Assembling and Validating Bioinformatic Pipelines for Next-Generation Sequencing Clinical Assays

Jeffrey A. SoRelle, MD; Megan Wachsmann, MD, MSc; Brandi L. Cantarel, PhD

• ***Context.***—**Clinical next-generation sequencing (NGS) is being rapidly adopted, but analysis and interpretation of large data sets prompt new challenges for a clinical laboratory setting. Clinical NGS results rely heavily on the bioinformatics pipeline for identifying genetic variation in complex samples. The choice of bioinformatics algorithms, genome assembly, and genetic annotation databases are important for determining genetic alterations associated with disease. The analysis methods are often tuned to the assay to maximize accuracy. Once a pipeline has been developed, it must be validated to determine accuracy and reproducibility for samples similar to real-world cases. In silico proficiency testing or institutional data exchange will ensure consistency among clinical laboratories.**

Objective.—**To provide molecular pathologists a step-by-step guide to bioinformatics analysis and validation design in order to navigate the regulatory and validation standards of implementing a bioinformatic pipeline as a part of a new clinical NGS assay.**

Data Sources.—**This guide uses published studies on genomic analysis, bioinformatics methods, and methods comparison studies to inform the reader on what resources, including open source software tools and databases, are available for genetic variant detection and interpretation.**

Conclusions.—**This review covers 4 key concepts: (1) bioinformatic analysis design for detecting genetic variation, (2) the resources for assessing genetic effects, (3) analysis validation assessment experiments and data sets, including a diverse set of samples to mimic real-world challenges that assess accuracy and reproducibility, and (4) if concordance between clinical laboratories will be improved by proficiency testing designed to test bioinformatic pipelines.**

(*Arch Pathol Lab Med.* 2020;144:1118–1130; doi: 10.5858/arpa.2019-0476-RA)

The rapid advancement in massively parallel sequencing/next-generation sequencing (NGS) technology and the commercial availability of NGS assay kits have rapidly advanced NGS-based clinical testing. The NGS-based clinical tests evaluate both somatic driver mutations in cancer, as well as germ-line mutations associated with congenital disease.

Next-generation sequencing has become an important diagnostic modality in oncology care by serving as a companion diagnostic to detect therapeutic and prognostic gene mutations. In the era of "personalized medicine," molecular testing is now well recognized as an important part of routine cancer care at major cancer centers.[1] Next-generation sequencing technology is also applied to the detection of inherited disease variants, both in affected and unaffected individuals (carriers). Because many inherited disease syndromes have multiple genes that could contribute to a similar phenotype, testing has expanded from single-gene sequencing to panel testing, with NGS being the standard.[2] When standard panels fail to determine a likely causative variant, whole-exome sequencing can be used. With whole-exome sequencing, for a specific individual, his or her biologic parents can be sequenced as well (trio testing) in order to minimize variants of uncertain significance. However, the processing of trio analysis presents its own bioinformatic challenges[3] because many diseases are complex and might have several variants that contribute to disease with small effect sizes. Additionally, because there are few functional annotations outside of the coding region, evaluating noncoding genetic changes is difficult without experimental validation.

During the past several years, many well-known institutions have published their development and validation of clinical oncology genomic tests for tumor mutation profiling, ranging from small gene panels to whole exomes.[4–14] Although some of these tests may overlap in basic NGS chemistry (amplicon versus hybridization-capture based) and the selected genes analyzed, ultimately each test varies in several preanalytic and/or postanalytic components. Kamps et al[15] provides an extensive review of NGS clinical oncology testing that encompasses far more than just DNA and RNA sequencing.[15] In order to streamline the variability in NGS oncology testing validation and reporting, the

Accepted for publication December 9, 2019.
Published online February 11, 2020.
From the Department of Pathology (SoRelle, Wachsmann), Bioinformatics Core Facility (Cantarel), and Department of Bioinformatics (Cantarel), University of Texas Southwestern Medical Center, Dallas.
Funding for this work was provided by Cancer Prevention and Research Institute of Texas (RP150596).
The authors have no relevant financial interest in the products or companies described in this article.
Corresponding author: Brandi L. Cantarel, PhD, Bioinformatics Core Facility, Department of Bioinformatics, UT Southwestern Medical Center, 5323 Harry Hines Boulevard, E4.350, MC 9365, Dallas, TX 75390-9365 (email: Brandi.Cantarel@UTSouthwestern.edu).

Table 1. Abbreviations of Technical Terms

Term	Abbreviation
Next-generation sequencing	NGS
Whole-exome sequencing	WES
Binary base call format	BCL file
Comma-separated file	CSV
Genome Research Consortium	GRC
Human genome version 19	hg19
University of California Santa Cruz	UCSC
Human Leukocyte Antigens	HLA
Sequence alignment map	SAM
Hierarchical indexing for Spliced Alignment of Transcripts 2	HiSAT2
Spliced Transcripts Alignment to a Reference	STAR
Binary-SAM	BAM
Polymerase chain reaction	PCR
Unique molecular identifier	UMI
Genome Analysis Toolkit	GATK
Empirical Bayesian mutation Calling	EBCall
Variant call format	VCF
Single-nucleotide variant	SNV
Insertions/deletions	indel
Copy number variation	CNV
Structural variation	SV
Control-FREE copy number and allelic content caller	Control-FREEC
Internal tandem duplication	ITD
Encyclopedia of DNA Elements	ENCODE
National Center for Biotechnology Information	NCBI
Exome Aggregation Consortium	ExAC
Genome Aggregation Database	gnomAD
Online Mendelian Inheritance in Man	OMIM
American College of Medical Genetics	ACMG
Genomics Evidence Neoplasia Information Exchange	GENIE
Clinical Interpretations of Variants in Cancer	CIVIC
Annotating principal splice isoforms	APPRIS
US Food and Drug Administration	FDA
Checkpoint kinase 2	CHEK2
Breast Cancer Type 1/2 Susceptibility Protein	BRCA1/2
Formalin-fixed, paraffin embedded	FFPE
National Institute of Standards and Technology	NIST
Genome in a Bottle	GIAB
Mutation allele frequency	MAF
Functional Analysis Through Hidden Markov Models	FATHMM
Knowledge-based mining platform for Genomic and Genetic studies using Sequence data	KGGSeq
Clinical Laboratory Improvement Amendments	CLIA
Fluorescence in situ hybridization	FISH
Health Insurance Portability and Accountability Act of 1996	HIPAA
Health Language 7	HL7
College of American Pathologists	CAP
Limit of Detection	LOD
Burrows-Wheeler Aligner	BWA

clinical and molecular diagnostic community established guidelines for the validation of NGS-based oncology panels as well as published standards and guidelines for interpretation and reporting of sequence variants in cancer.[16,17] Subsequently, guidelines from the Association of Molecular Pathology were released outlining recommendations for validating clinical bioinformatic pipelines.[18] Although these guidelines provide detailed recommendations, a user-friendly version would be helpful to walk through the process. This review will provide a step-by-step guide to navigate the many factors of bioinformatic analysis that affect NGS assay results, including unfamiliar abbreviations (Table 1).

We will describe the key elements for developing a bioinformatics workflow, how to validate a bioinformatics workflow, and how clinical NGS laboratories can ensure consistency across testing facilities. Considerations for the development of a bioinformatics workflow include: (1) choosing a human reference genome, (2) understanding the limitations of predicting copy number and structural variation, (3) choosing algorithms for identifying genetic variants, (4) evaluating publicly available annotation resources, and (5) determining filtering metrics for disease-causing variants (Tables 2 and 3).

RAW DATA PROCESSING AND MUTATIONAL PROFILING OF NGS DATA

In somatic testing, the quality of NGS sequence data is reliant on the quality of the sample, including tumor purity, DNA or RNA quality, sequence library complexity, and the efficiency of the hybridization baits. Although bioinformatics protocols cannot be altered to overcome the limitation of laboratory protocols, postanalytic filtering performed during report generation is reliant on tumor purity because it affects the ability to detect variants, especially if the purity is below the assay limit of detection. For example, in a sample estimated to be about 30% tumor, the frequency of driver mutations is expected to be about 15% mutation allele frequency. Limit of detection studies are performed at validation.

The NGS bioinformatics pipeline starts with raw sequence data that are produced by the sequencer and formatted by software provided from the sequencing vendor, such as Illumina. The pipeline will perform all of the necessary steps to predict the variants in the sample and annotate those variants with information about the gene, effect of the variant (missense, nonsense, splice-site, etc), the frequency, and finally its association with disease. This process aims to create a list of candidate variants, which are then manually curated for clinical actionability (Figure 1). Some clinical laboratories use ion torrent sequencing technology, which arrives as kits including accompanying software for analysis, with little manual work required. For this review, we focus on Illumina sequencing technology because it is more commonly used and there are many ways to analyze the data. Although Illumina has a platform for analysis called BaseSpace, there are still many choices for analysis for these types of data, whether the clinical laboratory uses BaseSpace, a cloud computing provider, or an internal computing infrastructure.

