# Assuring the Quality of Next-Generation Sequencing in Clinical Laboratory Practice

Next-generation Sequencing: Standardization of Clinical Testing (Nex-StoCT) Workgroup Principles and Guidelines

## Supplementary Guidelines

### Table of Contents





Nature Biotechnology: doi:10.1038/nbt.2403

## 1. Introduction

Next- generation sequencing (NGS) technology has expanded beyond research applications to deliver clinically actionable test results, for the diagnosis and treatment of rare diseases and cancer[1-5]. The utilization of NGS in clinical settings is driven by the comprehensive capacity for genomic analysis and the potential to consolidate single-gene diagnostic tests. The implementation of NGS technology in a clinical laboratory environment is complex, requiring significant infrastructure and expertise in clinical, scientific, and informatics specialties. Currently, laboratories lack uniform guidance on applying the technical aspects of quality management for test system validation, quality control (QC) and external quality assessment (EQA) or proficiency testing (PT) for NGS.

In the United States, diagnostic tests that are provided to clinical laboratories are regulated by the Food and Drug Administration (FDA). At this time, FDA has not developed guidance targeted to NGS but has engaged with other federal partners to develop a regulatory framework for NGS (meeting summary and webcast available at: http://www.fda.gov/MedicalDevices/NewsEvents/WorkshopsConferences/ucm255327.htm). To date, no NGS technologies have been approved or cleared by the FDA. These tests are currently developed in-house as laboratory-developed tests and regulated under the Clinical Laboratory Improvement Amendments (CLIA) regulations.[6]

To address these shortcomings, the Centers for Disease Control and Prevention (CDC) established a national workgroup of experts to develop



Nature Biotechnology: doi:10.1038/nbt.2403

recommendations for assuring the quality of NGS results. Here we describe the guidelines and recommendations of the workgroup.

## 2. Background

### a. Differences between NGS and Sanger sequencing

Capillary electrophoresis-based, semi-automated Sanger sequencing[7-9] is currently considered the gold standard for DNA sequencing. Sanger sequencing can produce long read-lengths with highly reliable results, but rapidly becomes cost- and time-prohibitive when larger expanses of the genome are targeted. Sanger sequence analysis is not practical to implement as a routine clinical service for either the human exome or whole genome. Analyses of gene panels are possible using Sanger sequencing, but the costs have been proportional to the numbers and size of the regions targeted. NGS allows gene panels to be sequenced at a lower cost, and provides the ability to perform rapid, large-scale exome and whole genome sequencing. The majority of currently available NGS sequencing instruments produce short read-lengths that require sophisticated alignment or assembly procedures to derive a reportable sequence result. The first wave of NGS platforms allow the simultaneous analysis of a large number of genomic regions at a lower per-base cost than Sanger sequencing (reviewed in Ref. 10-12). Although NGS provides several advantages over the Sanger method, the amount of data generated by NGS poses unprecedented informatics challenges for data analysis, management, and storage (reviewed in Ref. 13). Clinical laboratories should consider the cost, speed, and complexity of NGS data interpretation when contemplating the adoption of an NGS-based test. NGS tests require an informatics pipeline capable of accurately aligning raw data files to a



Nature Biotechnology: doi:10.1038/nbt.2403

reference sequence, calling and annotating sequence variants, determining which variants have clinical significance, and which variants require confirmatory testing. The number of variants identified in a patient's specimen is proportional to the size of the genomic region targeted and can quickly generate a data bottleneck for variant classification. Downstream pipelines and algorithms for clinical decision support using variant files as input are being developed[14,15] and tested in projects such as ClinSeq[16], but are not further discussed in this manuscript.

**b. NGS applications**

NGS can be utilized at several levels of complexity including targeted resequencing of multi-gene panels, whole exome sequencing (WES), and whole genome sequencing (WGS). Currently, targeted resequencing of multi-gene panels is the most widely adopted application of NGS offered in the clinical laboratory setting because it provides a cost-effective and comprehensive diagnostic approach to examine panels of genes for disease-associated sequence variations to answer specific clinical questions[5,11,17-20]. Targeted resequencing and WES require enrichment of genomic regions of interest prior to sequencing. Target enrichment can be performed using several strategies including PCR- based capture, molecular inversion probe-based capture, and hybrid capture methods (reviewed in Ref. 21). Whole exome sequencing, equivalent to a very large gene panel, involves the selective enrichment and sequencing of the majority of known protein-coding regions of the human genome (reviewed in Ref. 22), which contain approximately 85% of all variants currently known to contribute to human disease[12]. Whole genome sequencing differs from targeted



Nature Biotechnology: doi:10.1038/nbt.2403

sequencing and WES because it does not require target enrichment and allows the interrogation of both the protein-coding and non-protein-coding regions of the genome.

### c. NGS platform characteristics

Several commercial NGS platforms are currently available and the technology will continue to evolve. The first generation of technologies that have been integrated into clinical laboratory settings utilize clonally amplified DNA templates sequenced on a flow cell in a massively parallel fashion. These platforms and other technologies in development will not be extensively reviewed here (reviewed in Refs. 11, 12, 23-27). While these technologies utilize a variety of chemistries, including sequencing by synthesis and sequencing by ligation, the platforms share similar processing steps. First, DNA is fragmented and platform-specific oligonucleotide adaptors are added to repaired ends to generate a fragment library. The individual fragments are clonally amplified and then sequenced on a flow cell to generate luminescent or fluorescent images that are processed algorithmically into sequence reads[11]. More recently, semiconductor chips with sensors that detect hydrogen ions released by DNA polymerase during DNA synthesis have been used to enable DNA sequencing[27]; this platform is just beginning to be implemented in the clinical laboratory. Additional proprietary NGS platforms are available from outsourced service providers[28].

Most NGS platforms can sequence library fragments from both ends, referred to as paired-end sequencing. This process essentially doubles the amount of sequencing data, plus the expected distance between paired ends is known, thereby improving the alignment accuracy and detection of structural rearrangements, such as insertions and/or deletions (indels), and inversions. A complementary approach, known as mate-



Nature Biotechnology: doi:10.1038/nbt.2403

pair sequencing, is a modified paired-end strategy that permits the analysis of longer DNA fragments by NGS to improve the elucidation of structural rearrangements. Mate pairs are generated by circularizing DNA fragments of known size to a common linker, the fragments are cleaved at a known distance from the linker, and subsequently sequenced using the paired-end strategy[24,26].

**d. Data management and downstream informatics analysis**

A number of informatics tools are available for alignment and assembly of the millions of reads that are generated by NGS platforms. For most platforms, analysis of NGS data begins with the conversion of image files into base calls with their associated quality scores. Each platform uses its own algorithm to determine quality scores that are conceptually similar to Phred quality scores used in Sanger sequencing[11]. Next, individual reads are processed through quality filters which remove sequences that fall below a predetermined quality score and aligned to a reference sequence, or, when no reference genome sequence is available, used for *de novo* assembly[12]. A variety of computational methods have been developed to align short read sequence data to a reference sequence and these methods can be optimized, either globally or regionally, according to the specific characteristics of the sequence and variants being assessed (reviewed in Refs. 29,30). After reads are mapped to the reference sequence, variants are called when differences occur between a base call and its aligned position to the reference sequence. This process is referred to as the data analysis pipeline[13] and typically results in the generation of a standard file format for storing the sequence variation (single nucleotide polymorphisms [SNPs], indels, structural variation, etc.) with



Nature Biotechnology: doi:10.1038/nbt.2403

high level summaries and annotations of the analyzed sequence referred to as the variant call format (VCF)[31] or the genome variation format (GVF)[32].

Underlying the discussion of informatics for NGS is the need for a robust and sophisticated information technology infrastructure within any laboratory implementing NGS. Terabytes of data are typically produced that require significant storage capacity and computing power. Following analysis, laboratories retain data as a component of their quality management process. At the present time, it has not been resolved which level of data should be retained from a NGS analysis to be useful for reasonable interpretation while minimizing the significant storage costs[13]. One issue that has not been resolved is whether it is appropriate to simply store the list of variants, or VCF file, or retain sufficient information to support reanalysis of the data when elements of the informatics pipeline or reference sequence change. If the laboratory determines that retention of only the VCF file is not sufficient, the original genome sequence file (e.g., FASTQ format) that includes quality information or the alignment file (BAM format) could be retained. Regulatory standards for NGS data retention do not exist in 2012 but general guidance recommends that such data be retained as long as feasible and at least until the next PT challenge or use of an alternate approach to the independent assessment of test performance.[33]

Currently, there is no standard, streamlined data analysis pipeline and the data analysis process is customized based on the type of sequence variations targeted by the assay[13]. NGS chemistries are prone to introducing errors into individual reads. The error profile refers to the instrument-specific likelihood to make erroneous base-calls and is directly related to signal-to-noise considerations in each instrument's approach to



Nature Biotechnology: doi:10.1038/nbt.2403

chemistry and detection. For example, if errors are more likely to accumulate in reads from later sequencing cycles, informatics filters should be established that remove the later reads from sequence analysis to improve accuracy. Appropriate filters can only be designed when the source of false positives and the error model of the instrument data are understood. Errors in individual reads can be mitigated through the analysis of multiple overlapping reads[11]. The number of reads covering a given base position is described as depth of coverage, and this parameter contributes to the accuracy, sensitivity, and specificity of variant detection[34]. These types of errors are not observed in Sanger sequencing, which averages the errors of a large number of individual reactions to generate each base call in a single read. To accurately make a variant call using NGS, the variant should be present in multiple, overlapping individual reads, ideally derived from both DNA strands, which reduces the bias effects of sequence context (see section 4.c.i. for more detail). Sequence context can affect sequence analysis in several ways; two examples are GC bias and strand bias. Genomic regions with high GC content may prove difficult or impossible to enrich by target capture approaches and captured fragments can be difficult to sequence using NGS platforms. Therefore, it is difficult to obtain sequence information for the first exons of many genes, which are typically more GC rich than the other exons. Strand bias is observed either when the majority of sequence reads originate from only one DNA strand, or when variant bases occur preferentially on one strand compared to the other. Inaccurate base calls are more likely to cluster on one strand of the DNA[35]. Therefore, reads from both forward and reverse strands should be considered to make accurate variant calls, and reduce errors. The laboratory should develop appropriate filters that incorporate



Nature Biotechnology: doi:10.1038/nbt.2403

information about the distribution of reads on the forward and reverse strands to minimize errors due to strand bias. Some regions of the genome are not amenable to NGS, regardless of the depth of coverage achieved, due to genomic complexity (e.g. high GC content or areas with repetitive regions). In addition, the human genome reference sequence[36] poses difficulties for NGS read alignment because it is a composite representation of the human genome with data derived from several individuals and is not representative of all human genomes. Because it has not been possible to derive sequence from some regions of the genome, there are regions with no reference sequence available. At the present time, the analysis software is still evolving and a gold standard reference genome does not exist. Therefore, it is the responsibility of the laboratory director to assure that the informatics pipeline is properly validated and that there is an understanding of the types of variants that can or cannot be detected within the genomic region to be investigated.

