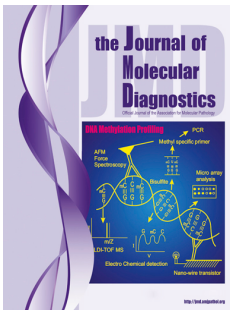
Describing the reportable range is important for reliable treatment decisions a multiple laboratory study for molecular tumor profiling using next-generation sequencing

Accepted Manuscript



Describing the Reportable Range Is Important for Reliable Treatment Decisions: A Multi-Laboratory Study for Molecular Tumor Profiling Using Next-Generation Sequencing

Véronique Tack, Lien Spans, Ed Schuuring, Cleo Keppens, Karen Zwaenepoel, Patrick Pauwels, Jeroen Van Houdt, Elisabeth M.C. Dequeker

PII: S1525-1578(17)30624-4

DOI: 10.1016/j.jmoldx.2018.06.006

Reference: JMDI 717

To appear in: *The Journal of Molecular Diagnostics*

Received Date: 22 December 2017

Revised Date: 14 May 2018

Accepted Date: 5 June 2018

Please cite this article as: Tack V, Spans L, Schuuring E, Keppens C, Zwaenepoel K, Pauwels P, Van Houdt J, Dequeker EMC, Describing the Reportable Range Is Important for Reliable Treatment Decisions: A Multi-Laboratory Study for Molecular Tumor Profiling Using Next-Generation Sequencing, *The Journal of Molecular Diagnostics* (2018), doi: 10.1016/j.jmoldx.2018.06.006.

Describing the reportable range is important for reliable treatment decisions: a multi-laboratory study for molecular tumor profiling using next-generation sequencing

Véronique Tack,* Lien Spans,† Ed Schuuring,‡ Cleo Keppens,* Karen Zwaenepoel,§ Patrick Pauwels,¶ Jeroen Van Houdt,‖ and Elisabeth M.C. Dequeker*,**

From the Biomedical Quality Assurance Research Unit,* Department of Public Health and Primary Care, and the Center of Human Genetics,† University of Leuven, Leuven, Belgium; the Department of Pathology,‡ University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; the Department of Pathology,§ University Hospital Antwerp, Edegem, Belgium; the Center for Oncologic Research (CORE),¶ University of Antwerp, Antwerp, Belgium; the Genomics Core,‖ and the Department of Medical Diagnostics,** University Hospital Leuven, Leuven, Belgium

Correspondence to: Prof. Dr. Elisabeth M.C. Dequeker

Department of Public Health and Primary Care, Biomedical Quality Assurance Research Unit, KU Leuven, Kapucijnenvoer 35 blok d, 3000 Leuven, Belgium

Email: els.dequeker@kuleuven.be.

Running title: Reportable range for careful therapy.

Disclosures: V.T. has received speaker fees from Qiagen. E.S. performed lectures for Illumina, Novartis, Pfizer, BioCartis; is consultant in advisory boards for AstraZeneca, Pfizer, Novartis, BioCartis; and received financial support from Roche, Biocartis, BMS, and Pfizer (all fees to the Institution). Horizon Discovery and Thermo Fisher Scientific provided DNA

control material. Illumina, Multiplicom, Roche, Thermo Fisher Scientific, and Qiagen provided reagents to the participating laboratories.

Abstract

As interpretation of next-generation sequencing (NGS) data remains challenging, optimization of the NGS process is needed to obtain correct sequencing results. Therefore, extensive validation and continuous monitoring of the quality is essential. NGS performance was compared to traditional detection methods and technical quality of nine NGS technologies was assessed. First, nine formalin-fixed, paraffin-embedded patient samples were analyzed by 114 laboratories using different detection methods. No significant differences in performance were observed between analyses with NGS and traditional techniques. Second, two DNA control samples were analyzed for a selected number of variants by 26 participants using nine different NGS technologies. Quality control metrics were analyzed from raw data files and a survey regarding routine procedures. Results showed large differences in coverages, but observed variant allele frequencies in raw data files were in line with pre-defined variant allele frequencies. Many false negative results were found due to low-quality regions, which were not reported as such. It is recommended to disclose the reportable range, the fraction of targeted genomic regions for which calls of acceptable quality can be generated, to avoid any errors in therapy decisions. NGS can be a reliable technique, only if essential quality control during analysis is applied and reported.

Introduction

In the past, to test a single gene, several PCR-based platforms or Sanger sequencing techniques were used for testing predictive and prognostic markers in cancer [1, 2]. Currently, the testing of several genes and multiple variant hotspots has become common practice in cancer treatment decision making [3]. The recruitment of clinical trial patients is increasingly based on confirmed variants and the knowledge of the molecular tumor spectrum [4]. However, it is difficult to respect the turnaround time to test multiple genes sequentially using Sanger sequencing. Moreover, the limited amount of available tumor tissue makes sequential analyses almost impossible. Thus, there is an increasing demand for methods that allow molecular testing of numerous variants simultaneously with low DNA input. Next-generation sequencing (NGS) can fulfil these requirements and is finding its way as primary technique for biomarker testing in tumor tissues in many laboratories [4-6].

Different NGS technologies are available for performing whole-genome, whole-exome, or targeted sequencing analysis. Today, the latter is the preferred option for oncology biomarker testing. Whole-genome or whole-exome sequencing are too expensive for routine practice as the sequence depth should be high enough for analysis of tumor tissue and there is a limited clinical actionability of most regions of the human genome [7]. In addition, formalin-fixed, paraffin-embedded (FFPE) tissue that contains fragmented DNA is not yet optimized to be used for whole-genome sequencing [8]. The NGS library preparation can be either PCR-based or capture-based and can be combined with different sequencing platforms [8-10].

Sequencing costs per sample are decreasing for NGS (http://www.genome.gov/sequencingcosts, last accessed March 23, 2017). Although targeted sequencing was first performed for a few thousand dollars, this is now available for a few hundred dollars per sample [11, 12]. The decreasing costs, the low turnaround time, the use of

FFPE-material and the broad coverage of clinically relevant genes will further encourage the use of targeted NGS in routine practice of molecular pathology laboratories.

The implementation of NGS also knows some limitations. It remains a challenge to handle the limited amount of DNA and the quality of DNA extracted from FFPE tissue in molecular pathology [13]. In addition, the interpretation of the results using bioinformatics becomes more complex. Many different quality control (QC) metrics can be applied to filter the huge amount of data and determine the correct variants in the sample. Reporting the correct genotype of a tumor is especially important in decisions for targeted therapy. For example, confirmation of activating *EGFR* variants in non–small-cell lung cancers is essential before therapy with EGFR tyrosine kinase inhibitors [14, 15].