Workflow Step 1: Input Data From Sequencer

Illumina sequencers automatically write raw data into binary base call format (BCL file). The BCL files are

Table 2. Bioinformatics Algorithms Strengths and Weaknesses[a]

Step	Algorithms	Use	Strengths	Weaknesses	Hyperlink
1	bcl2fastq	Convert sequencer files (BCL) to FASTQ	Works with Illumina products	Non-Illumina reagents and methods are not supported	https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html
2	TopHat2	Alignment of RNA	Widely used	Obsolete, was replaced by hisat2	See hisat2
2	STAR	Alignment of RNA	Widely used because it is faster than TopHat2	Less accurate for sequences with large variation compared to reference	https://github.com/alexdobin/STAR
2	HiSAT2	Alignment of RNA	More accurate for sequences with variation		https://ccb.jhu.edu/software/hisat2/index.shtml
3	BowTie	Alignment of DNA			http://bowtie-bio.sourceforge.net/index.shtml
3	BWA	Alignment of DNA	Alt-aware and more accurate for sequences with variation		http://bio-bwa.sourceforge.net/
3	samtools	Removing duplicates	Quick runtime	Duplicates can only be removed	http://samtools.sourceforge.net/
3	picard	Marking or removing duplicates	Widely used, duplicates can be marked or removed		https://broadinstitute.github.io/picard/
3	GATK	Germ-line variant calling	Highest sensitivity for indels, widely used for germ-line variant calling	Requires many steps for accurate calling, with many companion programs for filtering improving accuracy, slow runtime	https://software.broadinstitute.org/gatk/
3	Samtools	Germ-line variant calling	Highest sensitivity for SNVs, widely used, slow runtime, highest accuracy for indels	Low sensitivity for indels	http://samtools.sourceforge.net/
3	Strelka	Somatic and germ-line variant calling	High sensitivity and specificity		https://omictools.com/strelka-tool
3	Platypus	Germ-line variant calling	High sensitivity, fast runtime	Low specificity	https://omictools.com/platypus-tool
3	Freebayes	Somatic and germ-line variant calling	High sensitivity for low frequency variants, seen in mosaic samples (ie, tumors or subpopulations)	Low specificity	http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html
3	Mutect2	Somatic variant calling	Highest sensitivity for SNVs, widely used, slow runtime	Requires many steps for accurate calling, with many companion programs for filtering improving accuracy, slow runtime	
3	Varscan	Somatic and germ-line variant calling	Fast runtime	Produces many unique calls compared with 9 other callers	http://varscan.sourceforge.net/
3	EB-Call	Somatic variant calling		Does not provide results in VCF format	https://omictools.com/ebcall-tool
3	Virmid	Somatic variant calling	Overlapping call set with Mutect2, EBCall, and Strelka for somatic calls in 9-caller comparison, high specificity		https://bioinformaticshome.com/tools/descriptions/Virmid.html
3	Shimmer	Somatic variant calling	High specificity	Low sensitivity	https://omictools.com/shimmer-tool
4	OncoCNV	Copy number variation calling	Higher specificity than CNV-Kit	Lower sensitivity than CNV-Kit	https://omictools.com/oncocnv-tool
4	Control-FREEC	Copy number variation calling	Similar to OncoCNV		http://boevalab.inf.ethz.ch/FREEC/
4	CNV-Kit	Copy number variation calling	Higher sensitivity than OncoCNV	Lower specificity than OncoCNV	https://cnvkit.readthedocs.io/en/stable/quickstart.html
5	StarFusion	Gene fusion detection	High accuracy and low runtime compared with competitors		https://omictools.com/star-fusion-tool
5	PINDEL	Structural variant detection	High accuracy in detecting internal tandem duplications of FLT3		http://gmt.genome.wustl.edu/packages/pindel/

Abbreviations: BCL, binary base call format; BWA, Burrows-Wheeler Aligner; CNV, copy number variant; EB-Call, Empirical Bayesian mutation Calling; GATK, Genome analysis ToolKit; SNV, single-nucleotide variant; SV, structural variant; VCF, variant call file.

[a] All URLs have an access date of October 29, 2019.

Table 3. Considerations for Workflow Steps		
Step	**Feature**	**Considerations**
2	Genomic assembly (GRCh37, GRCh38)	Switching to a new genome can be onerous for a new lab, but the newest human genome assembly has been shown to be more accurate, thus reducing false-positive variants
2	Alignment algorithm	The human genome has many repeated sequences and these repeats cause misalignments and therefore errors in variant detection
2	Duplication	Errors in PCR can introduce false positives into the data set. These can be removed using algorithms for marking and removing these duplicates. They cannot be removed with amplicon-based panels
3	Variant calling type (somatic/germ line)	Germ-line variant detection algorithms are not designed to detect low-frequency variation. If normal tissue is available, algorithms for somatic variant calling have a higher sensitivity
3	Variant calling algorithm	Variant calling algorithms vary widely in sensitivity and specificity. Filters must be established experimentally to maximize sensitivity and specificity
3	Detecting artifacts	ML can be applied to detecting false positives, but it must be trained on validation data to maximize accuracy
4	CNV detection	The sensitivity of CNV calling is less than 90% for most clinical assays
5	SV detection	The sensitivity and specificity of SV calling are low for most clinical assays, because most clinical assays do not have uniform coverage of capture along the whole genome. Research group will use low-coverage whole genomes, which can be costly in a clinical assay
6	Variant annotation and effect	Defining the chromosomal locations of genes and exons is not trivial. Each gene annotation will have some inaccuracies. Choosing one over another might change a variant effect from an intron mutation to a coding mutation
6	Population studies	Population studies can help you filter out common variants, although some pathogenic variants can be in >1% of certain populations
6	Population studies	Some populations of people are not well represented in population databases
7	Identifying clinically actionable variants	There is no complete database of clinically actionable variants available publicly or commercially
7	Identifying clinically actionable variants	There is no database of FDA-approved drugs and their clinical and molecular indications
7	Identifying clinically actionable variants	Clinical trial databases are difficult to search, and matching of patients cannot be automated because most information is contained in paragraphs of text

Abbreviations: CNV, copy number variant; FDA, US Food and Drug Administration; ML, machine learning; PCR, polymerase chain reaction; SV, structural variant.

converted into FASTQ files using a program called bcl2fastq (provided by Illumina) in order to generate genomic sequence reads that are used by most analysis programs. FASTQ files are text files that contain a quality score (Phred) for each base that can then be used for sequencing alignment. A Phred quality score is calculated based on the probability that the base is incorrect, where a score of 20, the typical cutoff, represents a 1% probability of being incorrect; higher scores have higher quality. These quality scores are considered in downstream analysis so that bases with a higher quality score are given higher weight in genotype prediction. Multiplex sequencing allows for sample pooling within the same run. Each sample contains its own unique adapter sequence that can then be used to separate reads by sample/adapter in a process called demultiplexing. These adapter sequences, along with other information about the sample, such as sample ID and project name are passed to the bcl2fastq program in a comma separated (CSV) file. The automatically generated FASTQ files are now ready for bioinformatic analysis.[19]

Workflow Step 2: Alignment of DNA and RNA Onto a Human Reference Genome

Raw NGS data processing ensures each strand of sequenced DNA matches to its corresponding location in the genome. The accuracy of read alignment is dependent on the reference genome and alignment algorithm. There are currently 2 available versions of the reference human genome: GRCh37 (Genome Reference Consortium human) (hg19; human genome version 19), released in 2009, and GRCh38 (hg38), released in 2013. Prior to 2013, there were 2 slightly different human assembly versions released by the Genome Research Consortium (GRC) and UCSC (another curator of the human genome, University of California Santa Cruz). These assemblies (GRCh37 and hg19) varied in alternative chromosomes and scaffolds. Since the latest release, the GRC has been the primary source for human genome assemblies; therefore, the UCSC browser version (hg38) matches to avoid confusion. GRCh38 improved upon the earlier genome builds by (1) correcting errors, (2) filling nucleotides into ambiguous repeat regions (ie, centromeres) with model sequence, and (3) including alternative loci, which represent differences in human population, such as human leukocyte antigen (HLA) haplotypes. As a result of these improvements, the use of GRCh38 as the reference genome reduces errors in variant detection.[20,21] In an effort to further reduce false-positive rates, both genomes include decoy sequences, which act as a "sponge" for the most commonly misaligned reads. However, many clinical laboratories have been slow to adopt this new reference. This is likely due to several reasons, including the fact that many variant annotation databases themselves have not converted their positions to the new build. Furthermore, making the change of reference genome would be a considerable update, requiring a revalidation of the pipeline.

