Review of Clinical Next-Generation Sequencing

Sophia Yohe, MD; Bharat Thyagarajan, MD, PhD

• *Context.*—**Next-generation sequencing (NGS) is a technology being used by many laboratories to test for inherited disorders and tumor mutations. This technology is new for many practicing pathologists, who may not be familiar with the uses, methodology, and limitations of NGS.**

***Objective.*—To familiarize pathologists with several aspects of NGS, including current and expanding uses; methodology including wet bench aspects, bioinformatics, and interpretation; validation and proficiency; limitations; and issues related to the integration of NGS data into patient care.**

***Data Sources.*—The review is based on peer-reviewed literature and personal experience using NGS in a clinical setting at a major academic center.**

***Conclusions.*—The clinical applications of NGS will increase as the technology, bioinformatics, and resources evolve to address the limitations and improve quality of results. The challenge for clinical laboratories is to ensure testing is clinically relevant, cost-effective, and can be integrated into clinical care.**

(*Arch Pathol Lab Med.* 2017;141:1544–1557; doi: 10.5858/arpa.2016-0501-RA)

Next-generation sequencing (NGS) or massively parallel sequencing, a method of simultaneously sequencing millions of fragments of DNA (or complementary DNA), has been rapidly adopted in the clinical laboratory because of its ability to simultaneously analyze several genes or gene regions with a single test compared to traditional methods. As with any new technology, the use of NGS in the clinical laboratory has evolved and will continue to evolve over time. New applications for the technology continue to be developed, new bioinformatics and wet bench techniques are being developed to address current limitations and improve performance, and new knowledge regarding interpretation of rare variants is being accumulated. This article is an overview of clinical NGS, including recent trends as well as evolution that will likely occur in the near future. The review is based on peer-reviewed literature and personal experience using NGS in a clinical setting at a major academic center. The Molecular Diagnostics Laboratory at the University of Minnesota Medical Center, Fairview, has offered a capture-based NGS inherited disease test covering 568 genes since 2012 and expanded to 2484 genes in 2014.[1–4] In addition, since 2014 we have offered a 21-gene hotspot NGS panel for oncology (hematologic malignancies and solid tumors).[5] Our laboratory tests approximately 800 NGS inherited disease and 800 NGS oncology cases a year, and the 2 authors sign out approximately two-thirds of those cases. The first author also participates in a committee for a national pathology organization in which NGS-related issues have been discussed and addressed.[6,7]

CURRENT AND EXPANDING USES OF NGS

Next-generation sequencing is an established test method for germline (inherited) and somatic (acquired mutations) genetic mutations in many clinical laboratories. For inherited diseases, testing for germline mutations may include targeted panel, whole exome, whole genome, or mitochondrial DNA sequencing.[8,9] Targeted panel testing, which varies between laboratories, is possible for a wide variety of inherited disorders such as immune deficiencies, bone marrow failure syndromes, blindness, deafness, mitochondrial disorders, renal disorders, neurologic disorders, connective tissue disorders, cardiomyopathies, and cancer predisposition syndromes, among others.[10–17] Targeted panels for genes associated with a clinical phenotype are usually the first line of testing for inherited disorders, while whole exome sequencing is reserved for cases in which targeted testing has been uninformative.[18,19] Whole exome testing often involves testing the child and both parents (trio testing) to assist in the interpretation of variants.[19] In addition, NGS technologies are used in analyzing cell-free DNA in the prenatal setting.[20,21]

Targeted panels for cancer testing also vary between laboratories.[6] Targeted panels may be broad, including genes for both solid and hematologic malignancies, or may be more focused for a particular type of malignancy (such as myeloid neoplasms).[6] Any given gene within a panel may be completely sequenced or only partially sequenced (eg, hotspot regions). For both germline and somatic testing, it is important to know the content of the targeted panels when deciding on using a test. Whole exome and whole genome sequencing are not currently used clinically for oncology testing.

Several new applications for NGS have more recently moved into the clinical arena or are being actively researched for clinical use, including circulating tumor

Accepted for publication February 9, 2017.
Published as an Early Online Release August 7, 2017.
From the Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis.
The authors have no relevant financial interest in the products or companies described in this article.
Reprints: Sophia Yohe, MD, Divisions of Hematopathology and Molecular Pathology, Department of Laboratory Medicine and Pathology, University of Minnesota, MMC 609, 420 Delaware St SE, Minneapolis, MN 55455-0385 (email: yohe0001@umn.edu).

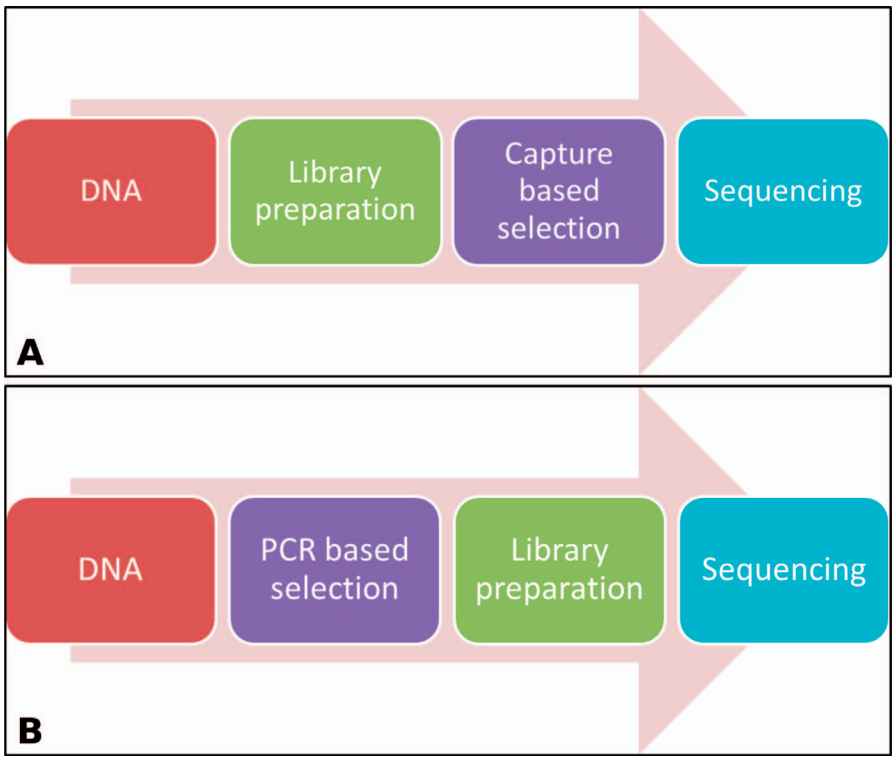

Figure 1. *A, Overview of wet bench steps for capture-based sequencing. DNA undergoes library preparation followed by capture-based selection before sequencing. B, Overview of wet bench steps for polymerase chain reaction (PCR)–based sequencing. The PCR selection step occurs before library preparation or may be combined with the library preparation step in PCR-based sequencing.*

DNA testing, human leukocyte antigen (HLA) typing, microbial analysis, RNA sequencing and expression, and methylation. Some of these new uses of NGS may be helped by the unique advantages of new instruments that are now available (see section on New Instrumentation). The use of NGS for HLA typing had some challenges to overcome: differentiating low-frequency alleles from high-frequency artifacts and distinguishing 2 similar alleles as 2 distinct alleles.[22,23] However, newer data analysis techniques such as stepwise threshold clustering have allowed NGS to be explored as a clinical option for HLA typing.[22,24] The use of NGS for identity testing using short tandem repeats (STRs) runs into the same problems as other repeat regions (see Difficult-to-Sequence Areas below); however, again newer data analysis techniques are making headway against this problem and may be applicable to other repeat regions.[25] Other uses for clinical NGS include pharmacogenetics, microbial sequencing, and advanced blood group typing (eg, type A1 versus type A2). Further discussion of these topics is beyond the scope of this article.

Cell-free DNA has been used for some time for prenatal testing; however, NGS of circulating tumor DNA (ctDNA), tumor-derived cell-free DNA, is a more recent development that is now clinically available.[20,26,27] This testing is often referred to as a *liquid biopsy*. The potential applications of sequencing ctDNA include screening or diagnosis of cancer, monitoring for progression or relapse, and guiding therapy for a patient with a known cancer diagnosis. Most research studies have evaluated the ability of ctDNA sequencing to detect somatic mutations in patients with known cancer and the ability to monitor disease.[26–29] Monitoring a known mutation by sequencing of ctDNA has been shown in several studies to correlate with relapse/progression of disease.[29,30] Also, using detection of mutations in ctDNA to help guide therapy for a patient with a known tumor has shown utility, for example, tyrosine kinase inhibitor response with epidermal growth factor receptor (*EGFR*)–activating mutations in lung cancer.[26,27,31,32] Although ctDNA may have lower sensitivity for detecting mutations than testing tumor tissue, the most common clinical applications of ctDNA appear to be for patients with metastatic cancers when there is insufficient tissue for testing and a repeated biopsy would be associated with significant morbidity and mortality, and when testing of ctDNA is a reasonable alternative.[26,27,31] Using ctDNA to screen for or diagnose early-stage cancers is more problematic.[33] Most studies of patients with known cancer have not contained normal controls, but a limited number of targeted sequencing studies have shown some degree of mutation detection (false positives), albeit usually at a low level, in normal controls.[28,34,35] Low sensitivity to detect early-stage cancer (false negatives) is another limitation. Studies have shown sensitivities in the 30% to 60% range for early-stage tumors, and some tumor types may have a higher false-negative rate, as ctDNA appears to be released owing to apoptosis and necrosis.[36–41] These false-positive and false-negative issues limit the practical use of ctDNA for early cancer diagnosis or screening.