## 3. Methods

A national workgroup was organized, and a two-day meeting was convened to initiate discussion of the issues and to develop consensus recommendations that are broadly applicable to both the current generation of NGS platforms and applications, as well as those anticipated in the near future. The meeting had forty-one participants with extensive knowledge and experience with NGS that included clinical laboratory directors, clinicians, platform and software developers, and informaticians. Individuals actively engaged in NGS guideline development from accreditation bodies (College of American Pathologists, CAP) and professional organizations (American College of Medical Genetics, ACMG; Association for Molecular Pathology, AMP; Clinical



Nature Biotechnology: doi:10.1038/nbt.2403

Laboratory Standards Institute, CLSI). Representatives from US government agencies (Food and Drug Administration, FDA; Centers for Medicare and Medicaid Services, CMS; National Institute of Standards and Technology, NIST; Nathional Institutes of Health; NIH; and CDC) also participated. The meeting consisted of plenary, roundtable, and workgroup sessions that were designed to facilitate discussion, foster collaboration, raise issues and build consensus among participants. Following the meeting, participants were engaged in teleconference meetings to complete the discussions that began in the face-to-face meeting. Discussions focused on NGS as applied to the clinical detection of constitutional germ line variants; therefore, consideration was given to the methods used to align sequence reads to the human reference genome build, not *de novo* assembly. To limit the scope of the meeting, the applications of NGS to infectious disease and oncology, as well as the use of NGS to detect large copy number variants (CNVs), structural variants, and mosaicism were not considered. Emphasis was placed on identification of platform-independent metrics to ensure quality of sequencing results; however, when necessary, performance characteristics unique to specific platforms were considered. Topics included elements of a quality management system related to the analytical process: test system validation, quality control (QC), proficiency testing (PT) or alternate assessment (AA) when PT is not available, and reference materials (RMs).

Discussions were focused on processes necessary to ensure the analytical validity of sequence results and the workgroup did not consider annotation or the clinical interpretation of test findings.



Nature Biotechnology: doi:10.1038/nbt.2403

**The following sections describe the results of the workgroup deliberations**

## 4. Validation

### a. Platform, test, and informatics pipeline validation

Prior to initiating patient testing, clinical laboratories must establish or verify the analytical validity of molecular genetic tests[6,33]. In the US, CLIA requires that laboratories intending to use tests approved by the Food and Drug Administration (FDA) verify the performance specifications established by the manufacturer (e.g., accuracy, precision, etc.) [6]. In contrast, CLIA requires establishment (or validation) of performance specifications for test system performance characteristics (accuracy, precision, reportable range, reference range, analytical sensitivity, analytical specificity, and other performance characteristics, as applicable) for clinical laboratory tests that are not cleared or approved by the FDA, i.e. laboratory-developed tests (LDTs).

The validation process may be divided into three interconnected components: platform validation, test validation, and informatics validation (Supplementary Fig. 1).

Platform validation is the process of establishing that the system can correctly identify each type of variant that the test is designed to detect. NGS technologies are relatively new and have multiple options for virtually every step in the complex workflow. Recommended performance specifications have not been established for each possible combination of assay and analysis tools, which makes it more difficult for the testing laboratory to validate the assay. During platform validation, performance specifications (the value(s) used to describe the quality of a test result) for all of the appropriate performance characteristics (accuracy, precision, etc.) of the sequencing platform and the analysis pipeline should be established within the clinical setting in which the testing



Nature Biotechnology: doi:10.1038/nbt.2403

is to be offered. All steps involved in the NGS assay, for example DNA isolation, enrichment methods, library preparation, and data analysis, should also be considered in the platform validation. This process must also include development of informatics thresholds for the alignment processes to flag reads that are not considered of high enough quality to make a reliable call[37]. Once parameters and performance of the individual parts of the test have been optimized and validated, changes that do not affect processes, such as replacement of a depleted reagent, do not necessitate a revalidation, but a confirmation that the performance specifications of the test are not affected using ongoing QC.

Once platform performance has been established, assays should be validated for their ability to identify variants in the specific regions of the genome under investigation; this process is referred to as test system validation. Test system validation should be an end-to-end validation, assessing the platform along with the unique content of each assay. For test system validation, a number of samples should be used to assess the performance of the assay. Patient samples with disease-associated sequence variation(s) should be used to the extent possible. Reference materials, such as gDNA from characterized cell lines that have similar sequence variations as those targeted by the assay can also be used (see detailed discussion of RMs in section 7). These latter variants may be nonpathogenic and should be located in genomic regions targeted by the assay.

The third component of clinical validation for an NGS assay is the validation of the informatics, or analysis pipeline (see section 4.d. for detailed description of assessment of analysis pipelines). Validation of the informatics pipeline serves to



Nature Biotechnology: doi:10.1038/nbt.2403

establish and document the software setting(s) that are appropriate for generating accurate sequence data and the capacity to detect variations within the targeted genomic region(s). The workgroup recommended that the informatics pipeline be optimized as a separate entity during test development to document processes unique to the relevant software components of the clinical NGS test.

Validation of the informatics component is necessary to ensure that the assay is capable of detecting all targeted variants within the genomic region that is interrogated. During the informatics validation, the performance specifications of the data analysis pipeline should be established using appropriate RMs that may include electronic reference data files that contain sequences that are simulated or based upon actual patient samples, or other RMs, such as characterized gDNA from cell lines.

It is important to describe the test characteristics in the method section of the clinical report that is given to the ordering physician. This report should contain information about what regions of the genome were sequenced and analyzed successfully and which were not. This is especially important when Sanger sequencing is not used to complete the regions that are not attainable using NGS. The report should include information that describes the test, including the genomic regions (genes, exons, etc.) that the test is designed to cover, as well as the test's capacity to detect different types of variants as a function of coverage (including the uniformity and average depth of coverage).

**b. Limitations: considerations for homologous sequences**

Before considering metrics and their application, it will be useful to consider some of the limitations of NGS for clinical applications. Homologous sequences can



Nature Biotechnology: doi:10.1038/nbt.2403

lead to false positive and false negative calls (for example through misalignment of reads from a pseudogene). Enrichment techniques are used when applying NGS to the analysis of exomes and gene panels. Hybridization capture and PCR-based target selection are the most common methods used for enrichment. Hybridization-based capture methods necessitate careful measures to minimize co-capture of non-target or homologous pseudogene sequences[11]. Usually it is not possible to isolate genes from their pseudogenes when using hybridization based methods. However, the targeted gene may be resolved from pseudogenes or other homologous regions by aligning with the whole genome as a reference and not just to the targeted region. This approach can help reduce interference from captured non-target sequences by aligning them to the correct location without forced, misalignment to the targeted region. These genes are also prone to gene conversion events that may make their interpretation difficult. PCR based target selection methods can be used to amplify only the target gene by designing primers specific for amplification of only the true gene target and not the pseudogene, when applicable. The entire sequenced region should be analyzed using software tools, such as BLAT (Blast-Like Alignment Tool[38]), to establish which regions are repetitive, have pseudogenes, or contain other properties that may need special attention when sequencing limited genomic regions or gene panels by NGS.

**c. Establishing performance specifications**

Regulatory requirements and quality management system standards[6,39] call for clinical laboratories to determine specifications for performance characteristics that include accuracy, analytical precision, analytical sensitivity, specificity, reportable range of test results, reference range and other characteristics of relevance as part of their



Nature Biotechnology: doi:10.1038/nbt.2403

validation process to ensure the analytical validity of test results[33]. Although these performance characteristics have been applied to Sanger sequencing[40], they do not readily translate to NGS. Therefore, a modified framework for considering these terms for NGS is presented (Supplementary Table 1).

### i. Accuracy

For nucleic acid sequence determination, accuracy can be established by determining the closeness of agreement between a measured value and the true value, which for NGS is the accepted reference sequence. Optimizing the accuracy of NGS requires consideration of several factors, some of which are unique to this type of testing. One parameter is the establishment of an adequate depth of coverage. An adequate depth of coverage threshold[34] necessary to make accurate variant calls should be established empirically during the validation of each NGS application. When establishing adequate depth of coverage, RMs or previously characterized samples may be used to define the depth at which additional coverage does not significantly improve upon the accuracy of the sequence. This can be done by analyzing coverage for a large number of variants included in the test validation and plotting the number of false positive and negative results as a function of coverage. The depth of coverage needed is dependent on the type of variation present in the sequence and its zygosity. In general, less coverage is needed to accurately detect homozygous SNPs than for heterozygous SNPs[41]. It is important to distinguish between a test's average coverage and uniformity of coverage. Average depth of coverage is the average number of overlapping reads within the total sequenced area. The uniformity of coverage is the distribution of coverage within specific targeted regions in which variant calling will



Nature Biotechnology: doi:10.1038/nbt.2403

occur. Although the average coverage may meet the laboratory established threshold required for accurate variant calling, the depth of coverage will vary across the genome, or targeted regions, resulting in variable accuracy across the genome. Uniformity of coverage should be measured by assessing coverage across the regions that are sequenced.

The allelic read percentage or allelic fraction defines the proportion of individual reads containing a variant needed to make a call. Ideally, homozygous variants would be expected to contain the variant in every read, or an allelic read percentage of 100, while a heterozygous variant should contain the variant in 50% of the reads, or an allelic read percentage of 50[11]. Amplification bias, alignment bias, and errors are inherent in the "random sampling" introduced by coverage; therefore, homozygous, hemizygous, and heterozygous variants exhibit a range of allelic read percentages. Establishing a threshold for variant calls should be defined empirically for each test. One approach is the use of synthetic controls with calculated variant percentages or using previously characterized human cell line DNA[42] to determine the observed variability during analysis. One important consideration for analysis of allelic read percentages is that duplicate reads (e.g., reads that are PCR duplicates or paired-end reads that have alignments beginning and ending at the same position) are generated by clonal amplification of the NGS library prior to sequencing. The number of duplicate reads may be high and their inclusion generates a risk of skewing the allelic fractions. For example, a possible consequence of this skewing would be a missed variant because fragments containing the nonpathogenic allele could be overrepresented. Therefore, duplicate reads (all but one with the highest quality score) should be removed during the



Nature Biotechnology: doi:10.1038/nbt.2403

alignment refining process. In addition, cutoffs should be defined for homozygous and heterozygous calls. For example, in a targeted PCR-based NGS test, Jones *et al.,(*5) observed that for all heterozygous variants, 23-74% of sequences contained the variant, while for all homozygous variants, 78-100% of sequences contained the variant. Based on these data, filters were established so that variant calls with an allelic read percentage of < 85% for homozygous variants and <40% for heterozygous variants were eliminated[5]. These cutoffs should be determined empirically for each assay to ensure that the minimum required depth of coverage to achieve the desired allelic read percentage for all regions is included in the test. When the established coverage threshold is not achieved in a region that requires analysis, the data should either be rejected with no results reported, flagged for further NGS analysis, or re-analyzed using an alternative analytically valid method (e.g. Sanger sequencing) before making a variant call in that region. A recent publication[34] indicated that even 30x average coverage may not be adequate to produce genotype calls with acceptably low error rates across a large portion of the genome. As previously mentioned, this is because not all regions of the genome will have the same amount of coverage. The average coverage threshold is typically established for all genomic region sequenced to achieve reliable base calling. Early adopters of NGS in the clinical setting have often established average coverage thresholds that range from 15X-100X, although this will be dependent on the assay design and technology. A separate threshold, termed the minimum base coverage threshold, should also be established to identify areas of low coverage in which a variant cannot be reliably called. For example, early adopters have established minimum base coverage thresholds of 15X[5] and 30X[18]. When the minimum coverage



Nature Biotechnology: doi:10.1038/nbt.2403

threshold of a targeted area in a gene panel assay is not achieved, or a specific region is problematic, an alternate method such as Sanger sequencing should be performed in place of or in parallel to NGS. For example, the first exon of many genes is often GC-rich, presenting an obstacle to reaching a desired coverage threshold[5].

For targeted panels, WES and WGS, the degree of coverage across the regions being sequenced should be comparable from run to run. The expected relative degree of coverage of each genomic region should be established during the validation and monitored with each patient run. Identification of a genomic region that is exhibiting unusual relative coverage does not mean that the entire data set should be rejected because the errors may be specific to that particular region and may indicate a localized change in coverage that needs to be further evaluated or a structural change such as a deletion; it may not indicate a systemic problem.