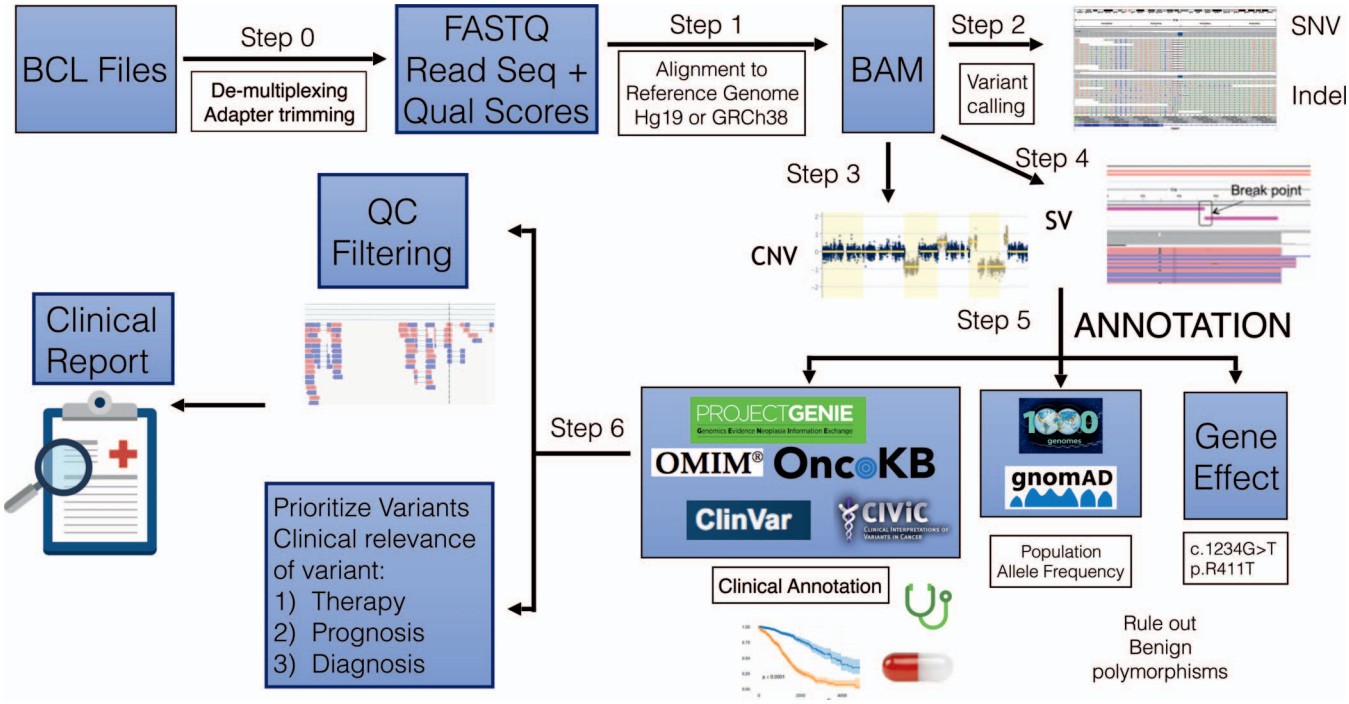

Figure 1. *Graphical overview of bioinformatic pipeline from data generation to generation of a clinical report. Step 0 occurs automatically by generating a BCL file from the instrument and demultiplexing multiple samples in a flow cell lane; because it requires no hands-on effort, it is pre–step 1. Step 1 takes FASTQ files that contain sequence and quality information and align the information to a reference genome to produce a BAM (Binary Alignment Map) file. In steps 2–4, the BAM file is processed by your choice of variant caller to yield information on single-nucleotide variant (SNV), indel, copy number variant (CNV), or structural variant (SV). Step 5 is a semimanual process where these variants are named and evaluated against publicly available databases for clinical significance. Lastly, variants must be manually evaluated for quality control (QC) and prioritization of variants.*

Algorithms are needed to actually perform the sequence alignment, and many are publicly available, with each having its own advantages and limitations. Short-read alignment (<200 bp mapping) to a reference genome is relatively easy, and several publicly available algorithms are available. For RNA (transcriptomics), mapping sequences over a large gap is necessary because of the splicing out of introns, therefore this requires splice-aware algorithms. These algorithms include TopHat2, HiSAT2 (Hierarchical indexing for Spliced Alignment of Transcripts 2), and STAR (Spliced Transcripts Alignment to a Reference), and they can map RNA reads from exon to exon across the spliced-out introns[22–24] (Table 2). However, in order to identify clinically significant chromosomal translocations, which involve even larger breaks, a modified version of STAR, STAR-Fusion, can be used[25] (Table 2). For variant calling, HiSAT2 generates more accurate alignment compared with other splice-aware aligners, likely because of its ability to model common variation as a graph reference.[26] Aligners for DNA, such as Bowtie2 and BWA (Table 2), are not splice aware, and thus they cannot split reads to improve the alignment.[27] For variant detection, BWA (Burrows-Wheeler Aligner) provides a more accurate alignment compared with Bowtie2, and therefore more accurate variant calls,[28] making it popular for clinical applications. BWA is an alt-aware alignment tool, meaning it can align to the alternative chromosomes and translate their location to chromosomes, which could explain its improved accuracy. Overall, these programs accurately place sequence data (alignment) onto the proper genome location (mapping) to produce sequence alignment map (SAM) or Binary-SAM (BAM) alignment files. SAM files are very large (gigabyte scale), whereas BAM files are a compressed format that is more efficient for software to process.

Sequencing data must also include quality metrics of the alignments, including: (1) depth and breadth of coverage, (2) mean mapping quality, and (3) mapping rate. The depth and breadth of coverage are the number of reads (depth) that cover each base on average and the percent of the targeted region covered by reads.[16] In cancer there are subpopulations of cells that might contain clinically actionable variants. Therefore, higher coverage sequencing is required to detect low-frequency variation. Because coverage is not even, it represents a wide distribution. If the limit of detection for the assay is 5% mutational allele frequency, 5 alternate reads will be required to detect this variant with 100 reads. However, in a sample with 100× median coverage, 50% of the positions have coverage of less than 100 reads. The targeted region of the clinical assay is the region of the genome, such as the exons of genes of interest, that the assay aims to detect. The mapping rate is the percent of the reads that map to the genome. A large number of unmapped reads can result from reads that are too short (poor quality marker) or from an insertion not in the primary assembly. The mean mapping quality reflects the probability that a read is misplaced (misaligned compared with the reference genome). Perfectly mapped reads have a mean mapping quality score of 60 ($1/10^{-6}$ = 0.0001% chance the alignment is incorrect), whereas a score below 30 is generally considered unacceptable ($1/10^{-3}$ = 0.1% chance the alignment is incorrect).

Gene panel sequencing involves either a targeted amplicon-based polymerase chain reaction (PCR) amplification or hybridization-based bait capture of genomic regions. In the hybridization-based bait capture assay, PCR duplicates arise and represent a technical artifact of the assay and not true biologic frequency. The PCR duplicates are identified as having the same start and stop sites. Typically, duplicates are marked for removal using programs like Samtools[29] and Picard[30,31] (Table 2). Introduction of short barcode sequences (unique molecular identifiers, 4- to 8-bp–long) reduces PCR error due to duplicates, especially for deep sequencing. In contrast, amplicon-based methods use primers to target regions, and duplicates cannot be removed.

Workflow Step 3: Identifying Single-Nucleotide, Insertion, and Deletion Variants in DNA

Variants in DNA are principally tested for inherited and acquired diseases. Germ-line variants are inherited and present in an individual's genome from conception. Variants that arise after conception are generally referred to as somatic. Because there are 2 copies of each chromosome, germ-line variants are designated "alternate alleles" and identified by comparison to the reference genome and calculated to be heterozygous—2 different nucleotides at the same allele—or homozygous—the same nucleotide at the same allele. Comparison of germ-line variants of an individual to those of his or her biologic parents is helpful for determining the significance of a variant. If an unaffected parent shares the same variant, it is likely not disease causing (benign), but if the variant is not in either parent and is new (de novo) there is a high likelihood of it being disease causing (pathogenic).[17] On the other hand, somatic mutations may represent a subpopulation of cells that are identified by comparison to a matched patient control (skin, saliva, or blood, depending on the type of malignancy, ie, skin for hematologic malignancy). Somatic variants are detected using the frequency of mismatches and gaps in the alignment at each base. For example, a germ-line variant is likely to be called if at a particular position 50% of the reads is an A, whereas the reference is a G.