Recommendations for using NGS in clinical practice describe various quality parameters that need to be taken into account, however, in most cases, no exact criteria for variant calling are given (www.wadsworth.org/sites/default/files/WebDoc/Updated%20NextGen%20Seq%20ONCO_Guidelines_032016.pdf, last accessed March 31, 2017) [16-18]. For instance, coverage, one of the most relevant technical variables in NGS, can help in troubleshooting errors and optimizing the laboratory's NGS workflow [10]. A tool to assure correct diagnostic results is participation in external quality assessment (EQA) schemes [19]. With participation in EQA, it can be verified whether the test method used and QC metrics used are valid and accurate, and reliable results are obtained.

This study assessed the performance of NGS compared to other variant-detection assays in an EQA scheme for somatic variant analysis in the *EGFR* gene. On the other hand, a multigene analysis of DNA control material was performed by 26 laboratories using nine different NGS

technologies. The technical quality within and between the technologies was compared by re-analyzing the raw data files and the performance of the laboratories was evaluated.

Materials and Methods

The study consisted of two parts. The first part was the analysis of patient FFPE samples for *EGFR* variants. The participants in this study were informed that the samples were part of the 2015 *EGFR* EQA scheme of the European Society of Pathology (ESP) (lung.eqascheme.org, accessed December 18, 2017; registration required for full laboratory participation.). This *EGFR* EQA scheme was open to all laboratories worldwide. The scheme was organized according to international guidelines [20]. To determine the performance of a laboratory, scores were assigned by two independent assessors by comparison of the participant's results with the validated results. Two points could be obtained for each correctly genotyped sample for a total of 18 points and points were deducted in case of a genotyping, technical, clerical, or nomenclature error. Genotyping errors were defined as false positive results, false negative results, or reporting of a wrong variant. A participant was seen as successful if at most one technical error and/or a nomenclature mistake, or no more than one genotyping error without nomenclature mistake was made. Nine FFPE patient samples were evaluated in the scheme (Table 1). For three of the samples, mock clinical information was provided by the organization, and participants needed to submit written diagnostic reports with molecular results and interpretation of these cases.

For the selection of the participants of the second part of this study, an invitation was sent to more than 600 institutes from the ESP database. A candidature form was filed by 98 laboratories and the final selection of 26 European laboratories was based on the used NGS methodology (panel and platform), accreditation status, NGS implementation date and number of samples tested per year to obtain a group with enough variation on these aspects (Table 2).

The DNA control material originated from two different manufacturers: sample A was the Quantitative Multiplex DNA reference standard from Horizon Discovery (Cambridge, UK) and sample B the Oncology Hotspot Control from Thermo Fisher (CA) (Table 1). For this study, 20 genes were selected for which the participants needed to report results: *AKT1, ALK, BRAF, CTNNB1, EGFR, ERBB2, ERBB4, FGFR2, FGFR3, KIT, KRAS, MAP2K1, MET, NRAS, PDGFRA, PIK3CA, PTEN, SMAD4, STK11,* and *TP53*. This list was selected based on the overlap between the targeted regions covered by the different NGS methodologies, the clinical relevance of markers and current availability of targeted therapies for biomarkers in these genes. The relevance (pathogenic versus benign) and position (intronic, exonic, splice site) of the variants in the selected genes was determined by the Biomedical Quality Assurance (BQA) unit (KU Leuven) with the Ingenuity Variant Analysis (IVA) software (build 4.3.20170418) from Qiagen (Valencia, CA).

The participants analyzed the DNA control material with their routine NGS methodology and were requested to report any variant above a variant allele frequency (VAF) of 1% in the 20 selected genes, regardless of the clinical relevance. A list with the identified variants, as well as the raw data files (BAM files) from the NGS analysis were submitted to the organisers of the study. In addition, questions were asked about the used QC metrics and the validation procedure of the participant's NGS technique. Not all questions in the survey were mandatory and more than one option could be chosen for several questions, hence why the sum of percentages is not equal to 100%. After the participants submitted their results, all further analyses, discussed below, were performed at the BQA Research Unit and the Center of Human Genetics of the KU Leuven. The limit of detection of the participant's technique was taken into account for the determination of correct or incorrect results.

All reported variants in the DNA samples were cross-checked with the variant list provided by the manufacturer to identify possible false positive or false negative results. The BED files with target definitions were provided by the companies of the NGS panels.

Technical information, such as total coverage and alternative allele coverage was calculated from the provided BAM files using bam-readcount version 0.8 software (https://github.com/genome/bam-readcount, last accessed May 2, 2017). A minimum mapping quality and base quality of 15 was applied. BED tools version 2.25.0 was used to analyze the observed coverages of participants [21, 22]. The observed VAFs were calculated from this dataset.

Statistical analysis was done with IBM SPSS (New York, NY) statistics version 22. Comparison of groups was done with a Chi squared test or with Fisher's exact test in case the expected frequency was less than 5 in more than 20%.

Results

The results of the analysis of nine FFPE samples from the 2015 ESP EQA scheme for *EGFR* variant analysis were evaluated. In this scheme, 33 of 114 participants (28.9%) used their routine NGS technique for *EGFR* variant analysis.

The average genotyping scores for laboratories using NGS (N = 33, NGS laboratories) and laboratories using another technique (N = 81, non-NGS laboratories) were 90.0% (16/18) and 87% (15.7/18), respectively. Only 79% (26/33) of the NGS laboratories and 64% (52/81) of the non-NGS laboratories made no genotyping errors (false positive results, false negative results, or reporting a wrong variant). More details on the genotyping errors can be found in Table 3. In addition, NGS laboratories tended to obtain the maximal score (18/18) more than non-NGS laboratories. However, the number of laboratories with a successful score was similar between both groups (Table 3). At samples level, the differences between the two

groups were smaller, but still more genotyping errors occurred when analyzing the samples with a non-NGS technology (6%) versus an NGS technology (4%). Detailed results on sample level showed more false negative results compared to false positive results or reporting a wrong variant (Table 3). Of the 33 NGS-laboratories, 12 used a laboratory-developed NGS panel. The average score of these 12 laboratories was 89% and four (33%) laboratories made a genotyping error. This lower average score was due to one laboratory that made four genotyping errors. More detailed information on the type of errors of the NGS laboratories versus the non-NGS laboratories can be found in Table 3. No significant differences could be observed between the two groups regarding the number of laboratories with major genotyping errors, technical errors or a maximum performance score (18/18) in the EQA scheme (P = 0.129, P = 0.447, and P = 0.193, respectively) [23].

