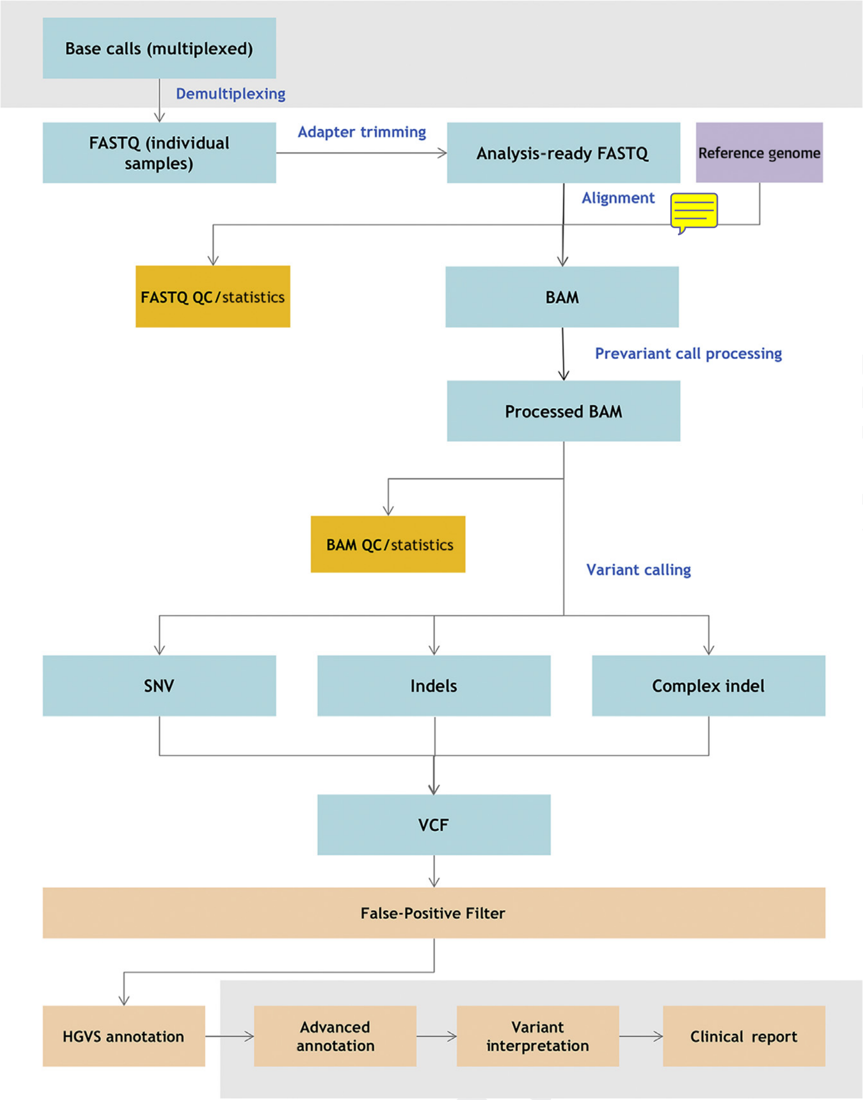


Figure 1 Next-generation sequencing (NGS) bioinformatics pipeline. The figure illustrates a bioinformatics pipeline and its components that are typically used for processing NGS data. The illustration may not represent nuances or additional algorithms that are specific to sequencing platforms. The components of the pipeline that overlap with the **gray shaded region** are out of scope of this guideline. BAM, binary alignment map; HGVS, Human Genome Variation Society; indel, insertion/deletion; QC, quality control; SNV, single-nucleotide variant; VCF, variant call format.

Sequence Alignment

Sequence alignment is the process of determining where each short DNA sequence read (each typically <250 bp) aligns with a reference genome (eg, the human reference genome used in clinical laboratories). This computationally intensive process assigns a Phred-scale mapping quality score to each of the short sequence reads, indicating the confidence of the alignment process. This step also provides a genomic context (location in the reference genome) to each aligned sequence read, which can be used to calculate the proportion of mapped reads and depth (coverage) of sequencing for one or more loci in the sequenced region of interest. The sequence alignment data are usually stored in a de facto standard binary alignment map (BAM) file format, which is a binary version of the sequence alignment/map format. The newer compressed representation or its encrypted version[6] is a viable alternative that saves space, although laboratories need to carefully validate variant calling impact if lossy (as opposed to lossless) compression settings are used in generating compressed representation of alignment/map/encrypted version of compressed representation of alignment/map files Q7 [European Nucleotide Archive, CRAM format specification Q8 version 3.0; *http://samtools.github.io/hts-specs/CRAMv3.pdf*, last accessed November 23, 2016].

Variant Calling

Variant calling is the process of accurately identifying the differences or variations between the sample and the reference genome sequence. The typical input is a set of aligned reads in BAM or another similar format, which is traversed by the variant caller to identify sequence variants. Variant calling is a heterogeneous collection of algorithmic strategies based on the types of sequence variants, such as single-nucleotide variants (SNVs), small insertions and deletions (indels), copy number alterations, and large structural alterations (insertions, inversions, and translocations). The accuracy of variant calling is highly dependent on the quality of called bases and aligned reads. Therefore, pre-variant calling processing, such as local realignment around expected indels and base quality score recalibration, is routinely used to ensure accurate and efficient variant calling. For SNVs and indels, the called variants are represented using the de facto standard variant call format (VCF; *https://samtools.github.io/hts-specs/VCFv4.3.pdf*, last accessed November 23, 2016). Alternative specifications exist for representing and storing variant calls [Genomic VCF Conventions, *https://sites.google.com/site/gvcftools/home/about-gvcf/gvcf-conventions*, last accessed November 23, 2016; The Sequence Ontology Genome Variation Format Version 1.10, *https://github.com/The-Sequence-Ontology/Specifications/blob/master/gvf.md*, last accessed November 23, 2016; The Human Genome Variation Society, Human Genome Variation Society (HGVS) Simple Version 15.11. 2016, *http://varnomen.hgvs.org/bg-material/simple*, last accessed November 23, 2016; Health GAfGa File Formats, *http://genomicsandhealth.org/working-groups/our-work/file-formats*, last accessed November 23, 2016].

Variant Filtering

Variant filtering is the process by which variants representing false-positive artifacts of the NGS method are flagged or filtered from the original VCF file on the basis of several sequence alignment and variant calling associated metadata (eg, mapping quality, base-calling quality, strand bias, and others). This is usually a postvariant calling step, although some variant callers incorporate this step as part of the variant calling process. This automated process may be used as a hard filter to allow annotation and review of only the assumed true variants.

Variant Annotation

Variant annotation performs queries against multiple sequence and variant databases to characterize each called variant with a rich set of metadata, such as variant location, predicted cDNA and amino acid sequence change (HGVS nomenclature), minor allele frequencies in human populations, and prevalence in different variant databases [eg, Catalogue of Somatic Mutations in Cancer, The Cancer Genome Atlas, Single-Nucleotide Polymorphism (SNP) Database, and ClinVar]. This information is used to further prioritize or filter variants for classification and interpretation.

Variant Prioritization

Variant prioritization uses variant annotations to identify clinically insignificant variants (eg, synonymous, deep intronic variants, and established benign polymorphisms), thereby presenting the remaining variants (known or unknown clinical significance) for further review and interpretation. Clinical laboratories often develop variant knowledge bases to facilitate this process.

Some clinical laboratories choose to apply hard filters on called variants on the basis of variant call metadata or from a data dictionary (variant filtering) as a component of the pipeline analysis software. Because its purpose is to hide certain variants from the view of the human interpreter, it is absolutely critical that filtering algorithms be thoroughly validated to ensure that only those variants meeting strict predefined criteria are being hidden from view. Otherwise, the human interpreter may miss clinically significant variants that may result in harming the patient.

As NGS technology evolves and its applications extend to multiple clinical areas, bioinformatics analysis algorithms and requirements and principles for NGS data analysis change, requiring active assessment and validation adjustments. In addition, customization of the bioinformatics pipelines on the basis of technology platform, assay type, and variant types should be carefully considered and validated.

NGS is attractive because it rapidly and effectively generates a high number of reads to cover target genomic regions. However, its accuracy is limited by read length and the ability to map reads to the reference genome, particularly in difficult regions (eg, low-complexity or high-repeat homopolymers and high GC content) where sequence alignment can be biased and error prone (ie, aligning the less complex and no-repeat reads preferentially over the tricky regions). Although these limitations might be acceptable for research projects using NGS for prescreening or discovery purposes, they are not suitable in clinical practice, where small errors may lead to severe consequences for patient diagnosis and/or treatments. Therefore, the bioinformatics pipeline component of a clinical NGS test needs to be rigorously validated to ensure it is accurate and reproducible, detects variants within the intended assay, and does so within the established limits of detection to ensure accurate reporting of analytic sensitivity and specificity.

Materials and Methods

Working Group Composition

The AMP Clinical Practice Committee, in collaboration with the Informatics Subdivision Leadership, gathered an expert working group consisting of members with expertise and experience in NGS testing for clinical patient care and in molecular bioinformatics, clinical informatics, and pipeline analysis. Members included practicing medical geneticists, molecular pathologists, and bioinformaticists from the United States. The AMP approved the appointment of the project chair and working group members. The AMP requested organizational representatives from the College of American Pathologists and the American Medical Informatics Association, who participated as full members of the AMP working group. All subject-matter expert working group members complied with the AMP conflicts of interest policy, which required disclosure of financial or other interests that may have an actual, potential, or apparent conflict throughout the project.

AMP Working Group Charge and Scope

This expert working group recommends factors and best practice guidelines for analytical validation of NGS bioinformatics pipelines for detection of SNVs, indels, and multinucleotide substitutions (delins in HGVS terminology) comprising a length of 21 bp or less from both somatic and germline human origin (herein referred to as small sequence variants). Small sequence variants were chosen for guideline development because they are the most commonly identified variants using NGS bioinformatics pipelines. Indels measuring up to 21 bp in length should be reliably detected by most NGS bioinformatics pipelines and comprise the vast majority of small insertions and deletions known to be clinically significant (eg, *EGFR* insertions and deletions in exons 18 to 21).[7] These guidelines encompass NGS bioinformatics analyses beginning from unaligned sequences (eg, FASTQ file) and ending with a list of variant calls (eg, VCF file) with basic annotation (ie, variant type and location and HGVS nomenclature) that is ready for further annotation and subsequent classification and interpretation by a laboratory professional (Figure 1).

Limitations of this Publication

These guidelines do not address clinical validation or clinical utility of an NGS test. Two published reports by the AMP have already addressed both clinical utility and NGS clinical validation.[8,9] For the purposes of this article, we begin with the assumption that the NGS clinical validation was performed in a manner consistent with these previously established guidelines. It also does not encompass advanced variant annotation, prioritization, and interpretation. For inherited conditions, this has been addressed by the Interpretation of Sequence Variants guideline from the American College of Medical Genetics and Genomics and the AMP.[10] The AMP Interpretation of Sequence Variants Working Group, composed of subject-matter experts from the AMP, American Society of Clinical Oncology, College of American Pathologists, and American College of Medical Genetics and Genomics, recently addressed these concepts in somatic conditions, resulting in publication of joint consensus guidelines.[11] This working group acknowledges that many NGS tests are now reporting mutational signatures (eg, microsatellite instability and tumor mutation burden) as biomarkers of response to checkpoint inhibition, on the basis of SNVs, indels, and other features, but does not specifically address these concepts.

In addition, these guidelines do not address the analytical validation of bioinformatics pipelines for large indels >21 bp in length, structural variants (inversions and translocations), gene fusion variants and translocations, gene expression variations, epigenetic variants, copy number alterations, and other variants not defined as SNVs or small indels (herein referred to as large variants). Bioinformatics pipelines designed to detect large variants may be different and less common than general purpose, small sequence variant calling algorithms. Although there is an increased interest in using NGS sequencing and bioinformatics pipelines for the detection of large variants, the experience related to performance characteristics in the bioinformatics community and specifically in the clinical domain is currently limited but rapidly evolving. As such, including guidelines for validation of large variant pipelines would significantly increase the degree of complexity of this project, and guidelines were recommended by the working group to be addressed independently. Although large variants are out of scope for the specific recommendations presented in this document, many of the high-level principles in these guidelines are appropriate for pipeline validation for large variant types. Until guidelines for large variants are available, this document serves as a framework for good pipeline validation principles and practice at the discretion of the laboratory director.