Algorithms for germ-line mutation detection for single-nucleotide variants (SNVs) and insertions/deletions (indels) include GATK (Genome Analysis ToolKit),[32] Samtools,[29] Platypus, Strelka,[33] and Freebayes (Table 2). Each of these methods uses a slightly different model for determining genotype likelihood that includes the quality scores of the sequences and of the alignment as well as the depth of sequencing and the location of alternate bases in reads. Therefore, each method has been shown to have differences in sensitivity and positive predictive value.[34–36] Approaches that combine these methods to identify variants with multiple levels of supporting evidence have been shown to improve accuracy.[36]

Somatic mutation detection compares the frequencies of alternative bases at each chromosomal position to a normal matched control in a tumor specimen, usually from the same person. Algorithms for somatic mutation detection for SNVs and indels include MuTect2,[37] VarScan,[38] EBCall (Empirical Bayesian mutation Calling),[39] Freebayes, Virmid,[40] and Shimmer[41] (Table 2). Like germ-line variant detection, these packages have slightly different models for determining a somatic mutation. Therefore, these packages can differ greatly, with little overlap among the various methods.[36,42] Recently, EBCall, Mutect2, Virmid, and Strelka have been shown to be the most reliable somatic variant callers for both medium- and high-coverage sequencing data of SNVs. EBCall demonstrated the highest sensitivity rate for indel identification.[42]

The main source of false-positive variant predictions stems from errors in the replication introduced during PCR and sequencing. Because of the biased nature of PCR, consistent errors called artifacts can occur. Machine learning has been applied to the discovery of sequencing artifacts in a modeling algorithm called Cerebro.[43] Cerebro uses various quality metrics of variant calls, including alignment and base quality scores, distributions of alternate and reference bases, and genomic context. As with variant calling, clinical laboratories must train the models on their data to achieve maximal sensitivity and specificity.

After variant calling software packages analyze the data, a file is produced called variant call format (VCF). This file contains the information about the chromosomal loci for the variant, the reference base(s), the alternate base(s), a score, statistical metrics on the call, and the genotype for each sample. There are many tools, such as vcftools[44] and SnpSift,[45] that can filter variants based on metrics contained in the VCF file.

Workflow Step 4: Identifying Copy Number Variation in DNA

Detecting copy number gains or losses is important in both germline and somatic testing. Although this can be accomplished by NGS technology, the detection of copy number variants (CNVs) is complicated by: (1) biases in bait capture,[46] leading to uneven coverage across exons, (2) sparse breadth of coverage across the chromosome because only targeted genes are captured, and (3) decreased sensitivity in heterogeneous tumor samples.[47] Similar to SNVs and indels, there are many different algorithms for detecting CNVs; these include VarScan2,[48] CNVKit,[49] Control-FREEC (Control-FREE Copy number and allelic content caller),[50] and OncoCNV[46] (Table 2). OncoCNV is designed for amplicon data and includes normalization methods specific for the biases in PCR methods. These algorithms rely on assessing the differences in read coverage in a particular region compared with the average of the surrounding regions and the allele frequency of the alternate allele. Because the coverage of targeted regions is not uniform for gene panels, the coverage is normalized to control samples without known variants. An independent comparison of these methods using simulated data showed that VarScan2 had more false positives compared with other methods, OncoCNV and Control-FREEC had similar performance, yet OncoCNV performed better with a panel of greater than 3 normal controls.[51] In this sample study, a comparison of accuracy of these methods on whole-exome data showed that CNVKit had a high sensitivity compared with OncoCNV, but a lower specificity. The accuracy of these methods is improved with high tumor cell percentage, low heterogeneity, and a panel of normal controls for comparison.

Workflow Step 5: Identifying Structural Variation in DNA and RNA

Determination of structural variants (SVs) has traditionally been performed by cytogenetics, but NGS clinical assays now detect some gene fusions because of advances in RNA-Seq. However, the prediction of SVs has a high false-

positive rate. This is because SV prediction relies on discordant read pairs, which only represent a small fraction of the total reads. Discordant read pairs are the read pairs that align to different chromosomes, for translocations or at distances different than expected by the average fragment sizes. Detecting translocations by DNA sequencing is challenging because the break points often lie in large intronic or intergenic regions that would require whole-genome sequencing to capture the break point. RNA-Seq overcomes the challenges associated with DNA sequencing by finding abnormally joined exons. Algorithms such as STAR-Fusion,[52] nFuse,[53] and EricScript[54] use read pairs aligned to different genes to identify translocations[55] (Table 2). STAR-Fusion has been shown to have high accuracy and lower runtime compared with its competitors.[55] Other clinically relevant structural variation includes large insertions, large deletion, and internal tandem duplication. Methods for identification of SVs include Pindel,[56] Lumpy,[57] and Delly[58] (Table 2). These methods use split reads (paired reads that align to 2 different regions) to identify structural variation, which work well with high-coverage data, as in most clinical applications. Lumpy and Delly also use discordant read pairs, reads that map to distances greater than expected for the average library size, to identify structural variation. Internal tandem duplicates for FLT3 (fms-like tyrosine kinase 3), a gene commonly mutated in acute myeloid leukemia, are challenging to detect by NGS methods, but Pindel has been shown to accurately identify them.[59] In general, the accuracy of SV algorithms decreases as the size of the SV increases. Bioinformatic algorithms along with RNA-Seq have advanced to allow SV detection, albeit with lower sensitivity and specificity than SNV detection.[60] Long-reads sequencing methods, including Pac-Bio and Oxford Nanopore, present the opportunity to improve SV prediction.[61] However, these methods are not adopted for clinical application because these techniques are expensive, they have a higher sequence error rate, and few bioinformatics tools exist to integrate these results with Illumina sequencing for clinical applications.

Workflow Step 6: Annotating Genetic Variation and Determining Variant Effects

Once a genetic variant has been identified, its significance in the context of a gene must be determined and this process is called annotation. Annotation depends on whether the variant lies within the transcribed portion of the gene. Thus, although GRCh38 is a reference genome, there are 2 reference transcriptomes: Gencode and RefSeq. Gencode was developed by The ENCODE project (Encyclopedia of DNA Elements), which is maintained by Ensembl, and RefSeq (Reference Sequence) was developed and is maintained by the National Center for Biotechnology Information. These databases differ mostly in their annotation of (1) noncoding genes and (2) alternatively spliced isoforms of coding genes. Through combinations of computational modeling and experimental evidence, the boundaries of transcribed exons are defined by Gencode[62] or RefSeq.[63] The annotation process determines the gene, the amino acid, and subsequent codon change for a specific variant. Other types of functional changes, such as those in promoters and regulatory regions, are more difficult to predict. Although most clinically significant genes are similar in RefSeq and Gencode, discrepancies do still exist largely because of the fact that RefSeq uses many experimentally modeled transcripts and Gencode relies more heavily on data derived from RNA-Seq experiments in order to define exonic boundaries.[64] Studies describing genetic test result differences between Gencode and REfSeq annotation are limited.[65] Resources such as APPRIS (annotating principal splice isoforms)[66] or transcript-level support can be used to determine the primary transcript for aid in reporting the most likely gene effect.

Many of the variants identified are actually relatively common polymorphisms that occur as a part of normal genetic variation. When a variant is present at more than 1% allele frequency within a population, this is evidence that the variant is likely benign. Conversely, pathogenic variants should confer reduced fitness and be less than 1%. However, there are several founder mutations close to 1% allele frequency that are still pathogenic within a population, such as CHEK2 (checkpoint kinase 2) c.1100delC (Europeans) or certain BRCA1/2 (Breast Cancer Type 1 Susceptibility Protein) mutations in Ashkenazi Jewish people.[67] Population sequencing of healthy individuals has been performed to address this very issue. The Exome Aggregation Consortium (ExAC; more than 60 000 healthy individuals) and genome Aggregation Database (gnomAD; with more than 126 000 individuals) performed whole-exome or whole-genome sequencing of healthy individuals to better estimate allele frequency of normal human variation. However, these databases are heavily populated by individuals of European ancestry. Although efforts to improve genetic diversity added many of Asian descent, there are still fewer individuals of Latino or African ancestry.[68] The 1000 Genomes project helps supplement this deficiency somewhat by determining the genetic variation in more than 3000 individuals from multiple, diverse subpopulations in Asia, Europe, Africa, and the Americas.[69] These population databases are a useful tool in first assessing whether a variant is a benign polymorphism or warrants further investigation.