In addition to the genotypes for each sample, laboratories also submitted diagnostic reports. Only three of the 33 NGS laboratories (9.1%) reported which exons were not informative enough for a reliable conclusion and one NGS laboratory reported that some exons were of sub-optimal quality, but did not state the exact exons. None of these laboratories had a false positive or a false negative result in the analysis of the patient samples.

Twenty-six institutes participated in the second part of the study, of which two institutes used two different NGS techniques (Table 4). Not all laboratories submitted the requested raw data files or the survey on QC metrics (Table 4). More than half of the participants were accredited (58%) according to a national or international standard. Half of them tested a limited number of samples with NGS in 2014 (0-249). Approximately a quarter tested 250 to 499 samples in 2014 and another quarter tested more than 500 samples. Most laboratories (58%) worked both in a clinical trial setting and in a diagnostic setting (Table 2). All institutes used amplicon-based panels in combination with sequencing-by-synthesis (Illumina – United States, MiSeq),

pyrosequencing (Roche – Switzerland, GS Junior), or Ion Torrent semi-conductor technology (Thermo Fisher, PGM or Ion Proton).

To compare the technical performance of nine different NGS protocols, the raw data files (BAM files) were re-analyzed at a central laboratory. Hereby, two important QC metrics, the coverage and VAF of variants in the different panels, were studied in detail. Only the NGS panels used by at least two participants were taken into account.

Figure 1A-D give an overview of the average coverage of the amplicon panels for each technology and participant. To make sure that reads were only counted once, the middle coordinate of each amplicon was used. As the TruSight Tumor panel (Illumina) uses two oligo pools for multiplex PCR, two pools were available for each participant. Participant 10 and Participant 11 performed a replicate analysis, hence the two results. Figure 1E-H give an overview of the uniformity of the coverages for each panel. For each amplicon of each sample, the expected number of reads and percentage of reads per amplicon were determined and the deviations from this number for each amplicon were used as a measure for the uniformity of the coverage. Participant 19 showed low coverage compared to the other participants. This laboratory used the Ion AmpliSeq Colon and Lung Cancer Panel (Thermo Fisher Scientific). Participant 28 used the TruSight tumor panel from Illumina and showed very low and uneven coverage. Participant 3 had all amplicons covered, although some at very low depth, explaining the high standard deviation. These analyses revealed a high variation in average coverage, both between the different technologies and within a specific technology (Figure 1).

Figure 2 visualizes the deviation of the observed VAF from the participant's panel from the expected VAF, as determined by the manufacturer of sample A. Ideally, the graph should be low and horizontal for a minimal deviation between the observed and the expected VAF. Due to the uneven coverage of Participant 28, and the very low coverage of Participant 19, these

laboratories were left out of the analysis. There was only one outlier left, due to the reporting of a variant at 28% instead of the expected 13%. The corresponding laboratory used the TruSeq Amplicon Cancer panel (Illumina). Except for these participants, the variation of the observed VAF was limited (< 3%), both within and between the different technologies. For most variants, the average observed VAF of the different participants is in accordance with the expected VAF per methodology. The Ion AmpliSeq Colon and Lung Cancer panel from Thermo Fisher has 5 of the 12 targeted variants with an average that differs more than 1% from the expected VAF, as determined by the manufacturer. None of these five show large systematic deviations in the other methodologies. For the other methodologies, one or two variants (≤ 15%) show an average VAF that differs more than 1% from the expected VAF. For sample B, only ranges of expected VAFs were provided by the manufacturer of the sample and no standard deviations could be calculated.

Subsequently, the BAM files of the two samples generated by the different technologies were analyzed to study the variants present in the aligned reads. Figure 3 gives an overview of the variants that should have been detected by the specific methodology, categorized according to the coverage limit and the minimal VAF of the specific participant. Mostly, the percentage of variants below the QC criteria of the participant varied between 10% and 30%. Some participants showed a higher percentage of variants below their QC criteria, however this was not in accordance with the other participants of that specific methodology. For example, Participant 19 showed no variants above the participant's QC criteria, which was not in line with the other participants using the Ion AmpliSeq Colon and Lung Cancer panel from Thermo Fisher Scientific. The TruSeq Amplicon Cancer panel from Illumina showed four variants without reads in all three participants. The other variants without coverage were sequenced and present in the BAM files by at least one other participant.

Following the technical analysis of the BAM files of the participants, the submitted list of the identified variants in both samples was analyzed in detail (Figure 4). For each participant, the percentage of reported and unreported variants was calculated. These variants were subsequently divided in different categories, based on the information in the BAM files of the participant. The limit of detection and the minimal coverage, as reported by the laboratory in the QC survey, were taken into account to assign these categories. Variants in the raw data files that met the QC criteria of the participant and were also reported by the participant in the variant list, or variants below the QC criteria and reported as such are seen as correctly reported variants. In case the coverage or limit of detection was below the participant's limit, this was considered as an incorrectly reported variant. Variants that did not fulfil the QC requirements should not be reported. For example, Participant 19 reported a high number of variants that did not meet QC criteria. Most participants obtained good results and reported more than 80% of the variants in a correct way.

In addition, for sample A, 14 laboratories had one or more false positive results. In total, 37 possible false positive variants were identified. The manufacturer of this sample could not confirm whether these variants were true variants or false positive variants. On average, their limit of detection for exome sequencing was 10% and the possible false positives were reported by the participants at a VAF lower than or around this limit. However, the observation that only three of the 37 variants were reported by two different laboratories and 31 by one laboratory argues that these variants are true false positives. For sample B, 11 laboratories had one or more false positive results and 36 false positive results were identified in total. Five variants were reported by two or more different laboratories. None of the false positive results were taken into account to assess the performance of the laboratories.

Since not all laboratories provided their BAM files, or the laboratory's NGS bioinformatics pipeline does not generate BAM files, some participants could only be evaluated based on the

reported list of identified variants (Figure 5). The GS Junior system from Roche generates sff files, which could not be further analyzed for their QC metrics [24]. For all these participants, only the limit of detection was taken into account to assign the categories and no conclusions were made regarding the quality of the reads (Figure 5). For sample A, most participants reported the correct variants, within their limit of detection. Only Participant 4b missed three of the 12 targeted variants. In sample B, more false negatives were present, but it was not clear whether it was due to low quality of the reads or not.