Systematic Evidence Review

The working group performed a systematic review of peer-reviewed published literature in a randomized and double-blinded manner to understand the existing practice regarding the Q10 bioinformatics pipeline validation process in the community. Criteria for inclusion and exclusion of articles were developed by the working group and are listed in Table 1. The medical [T1] literature was searched for publications specifically relating to analytical validation of bioinformatics pipelines by querying PubMed (National Library of Medicine, *http://www.ncbi.nlm.nih.gov/pubmed*, last accessed January 28, 2017). Individual queries submitted to PubMed are listed in Table 1. The queries were designed to maximize capture of articles meeting inclusion criteria and not meeting exclusion criteria (ie, relevant items). In information retrieval science, the terms precision and recall are used to assess the completeness of the information retrieval.[12] Precision is the positive predictive value of the items retrieved (ie, the percentage of relevant items retrieved of all items retrieved). Recall is the sensitivity of the items retrieved (ie, the percentage of relevant items retrieved of all relevant items that exist). The initial query retrieved many false positives (nonrelevant items) that required subsequent refinement of query language to improve precision and recall before manual review.[12] After query refinement, two members of the committee performed a title-only review of the retrieved items and excluded any that were related to nonvalidation activities or that used non-NGS methods or nonhuman samples. Duplicate items were also removed (Figure 2). The resulting items were prepared [F2] for phase 1 of manual review.

In phase 1 of the manual review process, the abstract and the title of each retrieved item were reviewed independently by two different members of the working group. Members of the working group were charged with determining whether the study in the article was relevant, meaning that the article met inclusion criteria and concomitantly did not meet any of the exclusion criteria based on the data in the abstract. When evaluation by the two reviewers was not congruent, a third

Table 1 Systematic Review Method

Description	Details
Individual PubMed queries conducted	Next-generation sequenc*[All Fields]; massively parallel DNA sequenc*[All Fields]; high-throughput nucleotide sequencing[MeSH Major Topic]
Filter terms applied to the initial list of articles from the PubMed queries	Valid*[All Fields] OR oncology[All Fields] OR clinical[All Fields] OR tumor [All Fields] OR tumour [All Fields] OR cancer [All Fields] OR neoplasm [All Fields] OR performance [All Fields] OR analysis [All Fields] OR characteristic [All Fields] OR evaluation [All Fields] OR targeted [All Fields]
Inclusion criteria	Study included analysis of each of the following: human DNA; NGS method; verification of detection of variants (somatic and/or germline) via orthogonal method; verification of detection of SNVs and/or small indels (≤21 bp) and/or multiple adjacent (complex) variants occurring within 21 bp of contiguous length
Exclusion criteria	Study included analysis of ONLY one or more of the following: nonhuman DNA or RNA; human RNA or proteins; non-NGS method(s); detection of variants with no verification of accuracy by an orthogonal method or a description of a process with no validation; verification of detection of large indels (>21 bp) and/or structural variants, including fusion variants and/or gene expression variants and/or epigenetic variants and/or CNAs and/or other variants outside of the inclusion criteria

*Wildcard character meaning that the characters starting at the * can be anything.

[All Fields], all PubMed fields were searched for the term indicated; CNA, copy number alteration; indel, insertion/deletion; [MeSH Major Topic], only the medical subject heading major topic in PubMed was searched for this term; NGS, next-generation sequencing; SNV, single-nucleotide variant.

working group member was assigned to review the abstract, also in a blinded manner, and generate a final decision. Q11

Abstracts identified as relevant in phase 1 were evaluated in phase 2. Phase 2 of the review process was similar to phase 1, with the following differences: The full text of the article was reviewed instead of the abstract, and the reviewers were asked to capture a list of consensus-derived predefined data elements for each article (Supplemental Table S2). The data elements were predefined in a survey using Survey Monkey (Palo Alto, CA). Only those articles that were deemed appropriate after phase 2 of the review process were included in the systematic review.

The working group met on a biweekly basis by conference call to review published evidence and develop the article. Several face-to-face working group meetings were held during the AMP 2015 and 2016 Annual Meetings. On the basis of the results of the systematic evidence review and the cumulative practice experience of the members of the working group, the guideline statements were developed by expert opinion consensus (majority vote) of the working group.

Results

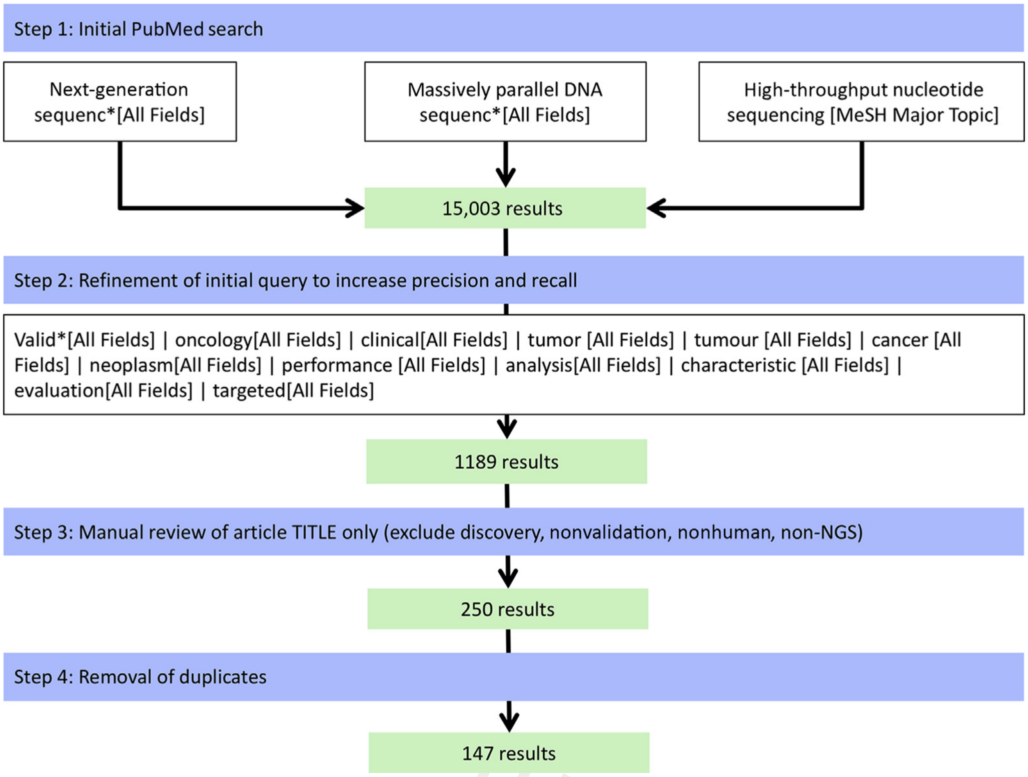
A total of 15,003 cumulative articles were retrieved from the initial search. Subsequent query refinement (Table 1), review of article titles, and removal of duplicates yielded 147 potential articles for phase 1 of the manual review process (Figure 2). Of 147 article abstracts manually reviewed in phase 1, 21 were deemed appropriate by consensus of the two initial reviewers. Five articles required a third review, and of these, three articles were deemed appropriate for inclusion. Therefore, a total of 24 articles were selected for phase 2 of the review. Fourteen articles were selected for systematic review after exclusion of nine articles during the phase 2 review (Supplemental Table S3).[13−26]

The systematic review revealed a clear absence of uniformity in the methods adopted for bioinformatics pipeline validation. Sequencing technologies used for validation of the assays included optical- and semiconductor-based platforms. Although most of the studies used a gene-panel approach (3 to 297 genes), one study validated a whole-exome sequencing assay. The number of samples used for validation ranged from 5 to 297. Cell lines and formalin-fixed, paraffin-embedded tissues were the most common sample types used for validation. Others included peripheral blood, bone marrow, body fluids, and frozen tissue. Minimum depth of coverage and allelic fraction ranged from 20× to 500× and from 2.5% to 10%, respectively. Reporting of performance characteristics (ie, analytic sensitivity, analytic specificity, and positive predictive value) was highly variable. Performance characteristics were not stratified by variant type in some of the studies. The reporting of CI for each Q12 of the performance characteristic measures was inconsistent among the reviewed studies.

The limited number of NGS pipeline validation studies and the high degree of variability between studies certainly reinforce the need for guidance in NGS bioinformatics pipeline validation. Consequently, levels of peer-reviewed evidence may not be associated with any of the guideline statements given later, and thus this initial published guideline is dependent on the expertise and experience of the members of the committee. It is hoped that subsequent validation studies are performed and published using this guideline as a framework, the results of which may be used to refine a later version of this standard as needed.

Guideline Statements and Discussion

Human variants of both somatic and germline origin are within scope because NGS tests developed for somatic variant detection also detect germline genetic variants occurring within the


Step 1: Initial PubMed search
Next-generation sequenc*[All Fields]
Massively parallel DNA sequenc*[All Fields]
High-throughput nucleotide sequencing [MeSH Major Topic]
15,003 results
Step 2: Refinement of initial query to increase precision and recall
Valid*[All Fields] | oncology[All Fields] | clinical[All Fields] | tumor [All Fields] | tumour [All Fields] | cancer [All Fields] | neoplasm[All Fields] | performance [All Fields] | analysis[All Fields] | characteristic [All Fields] | evaluation[All Fields] | targeted[All Fields]
1189 results
Step 3: Manual review of article TITLE only (exclude discovery, nonvalidation, nonhuman, non-NGS)
250 results
Step 4: Removal of duplicates
147 results

print & web 4C/FPO

Figure 2 Systematic literature review. The figure summarizes the strategy that was used to search the PubMed database for identifying articles for systematic literature review. The **asterisk** indicates a wildcard character meaning that the characters starting at the asterisk can be anything. [All Fields], all PubMed fields were searched for the term indicated; NGS, next-generation sequencing.

same regions. Many of the guideline statements and/or broader concepts from this guideline will be broadly applicable to NGS bioinformatics pipeline development, as determined by an experienced and qualified molecular laboratory professional. These guidelines should assist clinical laboratories in ensuring [T2] the quality and accuracy of NGS test–based results. Table 2 summarizes a complete list of the consensus guideline statements. The discussion presents each statement to assist in comprehension and implementation within clinical molecular diagnostic laboratory practice.

Recommendation 1: Clinical Laboratories Offering NGS-Based Testing Should Perform Their Own Validation of the Bioinformatics Pipeline

Most clinical laboratories performing NGS testing establish their bioinformatics pipelines by adopting algorithms and software tools that have been developed either in the academic community or by commercial vendors. In some cases, software is adopted as configured by the external source, or the laboratory may alter algorithm or software parameters for its particular application. Furthermore, in establishing a bioinformatics pipeline, a clinical laboratory may use more than one algorithm or software to generate a complete analysis pipeline. Examples would include the use of specific software for alignment, followed by different software for variant calling and variant annotation. Once the laboratory has set up its initial bioinformatics pipeline, it is imperative that the laboratory determines its performance characteristics for the specific clinical test purpose. This can be addressed by assessing performance in a pilot study using samples or reference materials with known variants. During this phase, the initial performance of the pipeline can be determined. The pipeline may perform adequately in its initial configuration. Alternatively, the pipeline may need refinement, and whether refinement can be achieved is dependent on whether relevant parameters can be altered in the software. Modifications of software may require programming expertise; in some commercial software, options for modifying parameters can be achieved through a graphical user interface. Through this iterative process, the performance and limitations of the pipeline are determined.