There are a number of clinical databases used to determine the clinical significance of genetic variation. The Online Mendelian Inheritance in Man (OMIM) database catalogs genes that are associated with mendelian genetic diseases, but many incomplete associations are common.[70] ClinVar is a clinical database that collects thousands of user-submitted variant classifications.[71] Previous critiques about the quality of interpretations have subsided with a large influx of interpretations that are frequently concordant.[72] Few labs submit supporting evidence, but the inclusion of specific data like PubMed citations is very helpful when coming to a conclusion on an interpretation. Most of the variant submissions in ClinVar are more relevant to inherited genetic disorders, but they can also be helpful in cancer. The final assessment of variant classification for inherited disorders must rely on the American College of Medical Genetics guidelines.[73]

Several databases specific to somatic genetics of cancer exist to aid in variant classification. Cosmic is a database of somatic mutations in cancer. Instead of classifying variants, this database displays histograms of amino acids frequently mutated within a gene for a variety of tumor types. Because gain-of-function mutations often drive cancers, mutational hotspots are seen with a high frequency, providing evidence for selection of these somatic mutations as cancer drivers (eg, *KRAS* commonly has activating mutations in codons 12, 13, and 61). Similar to the GnomAD project for health populations, the Genomics Evidence Neoplasia Information Exchange (GENIE) project has aggregated data from several

large cancer genome studies and has accumulated genetic and clinical data for more than 60 000 patient samples.[74] Because of its large sample size, cancer variants identified in clinical NGS testing can be compared to those in GENIE to determine if variants found within the same tumor type could be possible driver mutations. The Clinical Interpretations of Variants in Cancer (CIVIC) database is an expertly curated database that provides a literature review of specific genetic variations in cancer, including variant function, prognosis, and potential treatment.[75] OncoKB is a knowledge base that also provides annotation describing the biologic and clinical functions of genes and variants in cancer. OncoKB ranks clinical actionability by level of evidence sources, including US Food and Drug Administration (FDA) labeling, National Comprehensive Cancer Network guidelines, scientific literature, and expert curation.[76] Somatic mutations should receive a tier 1 to 4 assignment based on their therapeutic, prognostic, or diagnostic significance.[17]

In addition to manually curated databases of clinical significance and actionability, there are bioinformatically derived databases and tools that aim to predict whether a variant will lead to a loss-of-function mutation. These metrics use evolutionary metrics using (1) functional prediction, which uses sequence substitution matrices, similar to what is used for protein alignment algorithms; (2) conservation, which calculates a score derived from conservation across species; and (3) combinatorial method, which combines functional and conservation scores. A comparison of these scores found that FATHMM (Functional Analysis Through Hidden Markov Models)[77] and KGGSeq (Knowledge-based mining platform for Genomic and Genetic studies using Sequence data)[78] had the highest accuracy to classify variants.[79] M-CAP (Mendelian Clinically Applicable Pathogenicity score) is a metric that was developed specifically for clinical interpretation.[80] Although computationally derived metrics are useful in variant prioritization when experimental evidence is lacking, differing annotation among the various metrics can be difficult to interpret.

Workflow Step 7: Prioritizing Genetic Variation

Finally, once variants have been predicted, annotated, and classified, the list of candidate disease-causing mutations should be prioritized. First, poor-quality variants are filtered out, based on: (1) low read depth, usually less than 10; (2) low alternate read frequency, less than 30%; and (3) poor-quality scores, such as average mapping score or high strand bias. For mendelian genes, testing is usually performed on panels where all genes are related to the phenotype prompting testing (eg, epilepsy panel or maturity-onset diabetes of the young panel). However, when a whole-exome approach is used, parental testing is required (trio testing) to help rule out shared variants that are unlikely to be disease causative. The exception is when the proband inherits 2 loss-of-function variants for a recessive condition that would have no phenotype in the parents. When these variants are the same, a patient is homozygous, but if the variants are at different positions, this mode of inheritance is called compound heterozygous. Because whole-exome studies look at all genes, candidate mutations are further examined to see if the gene or the variants have a known association with the phenotype using the OMIM and ClinVar annotations. For most conditions, only variants that cause changes to coding in protein-coding genes are considered candidates for causing disease, because little is known about the functional impact of noncoding or synonymous coding changes.[81]

Compared with germ-line diagnostics, testing for genetic variants in cancer is primarily focused on finding targetable variants that can be treated therapeutically. Thus, the highest-priority variants are those that have FDA-approved drugs for the specific cancer type. A lower-level finding is a variant with an approved drug for a different cancer type. It was hoped that targeting a mutation in one cancer type would be analogous for other cancer types; however, the disparate results of *BRAF* inhibitors for V600E mutations in melanoma and colon cancer (80% versus 5% response rate) proved targeted therapy cannot be universally applied based on genetic findings. Thus, annotations from OncoKB, ClinVar, and GENIE can be used to determine the function of a cancer mutation, but they are not meant to replace manual curation that often involves literature searches. As with the example above, the significance of a variant in one cancer type cannot be extrapolated universally, and annotations must be in formed by the latest information in a field that is changing rapidly with new discoveries about genes, variants, and treatments. Lastly, any filtering of variants must be validated to determine how the filters affect sensitivity and specificity. The reduction of false positives can be detrimental if the sensitivity of clinically actionable variants also decreases.

VALIDATING THE BIOINFORMATICS PIPELINE

The purpose of validation is to determine the accuracy and reproducibility of the bioinformatics pipeline, in synthetic, reference, and clinical samples, similar to the type expected to be tested in the clinical setting (Table 4). Optimization of the pipeline should be performed during pipeline development and the workflows should be "lock-down" at the validation stage, and therefore not changed with different sample or variant types. Validation must be performed for a variety of sample and variant types. For example, if the expected sample types are blood, formalin-fixed, paraffin embedded (FFPE) tissue, and saliva, then the validation should include all of these samples' types with a variety of variant types expected to be examined in the assay, including SNVs, indels, CNVs, and SVs. Additionally, validation also ensures that there is no technical variation among lab equipment. Although the number of samples used in published validations has varied widely from 5 to nearly 300,[77] the samples should include at least 50 clinical samples and several references or synthetic samples that can test the sensitivity, specificity, limit of detection, and reproducibility of the assay.

There are 3 sections to the validation: preanalytic, analytic, and postanalytic, which are laid out in 5 steps (Figure 2). The preanalytic section of the validation (validation steps 1 and 2) ensures that the validation experimental design is consistent with the expected sample types (FFPE, blood, and bone marrow) and the expected variants the assay is designed to detect. The analytic section (validation step 3) ensures that the bioinformatics workflow performs as expected with the different variant types that will be reported in the assay. Finally, the postanalytic section (validation steps 4 and 5) ensures that data are transmitted, displayed, and stored properly.

Depending on the application (germ-line versus somatic sample testing), the minimum coverage may range from 50× to 500×. In general, the validation not only serves to ensure

Table 4. Dos and Don'ts		
Practical Tips	**DON'T**	**DO**
Hardware use during validation	DON'T perform validation studies on one computer system, then attach an ether cord to the data center server once assay is ready for "go-live"	DO use consistent hardware, OS, network configuration, and data workflow from validation onwards to assay implementation
Revalidation	DON'T skip revalidation of the bioinformatic pipeline even for minor changes	DO revalidate if even 1 line of code is changed DO revalidate even if just the order of algorithms or data flow is changed
Optimization	N/A	DO lock down the pipeline after it has become optimized before starting validation Optimization can use data from real sequenced or in silico data sets
Types of variants for validation	DON'T use too many simple variants for pipeline validation	DO include horizontally and vertically complex variants
Poorly covered regions	DON'T gloss over poorly covered regions	DO disclose which areas do not consistently meet quality standards

Abbreviation: N/A, not applicable.

that data can properly and faithfully transit from the sequencer through all networks, hardware, and software consistently for a variety of variant types, but also ensures that the bioinformatics pipeline is tuned to the clinical assay (Figure 2).

Validation Step 1: Reference Sample, Engineered Samples, and Synthetic Samples

Reference samples provide a truth set, allowing evaluation of sensitivity and positive predictive values for the clinical assay. These can be purchased commercially or developed by the clinical laboratory. For an NGS assay, the National Institute of Standards and Technology (NIST) reference sample, NA12878, is a gold standard sample that has been sequenced as part of the NIST Genome in a Bottle (GIAB) initiative and part of the Illumina Platinum Genome project.[87] This sample has been well studied and established as a standard of truth distributed by the GIAB consortium and Illumina. Using high-confidence regions of this genome, the sensitivity and specificity of germ-line