To collect information on the number of QC metrics in routine practice and the type of validation procedures, a survey was sent to all participants. For validation of their NGS methodology, 59% (13/21 laboratories) used reference material from Horizon Diagnostics and 19% (4/21) used reference material from Thermo Fisher Scientific. Other reference material was FFPE material (76%), cytology material (10%), blood (10%), clinical study samples (5%), and fresh frozen samples (5%). The laboratories used on average five QC metrics during the wet bench process of the NGS analysis and two during the bioinformatics phase for each analysis. In the sample preparation phase (DNA extraction, library preparation, and enrichment of the sample) of the NGS analysis, DNA concentration is an important QC metric for many laboratories. This metric is controlled in the different steps of sample preparation. Other QC metrics used by more than one laboratory were amplicon size, fragment analysis, or amplification of the PCR product. After extraction of the DNA, 70% of the participants monitored the DNA concentration. After sequencing, Phred score (37.0%), VAF (26%), and the coverage (26%) are the main QC metrics. The threshold for the VAF varies between 1% and 10%. The minimal coverage was calculated by different measures, per base, per exon or per sample, and shows high variation among the different participants.

Discussion

Many different workflows are available for NGS analysis. Optimization of various NGS processes is needed to obtain high-quality sequencing results [25, 26]. Interpretation of the results remains challenging and a lot of data are generated in the process [9]. To ensure that correct results are obtained at all times, extensive validation and continuous monitoring of the quality is essential [8, 27]. This study compared the performance of NGS technologies versus traditional detection methods and assessed the technical quality of nine routinely-used NGS technologies for variant detection in DNA control material.

The inclusion of the results of the participants of the ESP *EGFR* EQA scheme allows the comparison of NGS users versus non-NGS users (Table 3) (lung.eqascheme.org, last accessed December 18, 2017). The participating laboratories had different accreditation statuses or experience with NGS technology (Table 2). Although more NGS laboratories made technical errors, the percentage of technical errors for samples analyzed with NGS is similar to those analyzed with other techniques (Table 3). The majority of the laboratories using NGS had one technical error whereas more laboratories using other techniques had multiple technical errors. Although the differences between the NGS laboratories and non-NGS laboratories were small and not significant, there is a trend towards a better performance of NGS.

In the second part of the study, technical features of the different NGS technologies and the performance of the laboratories were analyzed. Based on the VAF of the variants in the DNA control sample A, the estimation of the VAF gave good results; the observed VAF from the raw data files corresponded well with the expected VAF. The discordant estimates of two participants were mainly due to uneven and/or low coverages. Therefore, to be able to detect all variants, including those with frequencies below 5%, the minimal coverages for different VAF should be determined during the validation process (www.wadsworth.org/sites/default/files/WebDoc/Updated%20NextGen%20Seq%20ONCO_Guidelin

es_032016.pdf, last accessed March 31, 2017) [25]. Once introduced in clinical context, this should be taken into account for the molecular interpretation and the diagnostic report.

The results of the DNA control material analyses (sample A and B) showed large differences between the NGS methodologies, especially in sample B. The percentage of correctly reported variants ranges between 58% and 100% for sample A and between 55% and 97% for sample B. There was no clear tendency that variants of a specific type (pathogenic, benign, uncertain, or unknown) were less reported by the participants than others. Reflecting this to routine practice, relevant pathogenic variants would also be missed, which could have a possible harmful effect on therapy decisions. Based on the analysis of the raw data files, the possible cause of false negative results showed similar trends within one technique. The analysis of sample B with the TruSight tumor panel from Illumina showed many results for which the observed VAF is below the expected VAF range and, consequently, lower than the corresponding limit of detection for that participant. Of all variants that should have been identified by this panel (N = 211), 26% had a coverage or a VAF that was lower than the limit for all four participants with this panel. For this fraction of the variants, the recipient of the report, including clinicians and laboratory personnel, would not have any idea that the results were of lower quality and thus not trustworthy. Continuous education may be needed for molecular biologists to include the reportable range and to inform clinicians on the importance of the reportable range, such that this information can be used correctly and efficiently.

This study emphasises the importance of providing information about the reportable range when describing NGS results in a tumor sample. The reportable range can be defined as the fraction of the targeted genomic regions for which calls of an acceptable quality can be generated [16, 28]. The importance of evaluating the reportable range of the NGS technique during the validation process was stressed by the Clinical Laboratory Improvement

Amendments (CLIA) from the United States in 2012 and was also emphasised in different recent guidelines [8, 10, 17]. There is a large influence on patient safety when reportable ranges are not added in the results. Figure 4 shows a large proportion of false negatives (not reported variants) of which the quality of the regions was not sufficient to be withheld during variant filtering (coverage or VAF below the participant's limit). This is especially the case for laboratories using the TruSight Tumor Panel from Illumina, which tend to produce more such regions than the other panels in this comparison. By indicating those regions in the report the clinician can take this into account for further interpretation of the results as some regions may be less covered or have a lower quality score than others and this can differ for each analysis. During the 2015 ESP *EGFR* EQA scheme, only three participants stated information on the non-informative exons or gave information about the coverage per exon and the minimal required coverage (one participant). Two of these laboratories were accredited and two of them had a university and research background. Despite the recommendations in guidelines and studies, mentioning the reportable range in the diagnostic report is not yet routine practice for laboratories. Evaluation of the reports in following years of EQA organisation by the ESP (2016 ESP *EGFR* EQA, 2016 Colon EQA, and 2017 Colon EQA) showed no large improvements (lung.eqascheme.org and kras.eqascheme.org, last accessed December 18, 2017; registration required for full laboratory participation.). Only one additional laboratory stated that the quality criteria were not met for a certain exon. It should be recommended to laboratories to follow the available guidelines to avoid any errors with an effect on therapy decisions. As different publications provide information about the reportable range, it should be verified in future EQA schemes if more laboratories adhere to the guidelines [8, 10].

The need to monitor the reportable range could also be part of the continuous validation [10]. In case a clinically important variant shows insufficient coverage on a regular basis, there is a need for revalidation of the technique.

Some false positive results were also reported. Confirmation of the variants by a secondary technique could be a solution to avoid such errors in the future [17, 29]. In routine practice, nine of the participating institutes (35%) used a secondary technique to confirm a positive result in diagnostic routine settings. These laboratories from six European countries had different degrees of experience in the number of samples tested per year and six were accredited. Three still reported a false positive result in this study. The practice for confirmatory testing used to be encouraged by guidelines and recommendations, but as NGS technology evolves it is no longer recommended for all cases [6, 10, 30]. Confirmatory testing should be applied in case of doubt, for example in variants with low frequencies or other unexpected results or in case a variant was identified outside the validated regions [6, 10, 30].