CURRENT METHODOLOGY OF CLINICAL NGS

Wet Bench Steps

Samples undergo DNA extraction, library preparation, target enrichment, and sequencing[42] (Figure 1, A and B).

DNA Extraction.—Almost all DNA extraction methods are acceptable. Extraction methods for formalin-fixed, paraffin-embedded (FFPE) tissue may need special care, and macrodissection or microdissection to enrich for tumor may be required for some cases.[43] DNA quantitation is performed by Qubit or Picogreen (Thermo Fisher Scientific, Waltham, Massachusetts) rather than by standard spectrophotometry.[44]

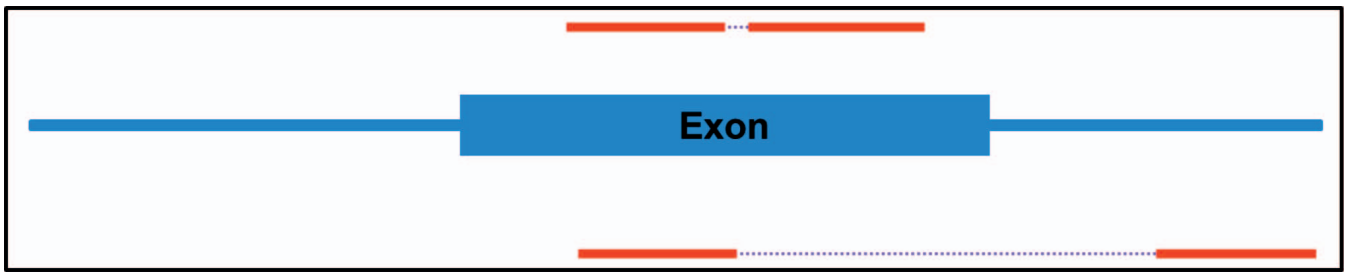

Figure 2. *Fragments with a short DNA insert size (top) are more likely to have both paired end reads (red bars) fall within the exon. Fragments with a long insert size are more likely to span the breakpoint of a rearrangement, which often occurs in the intron. Reprinted from Yohe SL. Hot topic spotlight—new frontiers in clinical next-generation sequencing. In: Cushman-Vokoun AM, Anderson WB, eds.* Precision Medicine Resource Guide. *Northfield, IL: College of American Pathologists; 2016:12. With permission from College of American Pathologists. Copyright 2016.*

Library Preparation.—Library preparation refers to the process of preparing DNA for use on a sequencer. Although many methods are available, they all result in breaking DNA into fragments and adding adaptors to the ends.[45–48] Adaptors may include molecular bar codes (to allow pooling of patient samples), universal polymerase chain reaction (PCR) primers, hybridization sequences to bind the DNA fragment to a surface, and recognition sites to initiate sequencing. The term *library* refers to these fragments of DNA with flanking adaptors that are ready for sequencing. The size of the DNA fragment between the adaptors is referred to as the *insert size*. The insert size may vary and there are different advantages to short insert sizes and long insert sizes. Shorter fragments are more likely to have both ends fall within an exon, which is often the area of interest, while longer fragments are more likely to have 1 end fall in an intron, which may increase detection of structural rearrangements if only exonic regions are being selected (Figure 2). For more details on structural rearrangements, please see Structural Variation and Copy Number Variation.

Target Enrichment.—The resulting library undergoes enrichment for both whole exome analysis and targeted testing or is sequenced directly for whole genome analysis. Enrichment may be performed by hybridization to complementary sequences (sequence capture) or by PCR.[49] Enrichment by PCR is generally combined with the library preparation step, as the primers that select the regions of interest may also contain the adaptor sequences. The choice of enrichment strategy is often dictated by the clinical use: sequence capture is preferred for large genomic regions, and PCR for smaller regions where greater enrichment is desired.[49]

Sequencing.—Most clinical sequencing is performed on 1 of 2 main types of instruments: Illumina sequencers (San Diego, California) including the HiSeq, MiSeq, and NexSeq; or the Ion Torrent series of machines including the IonPGM, IonProton, and IonS5 (Thermo Fisher Scientific). These 2 types of machines differ in their chemistry, detection methods, advantages, and disadvantages[42,50–54] (Table 1).

The first sequencing step for both the Illumina and Ion Torrent platforms is to immobilize each DNA fragment and clonally amplify it. Clonal amplification is needed to generate a large enough signal for detection. The Ion Torrent uses a bead emulsion for immobilization and clonal amplification, whereas the Illumina sequencers use a flow cell.[51,53,54] The flow cell or bead contains sequences that hybridize to part of the adaptor on the DNA fragments. The input DNA concentration is critical to ensure only 1 DNA fragment binds per bead and to ensure that the DNA fragments are well spaced out on the flow cell. The clonal amplification step creates a bead or cluster with approximately 1000 identical copies of a unique parent DNA molecule that are physically isolated from other molecules.[54] For the Ion Torrent the beads are then placed in a well (1 bead per well).[53]

Illumina sequencers use sequencing by synthesis with fluorescent detection[53,54] (Figure 3, A through D). All 4 fluorescently tagged nucleotides are added and compete for the next space. The complementary tagged nucleotide will bind but a blocker prevents addition of more than 1 nucleotide per round (reversible terminator chemistry). The remaining nonbound nucleotides are washed away. Laser excitation leads to a fluorescent emission that is recorded (simultaneously for each DNA fragment cluster). The fluorescent tag and blocker are cleaved, then the next round begins. In each round, 1 base pair is read from each DNA cluster. This process can be repeated on the opposite end of the DNA fragment, which is referred to as *paired end reads* (Table 2).

Ion Torrent sequencing is different, as only a single base is added in each round (eg, A in round 1, T in round 2).[53] When an added base is incorporated, a hydrogen ion is released, accompanied by a pH change that is detected for each bead within a well; if a base is not incorporated there is no voltage generated.[53,54] Incorporation of more than 1 of the same base leads to a proportionately higher voltage signal up to about 6 to 8 bases[51,55] (Figure 3, E). If more than 6 to 8 bases are incorporated, the signal is no longer

Table 1. Comparison of Illumina and Ion Torrent Platforms

Platform	Local Clonal Amplification	Detection	Read Length, bases	Pros	Cons
Illumina[a]	Flow cell	Fluorescent	100–300	Paired end reads	Errors in GC-rich regions
Ion Torrent[b]	Bead and emulsion	Ion (pH)	100–400	Short run time	Homopolymer errors
				Paired end reads[c]	Truncation errors

[a] Illumina, San Diego, California.
[b] ThermoFisher, Waltham, Massachusetts.
[c] Available on newer instruments only.