Sequence specific features such as the under-representation of GC-rich sequence reads, referred to as GC bias, can often reduce the uniformity of coverage in an NGS run[34]. Monitoring GC bias provides a measure of the uniformity of coverage across the genome or targeted area and should remain consistent between runs. The level of GC bias observed with an assay should be determined during validation, and should be monitored with every run as a QC measure. GC bias also provides information about the quality of the sample preparation and capture steps[34]. Laboratories should consider what would constitute a significant deviation in coverage that would warrant additional examination of the data and possibly its exclusion. In addition to the uniformity of coverage, an even distribution of forward and reverse reads should be achieved to avoid making errors due to strand bias, which is a common



Nature Biotechnology: doi:10.1038/nbt.2403

source of false positive calls. Strand bias in all targeted regions should be monitored during each run.

Another critical component that contributes to the confidence of the final sequence is the evaluation of per base quality scores (Q score) of overlapping sequencing reads. To assign each base a Q score, the quality of image files is evaluated by assessing the strength of a signal relative to the background across a read length (signal to noise ratio) and to neighboring bases. The base calls are assigned a Phred- scaled Q score that estimates the error probability for each base. For example, a Q score of 20 has a 1/100 likelihood of error[43]. There are no standards for deriving quality scores for NGS and they are not directly comparable among platforms. Therefore, it is critical that the performance of these scores be assessed during the validation; this is commonly done by generating a quantile-quantile plot in which a well-characterized sample is evaluated for the accuracy of calls made relative to the Q scores associated with those calls[44]. More accurate Q scores, or confidence scores, can be determined using base quality recalibration algorithms that correct for covariates such as confidence in alignment to the reference sequence, sequencing technology, machine cycle, dinucleotide context, depth of coverage, forward and reverse sequence balance, confidence based on the 5' or 3' location of the read, and the detection of a second allele[45,46].

During validation, an acceptable Q score required for each base in a read should be established and informatics filters should be used to remove reads containing poor quality bases before aligning to a reference sequence. Alternatively, when the 3' end alone has low Q scores, those ends can be trimmed before alignment. Tagging



Nature Biotechnology: doi:10.1038/nbt.2403

methods, referred to as indexing or barcoding, may be used to mark and track DNA fragments from multiple patient samples that are being sequenced on a single flow cell[47]. While indexing permits multiple samples to be assessed in each sequencing reaction and may provide a means to cost-effectively increase the number of samples assessed for precision, laboratories will have to determine how many samples can be pooled and still achieve the level of coverage necessary to make accurate variant calls.

### ii. Precision

For NGS applications, precision refers to the degree of agreement between replicate measurements of the same material. An adequate number of samples should be analyzed to establish precision by assessing reproducibility (between-run precision) and repeatability (within-run precision) during test validation. Repeatability can be established by sequencing the same samples multiple times under the same conditions and evaluating the concordance of variant detection and performance. Reproducibility assesses the consistency of results from the same sample under different conditions such as between different runs, different sample preparations, by different technicians, and using different instruments. A few early adopters of NGS in the clinical setting have established precision using three reference samples that were each sequenced 3-5 times in the same and in different runs (personal communications, Drs. Madhuri Hegde and Birgit Funke) and this is suggested as a minimum practice for establishing the precision of a platform. Quality control metrics, such as depth of coverage, uniformity of coverage and the transition/transversion ratio, should be determined during the validation, remain constant, conform to published values[46] and may provide supportive



Nature Biotechnology: doi:10.1038/nbt.2403

evidence for establishing precision (see section 5. for detailed discussion of quality control metrics).

### iii. Analytical sensitivity and analytical specificity

Traditionally, analytical sensitivity is defined as the proportion of biological samples that have a positive test result and are correctly classified as positive[48], or the lower limit of detection[33,49]. For both Sanger and NGS sequencing assays, the workgroup defined analytical sensitivity as the likelihood that an assay will detect a sequence variation when present within the analyzed genomic region (this value reflects a test's false negative rate). Analytical specificity is traditionally defined as the likelihood of a test to detect only the target analytes and not interfering substances[33]. The workgroup defined analytical specificity as the probability that an NGS assay will not detect sequence variation(s) when none are present within the analyzed genomic region (this value reflects a test's false positive rate). Currently, some laboratories establish specificity by calculating the number of false positives per assay run (or per genomic interval tested) (personal communication Dr. Birgit Funke). For NGS, these parameters can be established by comparing test results to a method that has been independently validated, such as Sanger sequencing or SNP array analysis. SNP arrays are most useful for assessing the detection of known SNPs in the genome and may serve as an effective independent technique to ensure adequate coverage of the genome for both whole genome[3] and exome analysis. For targeted gene panels, use of such arrays should be carefully considered to determine whether a sufficient number of useful SNPs are included. Concordance with SNP arrays only tests the performance for known SNPs, which generally do not include regions of the genome that are difficult to



Nature Biotechnology: doi:10.1038/nbt.2403

sequence. Discordance between SNP array data and NGS data may be resolved using Sanger sequencing; however, the number of SNPs to be resolved may be large and additional research is necessary to establish how many discordant calls require confirmation to produce statistically valid results.

For both next-generation and Sanger sequencing, it is impractical or impossible to evaluate analytical sensitivity and specificity with respect to the entire spectrum of disease-associated variants. Therefore, it is useful to establish these performance specifications using reference materials that contain both disease-associated and non-disease associated sequence variations. However, it is recommended that the more prevalent disease-associated sequence variations should be included in the analysis. It is also recommended that analytical sensitivity and specificity be established separately for each type of sequence variation because current test platforms and informatics pipelines exhibit differences in their capacity to detect different classes of genetic variations. For example, members of the working group have analyzed 40-71 positive, multiplexed samples that were previously characterized with an independent method to contain the desired variants[5], (personal communications, Drs. Madhuri Hegde and Birgit Funke). The number of samples required is defined by a) the need to validate the capacity of the platform to detect all types of relevant sequence variants (e.g., substitutions, indels) and b) the number of variants to establish an appropriate analytical sensitivity within an acceptable confidence interval. The number of samples needed is greater when no platform validation data exists. For example, 38 of the 71 validation samples used during one initial platform validation were chosen because there was an insertion or deletion present; 0/258 substitutions were missed (95% CI= 98.5-100%),



Nature Biotechnology: doi:10.1038/nbt.2403

(personal communication, Dr. Birgit Funke). Similar issues and challenges have been addressed with regard to cytogenetic microarray analysis (CMA), which examines the whole genome for constitutional cytogenetic abnormalities. For CMA, it is recommended that a minimum of 30 specimens with disease-associated chromosomal abnormalities are evaluated during test validation[50].

The mathematical relationship between depth of coverage and the probability for identifying the correct base should be considered when establishing sensitivity. While greater coverage increases the probability of calling a base correctly, there is also a practical upper limit to coverage that is platform specific. The result of insufficient coverage is a loss of statistical significance for making a reliable base call. The result of excessive coverage has not been well studied but the potential exists for amplification of systematic errors that can lead to an incorrect base call.

NGS is prone to both false positive and false negative results. The propensity for false positive or negative results should be established during the validation of the test to identify problematic regions of the genome which may require evaluation with an alternate analysis, such as Sanger sequencing. To ensure recognition of false positives during patient testing, the workgroup recommended confirmatory testing for all clinically actionable findings. While it is not practical to verify all negative findings for each patient's test, the laboratory should determine the false negative rate through test validation (which could be communicated to users as test limitations), and verify the accuracy of test results, including verification of the false negative rate, at least twice annually. Specific care should be taken to confirm the ability of the test to detect other, more difficult to detect genetic variations including mosaicism, indels and copy number



Nature Biotechnology: doi:10.1038/nbt.2403

changes, when these are relevant to a given disorder or indication. It may be necessary to test the sample using alternate methods to detect and/or confirm these alterations.

A loss of sensitivity and specificity may occur when coverage of a targeted sequence is below the criteria set during the validation; this is only true if low coverage regions are not completed by Sanger sequencing. The informatics pipeline can be used to identify whether a loss of sensitivity and specificity has occurred during a sequencing run by flagging genomic regions that fall below the required coverage threshold. Sensitivity and specificity will be affected by the quality of the sequence itself; therefore, evaluation of base quality scores and signal-to-noise ratios are also important. Stepwise approaches can be used during the informatics analysis; for example, include only reads that meet the established quality thresholds in the depth of coverage used to call a variant.

### iv. Reportable and reference ranges of test results

CLIA defines reportable range as "the span of test result values over which the laboratory can establish or verify the accuracy of the instrument or test system measurement response"[6]. For NGS, the workgroup defines reportable range as the portion of the genome for which sequence information can be reliably derived for a defined test system. The reportable range may not reflect a contiguous region of the genome, particularly for analysis of gene panels and the exome, but must be defined when establishing the test definition. There may be areas of the targeted region that cannot be sequenced reliably and therefore are excluded from the reportable range.

Reference range (or reference intervals) is defined as the range of test values expected for a designated population of persons[6]. For NGS, the workgroup defines



Nature Biotechnology: doi:10.1038/nbt.2403

reference range as the normal variation of sequence within the population that the assay is designed to detect. These include SNPs and other sequence variations such as transitions and transversions, indels, substitutions, expansions, short tandem repeats, single exon deletions, and structural variations within a specified region(s) of the genome that occur in the general population. Results that fall outside the reference range (e.g., detection of an indel not normally found in the sequenced region), may require additional investigation to establish the clinical significance. A caveat is that the distinction between a normal and disease-associated variation is not always well defined and in fact may vary among individuals and populations. Databases useful for understanding the spectrum of disease association for variants will be invaluable for making these determinations.

### d. Informatics: assessing the data analysis pipeline

The combination of informatics tools used for processing, aligning and detecting variants in NGS data is commonly referred to as the data analysis pipeline. Informatics software is rapidly evolving but there is no single program that can perform all applications necessary to detect each type of sequence variation. There are numerous programs designed to perform base calling, alignment, and variant calling (reviewed in Refs. 12, 20, 30). Different software tools must be applied to sequencing data to answer questions that are specific to a particular test. The data analysis pipeline established by the laboratory ultimately determines the types of variants that can be credibly called within the targeted genomic regions. There may be instances where different analyses are performed in parallel. For example, data may be analyzed using the same algorithm, but using different quality thresholds for specific regions, (e.g., if the



Nature Biotechnology: doi:10.1038/nbt.2403

target is a GC rich region or has repeats). Likewise, a single software setting is typically not ideal for optimal detection of different classes of sequence variations. For example, the efficient detection of SNPs can eliminate effective detection of indels. Detection of mid-sized indels (3-25 bp depending on the platform) is challenging due to the limitations associated with the use of platforms that produce short reads[18], but these challenges will likely be minimized as read-lengths are extended[51]. Large *de novo* indels, which are longer than the read-length, can also be found in the assembled sequence from the unmatched reads using secondary analysis[52]. Laboratories should consider establishing modular analysis pipelines in which different informatics tools are used to analyze the same data set. During the validation, laboratories should determine that a variant identified by the pipeline is actually present in the sequence and measure the concordance between NGS and the results from an alternate technique. This is a useful QC function during the early stages of implementing NGS into clinical laboratory practice.