Another strategy to avoid false positive and false negative results is to include reference standards in the NGS workflow [27]. It was not requested in this study if reference standards were used in routine practice. However, two thirds of the participants in the second part of this study included reference material in their initial validation, before implementation of the technique, which is already a good step towards improvement of the quality [10].

Conclusion

NGS performs well in clinical practice compared to traditional diagnostic testing methods. There is a trend towards a better performance of NGS, with less genotyping errors. Although the number of tested hotspots/genes analyzed in the same run increased significantly, there were still genotyping errors with the NGS technique, so it remains important to focus on the QC metrics of each analysis.

This study shows the importance of describing the reportable range in the report. The regions that cover clinical relevant variants for which no reliable calls were obtained should be reported clearly to the clinician who makes treatment decisions. At first sight, the NGS methodologies showed a moderate performance in identifying variants in DNA control material, but when the reportable range was taken into account in the evaluation, the performance increased to a higher level. Technical criteria must be available to determine a successful analysis and, where necessary, additional education should be provided to clinicians and NGS users to correctly use this information in their treatment decision making.

Acknowledgements

We thank Prof. Patrick Pauwels, Dr. Karen Zwaenepoel, Prof. Ed Schuuring, Dr. Nils ‘t Hart, Dr. Roberto Salgado, Dr. Javier Hernandez Losa, and Prof. Antonio Marchetti for collection, preparation, sharing, and validation of the ESP *EGFR* EQA samples; Lien Tembuyser for carefully reading the manuscript; the participating laboratories for analysis of the samples; Horizon Discovery and Thermo Fisher Scientific for providing the DNA control material; and Illumina, Multiplicom, Roche, Thermo Fisher Scientific, and Qiagen for providing reagents to the participating laboratories.

References

[1] McCourt CM, McArt DG, Mills K, Catherwood MA, Maxwell P, Waugh DJ, Hamilton P, O'Sullivan JM, Salto-Tellez M: Validation of Next Generation Sequencing Technologies in Comparison to Current Diagnostic Gold Standards for BRAF, EGFR and KRAS Mutational Analysis. PloS one 2013, 8:e69604.

[2] Martinez DA, Nelson MA: The Next Generation Becomes the Now Generation. PLOS Genetics 2010, 6:e1000906.

[3] Cagle PT, Raparia K, Portier BP: Emerging Biomarkers in Personalized Therapy of Lung Cancer. Advances in experimental medicine and biology 2016, 890:25-36.

[4] Renfro LA, An MW, Mandrekar SJ: Precision oncology: A new era of cancer clinical trials. Cancer letters 2016.

[5] Liu J, Hu J, Cheng L, Ren W, Yang M, Liu B, Xie L, Qian X: Biomarkers predicting resistance to epidermal growth factor receptor-targeted therapy in metastatic colorectal cancer with wild-type KRAS. OncoTargets and therapy 2016, 9:557-565.

[6] Mu W, Lu HM, Chen J, Li S, Elliott AM: Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. The Journal of molecular diagnostics : JMD 2016, 18:923-932.

[7] Hagemann IS, Cottrell CE, Lockwood CM: Design of targeted, capture-based, next generation sequencing tests for precision cancer therapy. Cancer Genetics 2013, 206:420-431.

[8] Deans ZC, Costa JL, Cree I, Dequeker E, Edsjo A, Henderson S, Hummel M, Ligtenberg MJ, Loddo M, Machado JC, Marchetti A, Marquis K, Mason J, Normanno N, Rouleau E, Schuuring E, Snelson KM, Thunnissen E, Tops B, Williams G, van Krieken H, Hall JA: Integration of next-generation sequencing in clinical diagnostic molecular pathology

laboratories for analysis of solid tumours; an expert opinion on behalf of IQN Path ASBL. Virchows Arch 2017, 470:5-20.

[9] Ross JS, Cronin M: Whole Cancer Genome Sequencing by Next-Generation Methods. American Journal of Clinical Pathology 2011, 136:527-539.

[10] Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, Temple-Smolkin RL, Voelkerding KV, Nikiforova MN: Guidelines for Validation of Next-Generation Sequencing-based Oncology Panels. The Journal of Molecular Diagnostics 2017, 19:341-365.

[11] Shao D, Lin Y, Liu J, Wan L, Liu Z, Cheng S, Fei L, Deng R, Wang J, Chen X, Liu L, Gu X, Liang W, He P, Wang J, Ye M, He J: A targeted next-generation sequencing method for identifying clinically relevant mutation profiles in lung adenocarcinoma. Scientific reports 2016, 6:22338.

[12] Gonzalez-Garay ML: The road from next-generation sequencing to personalized medicine. Personalized medicine 2014, 11:523-544.

[13] Betge J, Kerr G, Miersch T, Leible S, Erdmann G, Galata CL, Zhan T, Gaiser T, Post S, Ebert MP, Horisberger K, Boutros M: Amplicon sequencing of colorectal cancer: variant calling in frozen and formalin-fixed samples. PloS one 2015, 10:e0127146.

[14] Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW: Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. The New England journal of medicine 2004, 350:2129-2139.

[15] Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S: EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 2004, 304:1497-1500.

[16] Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, Race V, Sistermans E, Sturm M, Weiss M, Yntema H, Bakker E, Scheffer H, Bauer P: Guidelines for diagnostic next-generation sequencing. European journal of human genetics : EJHG 2016, 24:2-5.

[17] Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA, Agarwala R, Bennett SF, Chen B, Chin EL, Compton JG, Das S, Farkas DH, Ferber MJ, Funke BH, Furtado MR, Ganova-Raeva LM, Geigenmuller U, Gunselman SJ, Hegde MR, Johnson PL, Kasarskis A, Kulkarni S, Lenk T, Liu CS, Manion M, Manolio TA, Mardis ER, Merker JD, Rajeevan MS, Reese MG, Rehm HL, Simen BB, Yeakley JM, Zook JM, Lubin IM: Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol 2012, 30:1033-1036.

[18] Weiss MM, Van der Zwaag B, Jongbloed JD, Vogel MJ, Bruggenwirth HT, Lekanne Deprez RH, Mook O, Ruivenkamp CA, van Slegtenhorst MA, van den Wijngaard A, Waisfisz Q, Nelen MR, van der Stoep N: Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: a national collaborative study of Dutch genome diagnostic laboratories. Human mutation 2013, 34:1313-1321.