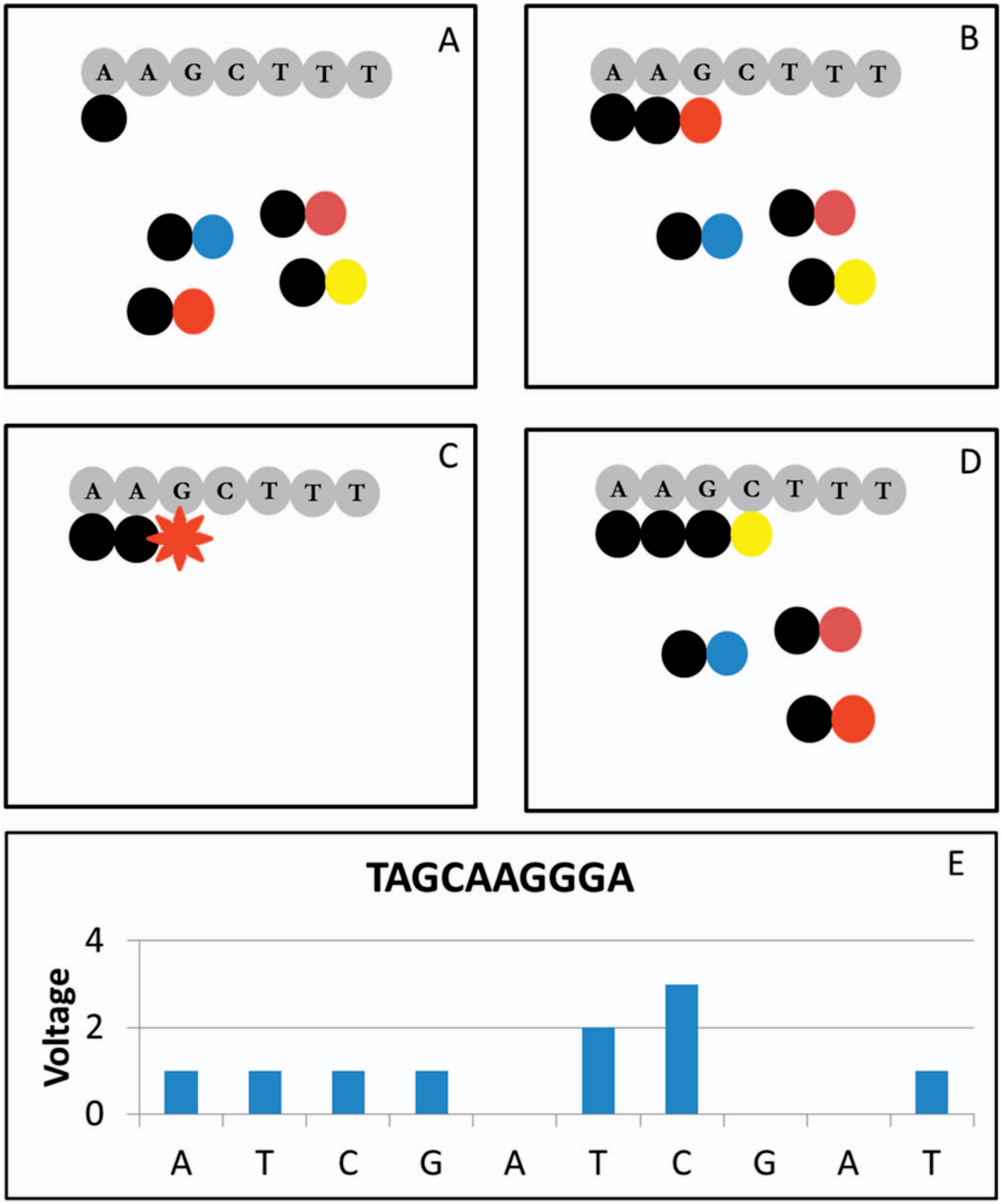

Figure 3. *Illustration of Illumina sequencing by synthesis (A through D) and Ion Torrent ion-based sequencing (E). A, Fluorescently tagged nucleotides (black circles with colored circles) compete for the next complementary space on the DNA strand (gray circles). B, When a fluorescently tagged nucleotide is incorporated, it blocks further addition of nucleotides. C, The flow cell is washed, removing additional fluorescently tagged nucleotides, and a laser signal leads to fluorescent emission. D, The fluorescent tag and blocker are removed and washed away, allowing incorporation of the next base during the next cycle. This occurs simultaneously for all DNA strands in a cluster and all clusters on the flow cell. E, In each cycle, a single base is added in a set pattern. For this example, the order of base additions is A, T, C, and G, which then repeats. If a base is incorporated an ion is released, leading to a pH (voltage) change that is proportional to the number of bases added in a row. Reprinted from Yohe SL. Hot topic spotlight—new frontiers in clinical next-generation sequencing. In: Cushman-Vokoun AM, Anderson WB, eds.* Precision Medicine Resource Guide. *Northfield, IL: College of American Pathologists; 2016:13–14. With permission from College of American Pathologists. Copyright 2016.*

Table 2. Definitions

Term	Definition
Alignment	To compare a sequence read to another sequence and determine where it belongs. There are 2 types of alignment: de novo assembly or resequencing.
De novo assembly	A sequence read is compared to all the other sequence reads of that sample to determine a consensus sequence.
Resequencing	A sequence read is compared to a reference sequence (eg, the reference human genome). Also referred to as *mapping*.
Bait	An artificial construct that is able to target the sequence of interest (eg, a complementary DNA or RNA sequence) and can be used to isolate that target sequence. Used for sequence capture target enrichment.
Demultiplex	Separate an individual sample's reads from the pooled reads of multiple samples by unique identifier codes that were attached before pooling.
Map/mapping	To compare a sequence read to a reference and determine where it belongs. See also Alignment, Resequencing.
Read	May refer to either the sequence result of a single base pair position or to the sequence result of a sequential length of base pair reads from a single clonally amplified DNA cluster.

proportional and the exact amount cannot be determined.[51,55]

Bioinformatics

The raw data reads from either type of instrument undergo a series of bioinformatics processes (also referred to as a *pipeline*) to ultimately deliver a variant call file (VCF)[56,57] (Table 3). These processes include demultiplexing (Table 2), quality analysis, mapping of the reads to a reference genome (resequencing), and variant identification/annotation.[43] Because of these specialized processes, dedicated bioinformatics personnel may be needed to set up and maintain a clinical NGS service.

The use of bar codes to tag a specimen's DNA fragments allows multiple samples to be pooled and sequenced together, thus decreasing the sequencing cost. However, this process requires a demultiplexing step in which all reads are sorted by bar code/sample before further analysis.[58] The demultiplexed file with raw reads is referred to as a *FASTQ file* (Table 3).

Following demultiplexing, the individual reads for a sample are mapped (Table 2) to a reference genome (BAM file Table 3) and any difference between the reference and the sequencing read is noted.[59] Identical (duplicate) reads are discarded for whole genome sequencing or sequence capture but not for amplicon-based sequencing. If multiple reads show the same difference, a variant is called (the threshold for the number or percentage of reads required is determined by the laboratory and should be validated). For example, a heterozygous single nucleotide variant (SNV) should be present in 50% of reads; however, in actual practice the range has been shown to vary as much as 23% to 74%.[60] The quality of signal for an individual base read and the mapping quality are also factors considered when calling a variant.[59] The output file that defines all the variants for a sample and their allelic fractions is referred to as a *variant call file*[59] (Table 3). This list of variants undergoes interpretation. The variant call file will contain all variants including common variants, although additional bioinformatics tools can be used to filter out variants meeting certain criteria (minor allele frequency above a threshold or variants previously identified as benign by a laboratory for example).

Before implementation, clinical NGS requires end-to-end validation from DNA extraction through the bioinformatics pipeline, and changes to both the wet bench or informatics portions of the test require revalidation (see Validation and Proficiency Testing sections).[61]

Interpretation of Variants

Variant interpretation is complex when applied to whole genes (as opposed to well-defined hotspots) and a large number of genes. The larger the area of the genome that is sequenced, the greater the probability of encountering rare or novel variants that will require interpretation. This has mainly been a problem in the inherited disease realm, but as oncology testing has moved away from hotspot testing to larger panels, the same issues have besieged it. Several laboratories sign out all or a subset of NGS cases in a consensus conference and share molecular data in molecular tumor boards.[62]

There are guidelines for the interpretation of germline variants put forth jointly by the American College of Medical Genetics (ACMG) (now the American College of Medical Genetics and Genomics), the Association of Molecular Pathologists, and the College of American Pathologists (CAP).[63] These guidelines assign strength of evidence for various criteria regarding a particular variant and rules for combining all the criteria to classify that variant as pathogenic, likely pathogenic, uncertain significance, likely benign, or benign.[63] The criteria include minor allele

Table 3. Next-Generation Sequencing File Types

File Type	Full Name	Description	Approximate File Size (Average Coverage 160×)	
			Exome	4800 Genes
FASTQ	Files with consensus assessment of sequence and variation	Raw sequencing data after demultiplexing	50 GB	18 GB
BAM	Binary version of sequence alignment/map	Sequencing data after alignment	16 GB	6 GB
VCF	Variant call file	File containing variants called relative to the reference	9.3 GB	3.5 MB

Abbreviations: GB, gigabytes; MB, megabytes.

Table 4. Public Databases Used in the Interpretation of Next-Generation Sequencing Data		
Type of Database	**Name of Database**	**Web Site[a]**
Population databases	Exome aggregation consortium (Exac)	http://exac.broadinstitute.org/
	gnomAD browser	http://gnomad.broadinstitute.org/
	1000 Genomes	http://www.internationalgenome.org/
	Exome server project	http://evs.gs.washington.edu/EVS/
Inherited disease databases	ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/
	dbSNP	https://www.ncbi.nlm.nih.gov/projects/SNP/
	NCBI[b] genetic testing registry	https://www.genetests.org
	Leiden open variant database (links to many locus-specific databases)	http://www.lovd.nl/3.0/home
Oncology databases	Catalogue of somatic mutations in cancer (COSMIC)	http://cancer.sanger.ac.uk/cosmic
	The cancer genome atlas (TCGA)	http://cancergenome.nih.gov/
	OncoKB (annotated TCGA data)	http://oncokb.org/#/
	dbSNP	https://www.ncbi.nlm.nih.gov/projects/SNP/
	JAX CKB	https://www.jax.org/clinical-genomics/ckb
	My cancer genome	https://www.mycancergenome.org/

[a] All Web sites accessed December 14, 2016.
[b] National Center for Biotechnology Information.

frequency from population databases and prevalence of a variant in affected individuals, segregation data, functional studies, the type of mutation and its predicted effect, similarity of the mutation to known mutations, computational models of effect, and inheritance factors.[63,64]

These guidelines have limitations and there is subjectivity to interpretation. For example, application of these guidelines to the same set of mutations between several different laboratories achieved a consensus classification 71% of the time after review and training on how to use the guidelines.[65] Another issue is that the population frequency criteria (absent or rare in population databases) can be problematic for recessive disorders, for variants with decreased penetrance or milder presentations, or in underrepresented ethnicities.[65] Population databases (Table 4) now contain information on more than 120 000 individuals, so rare pathogenic mutations may be present in a carrier state within the database.[66] These databases typically excluded patients with severe disease, but milder phenotypes or diseases with an older age of onset were not excluded.[66] Despite the limitations, these standards are a start and will allow comparisons between laboratories and for research studies. Similar criteria for somatic testing have only been available recently, and utility of these guidelines for standardizing somatic variant interpretation and reporting across laboratories remains to be evaluated.[67,68] Though there are some tools available to assist the implementation of these guidelines for variant classification, use of these guidelines is labor-intensive and we lack automated tools that can evaluate several of these criteria and support the process.[69]

As diagnostic panels increase in size, the likelihood of detecting incidental findings also increases, in particular with whole genome and whole exome testing. To fully realize the promise of precision medicine, these incidental findings would need to be incorporated into the clinical care of patients. For example, if a patient is found to have a pharmacogenetics variant leading to decreased metabolism of morphine, during whole exome testing, ideally that information would be available in the future if the patient ever required prescription pain medication. However, there are several issues surrounding reporting of incidental findings, not the least of which is ensuring that patient consent allows for the selection for return of all, some, or no incidental findings.[70,71] A patient may want some incidental results (eg, those that could affect response to a drug); however, he or she may not want other incidental results (eg, carrier status for a disease or a mutation for an adult-onset genetic disease that lacks effective treatment).[71] Having appropriate consents, ensuring that patients understand these consents, and then having the infrastructure in place to mask specific results for individual patients are all challenges to successful implementation.[72,73] Furthermore, there is the issue of which incidental findings are worth reporting from a medical standpoint (eg, should the variant that causes sensitivity/flushing to alcohol be reported).[73] In 2013 the ACMG recommended that a minimum set of 52 genes with high penetrance and available intervention be reported if those genes are analyzed; this list was updated to 59 genes in 2016.[74,75] These recommendations created substantial controversy surrounding the issue of informed consent and the patient's right to decline receiving incidental results and testing in minors, which have been incorporated into the updated ACMG recommendations.[76,77] Nevertheless, laboratory policy about how to handle incidental results usually takes these recommendations into consideration.