Recommendation 2: A Qualified Medical Professional with Appropriate Training in NGS Interpretation and Certification Must Oversee and Be Involved in the Validation Process

The interpretation of technically complex NGS data requires highly specialized personnel. Therefore, it is our recommendation that a medical molecular professional (eg, molecular laboratory director) with appropriate training and certification must be involved and oversee the analytical validation of the bioinformatics pipeline.[27] These medical molecular professionals should have adequate training and experience in sequence

Table 2 Consensus Recommendation Statements for NGS Bioinformatics Pipeline Validation

Recommendation number	Statement
1	Clinical laboratories offering NGS-based testing should perform their own validation of the bioinformatics pipeline
2	A qualified medical professional with appropriate training in NGS interpretation and certification must oversee and be involved in the validation process
3	Validation must be performed only after completion of design, development, optimization, and familiarization of the bioinformatics pipeline and its components
4	Bioinformatics pipeline validation should closely emulate the real-world environment of the laboratory in which the test is performed
5	Validation should include all individual components of the bioinformatics pipeline used in the analysis, and each component must be reviewed and approved by an appropriately qualified medical molecular professional and the laboratory director
6	The design and implementation of the bioinformatics pipeline must ensure the security of identifiable patient information and be compliant with all applicable laws at the local, state, and national levels
7	Validation of the NGS bioinformatics pipeline must be appropriate and applicable for the intended clinical use, specimen, and variant types detected of the NGS test
8	Laboratories must ensure that the design, implementation, and validation of the bioinformatics pipeline are compliant with applicable laboratory accreditation standards and regulations
9	The bioinformatics pipeline is part of the test procedure, and its components and processes must be documented according to laboratory accreditation standards and regulations
10	The identity of the sample must be preserved throughout each step of the NGS bioinformatics pipeline with a minimum of four unique identifiers, including a unique location identifier within the content of each data file read and/or generated by the pipeline
11	Specific quality control and quality assurance parameters must be evaluated during validation and used to determine satisfactory performance of the bioinformatics pipeline
12	The methods used to alter or filter sequence reads at any point in the bioinformatics pipeline before interpretation must be validated to ensure that the data presented for interpretation accurately and reproducibly represent the sequence in the specimen, and full documentation of these methods must be kept as part of the test documentation according to laboratory accreditation standards and regulations
13	Laboratories must include specific measures to ensure that each data file generated in the bioinformatics pipeline maintains its integrity and provides alerts for or prevents the use of data files that have been altered in an unauthorized or unintended manner
14	*In silico* validation can be used to supplement the validation of the bioinformatics pipeline but shall not be used in lieu of end-to-end validation of the bioinformatics pipelines using human samples
15	Validation of the bioinformatics pipeline must include confirmation of a representative set of variants with high-quality independent data; appropriate validation metrics by variant type should be reported
16	Clinical laboratories must ensure the accuracy of software-generated HGVS variant nomenclature and annotations and have an alert in place to indicate when the software-generated nomenclature and annotations need to be manually reviewed and/or corrected, and documentation of any corrections must be maintained
17	Supplemental validation is required whenever a significant change is made to any component of the bioinformatics pipeline

HGVS, Human Genome Variation Society; NGS, next-generation sequencing.

technologies, specifically NGS, and interpretation of sequence variations. Oversight encompasses directing the design, familiarization, optimization, and validation of the NGS analysis, including the NGS bioinformatics pipelines, and having the authority to review, approve, or reject the final validation results. Design, review, and approval of the various computational algorithms that compose an NGS bioinformatics pipeline may require specialized expertise that the medical molecular professional, at his or her discretion, may delegate to an individual with appropriate knowledge and experience with NGS pipelines in the clinical setting. However, the medical molecular professional approving the final validation results remains responsible for oversight of the entire analysis, including the pipeline components.

Recommendation 3: Validation Must Be Performed Only after Completion of Design, Development, Optimization, and Familiarization of the Bioinformatics Pipeline and Its Components

A bioinformatics pipeline is composed of a wide array of software algorithms to process raw sequencing data and generate a list of annotated sequence variants. Bioinformatics pipelines are either designed and developed by a vendor with or without customization by the laboratory or entirely developed by the laboratory. In the latter scenario, the laboratory must conceive the algorithmic approaches to process sequencing data, on the basis of available sequencing platform and intended clinical use, and choose appropriate bioinformatics pipeline components after thorough evaluation of available software (open source or proprietary) or generation of custom scripts. This phase of design and development should also include the choice of computer resources and operation environment. After the pipeline is developed or the laboratory opts to use a vendor-provided pipeline, the phase of optimization and familiarization (O&F) must follow.[9] During the familiarization process, the vendor-provided or custom-developed pipeline should be executed as a pilot to review the output at intermediate and final steps to systematically evaluate expected outcome, identify unanticipated errors, and determine scope of performance improvement. In addition, for the vendor-provided pipeline, software release notes, factory default settings for pipeline parameters, and QC cutoffs must be reviewed and documented. During the optimization phase, parameters for the bioinformatics pipeline may be modified or a newer software version may be used to achieve wanted performance. Data from sequencing of physical samples, well-characterized reference material, and *in silico* data sets may be used during the O&F phase. It is imperative that the pipeline validation is initiated only after completion of one or more rounds of O&F. During the process of pipeline validation, the configuration established and software versions used during the O&F phase must be locked down. If modifications to the pipeline are required, on the basis of the validation results, this resets the process to the O&F phase and, therefore, will require performing validation again.

Recommendation 4: Bioinformatics Pipeline Validation Should Closely Emulate the Real-World Environment of the Laboratory in which the Test Is Performed

The validation of the bioinformatics pipeline should be conducted in a manner that mimics how the NGS test will be performed for clinical patient care. Before validation of an NGS pipeline begins, the laboratory should become familiar with the overall NGS analysis. During the O&F phase, the design, workflow, components, and versions of the bioinformatics pipeline should be established to perform as it would be expected to perform in the clinical patient care setting (ie, after going live). The workflow and design of the bioinformatics pipeline before validation should, therefore, take into account the intended use of the test, the analytes to be examined, and the variants that are expected to be reproducibly and accurately detected. The location, versions, and backups of the hardware, software, transmission, and network resources to be used and by whom (user authorization) should also be considered.

For example, if the NGS pipeline is expected to be executed in production on a server running a specific operating system, database, application settings, network protocol, and set of pipeline algorithms in the institution's data center, then validation should be performed using software and hardware with the exact same configuration, versioning, and physical location as expected after go live (ie, in a clinical production environment). Performing validation with the bioinformatics pipelines running on a smaller server directly attached to the NGS instrument via Ethernet cable and then switching to the data center server configuration when the test is taken into production is not acceptable because of the number of computational inconsistencies in the hardware, operating system software, network configuration, and data workflow between the validation pipelines and the production pipelines. Similarly, using one set of pipeline algorithms during validation and then switching to different pipelines, altering the existing pipelines (even if just one line of computer code), or changing the order in which the algorithms or data flow is executed without performing an appropriate level of revalidation is also not acceptable for clinical patient care.

As with any computerized system, a procedure to perform disaster recovery should be developed and validated in the same manner that it would be expected to perform in production. Turnaround times should also be analyzed as part of the validation process to confirm that the pipeline algorithms, some of which can be long, are performing within expected and realistic parameters.

Recommendation 5: Validation Should Include All Individual Components of the Bioinformatics Pipeline Used in the Analysis, and Each Component Must Be Reviewed and Approved by an Appropriately Qualified Medical Molecular Professional and the Laboratory Director

A bioinformatics pipeline for NGS assay is most commonly executed within the performing laboratory. However, there is an increasing trend for clinical laboratories to outsource one or more components (algorithms) of the bioinformatics pipeline or the entire bioinformatics pipeline, including hardware infrastructure (storage and computer resources), to third-party service providers.[2] In such a scenario, the raw NGS sequencing data yield from the sequencers is transferred (uploaded) to the bioinformatics service provider that returns the identified variants and related metadata back to the clinical laboratory for interpretation and reporting. Although the bioinformatics pipeline is developed, validated, and hosted by the service provider, it is still an integral part of the NGS assay offered by the clinical laboratory (ie, performance of the bioinformatics pipeline will have direct impact on the overall performance of the clinical NGS assay). Therefore, the use of external bioinformatics pipeline services by a clinical laboratory must include the processes of optimization and familiarization and subsequent validation of the outsourced service by evaluating its performance characteristics in the context of the intended clinical use of the NGS assay. Before approval, the laboratory director should ensure that components of a pipeline or the entire pipeline, executed outside the physical location of the laboratory performing the test, is compliant with applicable law and accreditation standards. In addition, an appropriately qualified medical molecular professional satisfying the requirements of Recommendation 2 should ensure that each component of the bioinformatics pipeline, including those performed outside the physical location of the laboratory, is appropriate and included in the validation of the entire pipeline. The medical molecular professional may delegate review of pipeline components to someone with expertise in NGS bioinformatics pipelines as needed, but the medical molecular professional remains responsible for oversight of the entire analysis, including the pipeline components.

Recommendation 6: The Design and Implementation of the Bioinformatics Pipeline Must Ensure the Security of Identifiable Patient Information and Be Compliant with All Applicable Laws at the Local, State, and National Levels

Many countries have laws that restrict the use, disclosure, storage, and transport of patient information. In the United States, federal law has become increasingly restrictive on the use, disclosure, and security of individually identifiable health information, specifically protected health information, with additional specific restrictions regarding genetic information. These include, in chronological order, the following: the Health Insurance Portability and Accountability Act of 1996 (HIPAA; US Government Publishing Office, *https://www.gpo.gov/fdsys/pkg/PLAW-104publ191/html/PLAW-104publ191.htm*, last accessed March 23, 2017), the Health Insurance Reform: Security Standards (Final Security Rule; US Government Publishing Office, *https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/securityrule/securityrulepdf.pdf*, last accessed December 19, 2016), the Health Information Technology for Economic and Clinical Health Act (US Health and Human Services, *https://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf*, last accessed November 7, 2017), the Genetic Information Nondiscrimination Act (US Government Publishing Office, *https://www.hhs.gov/ohrp/regulations-and-policy/guidance/guidance-on-genetic-information-nondiscrimination-act/index.html*, last accessed December 19, 2016), and the HIPAA Omnibus Rule (US Government Publishing Office, *https://www.gpo.gov/fdsys/pkg/FR-2013-01-25/pdf/2013-01073.pdf*, last accessed December 19, 2016). The HIPAA Omnibus Rule made a change to HIPAA to indicate that genetic information, as defined under the Genetic Information Nondiscrimination Act, was included in the list of identifiers that had to be removed from protected health information to consider it deidentified data. This puts genetic information in the same class of sensitive identifiers as medical record number, patient name, or date of birth. The HIPAA Final Security Rule specifies administrative, physical, and technical safeguards for protected health information that are stored or transmitted electronically, and these include genomic data. Countries outside the United States also have privacy laws for health information, including the following: the Australian Privacy Act of 1988 (Office of the Australian Information Commissioner, *https://www.oaic.gov.au/privacy-law/privacy-act*, last accessed August 29, 2016), the European Union General Data Protection Regulation (EUR-Lex, *http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC*, last accessed August 29, 2016), and the Canadian Privacy Act and Personal Information Protection and Electronic Documents Act (Office of the Privacy Commissioner of Canada, *https://www.priv.gc.ca/resource/fs-fi/02_05_d_15_e.asp*, last accessed August 29, 2016).

Laboratories must be aware of and able to comply with local, state/province, and national health privacy laws, including privacy laws specifically directed toward genetic information. This may require consultation with the organization's legal and information technology representatives. Many vendor-based solutions were developed for research environments and may not comply with regulations pertaining to clinical patient data. Although uploading and managing such large data sets in the cloud is attractive both financially and administratively for many laboratories, such arrangements should be made with caution and only after consideration of legal compliance. For example, vendors of cloud infrastructure may provide HIPAA-compliant servers for added cost, but HIPAA compliance in these instances is limited to the physical safeguards of the servers. Laboratories in the United States are expected to ensure that the data are managed according to the administrative and technical safeguards as well (including data access audit trails). The physical location of the cloud-based servers is important for adherence to privacy laws in many countries. For example, data located on cloud servers in the United States are subject to discovery via the Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2001 (US Government Publishing Office, *https://www.congress.gov/bill/107th-congress/house-bill/3162*, last accessed November 7, 2017). Data located on cloud servers outside the United States may not be maintained in the HIPAA-compliant manner. Data maintained on any unsecured server may be at risk of a security breach. Several other articles[2,28,29] are available for review, which have more information on security and privacy of health information.

In addition to privacy of data, the definition of security of health data encompasses measures to ensure that those permitted to access data are able to on an ongoing basis. Precautions to prevent sudden and/or prolonged disruptions of access to data are required under the HIPAA Final Security Rule. Specifically, the HIPAA Final Security Rule requires all covered entities and business associations to have a data backup plan, disaster recovery plan, emergency mode operation plan, contingency operations for failed hardware, data backup and storage, and data integrity controls.

The Clinical Laboratory Improvement Amendments of 1988 (CLIA; Electronic Code of Federal Regulations Part 493—Laboratory Requirements: Clinical Laboratory Improvement Amendments of 1988, *https://www.ecfr.gov/cgi-bin/text-idx?SID=1248e3189da5e5f936e55315402bc38b&node=pt42.5.493&rgn=div5*, last accessed September 28, 2017) requires laboratories to be able to access validation information and test records for certain periods of time. Unlike the HIPAA Final Security Rule, CLIA is more specific about which specific files and pieces of data must be retained in a retrievable state, beginning with analytic systems records [CLIA 42 CFR § 1105(a)(3)]:

> (3) Analytic systems records. Retain quality control and patient test records (including instrument printouts, if applicable) and records documenting all analytic systems activities specified in §§493.1252 through 493.1289 for at least 2 years. In addition, retain the following: (i) Records of test system performance specifications that the laboratory establishes or verifies under §493.1253 for the period of time the laboratory uses the test system but no less than 2 years.