[19] Organisation for Economic C-o, Development: OECD guidelines for quality assurance in molecular genetic testing. Paris: OECD, 2007.

[20] van Krieken JH, Normanno N, Blackhall F, Boone E, Botti G, Carneiro F, Celik I, Ciardiello F, Cree IA, Deans ZC, Edsjo A, Groenen PJ, Kamarainen O, Kreipe HH, Ligtenberg MJ, Marchetti A, Murray S, Opdam FJ, Patterson SD, Patton S, Pinto C, Rouleau E, Schuuring E, Sterck S, Taron M, Tejpar S, Timens W, Thunnissen E, van de Ven PM, Siebers AG, Dequeker E: Guideline on the requirements of external quality assessment programs in molecular pathology. Virchows Arch 2013, 462:27-37.

[21] Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) 2010, 26:841-842.

[22] Quinlan AR: BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current protocols in bioinformatics 2014, 47:11.2.1-34.

[23] Keppens C, Tack V, Hart ‘t N, Tembuyser L, Ryska A, Pauwels P, Zwaenepoel K, Schuuring E, Cabillic F, Tornillo L, Warth A, Weichert W, Dequeker E for the EQA assessors expert group: A stitch in time saves nine: external quality assessment rounds demonstrate improved quality of biomarker analysis in lung cancer. Oncotarget 2018, 9:20524-38.

[24] Malde K: Flower: extracting information from pyrosequencing data. Bioinformatics (Oxford, England) 2011, 27:1041-1042.

[25] Aziz N, Zhao Q, Bry L, Driscoll DK, Funke B, Gibson JS, Grody WW, Hegde MR, Hoeltge GA, Leonard DG, Merker JD, Nagarajan R, Palicki LA, Robetorye RS, Schrijver I, Weck KE, Voelkerding KV: College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. Archives of pathology & laboratory medicine 2015, 139:481-493.

[26] Hinrichs JW, van Blokland WT, Moons MJ, Radersma RD, Radersma-van Loon JH, de Voijs CM, Rappel SB, Koudijs MJ, Besselink NJ, Willems SM, de Weger RA: Comparison of next-generation sequencing and mutation-specific platforms in clinical practice. Am J Clin Pathol 2015, 143:573-578.

[27] Froyen G, Broekmans A, Hillen F, Pat K, Achten R, Mebis J, Rummens J-L, Willemse J, Maes B: Validation and Application of a Custom-Designed Targeted Next-Generation Sequencing Panel for the Diagnostic Mutational Profiling of Solid Tumors. PloS one 2016, 11:e0154038.

[28] Hardwick SA, Deveson IW, Mercer TR: Reference standards for next-generation sequencing. Nature reviews Genetics 2017, 18:473-484.

[29] Chang F, Li MM: Clinical application of amplicon-based next-generation sequencing in cancer. Cancer Genetics 2013, 206:413-419.

[30] Strom SP, Lee H, Das K, Vilain E, Nelson SF, Grody WW, Deignan JL: Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. Genet Med 2014, 16:510-515.

Figure Legends

Figure 1: Average coverage per participant (**A-D**) and uniformity of coverage per participant (**E-H**). Sample A: Quantitative Multiplex DNA reference standard from Horizon Discovery. Sample B: Oncology Hotspot Control from Thermo Fisher Scientific. Part: participant. Only methodologies used by at least two participants are shown.

Figure 2: Deviation from expected variant allele frequency (VAF) for sample A (Quantitative Multiplex DNA reference standard from Horizon Discovery). Participants 19 (Ion AmpliSeq) and 28 (TruSight tumor) were left out of the analysis. TruSight tumor: TruSight Tumor panel from Illumina. TruSeq: TruSeq Amplicon Cancer panel from Illumina. Ion AmpliSeq: Ion AmpliSeq Colon, and Lung Cancer panel from Thermo Fisher. Actionable mut: Tumor Actionable mutations panel from Qiagen. The dotted lines represent the change in scale between the lower part and the upper part of each panel.

Figure 3: Quality of variants that should have been detected by the specific methodology. **A:** Sample A (Quantitative Multiplex DNA reference standard from Horizon Discovery). **B:** Sample B (Oncology Hotspot Control from Thermo Fisher). Tumor Actionable mutations panel from Qiagen. Ion AmpliSeq Colon and Lung Cancer panel, Oncomine Focus Assay and Oncomine Solid Assay from Thermo Fisher. Somatic 1 assay and Tumor Hotspot assay from Multiplicom (Niel, Belgium). TruSight Tumor panel and TruSeq Amplicon Cancer panel from Illumina. Part Participant. Filters used for analysis of the raw data: variant allele frequency and coverage limits as reported by the participant.

Figure 4: Percentage of correctly (C) and incorrectly (NC) reported variants by each participant. **A:** Sample A (Quantitative Multiplex DNA reference standard from Horizon Discovery). **B:** Sample B (Oncology Hotspot Control from Thermo Fisher). Tumor Actionable mutations panel from Qiagen (Hilden, Germany). Ion AmpliSeq Colon and Lung Cancer panel, Oncomine Focus Assay and Oncomine Solid Assay from Thermo Fisher. Somatic 1 assay and Tumor Hotspot assay from Multiplicom. TruSight Tumor panel and TruSeq Amplicon Cancer panel from Illumina. Filters used for analysis of the raw data: variant allele frequency and coverage limits as reported by the participant. Part- Participant, QC- quality control, LOD- limit of detection.

Figure 5: Variants reported by the participants for whom no BAM files were available. **A:** Sample A (Quantitative Multiplex DNA reference standard from Horizon Discovery). **B:** Sample B (Oncology Hotspot Control from Thermo Fisher). Tumor Actionable mutations panel from Qiagen. GS Junior system from Roche with in-house primers. Oncomine Solid Assay from Thermo Fisher. Somatic 1 assay from Multiplicom. Filters used for analysis of the raw data: variant allele frequency and coverage limits as reported by the participant. N- number of variants that should have been detected by the specific methodology, Part- Participant, R- reported by the laboratory, NR- not reported by the laboratory, LOD- limit of detection.

Table 1: overview of the material sent in both parts of the study.