The data analysis pipeline should undergo validation for the intended application because the software programs available to analyze NGS data use different algorithms that can cause variability in the reported sequence of a given sample. This validation should include consideration of systematic errors of the test platform. The software parameters should be manipulated during assay development to derive optimal settings for each type of variant the test is designed to detect. Quality thresholds should include metrics such as base calling quality, coverage, allelic read percentages, strand bias, and alignment quality. Analysis for each of these steps is currently software-dependent. Confidence scores can be calculated and assigned to each variant call to assess the



Nature Biotechnology: doi:10.1038/nbt.2403

quality of the read alignment used to generate the final sequence[34]. Sequence files used to validate the pipeline should be derived from samples (gDNA or engineered sequence files) with characterized sequence variations, including prevalent disease-associated sequence variations, and should evaluate the ability of the informatics pipeline to identify the targeted variations without generating false positive results. The final sequence should be compared to a reference sequence or to the results from analysis of the sample by an alternate method (e.g., Sanger sequencing, SNP array, etc.). Sequencing reads are typically aligned to the current build of the human reference genome[36], for exampleHG19, however this reference genome is derived from a small number of donors and is a very small sampling of human genetic variation. The human reference genome also contains rare and common disease risk variants which complicates the detection of these rare risk alleles. There are efforts to build a major allele reference sequence that should be considered for accurate, ethnically-concordant variant calling[53].

The NCBI reference genome is updated periodically, and when a new build of the reference genome is used, the data analysis pipeline should be revalidated to establish changes introduced with the new sequence data. The number of specimens that are required to validate the addition of a new reference build in a data analysis pipeline is an area that requires further consideration. Reads may be aligned to either the entire build of the reference genome, or to complementary targeted regions of the reference genome. Aligning to the entire human genome is more computationally intensive than aligning to a target region[11]. However, use of the whole genome as the reference can help reduce interference from captured non-target sequences by aligning reads to the



Nature Biotechnology: doi:10.1038/nbt.2403

correct location so they are not forced into alignment to targets. Resolving the co-capture of homologous regions or pseudogenes with NGS is problematic (see section 4.b. for considerations for homologous sequences). One approach to resolving homologous regions is the creation of an alignment to a reference sub-genome that can be modified to include the co-captured sequences. Sequence reads will align properly without using the entire genome, but this will not be practical for genes with highly homologous pesudogene sequences. The use of sequencing technologies that produce longer reads will help to minimize the computational intensity and decrease the mapping of reads to more than one location[49].

**e. Indications for repeating a validation**

Any changes to a clinical test, such as changes of instrumentation, specimen types, reagent replacement, software updates, or other modifications require that performance specifications be reestablished or otherwise shown to be unchanged. The extent of validation will depend on the extent of the change. For example, the laboratory should be able to determine that the performance of a new lot of reagent is identical to an older lot. For a more extensive change, such as the inclusion of new genes to an existing gene panel for NGS analysis, a broader revalidation is necessary to ensure the capability to detect new sequence variations without compromising the quality of the original assay. Frequent software and sequencing chemistry updates that require the reestablishment of performance specifications will present challenges for clinical laboratories. In these cases, it may only be necessary to reestablish performance specifications at or after certain steps in the process. For example, if only the informatics pipeline is altered, it may not be necessary to revalidate process steps



Nature Biotechnology: doi:10.1038/nbt.2403

prior to data analysis. Similar to the initial validation, the issue of how many samples must be evaluated to reestablish performance specifications must be considered. The number of samples selected should provide confidence in the test performance and results.

Laboratories certified under the CLIA regulations are required to perform calibration verification of test systems every six months or sooner if there is a reason, such as the physical transport of the test platform to a new location. Before calibration verification, the method for calibration needs to be established as part of establishing a new method along with QC procedures[6]. Traditionally, calibration is primarily applied to the biochemical analysis and test platform. Sequencing results are derived from involvement of both the sequencing steps and the informatics analysis, thus both processes are subject to calibration. These can be accomplished by sequencing a characterized RM and demonstrating instrument and software performance that is comparable to those specifications derived from the validation of the test platform and informatics pipeline.

## 5. Quality Control (QC)

### a. Process steps to be addressed

Quality control procedures must be implemented to monitor the performance of the analytical process. Control procedures are designed to detect immediate errors caused by test system failure, adverse environmental conditions, and operator performance, as well as to monitor the accuracy and precision of test performance over time[33]. Although sequence analysis is typically considered a qualitative assay, NGS



Nature Biotechnology: doi:10.1038/nbt.2403

has both qualitative and quantitative aspects that should be considered when devising effective controls and control procedures.

Quality control materials and metrics for NGS should be established during test validation. Each component of the NGS testing process, including DNA extraction, library preparation, DNA sequencing, and the informatics analysis pipeline should have established QC materials and metrics. It is not standard practice to include multiple positive controls with different variant types, such as those used in the initial validation of the assay, during each run due to the enormous cost and time involved. Including a single characterized external control with disease associated sequence variations to demonstrate that the procedure is working during each run of patient specimens may be sufficient. Ideally, a variety of controls should be utilized. With the recognition that all controls cannot be assessed during every run, a schedule may permit the rotation of a variety of controls that can be run over a reasonable timeframe to monitor performance.

Two general approaches for internal controls were suggested by the workgroup; the first involves the inclusion of a bar-coded gDNA RM or a non-human synthetic control nucleic acid material that is extrinsic to the sample (e.g., sequences that are spiked into the sample at the beginning of the sequencing process). Controls that are spiked into a sample will not serve as controls for the DNA extraction component of a NGS test and the use of a synthetic control is not representative of a patient sample due to its lower complexity and source. The effects of spiking in a QC sample, if any, should be determined during the test validation.

The second approach for internal controls utilizes a control sequence that is intrinsic to the sample, but not found in regions of the genome targeted by the test (e.g.,



Nature Biotechnology: doi:10.1038/nbt.2403

a highly-conserved housekeeping gene, or the mitochondrial genome). Even use of low polymorphic targets may be problematic at times, due to rare, uncharacterized sequence alterations. The mitochondrial genome can be used as an internal control that is integral to the sample[54]. One concern with using the mitochondrial genome is heteroplasmy, which occurs at a variable frequency of about 10%[55]; however, most individuals are homoplasmic in the regions that would be used to monitor platform performance. An additional challenge is the haploid nature of the mitochondrial genome, which does not reflect the true complexity of a diploid genome. The mitochondrial genome may also serve as a positive control for monitoring GC bias and depth of coverage, and may be included in the validation process and then analyzed with each patient sample as an internal QC. However, this approach has limitations with respect to the complexity, and copy number of the mitochondrial genome as compared to that of the nuclear genome.

Once the data analysis pipeline is validated, variation between runs should be minimal. Variation that falls outside the validated range should be investigated as it may indicate an inherent problem. Variability of the informatics pipeline should be monitored routinely. Combinations of spiked-in, synthetic, and actual sequence data are useful for ongoing quality assessment. Laboratories should ensure that appropriate QC procedures assess all aspects of the sequencing process, including sample performance on the machine, base calling, alignment, and variant calling.

**b. Measuring analytical performance during the run: metrics and their applications**



Nature Biotechnology: doi:10.1038/nbt.2403

The performance specifications established during the validation process should be used to monitor the quality of a run each time a sample is processed. Depth of coverage, uniformity of coverage, and base call quality scores are metrics that should be evaluated for each NGS assay, regardless of the application or platform. Additional performance metrics including GC bias, transition/transversion ratio, proportion of reads that map to a non-targeted region, first base read success, removal of duplicate reads, and monitoring the expected decline in signal intensity are also useful to evaluate platform performance (Supplementary Table 2 ). Meeting participants concluded that specific and generalized recommendations for ranges and thresholds associated with metrics, such as mapping quality, can not be established at this time because of inherent differences among applications, platforms, and informatics tools. Monitoring NGS assay performance metrics of control materials, such as characterized RMs and previously tested patient samples with disease-associated sequence variations, is required to verify the analytical quality of a sequencing run. If performance is not consistent with the profile established for these control materials, the accuracy of the run needs to be further investigated. Currently, the high cost and analysis time of NGS assays require interim review of certain metrics during the course of the procedure to ensure that the test is performing as expected. The run may be terminated prior to completion if one or more procedures (sometimes called "quality check point") fail, or if significant deviation from the specifications established during the validation procedure is detected. During the sequencing run, some platforms allow the assessment of one or more early base reads (e.g., 1st and 20th base read) to determine the early success of the run. Other important metrics that should be evaluated early in the sequencing



Nature Biotechnology: doi:10.1038/nbt.2403

process include: quality scores, coverage, GC content, and number of reads that pass other data analysis filters. In total, this level of evaluation addresses assay failures as well as procedural problems such as errors made during sample preparation and loading.

### c. Use of confirmatory testing

At this time, confirmatory testing of all clinically actionable variants detected by NGS is recommended because NGS is a relatively new technology, clinical laboratory experience is limited, and the error profiles of existing platforms are variable[5,18,56]. Many clinical laboratories use in-house developed informatics pipelines to identify the disease-associated sequence variants. Laboratory developed and publicly available DNA sequence databases are used for this purpose. This process ensures that variants are analyzed for their properties and effects on the coding sequence; recognized benign changes and system artifacts are not selected for confirmation. A general observation from the current panels offered by clinical laboratories is that the false positive rates hover between 1-3% of confirmed variants depending on the quality metrics of the run[5], (personal communication Dr. Madhuri Hegde). To ensure acceptable turnaround times for targeted, small panel testing, the laboratory should design and validate Sanger primers that amplify the genomic regions with the highest likelihoods of clinically significant variants. While Sanger sequencing is considered the gold standard for clinical sequencing, any analytically valid test, such as genotyping assays would be appropriate for confirmation of test results. Each NGS platform has unique systematic biases; therefore, with decreasing costs of NGS, sequence analysis using two different platforms with unique error profiles may prove more feasible. The combined use of



Nature Biotechnology: doi:10.1038/nbt.2403

WES with WGS[22] to detect variants and increase confidence of base calls may also become practical. It is important to note that some regions of the genome cannot be sequenced accurately using NGS, and these regions are also difficult to analyze using alternative methods.

The library preparation and enrichment steps are complex, multistep procedures that increase the possibility of sample mix-up. Therefore, it is critical for clinical laboratories to have a sample-tracking protocol in place. Sanger sequencing not only confirms the variant, but also provides a mechanism to ensure that no sample mix-up has occurred. Running a SNP array separately and comparing results with the WGS data is an alternative approach.

## 6. Strategies for Proficiency Testing (PT)

### a. Purpose of PT

The CLIA regulations mandate participation in PT programs for a specific set of test procedures and analytes, which currently do not include human molecular genetic tests. For each test subject to the regulations but with unspecified PT requirements, laboratories must verify test performance twice yearly[6]. For many tests, this is accomplished through participation in a formal PT program from an independent third party which provides blinded samples to laboratories on a periodic basis and collects and analyzes the results. Participants in PT programs test the PT samples in a manner similar to patient specimens using their standard laboratory methods and return the requested data, usually an analytical result and an interpretation, to the PT program. The PT program analyzes the results from all participants and returns a summary showing how the participant's results compared to those of its peer group, or to all



Nature Biotechnology: doi:10.1038/nbt.2403

participants. The participants are not individually identified in the summary report. Participation in PT permits laboratories to assess their ability to detect or measure the analytes of interest and provides an independent measure of laboratory performance compared to other sites using the same or different methods. Participation in PT also helps to identify analytical and interpretive errors and may also indicate problems with QC, calibration, or assay design.

In addition to formal PT, genetic testing laboratories can fulfill the requirement for independent verification of test performance by using alternate assessment (AA) procedures such as blinded inter-laboratory sample exchange, retesting of de-identified patient samples, and testing of split samples by two laboratories[57]. Proficiency Testing programs specific to NGS technologies do not currently exist, thus the workgroup discussed possible AA strategies and also considered how a NGS PT program might be structured.

**b. Alternate assessment (AA): Considerations for laboratories prior to availability of a PT program**

For NGS, several approaches may be taken to satisfy the AA requirement[57,58]. Alternative assessment exercises should assess the analytical processes associated with NGS as well as the pre- and post-analytical phases of testing. Alternate assessment schemes should account for variations between laboratories, such as targets and quality metrics, and should consider which aspects of the testing process can be reasonably compared among laboratories.