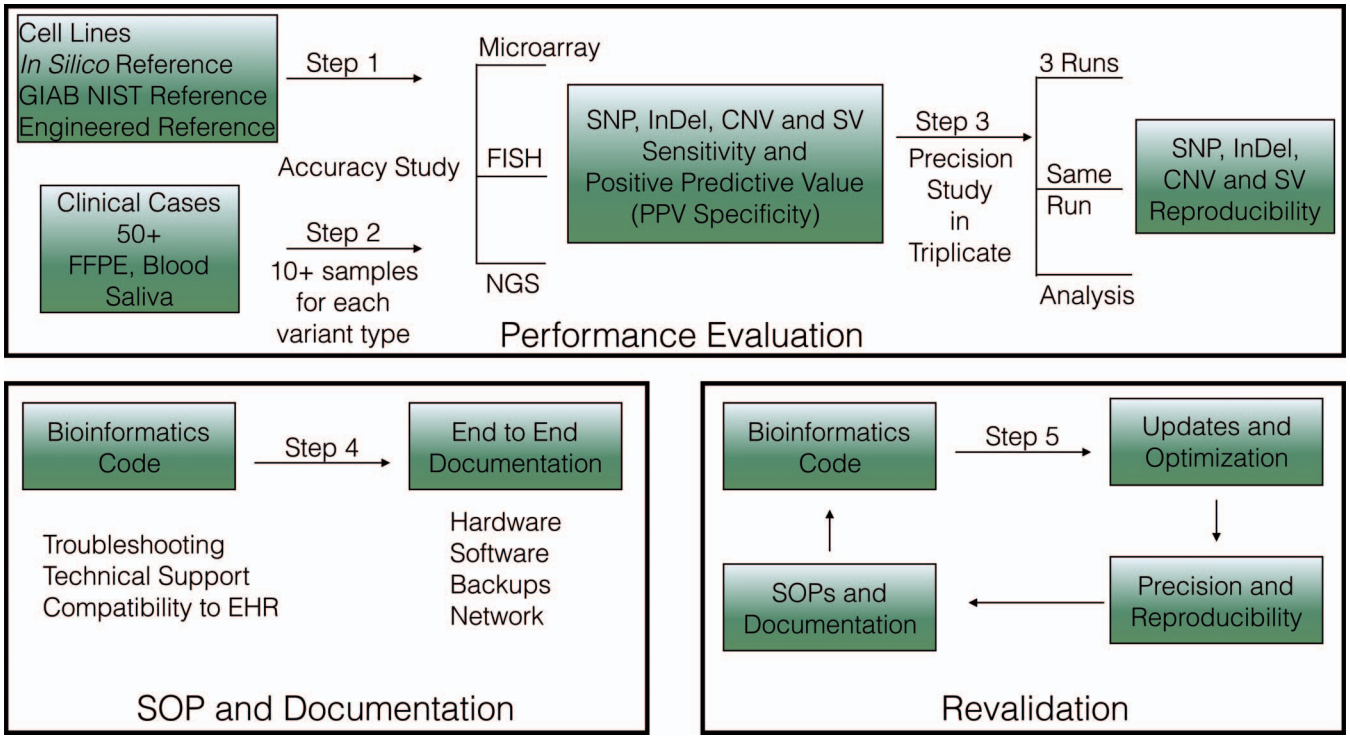

Figure 2. *Validation of a clinical bioinformatic pipeline begins with (A) performance evaluation using a combination of reference material and clinical cases to determine the accuracy and precision of variant calling compared with gold standard techniques. B, Standard operating procedures and documentation must be created for each step and each component of the process. C, Revalidation to improve limitations of the bioinformatics pipeline requires multiple samples to go through the same preceding process. Abbreviations: CNV, copy number variation; EHR, electronic health record; FFPE, formalin-fixed, paraffin embedded; FISH, fluorescence in situ hybridization; GIAB NIST, Genome in a Bottle National Institute of Standards and Technology; NGS, next-generation sequencing; SOP, standard operating procedures; SV, structural variant.*

variant detection can be evaluated for a bioinformatics pipeline.

In addition to GIAB, DNA or cell lines can be purchased by Corielle and run through the clinical assay. Cell lines can be fixed with FFPE, with resulting variants compared to that of native DNA in order to determine technical artifacts common from fixation (DNA fragmentation resulting in shorter read lengths).

Engineered reference samples that contain specific clinically relevant mutations (positive controls) can also be purchased for commercial entities, including Horizon Discovery and SeraCare. Engineered reference samples carrying distinct variants can be used to calculate accuracy in determining mutation allele frequency and lower limit of allele detection. Cell lines developed from specific diseases or known to carry specific mutations can also be used as reference material. The DNA from these cell lines can be mixed with the NA12878 GIAB sample at several ratios to determine the limit of detection of the bioinformatics pipeline. Mixed samples can be compared to pure NA12878 GIAB to evaluate the accuracy of a somatic mutation workflow for cancer assays. Commercial entities provide a list of verified mutations present in these samples. The allelic frequency is often determined by digital-droplet PCR but may estimate allele frequency differently than an NGS platform.

Synthetic samples can be created bioinformatically. Programs such as BamSurgeon can insert model SNVs and indels in BAM files at positions and frequencies specified by the user.[84,85] The bioinformatic pipeline is evaluated for the detection and accuracy of all variant types that are anticipated to be encountered by the assay. In the case of cancer bioinformatics pipelines, the original and modified BAM files can test the accuracy of detecting somatic mutations.

Validation Step 2: Clinical Samples

Concordance of results with clinical samples that have been tested by a Clinical Laboratory Improvement Amendments (CLIA)–accredited clinical laboratory are necessary to validate the entire assay, including the bioinformatics pipeline. Clinical samples should be the same sample type and have the same diversity of variation as the intended tested samples.

For germ-line testing, the typical sample types are blood or saliva. At minimum 10 to 20 samples containing each of the different expected variant types (CNVs, SNVs, indels, or SVs) are necessary to assess accurate variant detection. For rare diseases, parents and siblings are sequenced in order to determine disease-causing mutations. Validation of accurate genotype prediction for proband and relative is necessary to identify candidate disease causing mutations. For example, if both parents have no mutation where the proband has a new variant in a gene with dominant inheritance, the new variant would be classified as de novo, which increases the support for the variant being pathogenic according to American College of Medical Genetics criteria. Thus, bioinformatic pipelines for germ-line testing must be able to handle sequencing data from a proband and relatives.

In pan-cancer testing, FFPE, fresh blood, and bone marrow samples are the common sample types. The FFPE samples from solid tumors should include a variety of tissue sites, including breast, lung, skin, bladder, pancreas, adrenal, and kidney. The blood and bone marrow samples should include a diversity of diseases, including cases with a diagnosis of acute and chronic leukemias of myeloid or lymphoid origin. At minimum 10 to 20 samples for each sample type with a diversity of expected variant types, including SNVs, indels, CNVs, and gene fusions, are necessary to determine pipeline accuracy. Positive controls for gene fusions can come from samples with a positive fluorescence in situ hybridization result. Copy number changes can be confirmed by microarray or cytogenetic test results.

Validation Step 3: Precision Study

Reproducibility of the assay and bioinformatics pipeline is evaluated using technical replicates (1) prepared by different technologists, (2) run on all the different instruments that will be used for the clinical assay, (3) by run, (4) analyzed on different available hardware, and (5) analyzed by different bioinformaticians.

Reproducibility is demonstrated by a precision study consisting of intrarun and interrun reproducibility. Cases with a combination of SNV, SV, and indels are run in triplicate on the same sequencer run and in triplicate during 3 separate sequencing runs, with appropriate bar coding, to affirm the same results (>98% concordance) are obtained for all 3 specimens. The precision study should also assess reproducibility for any variable that might affect consistency of results. Laboratory scientists and bioinformaticians performing the assay should be compared to ensure reproducibility. Other factors, such as individual sequencing machines and computing hardware, should all be compared in the precision study to ensure concordant results are obtained from all instruments and hardware that will be used in the clinical assay.[18,82]

A special consideration in the precision of NGS assay studies is the reduction of recurrent false-positive variants. Any assay will have recurrent technical artifacts of sequencing that will appear in multiple runs. These variants can be filtered out by triplicate sequencing of control genomic DNA then identifying variants that should not occur and marking them as technical artifacts that can be bioinformatically filtered out.

Of note, sample quality is affected by many factors, including fixation time, contamination by necrotic tissue, and tumor cell percentage. Accuracy will be reduced in poor-quality specimens. Therefore, any tuning of the bioinformatic pipeline to reduce errors in samples of poor quality must be performed in the optimization stage before validation begins. This process will also determine the reportable and nonreportable regions of the assay.

There will always be limitations to any assay, and these must be disclosed as a part of the validation.[86] Sequencing can be difficult to perform in areas with GC-rich content, long homopolymer runs, or low complexity. Certain clinically relevant genes are known to be difficult to sequence, for example *CEPBA* and the *TERT* promoter.[82] It may not be the fault of the bioinformatic pipeline when sequencing chemistry yields poor coverage of an area, but proper disclosure of these limits must be described.