Another challenging area is determining what genes to test in a given clinical scenario. Although there are guidelines that define the common mutations or genes of interest (testing that is usually reimbursed), the literature and/or clinician interest may suggest other genes (testing that is usually not reimbursed) that may be medically useful.[78] Commercially and locally available panels often differ to some degree in the genes that are tested or the portions of the genes that are tested, and knowing the pros and cons associated with the different panels is challenging.[6] Databases or tools to assist with this selection process do not exist. Furthermore, multiple gene mutations in the same tumor that indicate differing prognosis or therapeutic response may be difficult to resolve. Lastly, oncology testing may identify possible germline mutations.[68,79] Although simultaneous testing of matched patient tumor and normal samples is performed in large research studies, in clinical laboratories this practice is difficult because of practical difficulties in obtaining a blood sample for germline testing from patients, and because it doubles the cost of testing, and is not reimbursed.[68] This is usually addressed by a disclaimer

or sometimes by follow-up testing of a germline sample in select cases.[68]

VALIDATION, PROFICIENCY TESTING, AND COST

Validation

It is critical to validate the entire NGS process from end to end (DNA extraction through the bioinformatics pipeline).[61] The validation process should demonstrate the ability to detect different genetic changes such as single nucleotide changes, insertions or deletions of varying sizes, and copy number variations or translocations if applicable. Validation should include patient samples with genetic variants detected by another methodology and may include commercial samples (HapMap or commercial controls); sample types (eg, FFPE, fine-needle aspiration, amniocytes) that will be run in clinical practice should be included as part of the validation. Similar to standard laboratory validations, sensitivity (false negatives), specificity (false positives), and reproducibility (including within run, between run, and with different operators) should be established for all assays. It is not feasible to evaluate these parameters for every possible mutation during validation, but common pathogenic variants should be included in the validation.[61] Limit of detection must also be assessed to establish the minimum amount of DNA needed for the assay and establish the minimum mutant allele frequency. This is especially important for any oncology assay in which tumor percentage and heterogeneity affect the allele frequency, but it is also relevant for the ability to reliably detect mosaicism in an assay to detect inherited disorders.[31,80,81]

During the validation, metrics should be defined to assess the quality of a test run and criteria for repeated testing established.[61,82–84] These metrics may include cutoffs for the insert sizes after library preparation; criteria for assessing adequate target enrichment; library concentration parameters for various steps; expected performance of controls; and metrics for sequencing performance such as clustering, base and mapping quality scores, error rates, GC bias, transition/transversion ratios, total number of sequencing reads, and coverage.[61,82–84] Identifying the need for repeat enrichment before sequencing may be time and cost-effective for a laboratory by avoiding wasted sequencing time and cost. For example, in our laboratory we run quantitative PCR against 3 targeted regions and 3 nontargeted regions for our inherited disease capture to ensure adequate enrichment. If this quality control fails, the sample(s) undergo recapture and reassessment before being sequenced.

Additionally, during validation criteria for supplementary testing should be established.[61] Supplementary testing may include genomic areas that are not reliably sequenced and confirmatory testing for certain variants that do not meet certain quality requirements.[4,85] There should be documentation of areas that cannot be reliably sequenced and a policy for addressing these areas (either as supplemental testing or a disclaimer in the report).[84] Like any test, NGS is subject to false positives, and the validation process should identify the metrics that require a confirmatory test to verify the presence of a variant identified by NGS.[4,61,85]

After the initial validation, any procedural changes require revalidation. Careful thought should be given to the initial design of an assay, as redesign requires revalidation. Changes that only involve the bioinformatics pipeline may be revalidated by using previous data sets and comparing the output of the old and new bioinformatics process. Changes in any wet bench process require end-to-end revalidation but may use fewer samples than the original validation.[61] The degree of change dictates how many samples should be assessed for a revalidation; a major change should evaluate more samples than a minor change.[61,84]

Proficiency Testing

Clinical Laboratory Improvements Amendments of 1988 require twice yearly proficiency testing (PT) for all clinical assays.[86] For tests that lack approved PT, the laboratory must verify the accuracy of the test twice yearly. These alternative assessments may include comparison to a national reference, interlaboratory exchange, or in some cases, intralaboratory verification.[86]

Ideally, PT materials would cover the assay from start (wet bench aspects) to finish (bioinformatics and interpretation). In addition, data files for testing only the bioinformatics to interpretation portion of the assay would be useful. The advantage of testing the bioinformatics portion is the ability to evaluate the bioinformatics process for multiple variants including variants of various sizes. One challenge to developing this type of PT is making a universal data file that can be recognized and tested through all the different platforms.[82,87] Proficiency testing materials may be analyte specific, which is not sufficient for NGS, or methods based on genomic DNA from individuals, genomic DNA from cell lines, or synthetic DNA.[82,87] Well-characterized materials are currently available from the Genetic Testing Reference Materials Program (GeT-RM) of the Centers for Disease Control and Prevention, from the Genome in a Bottle Consortium of the National Institute of Standards and Technology, and through the CAP Proficiency Testing Program. Estimation of tumor percentage is a necessary part of oncology NGS testing to determine if sufficient tumor is present for testing, and CAP also offers proficiency testing for this step.

Cost

The cost for the wet bench portion of NGS is primarily based on (1) library preparation (reagents, labor, necessary equipment), (2) selection strategy (PCR or capture), and (3) sequencer used. Library preparation costs vary widely depending on the method. Reagent costs are determined mainly by the commercial entities supplying the reagents and are often inversely proportional to the labor requirements. Cost of selection will vary depending on the selection strategy used (PCR versus capture), the amount of the genome being targeted (with custom capture–based products often being available in tiers), and labor and equipment required to perform the selection. Library preparation is combined with selection for PCR-based methods, which reduces the combined cost of these 2 steps. Cost of library preparation and selection may also be dependent on batching a fixed number of samples, which may be problematic for a clinical laboratory trying to maintain turnaround times. Lastly, sequencing costs are directly proportional to how much of the sequencer capacity is used for a sample and whether the sequencer is used to full capacity on a given run.

Although costs vary widely depending on the NGS design (size of sequenced area, depth of sequencing, size of sample batches, and scale of the sequencing operation), in general, for all designs, the higher the number of samples included in an analytic run, the lower the per-sample cost.

Laboratories can potentially reduce costs by streamlining workflow, choosing the most cost-effective library preparation, increasing sample volume, and if sample volume permits, automating library preparation. Instrument depreciation costs per sample are heavily dependent on instrument usage, and laboratories need to carefully evaluate sample volumes and instrument usage before deciding to purchase capital-intensive sequencing equipment. To minimize capital depreciation costs, we have adopted a model of sharing our high-throughput sequencer used for inherited disease cases with the University of Minnesota Genomics Center that uses the same instrument for research purposes. This increases the total number of samples analyzed on the instrument and dramatically reduces the capital depreciation costs for the clinical samples.

Because of the many variables that affect cost, it is hard to generalize, so we offer our experience of approximate costs for a large capture-based inherited disease panel and a small PCR-based oncology panel. For inherited disease panels, we typically sequence 9 samples and 1 control for 4800 genes (10.5 MB) on 2 lanes of a HiSeq2500 (2×100-bp run). Wet bench costs of sequencing these 9 samples to an average depth of 400× coverage is $12 145 ($1349 per sample). Library preparation accounts for 18% of the cost ($241 per sample), capture-based selection accounts for 18% of the cost ($244 per sample), and sequencing accounts for 64% of cost ($864 per sample). In addition, bioinformatics processing and use of commercial annotation and database software costs $200 per sample when averaged over our sample volume of 800 cases a year. Lastly, Sanger confirmation of NGS variants adds $50 to the total cost of the NGS assay. Thus, our total cost to run a large germline panel of 4800 genes is $1599 per sample if Sanger confirmation is required. By contrast, the wet bench cost for our small PCR-based oncology capture (13.8 kB) is lower, averaging $417 per sample. The breakdown of cost is as follows: 16% ($67) depreciation costs of our lower-throughput sequencer, 21% ($88) labor, and 63% ($263) reagents. However, the number of samples for a given run affects the per-sample cost, as the depreciation, labor, and a subset of reagent costs is divided over the number of samples.

Cost of validation also needs to be considered in developing an NGS assay and this can be a significant upfront cost. Our laboratory had access to sequencing instrumentation and some bioinformatics support, but our initial validation costs for our inherited disease assay were approximately $250 000 to $300 000 in 2012. A significant portion of this initial cost included development of infrastructure including bioinformatics infrastructure. With our infrastructure already in place and with the advancements in the field, subsequent validations of newer versions of the NGS assay have typically cost $50 000 to $70 000.

LIMITATIONS

Although there is a desire to use NGS as a single method to detect all clinically relevant genetic changes, significant limitations currently exist. These limitations include analytic sensitivity of mutation detection, areas of the genome that are difficult to sequence or analyze, limitations in the knowledge of how to interpret novel or rare mutations, limited ability to detect structural gene variation and copy number variation, and integration of genomic information into the medical care of patients. These limitations are discussed in more detail below.