	Part 1: samples of the 2015 *EGFR* EQA scheme of the ESP	
Variants in the samples (N = 9)	*EGFR* wild-type (five samples) *EGFR* c.2156G>C p.(Gly719Ala) *EGFR* c.2573T>G p.(Leu858Arg) *EGFR* c.2235_2249del p.(Glu746_Ala750del) *EGFR* c.2369C>T p.(Thr790Met) and c.2573T>G p.(Leu858Arg)	
	Part 2: sample A	Part 2: sample B
Material sent	1 µg	75 µg or 300 µg, dependent on participant's panel
Variants in sample	> 500 at different VAF	> 500 at different VAF
Variants selected for the study	16 variants in 10 genes	299 variants in 20 genes
Position of the selected variants	Exonic	Exonic, intronic, splice sites
Relevance of the selected variants	Pathogenic (100%)	Pathogenic (37.2%) Benign (8.64%) Uncertain (28.9%) Unknown (25.3%)

Reference sequence EGFR: NM_005228.4. ESP European Society of Pathology. sample A: Quantitative Multiplex DNA reference standard from Horizon Discovery (Cambridge, UK). Sample B: Oncology Hotspot Control from Thermo Fisher (CA). VAF variant allele frequency. As the target regions for each NGS panel differ, the actual percentages varied for each technique.

Table 2: Overview of the characteristics of the participants in both parts of the study.

Techniques	Number of NGS samples tested per year	Accredited	NGS implementation
		Number of laboratories	
Non-NGS based commercial (N = 59)	NA	Yes: 14 No: 45	NA
Non-NGS based laboratory developed (N = 22)	NA	Yes: 12 No: 10	NA
NGS based commercial (N = 29)	0-249: 12 250-499: 5 >500: 6 Missing: 6	Yes: 16 No: 13	Before 2014: 4 2014-2015: 18 2016: 1 Missing: 6
NGS based laboratory developed (N = 12)	0-249: 1 250-499: 2 >500: 0 Missing: 9	Yes: 7 No: 5	Before 2014: 1 2014-2015: 2 2016: 0 Missing: 9

For the accreditation status, different national and international standards were taken into account: ISO 15189 and ISO 17025 standards as recognized international accreditation standards, CAP 15189 (College of American Pathologists) as national accreditation standard and widely used national standards such as the national standard in the Netherlands (CCKL). Missing data arose because these questions were only asked in the survey of part 2 (analysis of DNA control material).

N: Total number of laboratories in that group; NGS: next-generation sequencing; NA: not applicable

Table 3: Comparison of the analysis results between NGS laboratories and non-NGS laboratories for the patient FFPE samples.

NGS vs Non-NGS	≥ 1 FP	≥ 1 FN	≥ 1 WM	≥ 1 Technical failure	Maximal score*	Successful score†
Number of participating laboratories (%) using:						
NGS (N = 33)	0 (0.0%)	5 (15.2%)	3 (9.1%)	10 (30.3%)	6 (18.2%)	17 (51.5%)
Non-NGS (N = 81)	8 (9.9%)	23 (28.4%)	7 (8.6%)	19 (23.5%)	7 (8.6%)	42 (51.9%)
Number of samples (%) analyzed using:						
NGS (N = 33)	0 (0.0%)	9 (3.3%)	4 (1.3%)	15 (5.1%)	NA	NA
Non-NGS (N = 81)	10 (1.4%)	30 (4.1%)	7 (1.0%)	38 (5.2%)	NA	NA

* Maximal score is defined as a 100% score (18/18).

† Successful score is defined as at most one test failure and/or a nomenclature mistake, or at most one genotype error without nomenclature mistake.

NGS: next-generation sequencing; FP: false positive; FN: false negative; WM: wrong mutation; NA: not applicable.

Table 4: Overview of the participants and the submitted information available for further analysis.

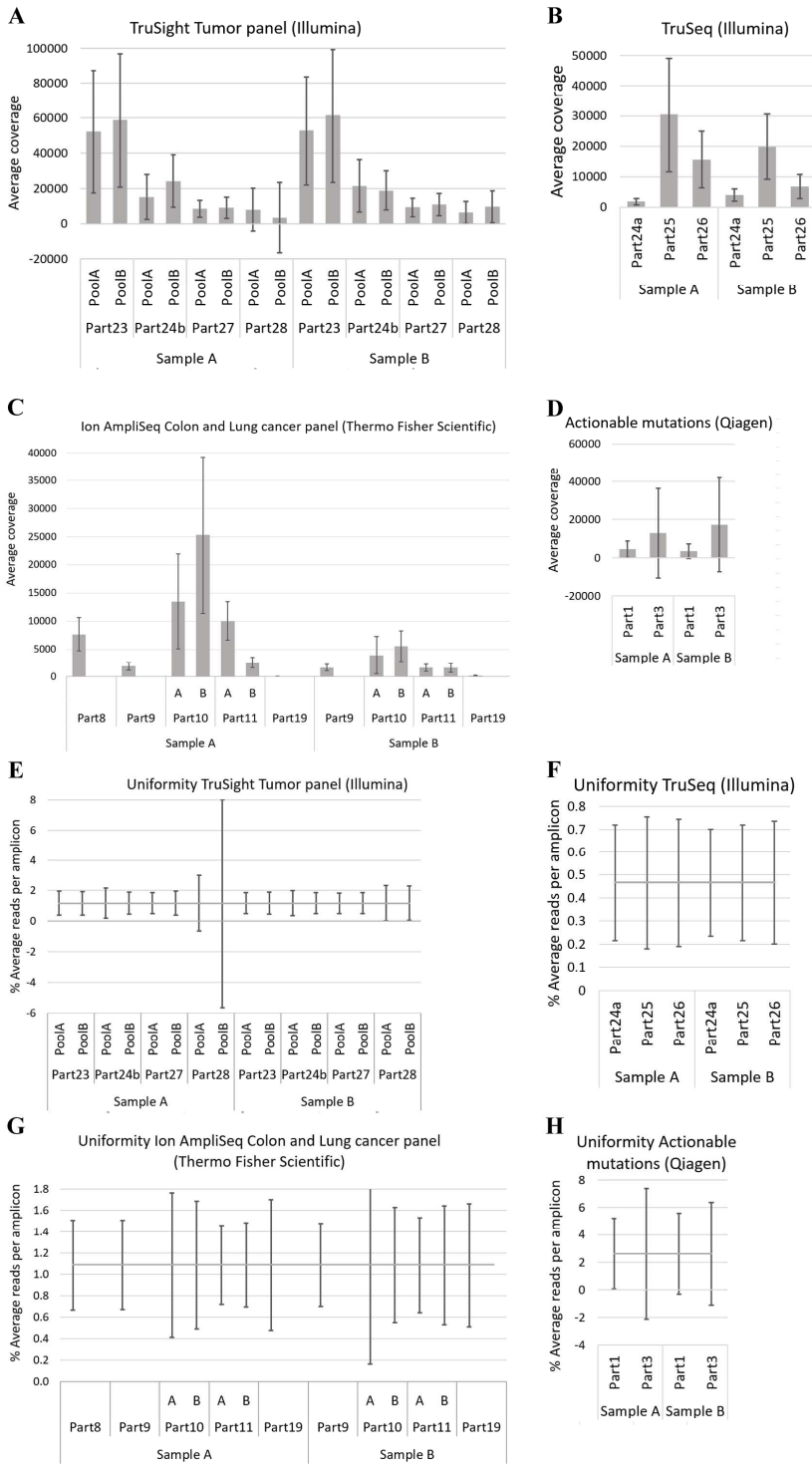
ID	Platform	Panel	Sample analyzed	QC metric survey	Raw data submitted
1	MiSeq (Illumina)	Tumor Actionable Mutations panel (Qiagen)	A and B	Yes	Yes
2			A and B	Yes	No
3			A and B	Yes	Yes
4a	GS Junior (Roche)	In-house panel*	A conclusive B inconclusive	No	NA
5			A and B	Yes	NA
6			A and B	Yes	NA
7			A and B	Yes	NA
8	PGM IonTorrent (Thermo Fisher)	Ion AmpliSeq Colon and Lung Cancer Panel (Thermo Fisher)	A conclusive B	Yes	Yes, Sample A
9			A and B	Yes	Yes
11			A and B	Yes	Yes
19			A and B	Yes	Yes
10	Ion Proton		A and B	Yes	Yes