Validation Step 4: Standard Operating Procedures and Documentation

Clinical laboratories have stringent documentation requirements to demonstrate compliance, and this includes the bioinformatic pipeline. Elements of documentation include the name of the pipeline, version number, developer, software and hardware, networks the pipeline

is connected to, backups (location and frequency), the system to transmit data, and technical support available for each component. Documentation for software should not only include the algorithm and code used but also the operating system, database where information is held, and transmission method.[82]

The pipeline is processing clinical data, and therefore Health Insurance Portability and Accountability Act (HIPAA) regulations apply. Samples should be identified with 4 unique patient identifiers, which is more than the usual 2 patient identifiers used in face-to-face interactions. Examples of essential identifiers include: (1) sample ID, (2) unique patient identifier, (3) run number, and (4) laboratory location. Although a laboratory might be inclined to use a specimen ID like S19-100, consider that this code is not unique, and many institutions may use such an identifier. When choosing the name, certain Health Language 7 (HL7) incompatible symbols (~| \ ^ & and #) should be avoided because HL7 is the required medium of transmitting patient care information.[18]

Validation Step 5: Updating and Revalidation

If updates to any software or any changes in any component of the pipeline are performed, the whole process must be validated again. This should be as straightforward as rerunning an appropriate number of previous runs through the upgraded pipeline and ensuring the results are equivalent. Often, an appropriate number of runs would be from 3 to 5, ensuring that a variety of mutation types are included (SNV, indel, SV, etc). As referenced below in Table 4, revalidation must be performed for simple changes, and this includes changes in even 1 line of code or the order of algorithms even if the algorithms themselves did not change. Thus, optimization must be heavily emphasized to avoid unnecessary revalidations.[18]

An important part of optimization is to challenge the system with a breadth of horizontally and vertically complex variant types, which could be encountered in clinical testing. Horizontally complex variants are 3 or more SNPs or indels (up to 21 bp in length) on the same read strand. Vertically complex variants are defined as 3 or more variants in the same region but on different read strands. Vertically complex variants could occur in the case of tumor heterogeneity (subclonal populations), mosaicism, compound heterozygotes, or sequencing artifacts.[18] Challenging the clinical bioinformatic pipeline with complex clinical samples or in silico–derived variants will produce a robust pipeline capable of detecting important variants without repetitive revalidations.

REPRODUCIBILITY ACROSS CLINICAL LABORATORIES AND PROFICIENCY TESTING

The CLIA law of 1988 originated out of a need to ensure laboratory standardization to ensure that a patient could go to any laboratory in the nation and receive a comparable result.[86] Proficiency testing with common material sent to all labs is one of the ways to confirm that all labs are reporting comparable results.

Even though clinical laboratories compare their results to results from other CLIA-validated laboratories in the validation process, often the full results cannot be compared because (1) most commercial laboratories only report the clinically actionable variants and not all variants observed, (2) each laboratory might be using different gene panels or capturing only a subset of the captured genes, and (3) each laboratory will have differences in their assay limit of detection, depth of coverage, and tested variant types. Discrepancies in bioinformatic analysis have been shown in 2 studies that performed interinstitutional FASTQ file exchanges.[88,89] Because of these challenges, the validation typically only examines differences in the sensitivity of compared laboratories.

For College of American Pathologists (CAP) accreditation, CAP-accredited laboratories are expected to perform proficiency testing at least twice yearly to ensure laboratories provide comparable results. NGS laboratories have proficiency tests (PTs) for the wet lab and dry lab: (1) DNA to test the entire lab workflow from sequence generation to variant prediction, and (2) sequencing results to test only the bioinformatic portion of the assay. Proficiency testing for the bioinformatic pipeline was deemed necessary because it is difficult to include a large range of complex variants in physical specimens.[90] The bioinformatic PT presents unique challenges because of the fact that bioinformatics workflows are designed for a specific assay and might not have the same level of accuracy with a different wet lab protocol, and this includes (1) intended sample types, (2) hybridization- versus amplicon-based gene panels, and (3) different sequencing platforms. For instance, if duplicate removal is part of the bioinformatic pipeline for a hybrid-based assay, all of the amplicons from a PT sample would be removed. CAP recognizes this challenge and will soon allow laboratories to send sequence files (FASTQ), which will be synthetically altered to include various frequencies of genetic mutation. Because this is a new type of PT, there will be implementation challenges as bioinformatics pipelines are tuned to the assay for which they were designed. Furthermore, complex in silico samples can be made with ease, which is discordant from the level of complexity tested in wet bench PT samples that focus on clinically relevant variants. These wet bench PT samples still involve bioinformatic analysis to determine results, which calls into question the necessity of additional bioinformatic proficiency testing.

As the CAP continues to improve upon the bioinformatic proficiency testing process, the regulations and guidelines that are monitored during inspection are also evolving for bioinformatics. The only professional society guidelines have been published by the Association for Molecular Pathology and are based mostly on expert opinion, with few empirical studies examining best practices of clinical NGS bioinformatics because it is still so new. The guideline and checklist both emphasize pipeline optimization before initiating validation, as is similar to regular assay validation. Furthermore, data storage and transfer must be consistent processes, with standard operating procedures written for these steps too. Because storage is becoming a growing cost for high-depth sequencing, there are different levels of storage that usually have a tradeoff between accessibility and cost (eg, Amazon Web Service has a deep freeze option that is much cheaper, but data access may take a few days). Lastly, the CAP inspection includes a review of bioinformatic processes, but the depth of the review will vary with the inspector's fluency in this area.

SUMMARY

Bioinformatics pipeline development for NGS is a rigorous process where all of the variables affecting accuracy

of the clinical assay must be taken into consideration along with the unique features of the wet-lab protocols. Because there is no standard protocol used by all clinical laboratories to generate sequence data from tissue, there is also no standard bioinformatics protocol to identify clinically actionable genetic variants. Regardless, there are essential elements for any bioinformatics pipeline, which include the reference genome, sequence alignment, variant detection (SNVs, indels, CNVs, and SVs), quality filtering, and clinical annotation. The bioinformatics pipeline validation is an integral part of the clinical assay validation. Validation examines the sensitivity, specificity, precision/reproducibility, and limit of detection of the whole clinical test, including the bioinformatics pipeline. Finally, proficiency testing is necessary to determine the performance of the test and ensure consistency across clinical laboratories.