Nature Biotechnology: doi:10.1038/nbt.2403

There are some drawbacks to AA. Sample exchange typically involves a small number of exchange partners; therefore, it does not allow performance comparison with a more diverse group using a variety of different methods. It also does not assure anonymity of the results of the partners unless a third party receives and interprets the results for them. If the exchange partners use the same technologies and methods, method-related analytical issues may not be identified. Finally, it may be difficult to resolve discrepancies when only two or a few laboratories are involved in the exchange. There are also drawbacks to blinded retesting of previously tested samples by the same laboratory. This method may not identify systematic errors and does not allow comparison of results to those of other laboratories who use different methods.

### c. How to provide PT for NGS

Traditionally, PT has been offered to assess tests for a defined genetic disorder, such as cystic fibrosis or fragile X syndrome. Some PT programs offer methods-based PT in which the ability of the laboratory to correctly execute a particular technique, such as Sanger sequencing or cytogenetic microarray analysis, is assessed independent of a particular disorder.

Laboratories offer NGS tests for different clinical indications. These tests target different genomic regions, and the test platforms and informatics pipelines vary between laboratories. This suggests that a methods-based approach designed to assess test performance independent of any specific indication for testing or condition will provide the best means to use PT for inter-laboratory comparison. A methods-based approach presumes that independent verification of the analytical accuracy correlates with the success of the laboratory performing the test for the detection of targeted sequence



Nature Biotechnology: doi:10.1038/nbt.2403

variations. One advantage of a methods-based approach is that the number of PT samples is minimized because the method, not its capacity to detect each targeted analyte, is being evaluated. This is an important consideration due to the cost and time needed to prepare the PT challenge, perform the laboratory testing, and analyze the results reported by all participants.

### d. Sources of PT samples

Many different types of samples, including characterized RMs, DNA derived from human cell lines, patients' samples (gDNA or whole blood), synthetic DNA, or electronic data can be used for PT. The advantages and disadvantages of each of these sample types are described in Supplementary Table 3.

Most PT programs that provide challenges for genetic testing currently distribute human cell line derived gDNA as PT samples[59-61]. These samples contain both normal and disease-associated sequence variations. They are characterized by several laboratories prior to distribution to assure their usefulness as PT samples. For existing disease-based PT challenges, this characterization is typically performed on a small region of the genome for a single gene or part of a gene which is targeted by the laboratory test. Genomic DNA from cell lines is readily available and many of these lines are well-characterized with respect to their intended PT use for different types of assays. However, because rearrangements may occur during the creation and passage of cell lines, the gDNA may not faithfully represent the genome from the original patient, so the sequence should be monitored each time new gDNA is prepared.



Nature Biotechnology: doi:10.1038/nbt.2403

Whole blood or gDNA from whole blood can also be used for PT. These sample types most accurately reflect clinical samples, which are a desirable characteristic for a PT sample[62]; however, this approach has several limitations. Blood from one patient is in limited supply and cannot be pooled with other samples but it will not contain rearrangements as a consequence of cell culture. Many PT programs do not distribute whole blood because of concerns related to sample stability and possible infectious agents.

NGS involves the manipulation of electronic sequence files; therefore, PT challenges designed to evaluate the ability of laboratories to correctly handle, align, analyze and interpret these files would be beneficial. Informatics tools used in NGS are continuously developed and tested and current software packages have significant differences in performance and accuracy due to their design and the complexity of the sequence being analyzed. An informatics PT challenge to evaluate the ability of the data analysis pipeline to detect variants comprising the variant spectrum targeted by a NGS test would be valuable. An actual or synthetic human genome dataset containing raw sequence data files with known variants can be used as the PT material for such an assessment activity. The sequence content of these files can be readily altered for the purpose of evaluating the effects on pipeline performance to align and call variants. FASTQ data files, text-based formats for storing sequence data with corresponding quality scores, may serve as a common file format and should be compatible with most data analysis pipelines (often specialized tools are needed for CSFASTA, which is a unique format using the SOLiD color-space method)[63]. Other file types such as SAM (Sequence Alignment/Map format)[64], BAM (Binary Alignment/Map format, a



Nature Biotechnology: doi:10.1038/nbt.2403

compressed binary version of SAM), or formats for describing only the sequence variants such as VCF (variant call format)[31] or GVF (Genome Variation Format)[32] may be useful to evaluate pipeline performance. Some file types may have limitations for PT, for example all file formats may not be interoperable across informatics pipelines, and VCF and GVF will not allow evaluation of earlier steps such as alignment or variant calling.

### e. Characterization of PT samples

DNA, blood or electronic sequence files to be used for PT should be characterized prior to shipment to participants[58,62,65]. DNA from characterized RMs or from patients whose genome has been extensively sequenced should be confirmed by an experienced laboratory to determine the identity of the sample and the suitable quality of the DNA for NGS analysis.

Samples without prior characterization will require more extensive study by one or more reliable reference laboratories before being used for PT. Possible approaches include whole genome or whole exome sequence analysis on one or more NGS platforms, secondary testing using an alternate method such as Sanger sequencing, SNP analysis, and additional characterization, which may include cytogenetic microarray analysis, and/or karyotype analysis. A consensus sequence from the results of a variety of NGS platforms and other analyses will help identify errors due to platform specific biases and difficult to sequence regions. Sequencing trios of family members (mother, father, and child) helps to increase confidence in variant calls, when such analyses are possible[53]. The consensus sequence of these PT samples can be



Nature Biotechnology: doi:10.1038/nbt.2403

updated as more data becomes available, especially as improved technologies and analysis software become available.

PT programs should recognize several issues as they characterize potential PT samples and when they evaluate the PT participant results. Results should be evaluated in accordance with stated limitations of the individual assay. For example, if the assay is not designed to detect indels, then the participating laboratory should not be penalized for the failure to detect this sequence variation in a PT sample. The sequence of the sample will vary depending on the NGS platforms and software analysis programs used to determine the sequence. The PT challenge should permit differences among laboratories for the genomic regions and types of variants targeted for testing. Different regions of the genome will have different degrees of certainty associated with their sequence, and some regions are more easily sequenced than others with various preparation and sequencing approaches[22]. The program should be mindful of this when evaluating participant results. Participants should return only those results that meet quality parameters established during their validation process and indicate those requested regions for which their assay is not designed or validated to detect. Additional considerations are needed to develop a PT program. For example, PT programs may need to supply a reference sequence to the participants for use when analyzing their data because laboratories may have validated their assay against a different reference.

**f. Possible approaches to provide PT for NGS**

**i. Use of DNA from well-characterized cell line as PT sample (wet laboratory challenge)**



Nature Biotechnology: doi:10.1038/nbt.2403

PT programs currently distribute purified gDNA from a cell line or patient's sample to the PT participants. For NGS, this approach would depend on the availability of well-characterized samples with good quality consensus sequence and regions of poor quality sequence clearly identified. Depending on the goals of the PT scheme and the clinical assays utilized by the participants, various types of data could be returned to the program for evaluation, for example, data derived from whole genome, whole exome or specific gene panels. Meta-data, such as the VCF file, from the analysis along with data files (FASTQ, SAM or BAM files) could also be returned to evaluate earlier steps of the data analysis procedures. In addition, participants may be required to interpret their findings and provide a report. The PT program will need to design evaluation protocols that consider the platform used and the stated performance specifications of each participant's NGS assay.

There are advantages to this type of PT scheme. Since purified gDNA is the starting point for the majority of NGS assays, use of gDNA as a PT sample can allow almost the entire analytical process, except the DNA extraction step, to be evaluated. In addition, gDNA is obtained from cell lines is renewable and samples can be sent to many participants over many PT challenges. Whole blood can also be used as a PT sample. This approach would allow the entire analytical process, including DNA extraction, to be evaluated. As mentioned previously, there are sample integrity and infectious disease issues associated with the use of whole blood, so most PT providers do not often utilize this sample type.

Laboratories and PT programs should consider several issues when deciding whether to use gDNA from a cell line as a PT sample. As previously mentioned, the



Nature Biotechnology: doi:10.1038/nbt.2403

genome of cell lines may not be stable over multiple passages. Care should be taken to verify the cytogenetic structure and sequence of each new batch of gDNA made from cell lines. Any genome may contain only a limited number of disease-associated variants, thus a single gDNA cannot be used to simultaneously evaluate the ability of a NGS assay to detect the complete spectrum of disease-associated variants. Many genes contain polymorphisms which may be useful as part of the assessment tool. PT programs should take into account that DNA purified by an outside vendor may not perform equivalently in all laboratory assays due to differences in methods for DNA isolation.

**ii. Use of electronic data as PT sample (dry laboratory challenge)**

Electronic data can be used as a PT sample to evaluate and compare the ability of the participant's data analysis pipeline to assemble and analyze a given sequence. Both actual and simulated data may be useful for this purpose. This approach, which is currently being used by CAP for Sanger sequencing challenges[59], has several advantages over the use of other sample types. Analysis of electronic PT samples does not require consumption of reagents and is less costly for the participant both in effort and time. This allows the PT provider to send more electronic samples per year, which increases the opportunities for a wider variety of challenges. It may also be possible to create electronic PT challenges that contain a broader and more defined spectrum of sequence variations in clinically important genes or regions that are more difficult to analyze. Such composite files can represent a combination of sequences that could not be replicated in a single gDNA sample. Programs should send electronic



Nature Biotechnology: doi:10.1038/nbt.2403

challenge files that are compatible with the data analysis pipeline, software capabilities, and test design of participating laboratories.

#### g. PT frequency and strategy

CLIA requires that laboratories evaluate the accuracy of testing at least twice a year for tests or analytes for which PT is not required, such as genetic tests. The College of American Pathologists (CAP) currently sends three gDNA samples twice a year for its molecular genetic PT surveys. Considering the time and cost involved in performing and analyzing results from a NGS test, provision of six gDNA PT samples per year may be a difficult financial and/or time burden for laboratories as well as for the PT programs.

Combining formal PT and AA may be a strategy that is less expensive and time consuming without sacrificing quality performance. It may be more practical to have PT programs provide a single challenge with the understanding that laboratories will also participate in a single AA event each year. This would allow laboratories to exchange samples with others performing similar tests, providing an opportunity to test samples with disease-causing variants in clinically relevant genes. It is proposed that each PT/AA event include the analysis of two samples. This approach minimizes both cost and burden to the PT programs and laboratories. It also meets the twice yearly requirement for external assessment of test performance under CLIA.