Analytic Sensitivity

The sensitivity of NGS for SNV detection is approximately 5% to 10%.[43,88,89] Although this sensitivity is acceptable for most inherited disease testing (it may not detect low levels of mosaicism), it limits testing in oncology for minimal residual disease, when a low tumor percentage is present, or to detect low-level mutation due to tumor heterogeneity. Possible reasons for this limited sensitivity include PCR noise compounded by C to T transversions with FFPE tissue, sequencing errors, and systematic errors.[90–92] The general pathologist should be aware that FFPE samples will have higher artifacts than fresh tissue samples; furthermore, small specimens (including cytology specimens) may have limited DNA that can affect testing by NGS methodologies.[44,92] Studies have shown that systematic errors lead to a 4% to 6% error rate; counterintuitively, the rate is higher with increasing coverage.[93] Systematic errors may be sequence-specific errors, errors at a particular location of the read (eg, the ends for Illumina sequencers), or related to the base pair content (GC rich for Illumina).[93–95] As neither PCR nor fixation causes insertions/deletions (indels), outside of repeat regions there is better sensitivity for detecting small indels than SNVs.[81]

There are 2 main methods for improving sensitivity; however, both of these methods decrease the number of useable reads and therefore will increase the sequencing cost to obtain a comparable coverage.[96] These methods are not in widespread clinical use at this time. The first method is the use of overlapping paired end reads. This method works only for areas where the paired ends overlap, therefore the DNA insert size must be the same size or smaller than the number of reads.[96] This technique is well suited to amplicon-based sequencing in which the DNA insert size/amplicon size can be strictly controlled.[96] In this situation, the DNA insert would be completely sequenced by both paired end reads (ie, the forward read and the reverse read). The sequence of these 2 reads should match and any base pair that does not match in both reads is discarded.

The second technique is to use random nucleotide tags, referred to as *unique identifiers* (UIDs) or primer IDs, as they are often incorporated into the PCR primer.[96,97] This method works with both sequence capture and amplicon-based DNA selection techniques. In this method, random nucleotide tags are added to DNA fragments, assigned if DNA was sheared, or incorporated during the first round or two of PCR for amplicon-based methods. Importantly, these steps occur before amplification and result in a DNA fragment with a random and unique nucleotide sequence at one or both ends. After amplification, multiple identical template molecules will be present and will be sequenced (Figure 4, A through D); therefore, duplicate reads must be retained during analysis.[96,97] All reads that map to the same location and have an identical UID are considered part of a UID family and will be analyzed as a group. Targeted areas should be covered by many different UID families. If a mutation is present in a majority (eg, >95%) of that UID family, the mutation is considered present and is considered 1 read.[96,97] This process is repeated for all other UID families.

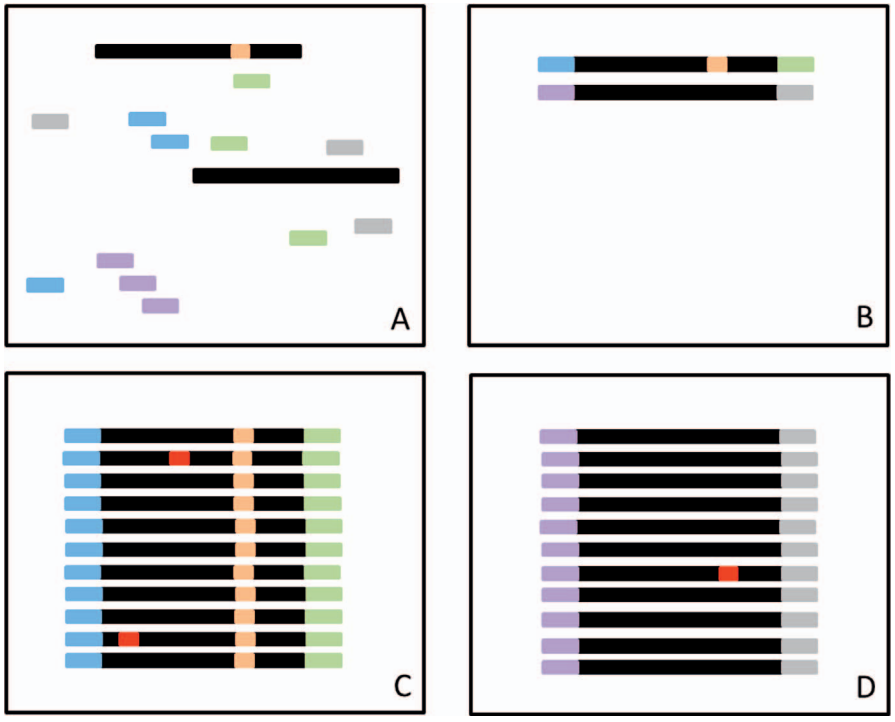

Figure 4. *A, Before amplification, random tags (short bars) are added to DNA fragments (black), some of which have a mutation (orange). B, The tags randomly attach to DNA fragments. C, During amplification some copies will develop an error (red). All fragments will be sequenced. Only mutations that are detected within a majority (eg, 95%) of all the sequencing reads with the same ID tag will be identified as true mutations. D, Mutations present in a minority of reads with the same ID tag are considered errors. Reprinted from Yohe SL. Hot topic spotlight—new frontiers in clinical next-generation sequencing. In: Cushman-Vokoun AM, Anderson WB, eds.* Precision Medicine Resource Guide. *Northfield, IL: College of American Pathologists; 2016:5. With permission from College of American Pathologists. Copyright 2016.*

Difficult-to-Sequence Areas

Homologous regions, repetitive regions, and GC-rich regions are not reliably interpretable by the current NGS platforms and standard bioinformatics algorithms. Homologous regions, including pseudogenes, are areas of the genome with high sequence similarity that may differ from the gene of interest by only a few base pairs. Fragments of DNA that are sequenced from the target gene and homologous regions may be so similar in sequence as to be indistinguishable; and the shorter the length of sequence, the more likely this is to occur. This is not a problem unique to NGS, as Sanger sequencing is also susceptible to inadvertent sequencing of homologous regions, and test design is important to mitigate the problem. In NGS analysis, fragments of DNA from a target gene and homologous regions will have poor mapping quality, and reads belonging to the homolog can be mismapped to the real gene and vice versa[98] (Figure 5). Mismapping may lead to both false-positive and false-negative calls (eg, mutations being missed and mutations being erroneously called). Many clinically relevant genes (such as *PMS2*, *STRC*) have pseudogenes, are challenging to interpret by NGS, and require specialized methods for target enrichment such as long-range PCR.[98,99] This problem may be solved by newer instruments that have much longer sequencing reads (see New Instrumentation section); however, in current practice, assessment of these areas requires traditional alternative methods.[100]

For repeat areas, unique sequence flanking the repeat is required to reliably map a sequencing read and determine the size of the repeat. Repetitive regions larger than the size of the DNA insert will not have a flanking sequence and therefore will not be accurately mapped. Smaller repeat sizes will have a unique flanking sequence on at least a proportion of the DNA fragments and therefore will map, although at lower coverage since some reads will not be informative. Even so, enumeration of the repeat size requires specialized bioinformatics algorithms, and errors still occur, which require interpretation. Sources of error include stutter (polymerase slippage leading to small shifts in repeat size) and PCR sequencing mistakes.[25] Homopolymers (ie, poly A or poly T) are difficult for the Ion Torrent sequencers, as the degree of change in voltage loses resolution above 6 to 8 base pairs. However, most testing of repeat areas (eg, trinucleotide disorders such as fragile X) continues to use traditional, established methods rather than NGS.

GC-rich regions appear to have higher background noise and lower quality of sequencing. In particular, Illumina sequencers give substitution errors in areas of high GC content and long G/C homopolymers.[94] GC-rich regions are known to form secondary structures, which may represent part, but perhaps not all, of the problem. There may also be accumulation of G or C fluorophores after washing or out-of-phase sequencing.[94]

Validation of an NGS assay should include assessment for areas that cannot reliably be genotyped by NGS methods, and at a minimum these areas should be documented.[101] Alternative testing strategies such as Sanger sequencing or long-range PCR may be possible for some regions.[98]

Limitation in Databases and Knowledge

Although the technical ability exists to perform whole genome analysis at a reasonable cost, especially for inherited disease, the ability to interpret all of those data lags behind. Sources to help with interpretation include databases (both publicly available, private, or laboratory-specific databases), genetic and medical knowledge, medical literature, patient information, clinical experience, and team discussion. There are different types of databases with different amounts of data within them. A layer-1 database or Clinical Genomic Variant Repository contains only sequence/variant information, a layer-2 database or Genomic Medical Data Repository contains sequence/variant