References

1. Meric-Bernstam F, Farhangfar C, Mendelsohn J, Mills GB. Building a personalized medicine infrastructure at a major cancer center. *J Clin Oncol*. 2013;31(15):1849–1857.
2. Yohe S, Hauge A, Bunjer K, et al. Clinical validation of targeted next-generation sequencing for inherited disorders. *Arch Pathol Lab Med*. 2015; 139(2):204–210.
3. Kothiyal P, Wong WSW, Bodian DL, Niederhuber JE. Mendelian inconsistent signatures from 1314 ancestrally diverse family trios distinguish biological variation from sequencing error. *J Comput Biol*. 2019;26(5):405–419.
4. Frampton GM, Fichtenholtz A, Otto GA, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol*. 2013;31(11):1023–1031.
5. Singh RR, Patel KP, Routbort MJ, et al. Clinical validation of a next-generation sequencing screen for mutational hotspots in 46 cancer-related genes. *J Mol Diagn*. 2013;15(5):607–622.
6. Rennert H, Eng K, Zhang T, et al. Development and validation of a whole-exome sequencing test for simultaneous detection of point mutations, indels and copy-number alterations for precision cancer care. *NPJ Genom Med*. 2016;1:pii: 16019.
7. Pritchard CC, Salipante SJ, Koehler K, et al. 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(1):56–67.
8. Tsongalis GJ, Peterson JD, de Abreu FB, et al. Routine use of the Ion Torrent AmpliSeq Cancer Hotspot Panel for identification of clinically actionable somatic mutations. *Clin Chem Lab Med*. 2014;52(5):707–714.
9. Cottrell CE, Al-Kateb H, Bredemeyer AJ, et al. Validation of a next-generation sequencing assay for clinical molecular oncology. *J Mol Diagn*. 2014; 16(1):89–105.
10. Singh RR, Patel KP, Routbort MJ, et al. Clinical massively parallel next-generation sequencing analysis of 409 cancer-related genes for mutations and copy number variations in solid tumours. *Br J Cancer*. 2014;111(10):2014–2023.
11. Cheng DT, Mitchell TN, Zehir A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J Mol Diagn*. 2015;17(3):251–264.
12. Kanagal-Shamanna R, Portier BP, Singh RR, et al. 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(2):314–327.
13. Luthra R, Patel KP, Reddy NG, et al. Next-generation sequencing-based multigene mutational screening for acute myeloid leukemia using MiSeq: applicability for diagnostics and disease monitoring. *Haematologica*. 2014; 99(3):465–473.
14. Stephens PJ, Greenman CD, Fu B, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*. 2011; 144(1):27–40.
15. Kamps R, Brandao RD, Bosch BJ, et al. Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification. *Int J Mol Sci*. 2017;18(2):pii: E308.
16. Jennings LJ, Arcila ME, Corless C, et al. 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(3):341–365.
17. Li MM, Datto M, Duncavage EJ, et al. 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(1):4–23.
18. Roy S, Coldren C, Karunamurthy A, et al. 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. *J Mol Diagn*. 2018;20(1):4–27.
19. Paszkiewicz KH, Farbos A, O'Neill P, Moore K. Quality control on the frontier. *Front Genet*. 2014;5:157.
20. Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*. 2017;109(2):83–90.
21. Pan B, Kusko R, Xiao W, et al. Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinformatics*. 2019;20(1):252.
22. Ballouz S, Dobin A, Gingeras TR, Gillis J. The fractured landscape of RNA-seq alignment: the default in our STARs. *Nucleic Acids Res*. 2018;46(10):5125–5138.
23. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36.
24. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. 2016;11(9):1650–1667.
25. Hsieh G, Bierman R, Szabo L, et al. Statistical algorithms improve accuracy of gene fusion detection. *Nucleic Acids Res*. 2017;45(13):e126.
26. Sahraeian SME, Mohiyuddin M, Sebra R, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat Commun*. 2017;8(1):59.
27. Hatem A, Bozdağ D, Toland AE, Çatalyürek ÜV. Benchmarking short sequence mapping tools. *BMC Bioinformatics*. 2013;14:184.
28. Otto C, Stadler PF, Hoffmann S. Lacking alignments?: the next-generation sequencing mapper segemehl revisited. *Bioinformatics*. 2014;30(13):1837–1843.
29. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079.
30. Ebbert MT, Wadsworth ME, Staley LA, et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*. 2016;17(suppl 7):239.
31. Broad Institute. Picard tools. 2016. https://broadinstitute.github.io/picard/.
32. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–1303.
33. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*. 2012;28(14):1811–1817.
34. Sandmann S, de Graaf AO, Karimi M, et al. Evaluating variant calling tools for non-matched next-generation sequencing data. *Sci Rep*. 2017;7:43169.
35. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci Rep*. 2015;5: 17875.
36. Cantarel BL, Weaver D, McNeill N, Zhang J, Mackey AJ, Reese J. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinformatics*. 2014;15(1):104.
37. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213–219.
38. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568–576.
39. Shiraishi Y, Sato Y, Chiba K, et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res*. 2013;41(7):e89.
40. Kim S, Jeong K, Bhutani K, et al. Virmid: Accurate detection of somatic mutations with sample impurity inference. *Genome Biol*. 2013;14(8):R90.
41. Hansen NF, Gartner JJ, Mei L, Samuels Y, Mullikin JC. Shimmer: detection of genetic alterations in tumors using next generation sequence data. *Bioinformatics*. 2013;29(12):1498–1503.
42. Krøigård AB, Thomassen M, Lænkholm AV, Kruse TA, Larsen MJ. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. *PLoS One*. 2016;11(3):e0151664.
43. Wood DE, White JR, Georgiadis A, et al. A machine learning approach for somatic mutation discovery. *Sci Transl Med*. 2018;10(457):pii: eaar7939.
44. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156–2158.
45. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;6:80–92.
46. Boeva V, Popova T, Lienard M, et al. Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. *Bioinformatics*. 2014;30(24):3443–3450.
47. Jacoby MA, Duncavage EJ, Walter MJ. Implications of tumor clonal heterogeneity in the era of next-generation sequencing. *Trends Cancer*. 2015; 1(4):231–241.
48. Reble E, Castellani CA, Melka MG, O'Reilly R, Singh SM. VarScan2 analysis of de novo variants in monozygotic twins discordant for schizophrenia. *Psychiatr Genet*. 2017;27(2):62–70.
49. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput Biol*. 2016;12(4):e1004873.

50. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*. 2012;28(3):423–425.
51. Rieber N, Bohnert R, Ziehm U, Jansen G. Reliability of algorithmic somatic copy number alteration detection from targeted capture data. *Bioinformatics*. 2017;33(18):2791–2798.
52. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol*. 2019;20(1):213.
53. McPherson A, Wu C, Wyatt AW, Shah S, Collins C, Sahinalp SC. NFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Res*. 2012;22(11):2250–2261.
54. Benelli M, Pescucci C, Marseglia G, Severgnini M, Torricelli F, Magi A. Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics*. 2012;28(24):3232–3239.
55. Kumar S, Razzaq SK, Vo AD, Gautam M, Li H. Identifying fusion transcripts using next generation sequencing. *Wiley Interdiscip Rev RNA*. 2016;7(6):811–823.
56. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;25(21):2865–2871.
57. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014;15(6):R84.
58. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;28(18):i333–i339.
59. Spencer DH, Abel HJ, Lockwood CM, et al. Detection of FLT3 internal tandem duplication in targeted, short-read-length, next-generation sequencing data. *J Mol Diagnostics*. 2013;15(1):81–93.
60. Mohiyuddin M, Mu JC, Li J, et al. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics*. 2015; 31(16):2741–2744.
61. Mizuguchi T, Suzuki T, Abe C, et al. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. *J Hum Genet*. 2019;64(5):359–368.
62. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. 2019; 47(D1):D766–D773.
63. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733–45.
64. Frankish A, Uszczynska B, Ritchie GRS, et al. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics*. 2015;16(suppl 8):S2.
65. Steward CA, Parker APJ, Minassian BA, Sisodiya SM, Frankish A, Harrow J. Genome annotation for clinical genomic diagnostics: strengths and weaknesses. *Genome Med*. 2017;9(1):49.
66. Rodriguez JM, Rodriguez-Rivas J, Di Domenico T, Vázquez J, Valencia A, Tress ML. APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res*. 2018;46(D1):D213–D217.
67. Lener MR, Kashyap A, Kluźniak W, et al. The prevalence of founder mutations among individuals from families with familial pancreatic cancer syndrome. *Cancer Res Treat*. 2017;49(2):430–436.
68. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285–291. doi:10.1038/nature19057
69. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74.
70. Chial H. Rare genetic disorders: learning about genetic disease through gene mapping, SNPs, and microarray data. *Nat Educ*. 2008;1(1):192.
71. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. 2016;44(D1): D862–8.
72. Henrie A, Hemphill SE, Ruiz-Schultz N, et al. ClinVar Miner: demonstrating utility of a Web-based tool for viewing and filtering ClinVar data. *Hum Mutat*. 2018;39(8):1051–1060.
73. Richards S, Aziz N, Bale S, et al. 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(5):405–424.
74. Litchfield K, Turajlic S, Swanton C. The GENIE is out of the bottle: landmark cancer genomics dataset released. *Cancer Discov*. 2017;7(8):796–798.
75. Griffith M, Spies NC, Krysiak K, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet*. 2017;49(2):170–174.
76. Chakravarty D, Gao J, Phillips S, et al. OncoKB: a precision oncology knowledge base. *JCO Precis Oncol*. 2017. doi: 10.1200/PO.17.00011.
77. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*. 2018;34(3):511–513.
78. Li MX, Kwan JSH, Bao SY, et al. Predicting Mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet*. 2013;9(1):e1003143.
79. Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet*. 2015;24(8):2125–2137.
80. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet*. 2016;48(12):1581–1586.
81. Isakov O, Perrone M, Shomron N. Exome sequencing analysis: a guide to disease variant detection. *Methods Mol Biol*. 2013;1038:137–158.
82. Schneider T, Smith GH, Rossi MR, Hill CE, Zhang L. Validation of a customized bioinformatics pipeline for a clinical next-generation sequencing test targeting solid tumor–associated variants. *J Mol Diagn*. 2018;20(3):355–365.
83. Genome in a bottle–a human DNA standard. *Nat Biotechnol*. 2015; 33(675). doi:10.1038/nbt0715-675a
84. Lee AY, Ewing AD, Ellrott K, et al. Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. *Genome Biol*. 2018;19(1):188.
85. Meng J, Chen YPP. A database of simulated tumor genomes towards accurate detection of somatic small variants in cancer. *PLoS One*. 2018;13(3): e0202982.
86. Kim J, Park WY, D Kim NK, et al. Good laboratory standards for clinical next-generation sequencing cancer panel tests. *J Pathol Transl Med*. 2017;51(3): 191–204.
87. Bachner P, Hamlin W. Federal regulation of clinical laboratories and the Clinical Laboratory Improvement Amendments of 1988–part II. *Clin Lab Med*. 1993;13(4):987–994.
88. Davies KD, Farooqi MS, Gruidl M, et al. Multi-institutional FASTQ file exchange as a means of proficiency testing for next-generation sequencing bioinformatics and variant interpretation. *J Mol Diagn*. 2016;18(4):572–579.
89. Duncavage EJ, Abel HJ, Merker JD, et al. A model study of in silico proficiency testing for clinical next-generation sequencing. *Arch Pathol Lab Med*. 2016;140(10):1085–1091.
90. Duncavage EJ, Abel HJ, Pfeifer JD. In silico proficiency testing for clinical next-generation sequencing. *J Mol Diagn*. 2017;19(1):35–42.