Q13

As with all clinical testing, the laboratory medical director and/or designee must determine which files and records for the test should be generated, stored, updated, and backed up, and the laboratory must have an established plan to recover from disaster, such as a catastrophic hardware or software failure. Laboratories should ensure that they follow their local, regional, and national laws with regard to the security, including accessibility, of health data, including NGS data. In countries where such regulations are lacking, the working group recommends following the US CLIA requirements for maintaining records and files at a minimum. Additional information regarding data management, storage, backup, and disaster recovery has been published elsewhere.[2]

Recommendation 7: Validation of the NGS Bioinformatics Pipeline Must Be Appropriate and Applicable for the Intended Clinical Use, Specimen, and Variant Types Detected by the NGS Test

Once the intended use, sample types, detectable variant types (eg, small sequence variants and insertions/deletions up to a specific number of base pairs), and variant allele fraction reference ranges have been determined for an NGS analysis and an appropriate system configuration has been designed, optimized, and approved by a medical molecular professional meeting the criteria in Recommendation 2, then the selection of samples for validation should begin.[9] These samples should be selected to verify the analytical sensitivity and positive predictive value of the pipeline(s). In addition, they must be appropriate for the intended use of the test, including the specimen types to be examined and the variant types that are expected to be reproducibly and accurately detected. Taken together, the combination of all samples used in the validation of the pipeline is the validation sample set.

Recommendation 15 describes specifics regarding the inclusion of representative variants in the validation sample set, and Recommendation 14 details the permissible use of *in silico* samples during validation. Recommendation 17 describes the requirement for appropriate revalidation when any component of the assay changes. These aspects of validating the test for its intended clinical use, specimen, and variant types will be covered in those recommendations.

LOD and Variant Allele Fraction Reference Ranges

The validation sample set should include representative samples with one or more variants having depth of coverage and allelic fractions at the test's intended limit of detection (LOD). For an NGS bioinformatics pipeline, the LOD is represented by two data points: the minimum required depth of coverage at the variant site and the minimum variant allele fraction, both of which have to be satisfied before a variant can be positively called. Depth of coverage may be affected by several variables in an NGS sequencing reaction, but local sequence context (particularly GC/AT ratio), insufficient nucleic acid, poor quality nucleic acid, and reaction inhibitors are common reasons for failure to achieve minimum depth of coverage. Variant allele fraction may also vary for several reasons, but the most common reason for low variant allele fraction is a low percentage of cells in the sample harboring the variant (eg, mostly normal tissue with a small focus of cancer). Therefore, validation samples should be included that are at the intended LOD for both depth of coverage and variant allele fraction to ensure that the test's performance characteristics are as expected. Validation samples intended for this purpose may naturally harbor the variant at low allelic fraction, or they may be artificially designed. Artificial design can occur by limiting the input quantity of nucleic acid to a lower depth of coverage or by diluting a sample harboring variants at a higher allele fraction with normal (reference) DNA to achieve a lower allele fraction. It is insufficient to set a minimum variant allele fraction when the minimum depth of coverage has not been specified and vice versa. Validation of the LOD is critical for all specimens but is especially critical when low-cellularity samples, such as pancreatic cysts or cerebrospinal fluid, are included in the validated sample types for the analysis.

If additional reference ranges have been set to determine germline heterozygosity, homozygosity, hemizygosity, or mosaicism, then the validation sample set should include clinically significant variants with variant allele fractions within each established reference range to ensure that these variants are detectable within the correct reference range for accurate interpretation.

Contiguous Genetic Regions

Contiguous genetic regions within the region of interest of an NGS assay may have different sequence contexts (eg, low complexity, GC-rich, and homopolymeric sequences). Such sequence characteristics pose challenges in adequately sequencing and detecting variants in those genetic regions. Therefore, the sequence quality for each contiguous genetic region included in the NGS assay should be analyzed across the cohort of validation samples to identify poorly sequenced regions. Distribution of quality metrics, such as depth of coverage, base quality, mapping quality, and strand bias, should be included when analyzing contiguous genetic regions. Variants representative of the different genetic regions in the assay, including poorly sequenced regions, should also be included to ensure that the pipeline(s) may accurately detect variants or confirm pipeline limitations. For example, GC-rich regions of the genome (eg, *TERT* promoter and *CEBPA*) are inherently difficult to sequence and frequently result in low coverage, lower base and mapping quality, and high strand bias that ultimately compromise the sensitivity of variant calling.[30] If bioinformatics algorithms for a given sequencing strategy are unable to reach the wanted level of analytic sensitivity and positive predictive value for variant calling in these regions, the laboratory must explicitly state the limitation of the offered NGS test and, if wanted, choose to validate an alternative sequencing method to address variant detection.

Horizontally and Vertically Complex Variants

Because this set of guidelines is specific to small sequence variants, the validation sample set should include SNVs, insertions up to and including 21 bp in length, deletions up to and including 21 bp in length, and horizontally and vertically complex variants. A horizontally complex variant is one in which two or more sequence alterations are present on the same read in close proximity such that they may represent a single complex variant. These variants are frequently represented as deletions-insertions and may result in ambiguous sequence description or HGVS nomenclature. A vertically complex variant occurs when three or more alleles are represented by different sequence reads, typically with or uncommonly without a reference (normal) allele, at the same genomic coordinate or set of coordinates. Variant calling and subsequent accurate variant representation of vertically complex variants are particularly challenging when indels represent one or more of the allelic states. Small sequence variants with vertical complexity may arise from compound heterozygosity, intratumoral genomic heterogeneity, germline mosaicism, and artifacts produced during NGS analysis. Several working group members reported nomenclature errors and false negatives in bioinformatics pipelines confronted with horizontally and vertically complex variants, respectively. Therefore, both types of complex variants are critical to include in the validation sample set to determine whether the pipeline is at risk for nomenclature errors and/or false negatives because of inappropriate filters built into the pipeline. Examples of horizontally and vertically complex variants are shown in Figures 3 and 4 and described in Supplemental Table S1. [F3] [F4]

Variants that Require Additional Algorithms

Some variant types are inherently difficult to detect using general purpose variant detection algorithms because of the complexity of the sequence change or the complexity of the region in the genome where the specific variant type occurs. Specific algorithms are often designed to increase the sensitivity of detecting such variants. For example, *FLT3* internal tandem duplication is a clinically significant variant in acute myeloid leukemia that requires use of specific algorithms outside of routine variant callers for detection.[5,31,32] Therefore, a minimum number of samples harboring such challenging variant types should be included to validate the specific component of the pipeline responsible for its detection.

Minimum Number of Wet Laboratory Samples to Include in the Validation Sample Set

Methods for determining the minimum number and type of so-called wet laboratory validation samples to include are specified within the AMP guidelines for validation of NGS-based oncology panels.[9] If the minimum number of wet laboratory samples is not sufficient to test adequately the pipeline's ability to detect the specified variant types and percentage reference ranges, then additional appropriate validation samples, commercially available reference materials, and/or verified *in silico* samples (see Recommendation 14) may be analyzed.

Minimum Number of Variants to Include in the Validation Sample Set

Although it is neither possible nor reasonable to validate every possible variant that may be detected in even a small NGS panel, it is critical to include enough of the right kinds of variants to ensure that all clinically significant variants that the assay is intended to discover will be detected (selected examples of complex clinically significant variants are described within Supplemental Table S4). For the variants in scope for this guideline, they fall into one of three categories: SNVs, indels, and other. The sequence variants in the last category cannot be classified as SNVs or indels, such as vertically complex variants that are ≤21 bp in contiguous size or variants that cannot be reliably detected using general purpose variant callers and require special algorithmic approaches (eg, *FLT3* internal tandem duplication mutation).

For each category of variant (SNV, indel, and other) that is classified as in scope for detection by the bioinformatics pipeline(s) being used in the analysis, the minimum number of variants in that same category that must be included in the validation sample set can be calculated using the calculation given later. This is the same calculation that is used to

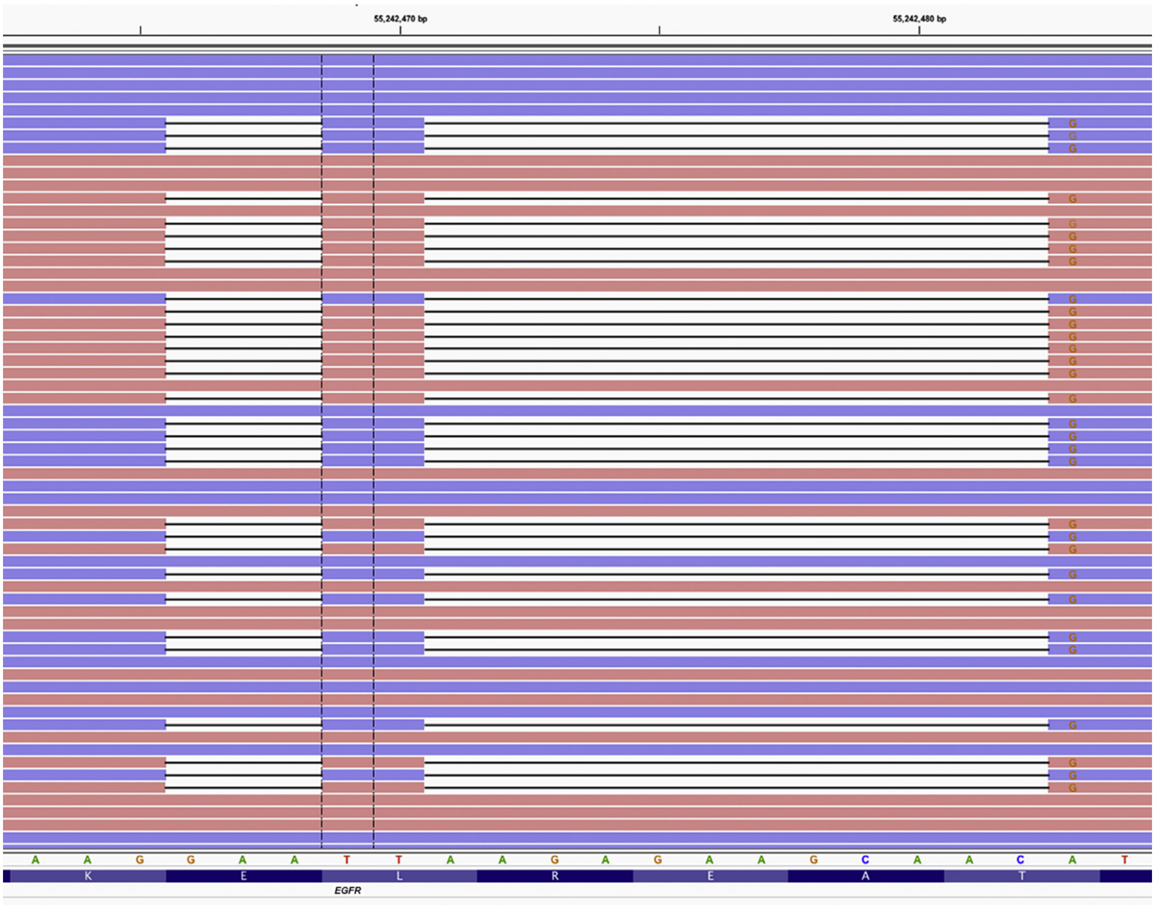


Figure 3 Example of a horizontally complex variant. An example of a horizontally complex variant in exon 19 of the *EGFR* gene (NM_005228.3: c.2236_2253delinsTTG: p.E746_T751delinsL). One deleted sequence is flanked by a deleted sequence on the left and a single-nucleotide change on the right. These individual variants (two deletions and one single-nucleotide substitution) represent a single haplotype, which is a well-characterized activating *EGFR* variant with therapeutic significance.

determine the number of wet laboratory samples in validation of an NGS analysis for cancer (see companion validation article[9]): $n = \ln(1 - CL)/\ln(P)$, where n is the number of variants to be included in the validation sample set, CL is the confidence level of detection, and P is the probability of detection (ie, reliability).

In an analysis that is intended to detect SNVs and indels in the regions analyzed with 95% confidence and 95% reliability, this means that at least 59 different SNVs and 59 different indels must be included in the validation sample set and correctly identified by the pipeline(s) to declare that the pipeline(s) is/are valid for such detection. If complex variants are also in scope for detection in the analysis, then at least 59 other (complex) sequence variants must be included and correctly identified in the validation sample set. Many, if not all, of these variants will have already been included in the validation sample set on the basis of the requirements described above to validate the sample types, limits of detection, reference ranges for allelic fractions, and contiguous genetic regions. However, if the number of variants in the validation sample set is fewer than the minimum criteria set by this calculation in one or more of the variant categories, then additional samples should be added to the validation sample set until the minimum is reached. Should there not be enough samples to validate fully an NGS pipeline for a particular variant category, or representative of a specific sequence context or another aspect of the test, the limitation(s) should be explicitly stated on the clinical report.