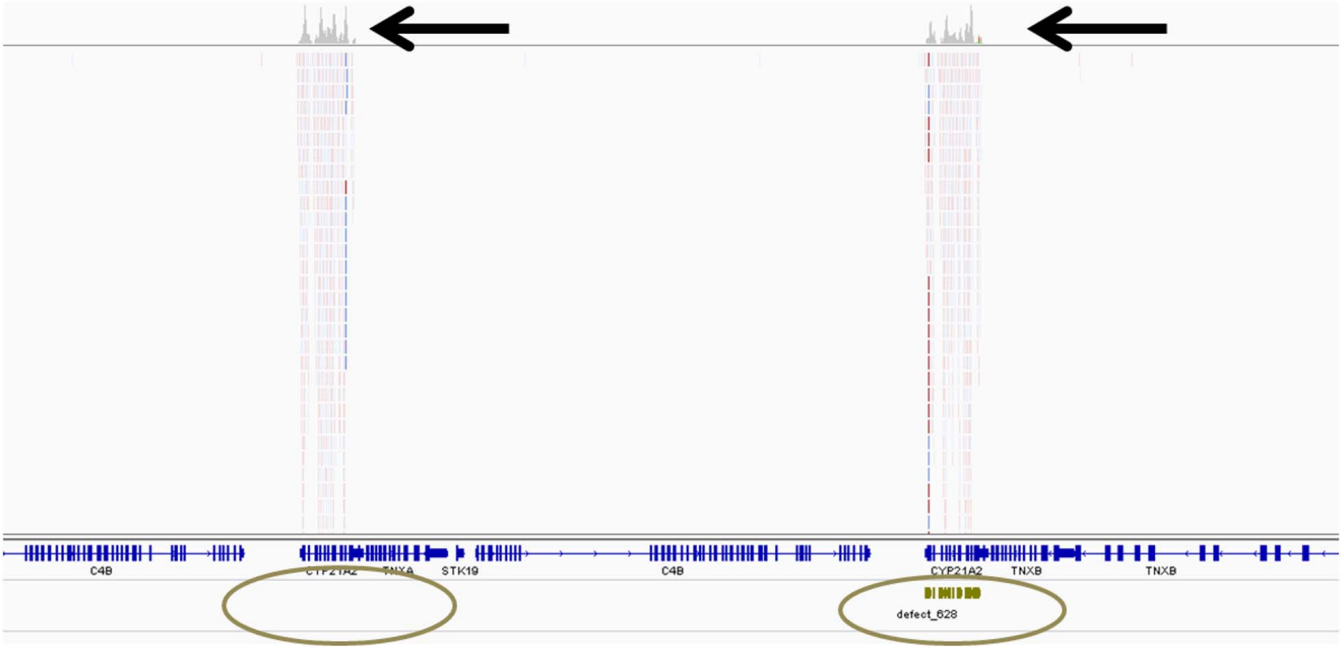

Figure 5. *The right side shows the* CYP21A2 *gene with baits (Table 2) designed for sequence capture (green bars). The left side shows the* CYP21A2 *pseudogene without baits. A similar number of sequencing reads are being mapped to the pseudogene as to the real gene; since the reads are so similar the actual source cannot be determined. These reads would have a low mapping quality score, as the reads are mapping to more than 1 location, as indicated by the faded colors. Black arrow: coverage (gray peaks), green circle: location of baits (if any). Cropped integrated genome viewer (IGV) screenshot (Broad Institute, Cambridge, Massachusetts).*[121,122] *Reprinted from Yohe SL. Hot topic spotlight—new frontiers in clinical next-generation sequencing. In: Cushman-Vokoun AM, Anderson WB, eds.* Precision Medicine Resource Guide. *Northfield, IL: College of American Pathologists; 2016:6. With permission from College of American Pathologists. Copyright 2016.*

information with clinical/phenotype data, and a layer-3 database or Genomic Medical Evidence Database contains the medical evidence of classification or association with the sequence/variant information.[7] Most databases contain data for either inherited disease or somatic mutations but not both with the exception of ClinVar and dbSNP (Table 4).

Although databases are extremely useful in the interpretation of variants, there are limitations of the current databases and no database is comprehensive or error free. Many databases lack assurance of the quality of sequence data or other data within the database.[7] Databases may not be up-to-date or may include conflicting data. Both the medical literature and databases must be used with caution, as some variants have been described as pathogenic by outdated criteria (ie, absent in 100 controls). Furthermore, there is limited knowledge about digenic or multigenic effects.

The significance of variants in introns or untranslated regions is often unknown, and rare or novel exon variants may also be difficult to interpret.[102] Novel or rare mutations that cause a frameshift or change an amino acid to a stop codon (stop loss or nonsense mutations) will generally be pathogenic if that mechanism has been described for the gene in question, but even then there may be exceptions.[63] Missense mutations are more difficult to interpret. Many factors are taken into consideration when interpreting these cases, including details about the specific mutation, details about mutations known to cause disease, similarity to known mutations, whether the mutation is in cis/trans with another known mutation or is de novo, presence/absence in other individuals (such as populations, normal controls, or affected and unaffected family members), and predicted protein effect (using in silico models).[63,103]

Structural Variation and Copy Number Variation

Next-generation sequencing performs reasonably well at detection of SNVs and small insertion/deletions (indels), but does not perform as well at detecting structural rearrangements or copy number variations (CNVs) especially when using enrichment to perform targeted NGS.[104,105] Additionally, detecting structural variation and CNVs requires different bioinformatics algorithms from SNV detection. Several clinical laboratories are currently using NGS data to detect CNVs and generally use a combination of 2 or more techniques.[5,104,105]

There are several techniques that have been used to detect CNVs, including depth of coverage (read depth), read pair, split pair, assembly based, or a combination of these techniques.[106] Clinical CNV analysis usually uses some combination of 2 or more of these techniques.[5,104,105] All the methods detect deletions better than duplications, cannot detect CNVs in repeat regions or difficult-to-map areas, and are limited by coverage (although the depth-of-coverage technique is more affected by coverage that the other techniques).[106] False positives are a problem, especially when applying CNV analysis over large areas of the exome, and have been reported to occur 10% to 89% of the time.[107] Recent advances that have incorporated machine-learning techniques hold the promise to reduce false-positive calls.[2] However, some areas of the genome are more prone to false positives than others.[101]

Using depth of coverage or read depth to detect CNVs works well with uniform sequencing, which is assumed for standard bioinformatics tools. These tools analyze for increased or decreased coverage to detect duplication/amplification or deletion, respectively. However, coverage varies between runs, within run, and between patients

especially when sequence enrichment is used, and spurious calls will be detected when sequencing is not uniform. With sequence enrichment, the pattern of coverage tends to be similar but the absolute coverage varies, requiring some sort of normalization. This may require comparison to a control as well as to control genes within the sample to normalize for the performance of the individual sample.[2] The advantage of the read depth technique is the ability to detect large CNVs and to predict the actual copy number; however, this method cannot detect the breakpoint or detect rearrangements.[106]

Read pair (or mate pair) analysis compares the distance of 2 ends of a read pair to the average insert size. Read pair analysis requires paired end reads, is limited by the insert size, and will only detect smaller CNVs.[106] One advantage of read pair analysis is that it detects both CNVs and rearrangements (translocations and inversions).[106] However, it will only detect duplications/amplifications smaller than the average insert size and deletions smaller than 1 kb, and it cannot accurately estimate the number of copies.[106]

Split pair (or split read) analysis looks specifically at paired reads where one of the paired reads fails to map or only maps partially. Split pair analysis also requires paired end reads, will only detect smaller CNVs, and does not perform well in regions of low complexity.[106] However, it can pinpoint the breakpoint and detect rearrangements.[106]

Finally, assembly-based analysis uses de novo alignment of the reads.[5] De novo alignment (Table 2) matches the individual reads to each other instead of to a reference genome. Because it is computationally intensive, this technique works better for small genomes, such as bacteria, but can be used clinically.[5]

INTEGRATION INTO THE MEDICAL CARE OF PATIENTS

There has been intense focus on the meaningful integration of genomics into patient care.[108] Many practical issues need to be solved for this to occur in a widespread manner. Issues include making reports understandable, interfacing genomic results with the electronic medical record (EMR), bioinformatics tools to help categorize variants, handling of incidental findings, and whether and how to offer genetic reevaluation.[109–111] Other issues include data storage, including what data to store (FASTQ, BAM, variant call file), how long to store data, and how to store large data sets securely. The CAP checklist for NGS provides guidance stating that some files must be stored for at least 2 years; these files should allow re-review of the case in the same manner that allowed generation of the original data.[84] Data storage and processing may be performed locally on a server or through a third party. Cloud companies now offer secure Cloud-based services and storage; however, it is incumbent on the health care facility to ensure services meet all HIPAA (Health Insurance Portability and Accountability Act) requirements for data transfer and storage.

Widespread meaningful integration of large-scale genomic data into the medical record, especially for smaller institutions, is still a challenge. Current laboratory information systems and EMRs can handle discrete data points with an associated interpretive or normal range and can handle interpretive text reports, but they are ill equipped to handle complex genomic data generated by whole genome, whole exome, and large targeted panel NGS.[112] Although laboratory information and EMR systems may evolve, at the current time and for the foreseeable future, ancillary systems are necessary to integrate large amounts of genomic data into the medical record.[112,113] However, implementation of these ancillary systems requires significant time and resources for information technology personnel, in addition to clinicians, laboratorians, pharmacists, and/or pathologists depending on the application.[113] Many of the places that have successfully implemented EMR integration with an ancillary genomic system are academic centers with genomic or information technology expertise and have implemented systems for a specific subset of genomic information (such as pharmacogenomics variants), with fewer institutions offering whole genome or exome testing.[108] Several new companies offer these ancillary systems. Some systems organize, annotate, track variants, and generate a report. These reports are usually pdf or text reports and do not have discrete fields that transfer to the EMR. Even more recently, a few companies have offered clinical decision support tools.

NEW INSTRUMENTATION

Two new sequencing instruments (sometimes referred to as *third-generation sequencers*) are currently available for research use that provide longer sequencing reads and are able to read the sequence of a single molecule: the PacBio SMRT (single molecule real time) (Menlo Park, California) and the Oxford Nanopore (Oxford, United Kingdom). These instruments use different underlying chemistry.[114,115] The PacBio SMRT uses multiple wells, each of which has a DNA polymerase affixed to the bottom with 1 long DNA fragment.[115] Each fluorescently tagged nucleotide (A, C, G, T) gives a different fluorescent signal when incorporated.[115] Illumination and detection occur from the bottom of the well and detection is sensitive enough to detect the single fluorescent signal that is released when a base pair is added to the DNA strand.[115] The Oxford Nanopore uses a protein pore inserted into a membrane.[114] A current is applied and flows through the pore between the 2 sides of the membrane.[114] As the structure (DNA or RNA strand) passes through the pore, the current changes and the degree of change correlates with the individual base (A, C, G, or T) and also correlates with the methylation status of C; therefore, methylation and hydroxymethylation can be detected.[114,116] The PacBio SMRT can also infer methylation status by analyzing the change in DNA polymerase kinetics (the time to incorporate a base and the time between incorporation of 2 bases).