### 7. Development and use of RMs

#### a. Importance of RMs for NGS

A RM is a material or substance, one or more of whose property values are sufficiently homogeneous and well established to be used for the calibration of a



Nature Biotechnology: doi:10.1038/nbt.2403

measuring system, the assessment of a measurement procedure, or for assigning values to materials[66]. Reference Materials can be considered a generic term that includes a number of different types of materials including certified or standard reference materials, QC materials and calibrators[67]. RMs are used by clinical laboratories for a variety of purposes including test development, test validation, QC, and PT or AA. Use of these materials is important for quality management of the analytical phase of the testing process and is required by regulation and recommended by professional organizations to help clinical laboratories develop and maintain well designed, accurate and reproducible assays[6,33,59,68-73]. RMs used in clinical assays must be appropriate for QC and other procedures designed to establish, monitor, and verify the reliability of the assay. Ideally, RMs should resemble actual patient specimens in order to accurately reflect testing conditions[49]. Laboratories should select a set of RMs containing most or all of the disease-associated sequence variations that the clinical assay is designed to detect. For broad application of NGS, (genome, exome, or large panels), this can be problematic because it is impractical to acquire or develop RMs possessing the full complement of disease-associated sequence variations that might occur over the large expanses of the genome that are targeted. This is not a new problem; obtaining a comprehensive set of control materials possessing the full complement of disease-associated variations can be a challenge for both targeted-variant and Sanger sequencing assays because of the large number of possible test results. To address this shortcoming, the use of RMs containing sequence variations (e.g., SNPS, indels, repeats, structural rearrangements, etc.) identical and/or similar to those for which the clinical assay was designed to detect is recommended. In



Nature Biotechnology: doi:10.1038/nbt.2403

taking this approach, there is recognition that the performance specifications attributed to targeted regions of the genome will be based on surrogate variants and not necessarily the disease-associated sequence variations within that region. RMs for NGS assays can be developed from a variety of materials, each with its own qualities, advantages and disadvantages which are summarized in Table 3. For human genetic testing, gDNA derived from blood or a cell line has proven useful for a broad range of clinical molecular assays. Manufactured or synthetic DNA, such as recombinant plasmids, oligonucleotides, or concatenated PCR products, may also be useful to incorporate into QC procedures with the recognition that these materials do not resemble gDNA and may not function properly in an assay designed to detect variants in patient samples[49,74]. Reference data, or electronic data files, containing sequence that is simulated or based upon actual patient samples can also serve as RMs. The electronic reference data may be used to assist with the validation of the informatics pipeline. Simulated sequence reads may be helpful in defining the performance limits of the informatics pipeline and when used in conjunction with actual patient samples contribute to a robust validation and establishment of performance specifications. These electronic files can be used for QC and PT/AA.

**b. Approaches to develop characterized RMs**

Characterized RMs can be developed from gDNA that is derived either from blood or from cell lines. Many of the publicly available cell lines, including some from the HapMap and 1000 Genome projects[75,76], have already been sequenced in a number of research and clinical laboratories using a variety of methods. These materials can be further characterized to ensure a high degree of confidence in the performance of NGS



Nature Biotechnology: doi:10.1038/nbt.2403

for the intended applications. The consensus sequence (or a list of sequence variants detected as compared to the current defined genome build) of each gDNA RM sample should be determined using as wide of a variety of sequencing technologies and informatics pipelines as possible to mitigate systematic biases introduced by a specific platform or analysis software. In addition, systematic biases for each platform and analysis algorithm should be characterized. It is not possible for the sequence of the entire genome to be determined to a high degree of accuracy at this time, thus regions with a lower degree of accuracy, or for which an accurate consensus sequence cannot be determined, should be identified and annotated. The gDNA RM samples should also be characterized by non-NGS methods such as SNP analysis, Sanger sequencing of specific regions, karyotype, and cytogenetic microarray analysis. Ongoing analysis of the gDNA RMs using updated software or improved NGS techniques should be performed periodically to further refine the sequence.

Synthetic DNA RMs may also be developed. While these materials might not contain human gDNA sequence, they could be designed to model characteristics of the human genome used to detect specific genetic markers. For example, synthetic constructs could contain different types of variants, including SNPs, indels, CNVs, CpG islands and repetitive sequences (e.g., homopolymers, tandem repeats, transposons, segmental duplications), as well as combinations of these variants. These constructs could be made in pairs to simulate a diploidy. Synthetic DNA RMs can be created by cloning gDNA into plasmid vectors, yeast artificial chromosomes (YACs), bacterial artificial chromosomes (BACs) or by creating desired DNA sequences synthetically. Synthetic materials need to be carefully characterized to not only confirm their



Nature Biotechnology: doi:10.1038/nbt.2403

sequence and physical properties, but also to assess performance in using a variety of enrichment and sequencing protocols. While synthetic constructs cannot perfectly model the human genome due to their limited complexity, they could be useful for answering specific questions about variant detection. In addition, they would provide a mechanism for blinded PT, since they could be spiked into samples to test variant detection and could be precisely engineered to assess specific questions related to the analysis when gDNA is not available. Synthetic constructs could be used for routine QC of each NGS run. They can be spiked into a sample library or sequenced in a separate lane to monitor base call error rate. However, care should be taken to establish that spiked-in synthetic oligos do not interfere with the analysis of the patient samples.

Electronic data files derived from biologic samples or from simulated data may be used to identify and monitor artifacts as well as assess accuracy and reliability in the sequence alignment, assembly, and/or variant calling of NGS data. The electronic data can include regions, such as repetitive sequences, that are challenging for the software to analyze. Use of electronic data files in this manner must be compatible with the sequencing platform's output and take into account characteristics such as read lengths and error profiles. As these characteristics change, the electronic files should be modified as well. In general, reference data generated *in silico* will be useful for testing informatics algorithms; however, they will not include the complexity of errors and biases present in genomic sequence data. Therefore, a combination of electronic and biochemical RMs may provide a robust framework for test validation, QC and other procedures such as PT used to establish and verify the reliability of the test system.

## 8. Conclusions



Nature Biotechnology: doi:10.1038/nbt.2403

The complexities of NGS technologies and data analysis make it especially difficult to adapt to accepted laboratory practices and to comply with regulatory, quality management system and other professional standards. This new area of clinical testing lacks uniform practices for quality management essential to ensure the analytical validity of test results. The workgroup raised multiple issues and offered suggestions for addressing challenges relevant to test validation, QC, reference material development, and independent measures of test performance including both PT and AA. Although NGS is rapidly evolving, the workgroup developed recommendations that will be useful to laboratories implementing and using this technology to help assure quality and meet regulatory and professional standards. The workgroup identified unresolved issues where additional data collection and analysis are needed to assure the quality of clinical NGS (Supplementary Table 4).

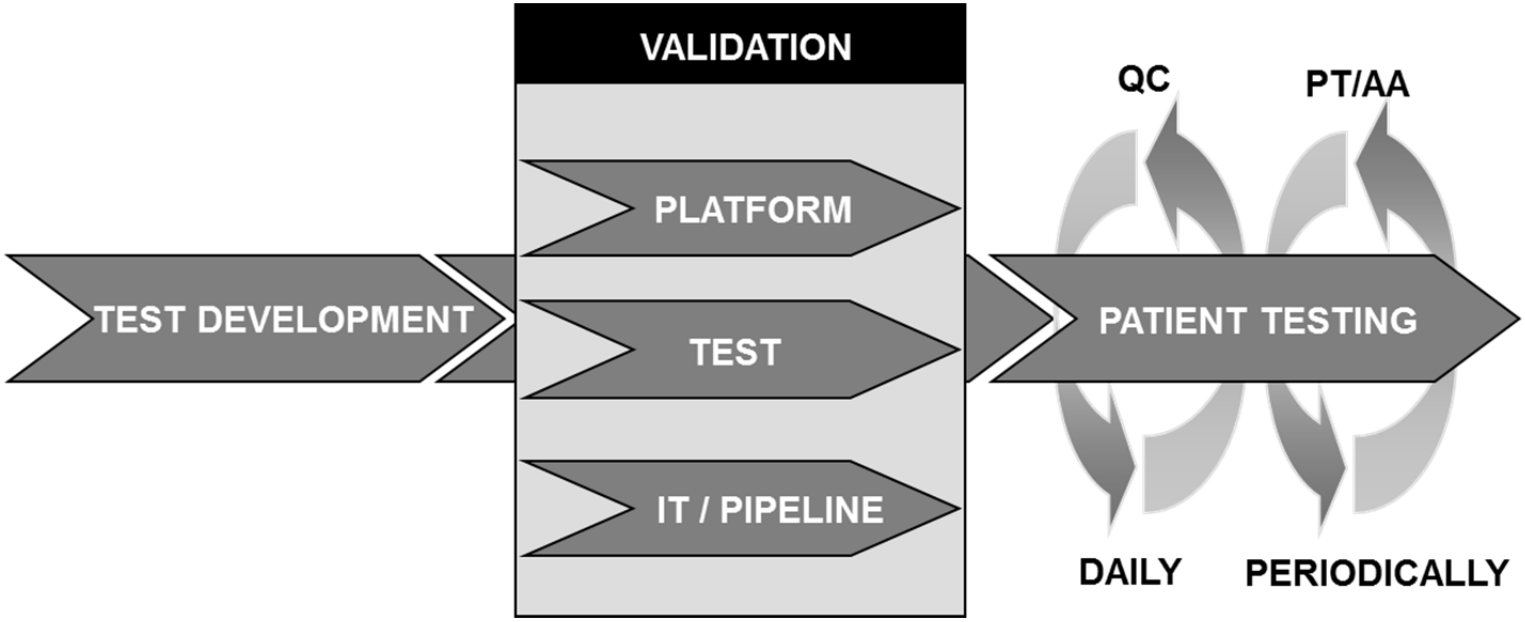
As experience is gained and the technology evolves, the expectation is that these and other practice recommendations will be reassessed. Additionally, clinical decision support systems need to be developed to assist the medical community with the interpretation of NGS results, which will be essential steps towards the realization of personalized genomics and medicine. It is important to maintain collaborations and ongoing discussions among laboratories, clinicians, manufacturers, service providers, software developers, professional organizations, and state and government agencies to ensure quality of DNA sequencing based tests.



Nature Biotechnology: doi:10.1038/nbt.2403

## 9. Figure and Tables

**Supplementary Figure 1. Implementation of clinical NGS testing.** Following test development, NGS assays are validated to establish performance specifications for certain performance characteristics. Ongoing quality control during patient testing ensures that the performance criteria established during validation are achieved. Proficiency testing or alternate assessment compares test performance among laboratories and is an important component of quality management. Abbreviations: QC, quality control; PT, proficiency testing; AA, alternate assessment.






Nature Biotechnology: doi:10.1038/nbt.2403

**Supplementary Table 1. CLIA regulatory standards and workgroup definitions for the validation and ongoing quality control of Sanger sequencing and NGS.**

| Analytical performance characteristics [a] | CLIA requirement [b] and interpretive guideline[c] | Implementation for Sanger sequencing | Implementation for NGS [d] |
|---|---|---|---|
| **Accuracy- "**Closeness of the agreement between the result of a measurement and a true value of the measurand"[39] or analyte. The measurand is "the particular quantity subject to measurement" [33,39]. | **CLIA requirement -**The laboratory is responsible for verifying (as required under §493.1253(b)(1)(i)(A)) or establishing (as required under (§493.1253(b)(2)(i) that the method produces correct results. | **Workgroup definition:** The degree of agreement between the nucleic acid sequences derived from the assay and a reference sequence.<br><br>**Reference material:** A reference sequence can be genomic or synthetic DNA that does or does not contain known sequence variants detectable by the assay. Normal samples and samples with known sequence variation are typically used for test validation and quality control.<br><br>**Considerations for establishing accuracy:**<br><br>**Redundancy (comparable to coverage):** Typically, laboratories will perform sequencing of both strands or independent replication of single strand sequencing.<br><br>**Quality Scores:** Phred scoring has become the de facto standard for reporting the quality of a base call. Peak height and shape are analyzed in | **Workgroup definition:** Same as for Sanger sequencing<br><br>**Reference material:** Same as for Sanger sequencing. Two types of reference materials may be available: 1) those containing a limited number of disease-associated variants present within the genomic regions targeted, and 2) those containing variants that are generally not disease-associated but are the same type of variant (e.g., SNPs, indels, etc.).<br><br>**Considerations for establishing accuracy:**<br><br>**Coverage:** Accuracy for NGS depends on sequence coverage or the number of times a base call is made at a given position within the region sequenced. See section 4.c.i. for detailed discussion of establishment of accuracy.<br><br>**Quality Scores:** Platform-specific algorithms which generate Phred-like quality scores for each base call; quality |