Failure to Detect a Variant in the Validation Sample Set

If the pipeline fails to detect any variant included in the validation sample set, then the causes of such failure should be investigated. A visual method of examining the pileups of reads in alignment files (eg, BAM) is required for validation and strongly recommended during routine sign out of NGS variants. An example of a visual viewer is the Integrated Genomics Viewer from the Broad Institute (Cambridge, MA).[33] These visual methods are critical for analysis during validation and are especially important for prospective evaluation of complex variants. If it is determined that the sample containing the variant has been compromised or would otherwise normally be considered inappropriate for testing as part of the routine clinical use of the analysis, then it is appropriate to replace that sample with another sample containing one or more variants of

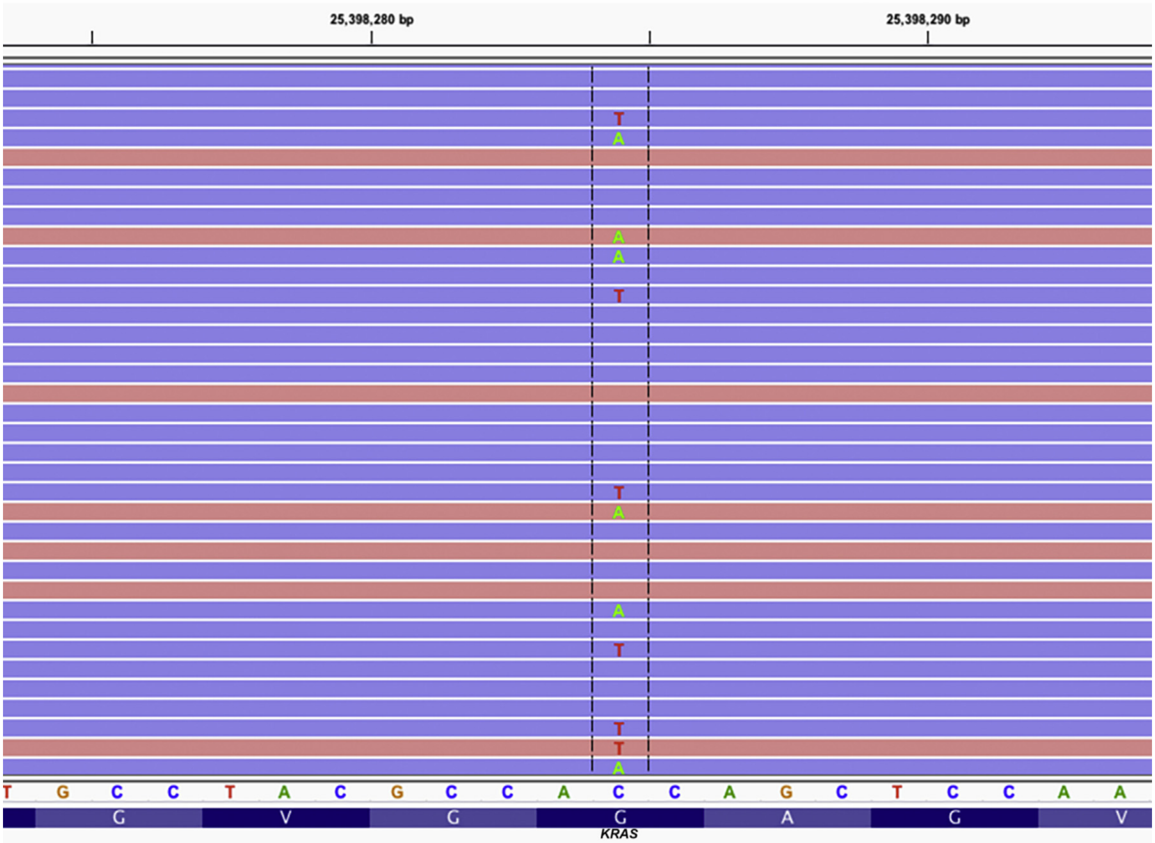


Figure 4 Example of a vertically complex variant. An example of a vertically complex variant in exon 2 of the *KRAS* gene (NM_004985.3: c.35G>A:p.G12D; Q27 NM_004985.3: c.35G>T:p.G12V). The representative image of the sequence pileup demonstrates three alleles (one reference and two alternative alleles) at the same genomic position.

equivalent type. If, however, the sample is not compromised (damaged) and would otherwise be considered appropriate for this test, then an investigation into the pipeline should be made. If changes to the pipeline or any other component of the analysis are required to keep the detection parameters the same, then see the subsection later regarding changes to test components. If changes are not made to the analysis, it is inappropriate to replace a valid sample in which a variant failed to be detected with a new sample containing a different variant. Failed detection of a variant in a valid sample during validation may indicate a flaw in the bioinformatics pipeline, and the options for resolution are either to change the pipeline design with subsequent full validation or to state explicitly such limitation(s) in the clinical report.

Recommendation 8: Laboratories Must Ensure That the Design, Implementation, and Validation of the Bioinformatics Pipeline Are Compliant with Applicable Laboratory Accreditation Standards and Regulations

Within the United States, clinical laboratory accreditation is required by the Centers for Medicare and Medicaid Services, along with certification in compliance with the CLIA. Laboratories may seek accreditation through deemed agencies that have developed requirements for NGS clinical testing that include specific requirements for NGS testing bioinformatics (eg, New York State Department of Health and College of American Pathologists).[34] Ongoing assessment of NGS test quality and performance is required once the test is in clinical service. Recommendations and additional details regarding best practices on these topics have been published elsewhere.[9]

Recommendation 9: The Bioinformatics Pipeline Is Part of the Test Procedure, and Its Components and Processes Must Be Documented according to Laboratory Accreditation Standards and Regulations

Documentation should be maintained that records the method, measurements, settings, parameters, specimen types, histopathologically confirmed tumor types, variant types, and final approval of the NGS pipeline validation to be used in the clinical laboratory.

For bioinformatics pipelines, documentation should specifically include the name, version number, developer, and technical support of each component of the pipeline, including the hardware, software, transmission system, backups, and networks. Software components of the pipeline include not only the individual computational algorithm software components but also the operating system(s), database(s), transmission software, and any other component that is used as part of the pipeline mechanism. If accessible, the source code of each software component should also be recorded, including the names and versions of the programming language(s) and development environment(s) used. For software and/or scripts developed and maintained by the laboratory, appropriate code repository tools (eg, GitHub, mercurial, and subversion) should be used to enforce version control and source code documentation.

This documentation serves three main purposes. It is required for compliance by most laboratory accreditation agencies, it assists information technology staff and bioinformatics professionals with troubleshooting problems, and, most important, it greatly reduces the recovery time after an unplanned system or hardware failure (disaster recovery).

Recommendation 10: The Identity of the Sample Must Be Preserved throughout Each Step of the NGS Bioinformatics Pipeline with a Minimum of Four Unique Identifiers, Including a Unique Location Identifier within the Content of Each Data File Read and/or Generated by the Pipeline

The common data file types used in most bioinformatics pipelines (eg, FASTQ, BAM/sequence alignment/map, and VCF) were developed for research and then moved into clinical use. These file types are de facto standards, and although they are commonly used, their specifications are not rigidly controlled by an international standards organization, unlike other international standards for communication of information (eg, Health Level 7, Logical Observations, Identifiers, Names, and Codes, and Systematized Nomenclature of Medicine—Clinical Terms). As such, multiple variations of sequence, alignment, and variant call format files exist, and deviations from published specifications for syntax are common. Unfortunately, none of these file formats requires inclusion of sample, patient, run, or test location information in the file's metadata or file name. In some cases, there is not any reference to the inclusion of sample identification within the file at all. Without a robust mechanism for identifying the sample (and thereby the patient), NGS results may be assigned to the wrong patient and/or wrong sample. Similarly, without robust identification of the run from which the data were generated, troubleshooting and differentiation of initial versus repeat analyses may be hindered or impossible. Identification of the patient as a separate identifier and the testing location are also important because many laboratories share a common accession number scheme, and sharing analysis files is becoming more common as NGS methods become more widespread. Despite the lack of requirements for patient, sample run, and testing location information in these file format specifications, laboratories must ensure that the identity of the sample is preserved throughout each step of the NGS bioinformatics pipeline with a robust unique identification system that allows tying all intermediate files to the individual sample, patient, run, and testing location. The following four identifiers, with exceptions noted later, must be included in all file types used in each of the common NGS bioinformatics file formats (FASTQ, sequence alignment/map/BAM, and VCF or equivalent): i) a unique sample identifier, ii) a unique patient identifier, iii) a unique run identifier, and iv) a laboratory location identifier.

Each of the identifiers should enable laboratory testing personnel to correctly identify the unique patient, the specimen (down to the aliquot) on which the test was performed, and the unique run that generated the data. If a laboratory uses a specimen/aliquot identifier that is globally unique, such that it would be impossible for another specimen anywhere in the world to have the same identifier, then such a global identifier may replace the requirement for separate patient, sample, and testing location identifiers. The run identifier, however, would still be required for reasons discussed later. The identifiers used must be present within the file's metadata, and the working group recommends that the identifiers are also present in the file name itself. Having the identification in both places will help with troubleshooting in the event that a file name or file data are changed in error.

The first identifier, the unique sample identifier, must uniquely identify the laboratory sample at the laboratory location represented in identifier 4. Similarly, the unique patient identifier should allow laboratory staff at the laboratory location represented in identifier 4 to uniquely identify the patient. Some laboratories perform a safety check by generating a genetic fingerprint for the patient (eg, using SNP genotyping technology) and later checking the NGS data for the patient against the fingerprint. If the genetic fingerprint has been statistically determined to be unique to an individual patient (provided that there is a way to differentiate the patient from any genetically identical siblings), and concise enough to fit into the data file, this data element is acceptable as a unique patient identifier. The unique run identifier should allow the staff at the testing laboratory to identify uniquely the instrument and run on which the sample was analyzed. It should also distinguish an original analysis of the sample from any repeated analyses that were performed on the same sample. In the common situation in which a laboratory's accession numbering system is shared by many laboratories, the laboratory location identifier allows for easy differentiation of in-house versus external samples with the same sample identifier. For example, probably >100 institutions have a case number of S17-123. Pairing a unique sample identifier with a unique patient identifier and laboratory location identifier allows developers to write code that prevents results going to the wrong patient's record. This will become more important as sample exchange and automation in the NGS laboratory become more widespread.

This guideline does not prescribe which specific identifiers should be used nor does it specify how these identifiers should be delimited. However, certain characters should be avoided in both identifiers and delimiters because they are function characters in Health Level 7, which is the international standard for exchange of information between health systems (Health Level Seven International, *http://www.hl7.org*, last accessed March 23, 2017). These

Table 3 Incorporating Four Unique Bioinformatic Identifiers into Sample Identifier Strings

Unique sample identifier	Unique patient identifier	Unique run identifier	Laboratory location identifier*	Delimiter	Final concatenated identifier string
S46-3986B1[†]	AH240963847[‡]	20170123120125-2	0115552438637	Underscore	AH240963847_ S46-3986B1_20170123120125-2_0115552438637
00101123123456	7283947	JKAInstB4thFlr00182	0115556837492	Double space	7283947 00101123123456 JKAInstB4thFlr00182 0115556837492
28394728	Germline fingerprint identifier	Z0G2489v22	0115553620421	Forward slash	28394728/Germline fingerprint identifier/Z0G2489v22/0115553620421

Examples of three identification strings that satisfy the recommendation to include a unique sample identifier, a unique patient identifier, a unique run identifier, and a laboratory location identifier. The laboratory location identifier must indicate the laboratory location where the unique sample identifier was generated. The unique patient identifier must correctly identify the patient at the laboratory location represented by the laboratory location identifier. All identifiers shown are fictional and do not represent any real patient, alive or deceased.

*The international telephone number for the laboratory.

[†]Case S46-3986 tissue block B1.

[‡]Patient location (eg, anywhere hospital) + medical record number.

function characters include the pike or pipe character (|), ampersand (&), carat (ˆ), tilde (∼), pound sign (#), and backslash (\). If the laboratory uses certain characters in any of the data elements expected to be used as identifiers, those characters should be avoided for use as a delimiter. For example, many accession numbers and case numbers use a hyphen (-), in which case a hyphen should be avoided as a delimiter. Examples of how different laboratories may satisfy the identification recommendation are shown in Table 3. This is up to the discretion of the individual laboratory until another standard becomes available for this purpose. Laboratories should consider the use of delimiters and prefixes to ensure correct computation and human readability of each of the file's identifiers.

Recommendation 11: Specific Quality Control and Quality Assurance Parameters Must Be Evaluated during Validation and Used to Determine Satisfactory Performance of the Bioinformatics Pipeline

Quality metrics used in NGS include both QC metrics and quality assurance (QA) metrics. Quality control metrics are computed from either control samples included in each step in the NGS process or a batch of patient samples processed together, whereas QA metrics are computed for each patient sample individually. During the optimization and familiarization processes, these QC and QA metrics should to be evaluated against wanted test performance (eg, sensitivity and positive predictive value) and required performance criteria (minimum and/or maximum values for each metric) established. These criteria should be confirmed during validation to deliver the required test performance. During routine testing, quality controls that fail to meet criteria are indicative of problems with a process step or a batch of samples, whereas QA metrics highlight sample-specific issues or limitations. A list of recommended quality metrics is provided in Table 4.