Neither instrument requires amplification steps and thus should reduce the background noise. Both instruments can perform long reads (14 000–40 000 for the PacBio SMRT and 8000–100 000 for the Nanopore), which can overcome issues with pseudogenes and repeat regions and may help with identifying RNA isoforms; however, both have high error rates.[115,117] The errors on the PacBio are random and therefore can be overcome by replicate sequencing of the same molecule and by using a consensus result.[118,119] The errors on the Nanopore are biased (meaning they occur in the same areas) and therefore cannot be overcome by replicate sequencing.[120] These instruments show promise and may address problems in many clinically relevant regions such as trinucleotide repeat regions, HLA, and homologous regions.[115] However, these sequencers have limited adoption in the clinical realm, possibly owing to their higher price and lower throughput and possibly owing

to the challenges of clinically validating an instrument with high intrinsic error rates.

SUMMARY

Next-generation sequencing is being implemented in clinical laboratories and the use will only increase as the technology, bioinformatics, and resources evolve to address the limitations, improve quality of results, and increase the number of clinically useful applications. Clinical NGS has expanded to detect SNVs as well as structural rearrangements and CNVs, to monitor circulating tumor DNA, and to analyze areas of the genome that previously were challenging for standard bioinformatics algorithms to manage. Further improvements will continue to occur; however, the challenge for clinical laboratories is to ensure testing is clinically relevant, cost-effective, and can be integrated into clinical care.

References

1. 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.

2. Onsongo G, Baughn LB, Bower M, et al. CNV-RF is a random forest-based copy number variation detection method using next-generation sequencing. *J Mol Diagn*. 2016;18(6):872–881.

3. Onsongo G, Erdmann J, Spears MD, et al. Implementation of Cloud based next generation sequencing data analysis in a clinical laboratory. *BMC Res Notes*. 2014;7:314.

4. Nelson AC, Bower M, Baughn LB, et al. Criteria for clinical reporting of variants from a broad target capture: NGS assay without Sanger verification. *JSM Biomar*. 2015;2:1005–1010.

5. Yang R, Nelson AC, Henzler C, Thyagarajan B, Silverstein KA. ScanIndel: a hybrid framework for indel detection via gapped alignment, split reads and de novo assembly. *Genome Med*. 2015;7:127.

6. Zutter MM, Bloom KJ, Cheng L, et al. The Cancer Genomics Resource List 2014. *Arch Pathol Lab Med*. 2015;139(8):989–1008.

7. Yohe SL, Carter AB, Pfeifer JD, et al. Standards for Clinical Grade Genomic Databases. *Arch Pathol Lab Med*. 2015;139(11):1400–1412.

8. Platt J, Cox R, Enns GM. Points to consider in the clinical use of NGS panels for mitochondrial disease: an analysis of gene inclusion and consent forms. *J Genet Couns*. 2014;23(4):594–603.

9. Pagon RA. GeneTests: an online genetic information resource for health care providers. *J Med Libr Assoc*. 2006;94(3):343–348.

10. Audo I, Bujakowska KM, Leveillard T, et al. Development and application of a next-generation-sequencing (NGS) approach to detect known and novel gene defects underlying retinal diseases. *Orphanet J Rare Dis*. 2012;7:8.

11. Nijman IJ, van Montfrans JM, Hoogstraat M, et al. Targeted next-generation sequencing: a novel diagnostic tool for primary immunodeficiencies. *J Allergy Clin Immunol*. 2014;133(2):529–534.

12. Esmail Nia G, Fadaee M, Royer R, Najmabadi H, Akbari MR. Profiling Fanconi anemia gene mutations among Iranian patients. *Arch Iran Med*. 2016; 19(4):236–240.

13. Artuso R, Fallerini C, Dosa L, et al. Advances in Alport syndrome diagnosis using next-generation sequencing. *Eur J Hum Genet*. 2012;20(1):50–57.

14. Weerakkody RA, Vandrovcova J, Kanonidou C, et al. Targeted next-generation sequencing makes new molecular diagnoses and expands genotype-phenotype relationship in Ehlers-Danlos syndrome. *Genet Med*. 2016;18(11): 1119–1127.

15. Liu YT, Lee YC, Soong BW. What we have learned from the next-generation sequencing: Contributions to the genetic diagnoses and understanding of pathomechanisms of neurodegenerative diseases. *J Neurogenet*. 2015;29(2–3): 103–112.

16. Voelkerding KV, Dames S, Durtschi JD. Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy: a paper from the 2009 William Beaumont Hospital Symposium on Molecular Pathology. *J Mol Diagn*. 2010;12(5):539–551.

17. De Keulenaer S, Hellemans J, Lefever S, et al. Molecular diagnostics for congenital hearing loss including 15 deafness genes using a next generation sequencing platform. *BMC Med Genomics*. 2012;5:17.

18. Sawyer SL, Hartley T, Dyment DA, et al. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. *Clin Genet*. 2016;89(3):275–284.

19. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med*. 2013;369(16):1502–1511.

20. McCullough RM, Almasri EA, Guan X, et al. Non-invasive prenatal chromosomal aneuploidy testing–clinical experience: 100,000 clinical samples. *PLoS One*. 2014;9(10):e109173.

21. Lefkowitz RB, Tynan JA, Liu T, et al. Clinical validation of a noninvasive prenatal test for genomewide detection of fetal copy number variants. *Am J Obstet Gynecol*. 2016;215(2):227.e1–227.e16.

22. Stutz WE, Bolnick DI. Stepwise threshold clustering: a new method for genotyping MHC loci using next-generation sequencing technology. *PLoS One*. 2014;9(7):e100587.

23. Weimer ET, Montgomery M, Petraroia R, Crawford J, Schmitz JL. Performance characteristics and validation of next-generation sequencing for human leucocyte antigen typing. *J Mol Diagn*. 2016;18(5):668–675.

24. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*. 2014;30(23):3310–3316.

25. Van Neste C, Vandewoestyne M, Van Criekinge W, Deforce D, Van Nieuwerburgh F. My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing. *Forensic Sci Int Genet*. 2014;9:1–8.

26. Thompson JC, Yee SS, Troxel AB, et al. Detection of therapeutically targetable driver and resistance mutations in lung cancer patients by next-generation sequencing of cell-free circulating tumor DNA. *Clin Cancer Res*. 2016;22(23):5772–5782.

27. Lanman RB, Mortimer SA, Zill OA, et al. Analytical and clinical validation of a digital sequencing panel for quantitative, highly accurate evaluation of cell-free circulating tumor DNA. *PLoS One*. 2015;10(10):e0140712.

28. Leary RJ, Sausen M, Kinde I, et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl Med*. 2012;4(162):162ra154.

29. Forshew T, Murtaza M, Parkinson C, et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci Transl Med*. 2012;4(136):136–168.

30. Yamada T, Iwai T, Takahashi G, et al. Utility of KRAS mutation detection using circulating cell-free DNA from patients with colorectal cancer. *Cancer Sci*. 2016;107(7):936–943.

31. Li Z, Zhang Y, Bao W, Jiang C. Insufficiency of peripheral blood as a substitute tissue for detecting EGFR mutations in lung cancer: a meta-analysis. *Target Oncol*. 2014;9(4):381–388.

32. Ai B, Liu H, Huang Y, Peng P. Circulating cell-free DNA as a prognostic and predictive biomarker in non-small cell lung cancer. *Oncotarget*. 2016;7(28): 44583–44595.

33. Hofman P. Liquid biopsy for early detection of lung cancer. *Curr Opin Oncol*. 2017;29(1):73–78.

34. Fernandez-Cuesta L, Perdomo S, Avogbe PH, et al. Identification of circulating tumor DNA for the early detection of small-cell lung cancer. *EBioMedicine*. 2016;10:117–123.

35. Chan KC, Jiang P, Zheng YW, et al. Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clin Chem*. 2013;59(1):211–224.

36. Chen KZ, Lou F, Yang F, et al. Circulating tumor DNA detection in early-stage non-small cell lung cancer patients by targeted sequencing. *Sci Rep*. 2016; 6:31985.

37. Guo K, Zhang Z, Han L, et al. Detection of epidermal growth factor receptor mutation in plasma as a biomarker in Chinese patients with early-stage non-small cell lung cancer. *Onco Targets Ther*. 2015;8:3289–3296.

38. Uemura T, Hibi K, Kaneko T, et al. Detection of K-ras mutations in the plasma DNA of pancreatic cancer patients. *J Gastroenterol*. 2004;39(1):56–60.

39. Brychta N, Krahn T, von Ahsen O. Detection of KRAS mutations in circulating tumor DNA by digital PCR in early stages of pancreatic cancer. *Clin Chem*. 2016;62(11):1482–1491.

40. Jahr S, Hentze H, Englisch S, et al. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res*. 2001;61(4):1659–1665.

41. Bettegowda C, Sausen M, Leary RJ, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med*. 2014;6(224): 224ra24.

42. Rizzo JM, Buck MJ. Key principles and clinical applications of "next-generation" DNA sequencing. *Cancer Prev Res (Phila)*. 2012;5(7):887–900.

43. de Leng WW, Gadellaa-van Hooijdonk CG, Barendregt-Smouter FA, et al. Targeted next generation sequencing as a reliable diagnostic assay for the detection of somatic mutations in tumours using minimal DNA amounts from formalin fixed paraffin embedded material. *PLoS One*. 2016;11(2):e0149405.

44. Sah S, Chen L, Houghton J, et al. Functional DNA quantification guides accurate next-generation sequencing mutation detection in formalin-fixed, paraffin-embedded tumor biopsies. *Genome Med*. 2013;5(8):77.

45. Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS One*. 2011;6(11):e28240.

46. Kim H, Bartsch MS, Renzi RF, et al. Automated digital microfluidic sample preparation for next-generation DNA sequencing. *J Lab Autom*. 2011; 16(6):405–414.

47. Head SR, Komori HK, LaMere SA, et al. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*. 2014;56(2): 61–64, 66, 68, passim.