Nature Biotechnology: doi:10.1038/nbt.2403

| | | | |
|---|---|---|---|
| | | conjunction with "lookup" tables to give a base-specific quality score. CLSI document MM-09A recommends a Phred score of 40 or higher, (or a 1/10,000 likelihood of error) as an acceptable quality score. The quality score for a given base can be modified by local sequence context. Quality scores are comparable across Sanger sequencing platforms. | scores cannot be directly compared among platforms. |
| **Precision-** "Closeness of agreement between independent test results obtained under stipulated conditions" [39]. Precision is typically determined by assessing repeatability and reproducibility[33,39]. | **CLIA requirement -** The laboratory is responsible for verifying (as required under §493.1253(b)(1)(i)(B)) or establishing (as required under §493.1253 (b)(2)(ii) the precision of each test system by assessing day-to-day, run-to-run, and within-run variation, as well as operator variance. | **Workgroup definition:** degree to which a repeated measurement gives the same result.<br><br>**Repeatability (within-run precision)-** degree to which the same sequence is derived when sequencing a reference sample multiple times, under the same conditions.<br><br>**Reproducibility (between-run precision)** – degree to which the same sequence is derived when performed by multiple operators using more than one instrument. Samples may be shared with another laboratory for the same testing to determine reproducibility.<br><br>**Considerations for establishing precision:** The availability of reference materials containing sequence variations targeted by the assay may be limited. | **Workgroup definition:** Same as for Sanger sequencing.<br><br>**Repeatability-** Same as for Sanger sequencing.<br><br>**Reproducibility –** Same as for Sanger sequencing.<br><br>**Considerations for establishing precision:** Same as for Sanger sequencing. Also, because of the larger expanse of DNA analyzed, only a limited number of samples can be sequenced and compared. For this reason, other parameters, such as the measurement of the distribution of coverage across the targeted region are useful for establishing repeatability and reproducibility as a different but related parameter. |



Nature Biotechnology: doi:10.1038/nbt.2403

| | | | |
|---|---|---|---|
| **Analytical sensitivity-** "the proportion of biological samples that have a positive test result or known mutation and that are correctly classified as positive" [6,48], or "the ability to detect the lower limit of detection"[49]. | **CLIA requirement -** §493.1253(b)(2)(iii): The laboratory is responsible for determining the lowest concentration or amount of the analyte or substance that can be measured or distinguished from a blank, i.e., minimum detection limits (limit of detection or limits of quantification) or how much of the analyte must be present to be measured. | **Workgroup definition:** The likelihood that the assay will detect the targeted sequence variations, if present. | **Workgroup definition:** Same as for Sanger sequencing.<br><br>**Considerations for establishing analytical sensitivity:**<br>Sensitivity may vary based on coverage, the type of sequence variation and the sequence context. Sensitivity should be assessed at a given coverage threshold across the genomic regions targeted for analysis. |
| **Analytical specificity –** generally defined as the ability of a test to detect only the target analytes and not interfering substances[33].<br><br>"ACMG, CAP and CLSI define analytical specificity as the ability of a test to distinguish target sequences, alleles, or mutations from other sequences of alleles in the specimen or genome being analyzed" [33]. | **CLIA requirement –** §493.1253(b)(2)(iv): The laboratory must determine the extent to which the method measures the analyte for which it is reporting results.<br><br>Interfering Substances- The laboratory must document information regarding interfering substances from product information, literature, or its own | **Workgroup definition:** The probability that the assay will not detect a sequence variation when none are present. The false positive rate is a better measure for sequencing assays. | **Workgroup definition:** Same as for Sanger sequencing.<br><br>**Considerations for establishing analytical specificity:**<br>Specificity may vary based on coverage, the type of sequence variation, and the sequence context. Specificity should be assessed for the full workflow at a given coverage threshold across the genomic regions targeted for analysis. |



Nature Biotechnology: doi:10.1038/nbt.2403

| | | | |
|---|---|---|---|
| | testing. | | |
| **Reportable range-** "the span of test result values over which the laboratory can establish or verify the accuracy of the instrument or test system measurement response"[6]. | **CLIA requirement[6] -** §493.1253(b)(1)(i)(C): The laboratory is responsible for verifying the reportable range of patient test results for each test system. §493.1253(b)(2)(v): The laboratory is responsible for establishing the upper and lower limits of the test system. | **Workgroup definition:** The region of the genome in which sequence of an acceptable quality can be derived by the laboratory test.<br><br>**Considerations for establishing reportable range:** Reportable range can include a region internal to a gene, the coding portion of a gene, or other regions encompassing a limited number of genes. | **Workgroup definition:** Same as for Sanger sequencing. The region sequenced can include large regions with multiple genes, exomes, or the portion of the whole genome for which sequence information can be derived.<br><br>**Considerations for establishing reportable range:** The regions of interest should be defined and areas of difficulty located (e.g. repeat regions, indels, allele drop-outs, etc.). Biases that are introduced by capture-based or enrichment methods should be identified. |
| **Reference range or reference interval (normal values)-** " the range of test values expected for a designated population of persons"[6]. | **CLIA requirement -** §493.1253(b)(2)(vi): The laboratory must establish a reference range that is appropriate for the laboratory's patient population §493.1253(b)(1)(ii): Verify that the manufacturer's reference intervals (normal values) are appropriate for the laboratory's patient population. | **Workgroup definition:** The spectrum of sequence variations that occur in an unaffected population from which the patient specimen is derived. Test results outside this range may be clinically significant. | **Workgroup definition:** Same as for Sanger sequencing. |



Nature Biotechnology: doi:10.1038/nbt.2403

[a] In addition to those listed here, any other performance characteristics required or necessary for test performance should also be established[6].

[b] Standards setting and professional organizations, such as ISO, CAP, ACMG, and CLSI, and regulatory agencies, such as state organizations and CMS and require or recommend that clinical laboratories establish or verify accuracy, precision, analytical sensitivity, analytical specificity, reportable range, reference range, and additional necessary performance characteristics.

[c] Interpretive Guidelines Pertaining to Analytical Performance Characteristics: Centers for Medicare & Medicaid Services. Appendix C: Survey procedures and interpretive guidelines for laboratories and laboratory services. Baltimore, MD: Centers for Medicare & Medicaid Services; 2011. Available at http://www.cms.hhs.gov/clia/03_interpretive_guidelines_for_laboratories.asp.

[d] Determining a sequence using a NGS assay is a multistep process that requires sample preparation, sample analysis, generation of sequence read files, and application of informatics for derivation of the final sequence. Informatics pipelines typically include derivation of base calls, alignment, assembly, and variant calling. Each of these processes requires validation, quality assurance and control procedures.



Nature Biotechnology: doi:10.1038/nbt.2403

**Supplementary Table 2: Metrics needed to evaluate analytical performance of NGS sequencing run: considerations for validation and ongoing quality control.**

| Quality metric | Considerations for validation | Considerations for ongoing quality control |
|---|---|---|
| Depth of coverage | The depth of coverage characteristic of a particular region under standard assay conditions (coverage threshold) should be established.<br><br>It is critical that adequate coverage be defined to achieve adequate sensitivity and specificity in the region(s) of interest. | When the coverage threshold is outside the validated range, that region should be subjected to analysis by an alternate method (e.g., Sanger sequencing) or require additional evaluation. |
| Uniformity of coverage | The coverage across the targeted regions that must be achieved to produce reliable sequencing results should be established. | The uniformity of coverage should be monitored and compared to that established during the validation.<br><br>When the expected coverage uniformity profile is not consistent with the profile established during validation, this may indicate errors in the testing process. |
| GC bias | GC content affects the efficiency of the sequencing reactions and will affect the uniformity of coverage of the targeted region. The amount of GC bias in all parts of the genome included in the assay should be determined during validation. | GC bias should be monitored with every run to detect changes in test performance. |
| Transition/ transversion ratio | The ratio of transitions to transversions (Ti/Tv) should be comparable to published values. | The Ti/Tv ratio should be monitored with every sample to detect a change in test performance.<br><br>When the Ti/Tv ratio is lower or higher than expected, this is an indication that the quality of base calls was low, and potentially contains errors. |
| Base call quality scores | An acceptable raw base call quality score threshold should be established during validation.<br><br>Informatics filters should be established to eliminate any reads | Quality of the signal- to- noise ratio should be monitored by examining the quality scores and quality of signal- to-noise ratio across a read for each run. |



Nature Biotechnology: doi:10.1038/nbt.2403

| | | |
|---|---|---|
| | with raw base calls lower than the established quality score.<br><br>In long-read technologies when detection of larger indels is of interest, alignments can tolerate lower base call quality because the sequence length and accuracy at the base level is less critical. | Quality scores among existing instruments are not readily comparable from one to another. |
| Mapping Quality | Mapping Quality is a measure of the uncertainty that a read is mapped properly to the genomic position.<br><br>During the validation, it should be demonstrated that the test only analyzes the reads that map only to the specific regions targeted in the test.<br><br>Informatics filters should be established to eliminate any reads that map to non-targeted regions and remove duplicate reads | The proportion of reads that do not map to target regions should be monitored during each run. When reads do not match to the reference sequence, this is an indication that the sample is not performing within normal parameters and those reads should be excluded from analysis.<br><br>For applications that involve enrichment steps, poor mapping quality may be a result of non-specific amplification, capture of off target DNA, or contamination. |
| Removal of duplicate reads | Informatics filters should be established to eliminate duplicate reads resulting from clonal amplification (all but one with the highest quality score) during alignment. | This should be monitored to prevent skewing of allelic fractions. |
| First base read success - only applicable for limited platforms | Some platforms allow the early intra-assay evaluation of sequence reads to determine quality scores and the number of reads that pass established quality filters. The number of reads that pass the established quality filters should be established during assay validation. | Evaluation of quality scores and the number of reads that pass the established quality filters early in the sequencing process can be used to monitor for contamination, confirm proper sample loading, and ultimately assess the likelihood of a successful run.<br><br>Some platforms will allow a run to be prematurely terminated if it is not meeting established quality parameters. |
| Decline in signal intensity | During assay validation the expected signal intensity across a read should be evaluated to establish the normal performance ranges and expected decline in signal intensity. Signal intensity across a read length will be platform dependent. | The expected decline in signal intensity should be monitored for each run. A sudden reduction or increase in signal intensity indicates an error in the sequencing chemistry. |



Nature Biotechnology: doi:10.1038/nbt.2403

**Supplementary Table 3: Reference materials for NGS: advantages and disadvantages**

| Type of Material | Advantages | Disadvantages |
|---|---|---|
| Genomic DNA from blood | • Most similar to patient's sample<br>• Will work well in many assays<br>• May have known variant(s) | • Not necessarily renewable<br>• Limited amount of DNA |
| Genomic DNA from cell line | • Renewable<br>• Large supply of DNA<br>• Similar complexity to patient's DNA<br>• Compatible with many assays<br>• May have known variant(s) | • May have rearrangements or loss of DNA<br>• May be heterogeneous due to clonal populations that arise during cell line maintenance<br>• Possible genomic instability over time |
| Synthetic DNA | • Can synthesize a broad range of sequences and variations<br>• Can make sequence templates with complex regions, e.g. deletions and duplications<br>• Can manufacture large amounts of material | • Does not represent human genome<br>• May not perform as human DNA due to differences in sequence complexity<br>• Will not cover all regions of the genome<br>• May exhibit higher variant calls than natural DNA due to errors in synthesis |
| Electronic Reference data files | • Can engineer sequence files with any characteristic<br>• Can be used to assure software performance | • Reference only for data analysis steps (not chemistry)<br>• Must mimic output data from evolving sequencing technologies<br>• Requires many reference data sets to mimic many types of sequence data<br>• Data files may not be interoperable among different platforms |