Both *in silico* and wet laboratory approaches can be used to probe these metrics and establish specific thresholds. These thresholds will vary, depending on the specific methods used and clinical use of the test. Laboratories should consider establishing multiple threshold levels for each metric: for example, at a warning level, results would be flagged for additional review, with a pass/fail decision made by an appropriate technical expert (following a process defined in the laboratory standard operating procedures). At a different level, the results would be failed outright.

Laboratories should establish standard operating procedures describing actions to be taken when QC and/or QA metrics fail to meet performance requirements. In addition to any corrective actions indicated, these standard operating procedures should guide whether the NGS results may be included in a test report or not and (if included) whether additional limitations need to be communicated on the test report(s). The appropriate actions can vary, depending on the specific metric(s) involved, the results, and the clinical goal of the test. Laboratories should continuously monitor these metrics and have systems in place to track them over time, because trends in these metrics can indicate an emerging issue with an NGS process that has not yet manifested itself in failed tests.

We highlight certain specific metrics in the text that follows.

Read coverage is an important and common NGS quality metric, although the specifics of how coverage is calculated matter greatly. For example, the minimum coverage achieved across targeted positions can be a much more important determinant of sensitivity compared with the average coverage across targets. Coverage should be calculated on a per-position basis, rather than per locus, with summary statistics chosen in such a way that diagnostically critical regions are highlighted. Summaries such as 95% of bases had 50× or higher coverage can obscure the fact that a critical region (given the patient's indication) had low coverage and, thus, impaired sensitivity. With hybridization-based capture and whole-genome sequencing, PCR duplicates should be removed from coverage calculations. Ideally, this should be done with amplicon sequencing as well, although specific biochemical methods to enable this (eg, incorporation of random tags) are not part of all commercial kits. Appropriate actions in response to low coverage metrics should be considered carefully: although sensitivity can be affected, variants detected are still likely reportable true positives, particularly if they are confirmed using an orthogonal method.

Flags should be set to review data more closely when vertically or horizontally complex variants are detected within a defined genomic window. Algorithms that determine the HGVS representation of variants frequently misname (ie, use a noncannonical representation for) indels and horizontally complex variants. These HGVS strings should be reviewed by the interpreting professional before release and edited if necessary. Similarly, some pipeline algorithms have been known to filter out two or more variants on different reads but at the same coordinate position (ie, vertically complex variants). This is a potentially dangerous situation when one of the variants is clinically significant and true and the others are artifacts. Rather than filtering such vertical complexity out, these should be flagged for additional review by the interpreting professional.

Identification of possible sample contamination or cross contamination is another quality metric that is particularly valuable for germline tests of medium- to large-sized panels. For example, one can look for low-frequency alleles at otherwise homozygous variants, particularly benign SNPs.[35–37] Although these methods can be adapted to somatic tests, establishing criteria is more difficult in this setting because sample heterogeneity is expected and test targets can be small. An alternative method of detecting cross contamination is to use sample-specific spiked controls.[38,39] For genome and exome sequencing, laboratories may consider adding additional quality metrics (eg, number of transitions/number of transversions, heterozygous/homozygous, and SNP/indel ratios).

Recommendation 12: The Methods Used to Alter or Filter Sequence Reads at Any Point in the Bioinformatics Pipeline before Interpretation Must Be Validated to Ensure That the Data Presented for Interpretation Accurately and Reproducibly Represent the Sequence in the Specimen, and Full Documentation of These Methods Must Be Kept as Part of the Test Documentation according to Laboratory Accreditation Standards and Regulations

The NGS method, by its nature, can generate a significant amount of poor, failed, or false-positive sequences that are not usually seen in other molecular testing methods. One purpose of a bioinformatics pipeline is to filter sequences from view that do not meet the laboratory's predefined quality standards. At the same time, the bioinformatics pipeline must preserve and accurately represent the sequence in the specimen for the variants that the assay has been designed to detect. Bioinformatics pipelines that have been improperly designed and validated are at increased risk of filtering out sequences that are true positives and true negatives, thereby increasing the risk of an erroneous false-positive or false-negative interpretation. Either of these errors may be disastrous for a patient.

Both proprietary and open-source bioinformatics pipelines usually come preconfigured with several filtering and alteration algorithms. These algorithms may be optional but turned on by default (ie, configurable), or they may be mandatory and nonconfigurable. During optimization and familiarization and definitely before validation begins, the laboratory must carefully evaluate the actions of each filtering or alteration algorithm on the sequence data, not only for what the algorithms do but also for what wanted alterations and filters they do not provide. This includes knowing the criteria that trigger sequence data to be valid, clipped, trimmed, filtered, aligned, mapped, and barcoded. If unwanted mandatory algorithms are present, then the laboratory must determine whether a different pipeline should be used altogether. For configurable algorithms, the laboratory must determine whether the default parameters are acceptable, if an alternate set of parameters is more wanted, or if the algorithm should be used at all. If additional algorithms are wanted, then these should be chosen and implemented with care. The function of computational algorithms that alter or filter sequence data will likely change, depending on at what point in the pipeline the algorithm is set to run. For example, an algorithm that trims the last 30 bp of sequence to remove adapters, barcodes, or indexes may remove 30 bp of the patient's sequence on that read if it is called at the wrong time in the pipeline.

If any quality metric, including but not limited to those referenced in Recommendation 11, is used as a criterion for preinterpretation filtering of data, the filtering algorithm must be validated to ensure that it is performing as expected and not filtering out data that contribute to an accurate representation of the sequence in the specimen.

Algorithms may filter false-positive variants that have some special considerations. Two kinds of errors contribute the most false-positive SNPs: a coordinate position, which is heterozygous and is falsely called homozygous because of an ignoring a reference allele error; or a coordinate position, which is homozygous and is falsely called heterozygous because of an adding reference allele error. The frequency of these errors depends on the variant caller used. These errors should be properly characterized during pipeline optimization and familiarization. Other quality metrics for alignment, coverage, read depth, quality scores, variant allele fractions, segmental GC content, presence of homopolymers, and others around the variant site can help identify false-positive calls. During validation, laboratories should develop methods, using these metrics, to assess systematically the accuracy of the variant calls and flag potential false positives for review.

Table 4 Recommended Quality Metrics for Clinical Bioinformatics Pipelines

Category	Use	Quality metric	Performance criteria*	Used for
Preanalytical	REQ	% of nucleated cells that are tumor cells	Min	Tumor samples
Sample	REQ	DNA concentration	Min, Max	All sample types
Sample	REQ	DNA fragment size	Min, Max	All sample types
Sample	REQ	Library DNA quantification	Min	All sample types
Run metrics*	REQ	Cluster density	Min, Max	All sample types on Illumina platforms that include this metric by default
Run metrics*	REQ	% of bases higher than the minimum Phred score of all bases called	Min	All sample types
Run metrics	REQ	Demultiplexing success (ie, all molecular identifiers present and no unexpected molecular identifiers detected)	Pass/fail	All sample types when multiplexing is used
Run metrics*	REQ	% of reads passing a minimum Phred score criterion (eg, 99% of bases at Q30 or higher)	Min	All sample types
Read filters*	REQ	Mapping quality	Min	All sample types
Mapping*[†]	REQ	Mean on-target coverage of reads	Min	All sample types
Mapping[†]	REQ	% of targeted bases with coverage greater than a specified minimum	Min	All sample types
Mapping[†]	REQ	% of bases exceeding the minimum Phred score mapped on target	Min	All sample types
Mapping[†]	OHR	% of aligned bases exceeding the minimum Phred score that disagree with reference	Max	Samples for germline analysis only
Mapping[†]	OHR	AT/GC bias	Max	All sample types
Mapping[†]	REQ	Mean insert size (bp)	Min, Max	All sample types for hybrid capture methods only
Mapping[†]	REQ	% PCR duplicates	Max	All sample types using non—amplicon-based sequencing
Per variant	REQ	Depth of coverage at variant's position	Min	All sample types
Per variant	REQ	Quality score	Min	All sample types
Per variant	Opt	Number of germline SNVs	Min, Max (may have to have separate criteria for different ethnicities)	All sample types
Per variant	REQ	Allele fraction	Min	All sample types
Per variant	REQ	Strand bias	Max	All sample types
Per variant	Opt	Haplotype bias	Max	All sample types
Per variant	REQ	Number of distinct vertical variants at the same position	≤2	All sample types
Per variant	REQ	Number of distinct horizontal variants within a prescribed cluster window size (bp)	≤1	All sample types
QC[†]	OHR	Estimate of % contamination from another sample	Max	Samples for germline analysis only (optional for tumor samples)
QC[†]	Opt	Fingerprint genotypes match NGS results	Yes (no requires investigation/explanation)	All sample types
QC[†]	REQ	Observed sex matches reported sex	Yes (no requires investigation/explanation)	All sample types if X/Y chromosomes are included in assay
QC[†]	Opt	% of bases called that are variants	Min, Max	Samples for germline analysis only (optional for tumor samples)
QC[†]	Opt	SNP/indel ratio	Min, Max	All sample types
QC[†]	Opt	Ti/Tv ratio	Min, Max	All sample types

(*table continues*)

Table 4 *(continued)*

Category	Use	Quality metric	Performance criteria*	Used for
QC[†]	Opt	Ratio of heterozygous/homozygous variants	Min, Max	Samples for germline testing only
QC[†]	Opt	Coverage profile compared with controls	Goodness-of-fit test	Critical for copy number analysis but also useful for assay QC

*Each quality metric should have performance criteria (thresholds) established to determine the validity and accuracy of the assay. The criteria indicated in this column are required to be established and used when the metric itself is required for the sample type indicated. Q30

[†]Evaluation of the quality metric per sample is needed.

Indel, insertion/deletion; Max, maximum threshold (value above which the sample or metric is considered unacceptable or failed); Min, minimum threshold (value below which the sample or metric is considered unacceptable or failed); NGS, next-generation sequencing; OHR, optional but highly recommended; Opt, optional; Q30, ■ ■ ■; QC, quality control; REQ, required for the sample types indicated; SNV, single-nucleotide variant; Ti/Tv, number of transitions/number Q31 of transversions.

In conclusion, all filters and alterations of sequence data should be understood and evaluated by the laboratory and should be kept, changed, bypassed, or inactivated as deemed appropriate by the laboratory before being used in a bioinformatics pipeline that is used for clinical patient care. The parameters and intended functions of each algorithm used for filtering and/or altering sequence data should be recorded along with all components of the bioinformatics pipeline in the laboratory's test procedure.

Recommendation 13: Laboratories Must Include Specific Measures to Ensure That Each Data File Generated in the Bioinformatics Pipeline Maintains Its Integrity and Provides Alerts for or Prevents the Use of Data Files that Have Been Altered in an Unauthorized or Unintended Manner

Computer security methods should be implemented to prevent the intentional or unintentional unauthorized modification or deletion of data. Traceability features should be implemented to know which software (including specific version, parameters, and, if applicable, operator) generated or modified any given data file used in clinical reporting.

In many NGS workflows, large data files are moved from computer to computer or to/from remote servers. Depending on the method used, these transfers may fail silently, resulting in a truncated or modified data file on the receiving system. Causes can include a network transmission failure or a disk-full error. Some file formats, including FASTQs and VCFs, do not contain an explicit end-of-file flag, thus making it possible to analyze incomplete files without any indication to the end user that this has happened. Further processing of such truncated data files does not always cause apparent errors and, therefore, increases the risk of incorrect result interpretation and reporting. Laboratories must implement methods to ensure that files are transferred completely and with intact unmodified data. Checking exact file sizes (in bytes) will detect some errors but is not sufficient as a data integrity check. Laboratories must use a hash/checksum method (eg, MIT Laboratory for Q15 Computer Science and RSA Data Security, Inc., The MD5 Message-Digest Algorithm, *https://www.ietf.org/rfc/rfc1321.txt*, last accessed September 26, 2017; or National Institute of Standards and Technology Computer Security Resource Center, Hash Functions, *https://csrc.nist.gov/Projects/Hash-Functions/publications*, last accessed September 26, 2017) to detect file integrity and completeness. Although pipeline scripts may use underlying tools that automatically test whether transfers are completed, these may not actually examine the return codes that indicate success or failure of these tools. Any scripts used in data analysis should check to determine whether the entire file has been analyzed before returning data or moving to the next step of the pipeline, because system errors may result in aborted processes. When obtaining software or scripts from a third party, laboratories must confirm their analytic accuracy and ability to detect and notify the end user about all of the errors previously described.