48. Garcia-Garcia G, Baux D, Faugere V, et al. Assessment of the latest NGS enrichment capture methods in clinical context. *Sci Rep*. 2016;6:20948.

49. Samorodnitsky E, Jewell BM, Hagopian R, et al. Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. *Hum Mutat*. 2015;36(9):903–914.

50. Salipante SJ, Kawashima T, Rosenthal C, et al. Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Appl Environ Microbiol*. 2014;80(24):7583–7591.

51. Quail MA, Smith M, Coupland P, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*. 2012;13:341.

52. Liu L, Li Y, Li S, et al. Comparison of next-generation sequencing systems. *J Biomed Biotechnol*. 2012;2012:251364.

53. Mardis ER. Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif)*. 2013;6:287–303.

54. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333–351.

55. Loman NJ, Misra RV, Dallman TJ, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*. 2012;30(5): 434–439.

56. Magi A, Benelli M, Gozzini A, Girolami F, Torricelli F, Brandi ML. Bioinformatics for next generation sequencing data. *Genes (Basel)*. 2010;1(2): 294–307.

57. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011; 43(5):491–498.

58. Tu J, Ge Q, Wang S, et al. Pair-barcode high-throughput sequencing for large-scale multiplexed sample analysis. *BMC Genomics*. 2012;13:43.

59. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43:11.10.1–33.

60. Jones MA, Bhide S, Chin E, et al. Targeted polymerase chain reaction-based enrichment and next generation sequencing for diagnostic testing of congenital disorders of glycosylation. *Genet Med*. 2011;13(11):921–932.

61. Rehm HL, Bale SJ, Bayrak-Toydemir P, et al. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med*. 2013;15(9):733–747.

62. Roychowdhury S, Iyer MK, Robinson DR, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med*. 2011;3(111):111–121.

63. 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.

64. Duzkale H, Shen J, McLaughlin H, et al. A systematic approach to assessing the clinical significance of genetic variants. *Clin Genet*. 2013;84(5): 453–463.

65. Amendola LM, Jarvik GP, Leo MC, et al. Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the Clinical Sequencing Exploratory Research Consortium. *Am J Hum Genet*. 2016;98(6): 1067–1076.

66. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285–291.

67. Ritter DI, Roychowdhury S, Roy A, et al. Somatic cancer variant curation and harmonization through consensus minimum variant level data. *Genome Med*. 2016;8(1):117.

68. 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.

69. Kleinberger J, Maloney KA, Pollin TI, Jeng LJ. An openly available online tool for implementing the ACMG/AMP standards and guidelines for the interpretation of sequence variants. *Genet Med*. 2016;18(11):1165.

70. Matthijs G, Souche E, Alders M, et al. Guidelines for diagnostic next-generation sequencing. *Eur J Hum Genet*. 2016;24(1):2–5.

71. Blackburn HL, Schroeder B, Turner C, Shriver CD, Ellsworth DL, Ellsworth RE. Management of incidental findings in the era of next-generation sequencing. *Curr Genomics*. 2015;16(3):159–174.

72. Berg JS, Adams M, Nassar N, et al. An informatics approach to analyzing the incidentalome. *Genet Med*. 2013;15(1):36–44.

73. Berg JS, Khoury MJ, Evans JP. Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. *Genet Med*. 2011;13(6):499–504.

74. Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med*. 2013;15(7):565–574.

75. Kalia SS, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics [published online ahead of print November 17, 2016]. *Genet Med*. doi:10.1038/gim.2016.190.

76. Anderson JA, Hayeems RZ, Shuman C, et al. Predictive genetic testing for adult-onset disorders in minors: a critical analysis of the arguments for and against the 2013 ACMG guidelines. *Clin Genet*. 2015;87(4):301–310.

77. Evans BJ. Minimizing liability risks under the ACMG recommendations for reporting incidental findings in clinical exome and genome sequencing. *Genet Med*. 2013;15(12):915–920.

78. Bejar R. Implications of molecular genetic diversity in myelodysplastic syndromes. *Curr Opin Hematol*. 2017;24(2):73–78.

79. Catenacci DV, Amico AL, Nielsen SM, et al. Tumor genome analysis includes germline genome: are we ready for surprises? *Int J Cancer*. 2015;136(7): 1559–1567.

80. 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.

81. 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.

82. Gargis AS, Kalman L, Berry MW, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol*. 2012; 30(11):1033–1036.

83. Aziz N, Zhao Q, Bry L, et al. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch Pathol Lab Med*. 2015;139(4):481–493.

84. College of American Pathologists-Commission on Laboratory Accreditation. Molecular Pathology Checklist. Northfield, IL: College of American Pathologists; August 17, 2016.

85. Baudhuin LM, Lagerstedt SA, Klee EW, Fadra N, Oglesbee D, Ferber MJ. Confirming variants in next-generation sequencing panel testing by Sanger sequencing. *J Mol Diagn*. 2015;17(4):456–461.

86. Centers for Medicare and Medicaid Services. Clinical Laboratory Improvement Amendments of 1988—Laboratory Requirements. 42 CFR §493.1443–1495.

87. Duncavage EJ, Abel HJ, Pfeifer JD. In silico proficiency testing for clinical next-generation sequencing. *J Mol Diagn*. 2016;19(1):35–42.

88. Froyen G, Broekmans A, Hillen F, et al. Validation and application of a custom-designed targeted next-generation sequencing panel for the diagnostic mutational profiling of solid tumors. *PLoS One*. 2016;11(4):e0154038.

89. D'Haene N, Le Mercier M, De Nève N, et al. Clinical validation of targeted next generation sequencing for colon and lung cancers. *PLoS One*. 2015; 10(9):e0138245.

90. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res*. 2014;322(1):12–20.

91. Costello M, Pugh TJ, Fennell TJ, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res*. 2013; 41(6):e67.

92. 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(5):623–633.

93. Wall JD, Tang LF, Zerbe B, et al. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res*. 2014;24(11):1734–1739.

94. Shin S, Park J. Characterization of sequence-specific errors in various next-generation sequencing systems. *Mol Biosyst*. 2016;12(3):914–922.

95. Chen G, Mosier S, Gocke CD, Lin MT, Eshleman JR. Cytosine deamination is a major cause of baseline noise in next-generation sequencing. *Mol Diagn Ther*. 2014;18(5):587–593.

96. Zhang TH, Wu NC, Sun R. A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing. *BMC Genomics*. 2016;17:108.

97. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A*. 2011;108(23):9530–9535.

98. Mandelker D, Amr SS, Pugh T, et al. Comprehensive diagnostic testing for stereocilin: an approach for analyzing medically important genes with high homology. *J Mol Diagn*. 2014;16(6):639–647.

99. Li J, Dai H, Feng Y, et al. A comprehensive strategy for accurate mutation detection of the highly homologous PMS2. *J Mol Diagn*. 2015;17(5):545–553.

100. Vona B, Hofrichter MA, Neuner C, et al. DFNB16 is a frequent cause of congenital hearing impairment: implementation of STRC mutation analysis in routine diagnostics. *Clin Genet*. 2015;87(1):49–55.

101. Mandelker D, Schmidt RJ, Ankala A, et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet Med*. 2016;18(12):1282–1289.

102. Plon SE, Eccles DM, Easton D, et al. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat*. 2008;29(11):1282–1291.

103. Richards CS, Bale S, Bellissimo DB, et al. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet Med*. 2008;10(4):294–300.

104. Pugh TJ, Amr SS, Bowser MJ, et al. VisCap: inference and visualization of germ-line copy-number variants from targeted clinical sequencing data. *Genet Med*. 2016;18(7):712–719.

105. Nord AS, Lee M, King MC, Walsh T. Accurate and exact CNV identification from targeted high-throughput sequence data. *BMC Genomics*. 2011;12:184.

106. Zhao M, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. 2013;14(suppl 11):S1.

107. Pinto D, Darvishi K, Shi X, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol*. 2011;29(6):512–520.

108. Green RC, Goddard KA, Jarvik GP, et al. Clinical Sequencing Exploratory Research Consortium: accelerating evidence-based practice of genomic medicine. *Am J Hum Genet*. 2016;98(6):1051–1066.

109. Bowdin S, Gilbert A, Bedoukian E, et al. Recommendations for the integration of genomics into clinical practice. *Genet Med*. 2016;18(11):1075–1084.

110. Rasmussen LV, Overby CL, Connolly J, et al. Practical considerations for implementing genomic information resources: experiences from eMERGE and CSER. *Appl Clin Inform*. 2016;7(3):870–882.

111. Vassy JL, Lautenbach DM, McLaughlin HM, et al. The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine. *Trials*. 2014;15:85.

112. Starren J, Williams MS, Bottinger EP. Crossing the omic chasm: a time for omic ancillary systems. *JAMA*. 2013;309(12):1237–1238.

113. Herr TM, Bielinski SJ, Bottinger E, et al. Practical considerations in genomic decision support: the eMERGE experience. *J Pathol Inform*. 2015;6:50.

114. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol*. 2016;17(1):239.

115. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*. 2015;13(5):278–289.

116. Schreiber J, Wescoe ZL, Abu-Shumays R, et al. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proc Natl Acad Sci U S A*. 2013;110(47):18910–18915.

117. Miyamoto M, Motooka D, Gotoh K, et al. Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics*. 2014;15:699.

118. Koren S, Schatz MC, Walenz BP, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol*. 2012;30(7):693–700.

119. Jiao X, Zheng X, Ma L, et al. A benchmark study on error assessment and quality control of CCS reads derived from the PacBio RS. *J Data Mining Genomics Proteomics*. 2013;4(3).

120. Li C, Chng KR, Boey EJ, Ng AH, Wilm A, Nagarajan N. INC-Seq: accurate single molecule reads using nanopore sequencing. *Gigascience*. 2016;5(1):34.

121. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14(2):178–192.

122. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24–26.