Nature Biotechnology: doi:10.1038/nbt.2403

**Supplementary Table 4: Areas where additional data collection and analysis are needed to assure acceptable performance specifications for NGS**

| Area | Description |
|---|---|
| Terminology | Regulatory, accrediting and professional organizations should agree on the definition of performance characteristics (accuracy, precision, etc.) as applied to NGS. |
| Precision | Statistically robust and cost-effective procedures need to be identified for establishing the precision of NGS tests. |
| Resolving discordance | A process for determining acceptable discordance among analogous techniques needs to be established. Multiple factors must be considered including the quality of the reference material and sequence as well as the NGS platform and alternate techniques that are used. |
| Reference sequences | A process for establishing a reference sequence applicable to a clinical sample to be used as a reference material. |
| Assessment of test performance | Studies to evaluate approaches to the independent assessment of test performance. Within this manuscript, a combination of PT and alternate assessment is proposed. |
| Daily controls | CLIA regulations require that the laboratory must include daily testing of negative and positive control materials for qualitative procedures, once each day patient specimens are tested. Data should be evaluated to determine whether analysis of a reference sample, containing both disease-associated and naturally occurring variants, can suffice to detect errors and meet the intent of this requirement. |
| Assessment of informatics pipelines | Models for establishing electronic data files useful for assessing the informatics pipeline in one or multiple laboratories need to be developed and evaluated in the clinical laboratory environment. A suitable series may have a common base sequence but contain alterations of increasing complexity designed to test the limits of an informatics pipeline in detecting the targeted sequence variations. Such an electronic file set may be a useful tool for test validation and inter-laboratory PT. |



Nature Biotechnology: doi:10.1038/nbt.2403

## 10. References


1. Bick, D.P. & Dimmock, D.P. Whole exome and whole genome sequencing. *Curr Opin Pediatr.* **23**, 594–600 (2011).

2. Worthey, E.A. *et al*. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.* **13**, 255–262 (2011).

3. Welch, J.S., *et al.* Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA* **305**, 1577–84 (2011).

4. Link, D.C. *et al*. Identification of a Novel *TP53* Cancer Susceptibility Mutation Through Whole-Genome Sequencing of a Patient With Therapy-Related AML. *JAMA* **305**, 1568–1576 (2011).

5. Jones, M.A., *et al*. Targeted polymerase chain reaction-based enrichment and next generation sequencing for diagnostic testing of congenital disorders of glycosylation. *Genet. Med.* **13**, 921–932 (2011).

6. Centers for Medicare and Medicaid Services. US Department of Health and Human Services. Part 493—Laboratory Requirements: Clinical Laboratory Improvement Amendments of 1988. 42 CFR §493.1443-1495.

7. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Nati. Acad. Sci. USA* **74**, 5463–5467 (1977).

8. Swerdlow, H., Wu, S.L., Harke, H. & Dovichi, N.J. Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *J. Chromatogr*. **516**, 61–67 (1990).

9. Hunkapiller, T., Kaiser, R.J., Koop, B.F. & Hood, L. Large-scale and automated DNA sequence determination. *Science* **254**, 59–67 (1991).

10. Tucker, T., Marra, M. & Friedmanm J.M. Massively parallel sequencing: the next big thing in genetic medicine. *Am. J. Hum. Genet.* **85**, 142–154 (2009).

11. Voelkerding, K.V., Dames, S. & Durtschi, J.D. 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.* **12**, 539–551 (2010).

12. Su, Z. *et al.* Next-generation sequencing and its applications in molecular diagnostics. *Expert. Rev. Mol. Diagn.* **11**, 333–343 (2011).




Nature Biotechnology: doi:10.1038/nbt.2403

13. Sboner, A., Mu, X.J., Greenbaum, D, Auerbach, R.K., & Gerstein, M.B. The real cost of sequencing: higher than you think! *Genome Biol.* **12**, doi:10.1186/gb-2011-12-8-125 (2011).

14. Wang, K., Li, M., & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high- throughput sequencing data. *Nucleic Acids Res.* **38**, doi: 10.1093/nar/gkq603 (2010).

15. Yandell, M. *et al*. A probabilistic disease-gene finder for personal genomes. *Genome Res.* **21**, 1529–1542 (2011).

16. Biesecker, L.G. *et al.* The ClinSeq Project: Piloting large-scale genome sequencing for research in genomic medicine. *Genome Res*. **19**, 1665–1674 (2009).

17. Anderson, M.W. & Schrijver, I. Next Generation DNA Sequencing and the Future of Genomic Medicine. *Genes* **1**, 38-69 (2010).

18. Gowrisankar, S. *et al.* Evaluation of Second-Generation Sequencing of 19 Dilated Cardiomyopathy Genes for Clinical Applications. *J. Mol. Diagn.* **12**, 818–827 (2010).

19. Bell, C.J. *et al*. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med.* **3**, doi:10.1126/scitranslmed.3001756 (2011).

20. Zhang, J., Chiodini, R., Badr, A., & Zhang, G. The impact of next-generation sequencing on genomics. *J Genet Genomics.* **38**, 95–109 (2011).

21. Mamanova, L., *et al*. Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7**, 111–118. (2010).

22. Clark, M.J. *et al*. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol.* **29**, 908–914 (2011).

23. Mardis, E.R. Next-generation DNA sequencing methods. Annu. Rev. *Genomics Hum. Genet*. **9**, 387–402.

24. ten Bosch, J.R., & Grody, W.W. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *J. Mol. Diagn.* **10**, 484–492 (2008).

25. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol*. **26**, 1135–1145 (2008).

26. Metzker ML. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).

27. Rothberg, J.M. *et al*. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).

28. Drmanac R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).

29. Flicek, P. & Birney, E. Sense from sequence reads: methods for alignment and assembly. *Nat. Methods* **6**, S6–S12 (2009).

30. Nagarajan, N. & Pop, M. Sequencing and genome assembly using next-generation technologies. *Methods Mol. Biol.* **673**, 1**–**17 (2010).

31. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

32. Reese, M. *et al.* A standard variation file format for human genome sequences. *Genome Biol.* **11**, doi:10.1186/gb-2010-11-8-r88 (2010).

33. Chen, B. *et al.* Good laboratory practices for molecular genetic testing for heritable diseases and conditions. *MMWR* **58**, 1–29 (2009)

34. Ajay, S.S., Parker, S.C.J., Ozel Abaan, H., Fuentes Fajardo, K.V. & Margulies, E.H. Accurate and comprehensive sequencing of personal genomes. *Genome Res.* **9**, 1498–1505 (2011).

35. Bansal. V. A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics* 26, i318–i324 (2010).

36. NCBI Human Genome Resources. http://www.ncbi.nlm.nih.gov/genome/guide/human/, (2012).

37. Schatz, M.C., Delcher, A.L. & Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20**, 1165–1173 (2010).

38. Kent, W.J. BLAT--The BLAST-Like alignment tool. *Genome Res*. **12**, 656–664 (2002).

39. International Organization for Standardization: ISO 15189 medical laboratories – particular requirements for quality and competence. (2007).

40. Pont-Kingdon, G. *et al*. Design and analytical validation of clinical DNA sequencing assays. *Arch. Pathol. Lab. Med.* **136**, 41–46 (2012).

41. McKernan, K.J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).

42. Harismendy, O. *et al*. Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. *Genome Biol*. **12**, R124 (2011).

43. Voelkerding, K.V., Dames, S.A., & Durtschi, J.D. Next-generation sequencing: from basic research to diagnostics. *Clin Chem*. **55**, 641–658 (2009).

44. Rougemont, J. *et al.* Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics* **9**, doi:10.1186/1471-2105-9-431 (2008).

45. Li, H. & Homor, N. A survey of sequence alignment algorithms for next-generation sequencing. *Breif Bioinform.* **11**, 473–483 (2010).

46. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

47. Meyer, M., Stenzel, U. & Hofreiter, M. Parallel tagged sequencing on the 454 platform, *Nature Protocols* **3**, 267–278 (2008).

48. American College of Medical Genetics. ACMG standards and guidelines for clinical genetic laboratories. Bethesda, MD: American College of Medical Genetics; 2008. Available at http://www.acmg.net, (2012).

49. Clinical and Laboratory Standards Institute. Verification and validation of multiplex nucleic acid assays; approved guideline, MM17-A (2008).

50. Shaffer, L.G. *et al.* Microarray analysis for constitutional cytogenetic abnormalities. *Genet. Med.* **9**, 654–662 (2007).

51. Schuster, S.C. *et al*. Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943–947 (2010).

52. Chou, L.S., Liu C.S., Boese, B., Zhang, X. & Mao, R. DNA sequence capture and enrichment by microarray followed by next-generation sequencing for targeted resequencing: neurofibromatosis type 1 gene as a model. *Clin. Chem.* **56**, 62–72 (2010).

53. Dewey, F.E. *et al*. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet*. **7**, doi:10.1371/journal.pgen.1002280 (2011).

54. Durtschi, J.D., Margraf, R.L., Ridge, P.G.& Voelkerding, K.V. Evaluation of Mitochondrial DNA Alignment Data in Illumina HiSeq Whole Genome Sequencing for Quality Control Monitoring. 61st Annual Meeting of The American Society of Human Genetics. Abstract 855T, (2011).

55. Li, M. *et al*. Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am. J. Hum. Genet.* **87**, 237–249 (2010).

56. Lam, H.Y.K. *et al.* Performance comparison of whole-genome sequencing platforms. *Nature Biotechnol*. **30**, 78–82 (2012).

57. Clinical and Laboratory Standards Institute. Assessment of laboratory tests when proficiency testing is not available; approved guideline, GP29-A2 (2008).

58. Clinical and Laboratory Standards Institute. Using Proficiency Testing to Improve the Clinical Laboratory; approved guideline, GP27-A2 (2007).

59. College of American Pathologists - CAP Home, www.CAP.org, (2012).

60. United Kingdom National External Quality Assessment Service http://www.ukneqas.org.uk/content/PageServer.asp?S=99359106&C=1252&CID=1&type=G, (2010).

61. CF Network. http://cf.eqascheme.org/, (2011).

62. Clinical and Laboratory Standards Institute. Proficiency testing (external quality assessment) for molecular methods; approved guideline, MM14-A (2007).

63. Cock P.J.A. *et al*. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771 (2010).

64. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

65. International Organization for Standardization: ISO/IEC. FDIS 17043 - Conformity assessment – General requirements for proficiency testing. (2008).

66. International Organization for Standardization: ISO 15195:2003 - Laboratory medicine – Requirements for reference measurement laboratories. (2003).

67. Emons, H., Fajgelj, A., van der Veen, A.M.H., & Watters, R. New definitions on reference materials. *Accred. Qual. Assur.* **10,** 576–578 (2006).

68. Association for Molecular Pathology statement: recommendations for in-house development and operation of molecular diagnostic tests. *Am. J. Clin. Pathol.* **111**, 449–463 (1999).

69. Chen. B. *et al.* Developing a sustainable process to provide quality control materials for genetic testing. *Genet. Med.* **7**, 534–549 (2005).



Nature Biotechnology: doi:10.1038/nbt.2403

70. Clinical and Laboratory Standards Institute. Molecular Diagnostic Methods for Genetic Diseases; approved guideline, MM01–A2 (2006).

71. CLIA, Washington State Legislature, http://www.doh.wa.gov/hsqa/fsl/lqa_home.htm, accessed March 23, 2012.

72. New York State Clinical Laboratory Evaluation Program, http://www.wadsworth.org/labcert/clep/clep.html, accessed March 23, 2012.

73. American College of Medical Genetics, Standards and Guidelines for Clinical Genetics Laboratories 2006 Edition, http://www.acmg.net/Pages/ACMG_Activities/stds-2002/g.htm, accessed March 23, 2012.

74. Strom, C.M. *et al.* Technical validation of a multiplex platform to detect thirty mutations in eight genetic diseases prevalent in individuals of Ashkenazi Jewish descent. Genet. Med. 7, 633–639 (2005).

75. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

76. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).



Nature Biotechnology: doi:10.1038/nbt.2403