Recommendation 14: *In Silico* Validation Can Be Used to Supplement the Validation of the Bioinformatics Pipeline but Shall Not Be Used in Lieu of End-to-End Validation of Bioinformatics Pipelines Using Human Samples

In the ideal scenario, laboratories would be able to source a diversity of nucleic acid samples that harbored a representative spectrum of the variants that the test was designed to detect. These would be used to assess rigorously the ability of the bioinformatics pipeline during validation. In reality, it is difficult to source this diversity of samples. Laboratories take several approaches to sourcing samples, including using existing samples characterized by orthogonal techniques, sharing samples with other laboratories, and acquiring cell lines and other reference materials from repositories and commercial sources. The samples constitute the foundational material used during validation. Both for the O&F phase and to augment and supplement the validation of bioinformatics pipelines, some laboratories are using sequence data sets in which variants have been artificially introduced by custom algorithms. The research community initially used this *in silico* manipulation of sequence data sets as a way to assess and refine bioinformatics algorithms for sequence alignment and variant calling.[40] Two variations on extending this concept to proficiency testing have been reported in the literature. First, Davies et al[41] reported on a multi-institutional exchange of FASTQ files as an approach for alternative proficiency testing. In their report, focused on detection of variants in solid tumor samples, all laboratories used the same library preparation kit in combination with the same instrument to generate FASTQ files that were shared between participating laboratories. The participating laboratories analyzed the FASTQ files through their respective bioinformatics pipelines, and the results obtained were comparatively analyzed. The study showed a high degree of concordance for identification of single-nucleotide variants at varying allele fractions, whereas a lower degree of concordance was demonstrated for the identification of insertions and deletions. Second, Duncavage et al[42] reported on a pilot study using *in silico* manipulated FASTQ files generated from two commercially available reagent sets for NGS of solid tumors. In this study, a variety of sequence variants were introduced into FASTQ files at varying allele frequencies and distributed to the pilot laboratories, with results demonstrating that laboratories were able to process the files and identify variants, establishing the feasibility of the approach. Although the studies by Davies et al[41] and Duncavage et al[42] focused on proficiency testing, the same concepts can be applied during validation of the bioinformatics portion of the overall NGS test. It is critical to reiterate that validations should be based on real-world samples, because they are the closest surrogate to prospective samples that will be encountered during postvalidation clinical testing. Acknowledging this, the use of *in silico*—manipulated sequence data sets represents a supplement to the foundational validation. Such data sets can be generated to reflect a greater diversity of variants than the laboratory has been able to source through physical samples. They can be used to assess further the limits and capabilities of the bioinformatics pipeline in the O&F phase, and they can be further leveraged to augment the validation. Not all laboratories will be able to incorporate *in silico*—manipulated sequence data sets, because it requires considerable bioinformatics expertise to generate and use the data sets.

When this guideline was written, the use of *in silico* samples in any component of NGS validation was still relatively new. Using the same *in silico* sample in different bioinformatics pipelines has been challenging, particularly with some proprietary and vendor-supplied pipelines, because of the way in which the pipelines were constructed. However, a study of a proficiency testing model indicated that the use of *in silico* samples was feasible in certain settings.[42] To be used in a validation sample set, *in silico* samples must be properly verified as accurately representing the variants and allele fractions that they are intended to represent (herein referred to as verified *in silico* samples) in the same way that wet laboratory validation samples must be verified for accuracy before use. The working group anticipates that the quality and use of *in silico* samples will increase rapidly over time. Consequently, this guideline does not specify a maximum number of *in silico* samples that can be used in a validation sample set provided the minimum criteria for wet laboratory samples have been met, as described in Recommendation 7. As with all validation samples, laboratory medical directors and molecular medical professionals, as referenced in Recommendation 2, must use sound medical judgment and good clinical practice when deciding to include any validation sample in a validation sample set.

Recommendation 15: Validation of the Bioinformatics Pipeline Must Include Confirmation of a Representative Set of Variants with High-Quality Independent Data; Appropriate Validation Metrics by Variant Type Should Be Reported

Bioinformatics methods should be validated using a representative population of variants (see Recommendation 7) selected to match the expected prevalence of clinically significant variants in patients. Whenever possible, the reference (comparator) data used for validation should be generated using a high-quality clinically validated method. Moreover, reference data produced using an orthogonal (ie, different) method are preferred, because a highly similar reference method may have the same systematic false-negative and false-positive properties as the test itself. If a highly similar method is the only option, then the limitations of the reference method must be reflected in the specifications of the new test.

A medical molecular professional meeting the criteria in Recommendation 2 must assess the entire validation sample set and ensure that, in addition to the guidance given above, samples with variants have been included that adequately represent the expected types and clinical significance of those anticipated to be seen when the assay is being used clinically. For example, if clinically significant variants are expected to be detected, then a representative portion of these should be included in the validation sample set. Similarly, the samples and allele fractions should be representative of what is expected. Testing of only benign polymorphisms or normal individuals in this situation would be inappropriate.

If a laboratory is using an NGS assay with short reads (eg, <150 bp in length), then laboratories should be aware that insertions and deletions of up to 21 bp may be difficult to detect. Laboratories must include insertions and deletions of varying length up to and including the maximum number of base pairs that they expect to reliably detect in the validation sample set. Laboratories should report only those indel lengths that the assay is able to accurately and reproducibly detect. For example, if a laboratory is not able to reliably detect insertions and deletions of 21 bp with the test, then the laboratory should ascertain what maximum indel length can be accurately and reproducibly detected with the analysis. This maximum indel length must be included in descriptions of the performance characteristics of the assay in the patient's report and in other areas where performance characteristics are described. Accuracy metrics should always be reported separately for different variant types (eg, single-nucleotide variants, indels, and complex variants). For each variant type, all of the following numbers and calculations must be reported using standard methods: true positives, true negatives, false positives, false negatives, analytic sensitivity (alias positive percentage agreement), and analytical positive predictive value.[9] Overall percentage accuracy (alias overall percentage agreement) is the percentage of samples that resulted as true positives and true negatives and should also be calculated.

Sensitivity should be measured using reference variant calls that were produced on the same sample using a test that was performed independently from the test being validated. In other words, it is not appropriate to assess the sensitivity of a result by performing a confirmatory test for only those variants that were discovered in the test being validated because false negatives would not be discovered using this method. However, such data are appropriate for measuring analytical positive predictive value and overall percentage agreement.

CIs for each type should always be included and computed using an appropriate method. For example, methods such as the Wilson score approach to computing intervals can be more appropriate than the traditional (Wald) method, particularly when accuracies are high (ie, error rates are near zero), consistent with other published recommendations.[43,44]

Specificity may be a problematic metric for NGS tests, because the actual true negatives are not always well known in a validation study. More problematically, the inclusion of many negative samples/sites in a study can skew the specificity metric. The basis of measuring specificity is also not uniform, because some studies report specificity per gene, some per variant site (in any sample or in a population database), and some per base pair. These issues can lead to unintuitive (eg, 99.9999%) rates that are not comparable across different assays or studies.

Recommendation 16: Clinical Laboratories Must Ensure the Accuracy of Software-Generated HGVS Variant Nomenclature and Annotations and Have an Alert in Place to Indicate When the Software-Generated Nomenclature or Annotations Need to Be Manually Reviewed and/or Corrected, and Documentation of Any Corrections Must Be Maintained

One of the critical steps in clinical validation of bioinformatics pipelines is to compare the variants detected by the NGS test with the orthogonal method or clinically validated reference sample. In clinical laboratories, HGVS nomenclature is a de facto standard for specifically describing variants (HGVS, *http://varnomen.hgvs.org*, last accessed March 23, 2017), and algorithms to generate HGVS nomenclature are included in many pipeline software components. Laboratories are increasingly using software that annotates VCF data with HGVS nomenclature as one of the several annotation metadata. There are challenges related to the accuracy of the rendered nomenclature, particularly with indels and complex variants.

Variant Representation [Left (5′) versus Right (3′) Alignment]

Q17

The current most commonly used VCF specification denotes the starting position of an insertion or deletion using the genomic coordinate associated with the left-most (5′) nucleotide base (left justification or left aligned). By contrast, the HGVS nomenclature system, which is widely accepted by most clinical laboratories for reporting variants, specifies that the right-most (3′) nucleotide base position that is possibly associated with the insertion or deletion is arbitrarily assigned to have been changed (right justification, right aligned, or 3′ rule).[45] Because automated algorithms that generate HGVS nomenclature use the coordinates found in the VCF file, this process runs the risk of generating incorrect HGVS nomenclature. For example, algorithms that do not account for conversion of left alignment to right alignment will result in misrepresentation of the variant. Indels may span adjacent codon triplets or truly be discrete variants that automated annotation algorithms may misclassify. It is essential during assay optimization and subsequent validation to ensure that the annotation software performs the correct conversion of genomic coordinate variant representation in the VCF file to HGVS nomenclature.

Variant Normalization

Q18

A normalized variant representation in a VCF file requires that it be parsimonious and left aligned (Supplemental Table S1).[46] This has potential impact on downstream HGVS nomenclature and further annotations. A nonnormalized variant may yield incorrect HGVS nomenclature and, therefore, the high likelihood of misinterpretation and incorrect reporting. It is, therefore, important that during validation that a variant normalization algorithm be incorporated in either general-purpose annotation software or automated HGVS generators or as a discrete component of the bioinformatics pipeline.

Choice of Transcript

Q19

HGVS nomenclature is generated against a given transcript sequence. Genes having more than one transcript may or may not result in different HGVS nomenclature for the same variant. Therefore, during optimization and subsequent validation, it is critical that the laboratory review and select the most appropriate transcript for each gene in an NGS assay. The accession and version of the chosen transcript should be documented, such that the HGVS nomenclature is generated against the chosen transcript consistently.

The laboratory should exercise additional review of the nomenclature generated for insertions, deletions, and horizontally complex variants, taking into account the required conversion described previously. Vertically complex variants may also pose challenges for nomenclature assignment algorithms. The working group recommends that automated variant annotation algorithms include mechanisms to alert users that HGVS nomenclature may need additional review. Clinical laboratories should ensure accuracy of the HGVS nomenclature generated by any software or custom code as part of the validation process. In addition, manual correction of software-generated HGVS nomenclature should be documented by the laboratory for QA and quality improvement purposes.

Q20 Other Variant Annotations

In addition to HGVS nomenclature annotations, variants may be annotated by software to indicate their suspected origin (germline versus somatic) and/or their potential clinical significance.[9,11] As with HGVS nomenclature, automated annotation of any type must be thoroughly validated, and the professional interpreting the test must always have the capability of examining the accuracy of such an annotation and to override it when necessary. Automated annotations of variant-derived tumor samples, which attempt to determine germline versus somatic origin, should never be used in lieu of manual examination by the interpreter. Germline classification of variants should only be assigned in such samples when a germline sample separate from and devoid of tumor has been analyzed in conjunction with the tumor sample itself. This is because tumors cannot only gain somatic variants but lose germline variants and because most tumors contain some portion of nontumor germline tissue. Allele fraction is not a reliable method of automated classification in such samples because variants with approximately 50% or 100% allele burden can be of somatic origin, whereas variants of allele fractions outside of this range may be of germline origin. Annotations indicating potential clinical significance in both germline and tumor samples should also require manual review by the interpreter and allow the interpreter to override the annotation. This is because the same somatic variant can have varying clinical significance, depending on the tissue or tumor of origin, and because the clinical significance of both germline and somatic variants can be altered by their allele fraction and the presence of other variants.

Recommendation 17: Supplemental Validation Is Required whenever a Significant Change Is Made to Any Component of the Bioinformatics Pipeline

Supplemental validation of a bioinformatics pipeline is required when components of the pipeline are modified. If a change to the intended use of the test, the bioinformatics pipeline, the specimen types, or the variant types is planned, then a new validation appropriate to the new parameters should be designed, optimized, performed, and approved before providing the test for clinical patient care.[9] It would be impossible for the working group to list all of the possible ways that a pipeline may change. Also, the scope and impact of each change are entirely dependent on the type of change, the environment in which the change occurs, and the impact to the overall results of the pipeline. Therefore, a qualified molecular professional, as described in Recommendation 2, is expected to fully understand every change being made to the pipeline and to determine the appropriate amount and extent of revalidation that are required for the specific change being implemented. What is important in this overall process is that the primary pipeline's and each future modified pipeline's need must be identified uniquely and documented, according to Recommendation 9 in this guideline.

Conclusion

The NGS method provides unique advantages for detection of multiple genetic alterations using a single platform and is rapidly becoming a method of choice for somatic and germline variant detection by clinical laboratories. However, this new method is challenging and requires thorough analytical validation to ensure the high quality of sequencing results. This first version of Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines constitutes a combination of evidence-based and expert opinion recommendations for analytical validation of NGS bioinformatics pipelines used for clinical detection of SNVs, small indels ($\leq$21 bp), and multiple adjacent (complex) variants occurring within 21 bp of contiguous length in the clinical setting. A broad spectrum of topics, including NGS bioinformatics pipeline overview, design, development, validation, implementation, and quality control metrics, have been addressed, with general principles applicable to both germline and somatic variant detection pipelines. This document summarizes the current knowledge about NGS bioinformatics pipelines in the field of molecular diagnostics, exposes challenges of this technology, emphasizes the role of the molecular medical professional, and provides guidance on how to ensure high-quality bioinformatics are developed and implemented for high-quality patient care.

Disclaimer

The Association of Molecular Pathology (AMP) Clinical Practice Guidelines and Reports are developed to be of assistance to laboratory and other health care professionals by providing guidance and recommendations for particular areas of practice. The guidelines or report should not be considered inclusive of all proper approaches or methods or exclusive of others. The guidelines or report neither guarantees any specific outcome nor establishes a standard of care. The guidelines or report is not intended to dictate the treatment of a particular patient. Treatment decisions must be made on the basis of the independent judgment of health care providers and each patient's individual circumstances. The AMP makes no warranty, express or implied, regarding the guidelines or report and specifically excludes any warranties of merchantability and fitness for a particular use or purpose. The AMP shall not be liable for direct, indirect, special, incidental, or consequential damages related to the use of the information contained herein.

Supplemental Data

Supplemental material for this article can be found at *https://doi.org/10.1016/j.jmoldx.2017.11.003*.

References

1. Metzker ML: Sequencing technologies: the next generation. Nat Rev Genet 2010, 11:31–46
2. Roy S, LaFramboise WA, Nikiforov YE, Nikiforova MN, Routbort MJ, Pfeifer J, Nagarajan R, Carter AB, Pantanowitz L: Next-generation sequencing informatics: challenges and strategies for implementation in a clinical environment. Arch Pathol Lab Med 2016, 140:958–975
3. Luscombe NM, Greenbaum D, Gerstein M: What is bioinformatics? a proposed definition and overview of the field. Methods Inf Med 2001, 40:346–358
4. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM: The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 2009, 38:1767–1771
5. Van Deusen B, Bessette M, Johnson L, Berlin A, Banos M, Griffin L, Reckase E, Stahl J, Licon A, Kudlow BA: Comprehensive detection of driver mutations in acute myeloid leukemia including internal tandem duplications with anchored multiplex PCR and next-generation sequencing. Blood 2016, 128:5251
6. Huang Z, Ayday E, Lin H, Aiyar RS, Molyneaux A, Xu Z, Fellay J, Steinmetz LM, Hubaux J-P: A privacy-preserving solution for compressed storage and selective retrieval of genomic data. Genome Res 2016, 26:1687–1696
7. Li AR, Chitale D, Riely GJ, Pao W, Miller VA, Zakowski MF, Rusch V, Kris MG, Ladanyi M: EGFR mutations in lung adenocarcinomas: clinical testing experience and relationship to EGFR gene copy number and immunohistochemical expression. J Mol Diagn 2008, 10:242–248
8. Joseph L, Cankovic M, Caughron S, Chandra P, Emmadi R, Hagenkord J, Hallam S, Jewell KE, Klein RD, Pratt VM, Rothberg PG, Temple-Smolkin RL, Lyon E: The spectrum of clinical utilities in molecular pathology testing procedures for inherited conditions and cancer: a report of the Association for Molecular Pathology. J Mol Diagn 2016, 18:605–619
9. Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, Temple-Smolkin RL, Voelkerding KV, Nikiforova MN: Guidelines for validation of next-generation sequencing–based oncology panels: a joint consensus recommendation of the Association for Molecular Pathology and College of American Pathologists. J Mol Diagn 2017, 19:341–365
10. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015, 17:405–423
11. Li MM, Datto M, Duncavage EL, Kulkarni S, Lindeman NI, Roy S, Tsinberidou AM, Vnencak-Jones CL, Wolff DJ, Younes A, Nikiforova M: Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn 2017, 19:4–23
12. Manning CD, Raghavan P, Schütze H: Introduction to Information Retrieval, 2008; 2008. pp. 496 Q21
13. Singh RR, Patel KP, Routbort MJ, Reddy NG, Barkoh BA, Handal B, Kanagal-Shamanna R, Greaves WO, Medeiros LJ, Aldape KD, Luthra R: Clinical validation of a next-generation sequencing screen for mutational hotspots in 46 cancer-related genes. J Mol Diagn 2013, 15:607–622
14. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al: Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol 2013, 31:1023–1031
15. Pritchard CC, Salipante SJ, Koehler K, Smith C, Scroggins S, Wood B, Wu D, Lee MK, Dintzis S, Adey A, Liu Y, Eaton KD, Martins R, Stricker K, Margolin KA, Hoffman N, Churpek JE, Tait JF, King MC, Walsh T: Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. J Mol Diagn 2014, 16:56–67
16. Spencer DH, Tyagi M, Vallania F, Bredemeyer AJ, Pfeifer JD, Mitra RD, Duncavage EJ: Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. J Mol Diagn 2014, 16:75–88
17. Cottrell CE, Al-Kateb H, Bredemeyer AJ, Duncavage EJ, Spencer DH, Abel HJ, Lockwood CM, Hagemann IS, O'Guin SM, Burcea LC, Sawyer CS, Oschwald DM, Stratman JL, Sher DA, Johnson MR, Brown JT, Cliften PF, George B, McIntosh LD, Shrivastava S, Nguyen TT, Payton JE, Watson MA, Crosby SD, Head RD, Mitra RD, Nagarajan R, Kulkarni S, Seibert K, Virgin HW, Milbrandt J, Pfeifer JD: Validation of a next-generation sequencing assay for clinical molecular oncology. J Mol Diagn 2014, 16:89–105
18. Linderman MD, Brandt T, Edelmann L, Jabado O, Kasai Y, Kornreich R, Mahajan M, Shah H, Kasarskis A, Schadt EE: Analytical validation of whole exome and whole genome sequencing for clinical applications. BMC Med Genomics 2014, 7:20
19. Chong HK, Wang T, Lu H-M, Seidler S, Lu H, Keiles S, Chao EC, Stuenkel AJ, Li X, Elliott AM: The validation and clinical implementation of BRCAplus: a comprehensive high-risk breast cancer diagnostic assay. PLoS One 2014, 9:e97408
20. Lin M-T, Mosier SL, Thiess M, Beierl KF, Debeljak M, Tseng L-H, Chen G, Yegnasubramanian S, Ho H, Cope L, Wheelan SJ, Gocke CD, Eshleman JR: Clinical validation of KRAS, BRAF, and EGFR mutation detection using next-generation sequencing. Am J Clin Pathol 2014, 141:856–866
21. Ong M, Carreira S, Goodall J, Mateo J, Figueiredo I, Rodrigues DN, Perkins G, Seed G, Yap TA, Attard G, de Bono JS: Validation and utilisation of high-coverage next-generation sequencing to deliver the pharmacological audit trail. Br J Cancer 2014, 111:828–836
22. Choudhary A, Mambo E, Sanford T, Boedigheimer M, Twomey B, Califano J, Hadd A, Oliner KS, Beaudenon S, Latham GJ, Adai AT: Evaluation of an integrated clinical workflow for targeted next-generation sequencing of low-quality tumor DNA using a 51-gene enrichment panel. BMC Med Genomics 2014, 7:62
23. Sutton LA, Ljungström V, Mansouri L, Young E, Cortese D, Navrkalova V, Malcikova J, Muggen AF, Trbusek M, Panagiotidis P, Davi F, Belessi C, Langerak AW, Ghia P, Pospisilova S, Stamatopoulos K, Rosenquist R: Targeted next-generation sequencing in chronic lymphocytic leukemia: a high-throughput yet tailored approach will facilitate implementation in a clinical setting. Haematologica 2015, 100:370–376
24. Spencer DH, Sehn JK, Abel HJ, Watson MA, Pfeifer JD, Duncavage EJ: Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. J Mol Diagn 2013, 15:623–633

25. Kanagal-Shamanna R, Portier BP, Singh RR, Routbort MJ, Aldape KD, Handal BA, Rahimi H, Reddy NG, Barkoh BA, Mishra BM, Paladugu AV, Manekia JH, Kalhor N, Chowdhuri SR, Staerkel GA, Medeiros LJ, Luthra R, Patel KP: Next-generation sequencing-based multi-gene mutation profiling of solid tumors using fine needle aspiration samples: promises and challenges for routine clinical diagnostics. Mod Pathol 2014, 27:314–327
26. Sakai K, Tsurutani J, Yamanaka T, Yoneshige A, Ito A, Togashi Y, De Velasco MA, Terashima M, Fujita Y, Tomida S, Tamura T, Nakagawa K, Nishio K: Extended RAS and BRAF mutation analysis using next-generation sequencing. PLoS One 2015, 10:e0121891
27. Schrijver I, Farkas DH, Gibson JS, Lyon E; AMP Executive Committee: The evolving role of the laboratory professional in the age of genome sequencing: a vision of the Association for Molecular Pathology. J Mol Diagn 2015, 17:335–338
28. Cucoranu IC, Parwani AV, West AJ, Romero-Lauro G, Nauman K, Carter AB, Balis UJ, Tuthill MJ, Pantanowitz L: Privacy and security of patient data in the pathology laboratory. J Pathol Inform 2013, 4:4
29. Yohe SL, Carter AB, Pfeifer JD, Crawford JM, Cushman-Vokoun A, Caughron S, Leonard DG: Standards for clinical grade genomic databases. Arch Pathol Lab Med 2015, 139:1400–1412
30. Griffith M, Miller CA, Griffith OL, Krysiak K, Skidmore ZL, Ramu A, Walker JR, Dang HX, Trani L, Larson DE, Demeter RT, Wendl MC, McMichael JF, Austin RE, Magrini V, McGrath SD, Ly A, Kulkarni S, Cordes MG, Fronick CC, Fulton RS, Maher CA, Ding L, Klco JM, Mardis ER, Ley TJ, Wilson RK: Optimizing cancer genome sequencing and analysis. Cell Syst 2015, 1:210–223
31. Au CH, Wa A, Ho DN, Chan TL, Ma ES: Clinical evaluation of panel testing by next-generation sequencing (NGS) for gene mutations in myeloid neoplasms. Diagn Pathol 2016, 11:11
32. Spencer DH, Abel HJ, Lockwood CM, Payton JE, Szankasi P, Kelley TW, Kulkarni S, Pfeifer JD, Duncavage EJ: Detection of FLT3 internal tandem duplication in targeted, short-read-length, next-generation sequencing data. J Mol Diagn 2013, 15:81–93
33. Thorvaldsdóttir H, Robinson JT, Mesirov JP: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013, 14:178–192
34. MM09-A2, Nucleic Acid Sequencing Methods in Diagnostic Laboratory Medicine; Second. CLSI, 2014 Q22
35. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM: Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 2012, 91:839–848
36. Li M, Stoneking M: A new approach for detecting low-level mutations in next-generation sequence data. Genome Biol 2012, 13:R34
37. Sehn JK, Spencer DH, Pfeifer JD, Bredemeyer AJ, Cottrell CE, Abel HJ, Duncavage EJ: Occult specimen contamination in routine clinical next-generation sequencing testing. Am J Clin Pathol 2015, 144:667–674
38. Quail MA, Smith M, Jackson D, Leonard S, Skelly T, Swerdlow HP, Gu Y, Ellis P: SASI-Seq: sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. BMC Genomics 2014, 15:110
39. Davis JC, Shon J, Wong DT, Jaffe S, McEvoy JM: A DNA-based biological sample tracking method. Cell Preserv Technol 2005, 3:54–60 Q23
40. Escalona M, Rocha S, Posada D: A comparison of tools for the simulation of genomic next-generation sequencing data. Nat Rev Genet 2016, 17:459–469
41. Davies KD, Farooqi MS, Gruidl M, Hill CE, Woolworth-Hirschhorn J, Jones H, Jones KL, Magliocco A, Mitui M, O'Neill PH, O'Rourke R, Patel NM, Qin D, Ramos E, Rossi MR, Schneider TM, Smith GH, Zhang L, Park JY, Aisner DL: Multi-institutional FASTQ file exchange as a means of proficiency testing for next-generation sequencing bioinformatics and variant interpretation. J Mol Diagn 2016, 18:572–579
42. Duncavage EJ, Abel HJ, Merker JD, Bodner JB, Zhao Q, Voelkerding KV, Pfeifer JD: A model study of in silico proficiency testing for clinical next-generation sequencing. Arch Pathol Lab Med 2016, 140:1085–1091
43. Clinical Laboratory Standards: Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures; Approved Guideline, ed 2. Wayne, PA, CLSI Doc. EP17-A2, 2012 Q24
44. User Protocol for Evaluation of Qualitative Test Performance: Approved Guideline—Second Edition EP12-A2. Wayne, PA, CLSI, 2008 Q25
45. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, Mcgowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PEM: HGVS recommendations for the description of sequence variants: 2016 update. Hum Mutat 2016, 37:564–569
46. Tan A, Abecasis GR, Kang HM: Unified representation of genetic variants. Bioinformatics 2015, 31:2202–2204