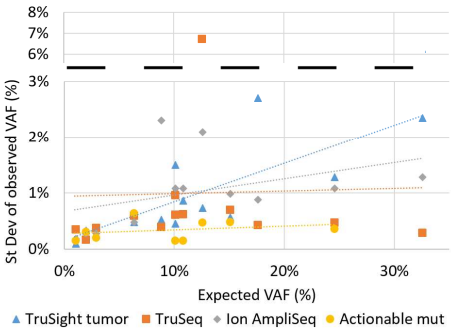
	(Thermo Fisher)				
12	PGM IonTorrent (Thermo Fisher)	Oncomine Focus Assay (Thermo Fisher)	A and B	Yes	Yes
13	PGM IonTorrent (Thermo Fisher)	Oncomine Solid Tumor DNA kit (Thermo Fisher)*	A and B	Yes	No
14			A and B	Yes	Yes
4b			A and B	Yes	No
15	MiSeq (Illumina)	Somatic 1 MASTR assay (Multiplicom)	A and B	Yes	Yes
16			A and B	Yes	No
17			A and B	Yes	No
18			A conclusive B no result	Yes	No
22	MiSeq (Illumina)	Tumor hotspot (Multiplicom)	A and B	Yes	Yes
24a	MiSeq (Illumina)	TruSeq Amplicon Cancer panel (Illumina)*	A and B	Yes	Yes
25	MiSeq dx (Illumina)		A and B	Yes	Yes
26	MiSeq (Illumina)		A and B	Yes	Yes
23	MiSeq (Illumina)	TruSight tumor panel (Illumina)*	A and B	Yes	Yes
24b			A and B	Yes	Yes
27			A and B	Yes	Yes
28			A and B	Yes	Yes

*Two institutes participated with two different techniques.

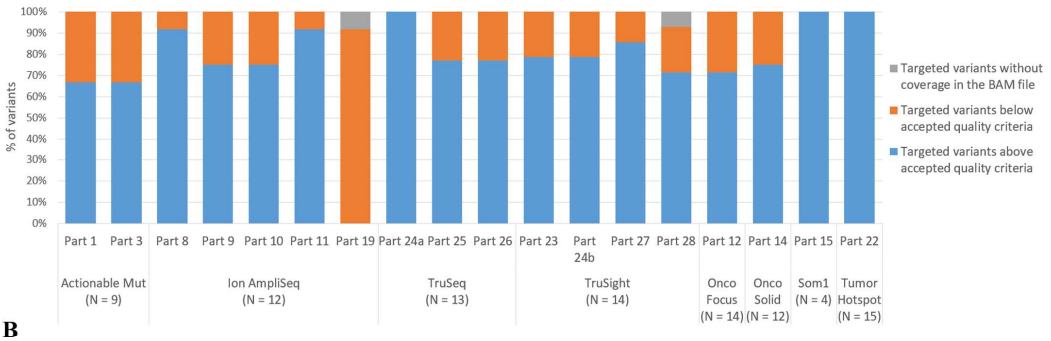
Sample A: Quantitative Multiplex DNA reference standard from Horizon Discovery. Sample B: Oncology Hotspot Control from Thermo Fisher.

NA-not applicable.

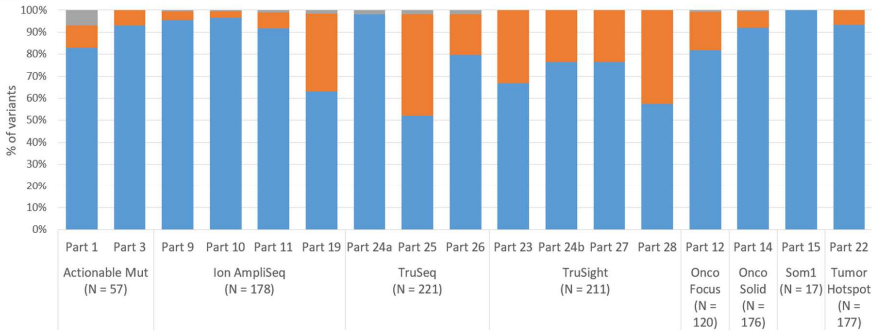

A
TruSight Tumor panel (Illumina)
Average coverage
PoolA
PoolB
Part23
Part24b
Part27
Part28
Sample A
Sample B
B
TruSeq (Illumina)
Average coverage
Part24a
Part25
Part26
Sample A
Sample B
C
Ion AmpliSeq Colon and Lung cancer panel (Thermo Fisher Scientific)
Average coverage
A
B
Part8
Part9
Part10
Part11
Part19
Sample A
Sample B
D
Actionable mutations (Qiagen)
Average coverage
Part1
Part3
Sample A
Sample B
E
Uniformity TruSight Tumor panel (Illumina)
% Average reads per amplicon
F
Uniformity TruSeq (Illumina)
% Average reads per amplicon
G
Uniformity Ion AmpliSeq Colon and Lung cancer panel (Thermo Fisher Scientific)
% Average reads per amplicon
H
Uniformity Actionable mutations (Qiagen)
% Average reads per amplicon

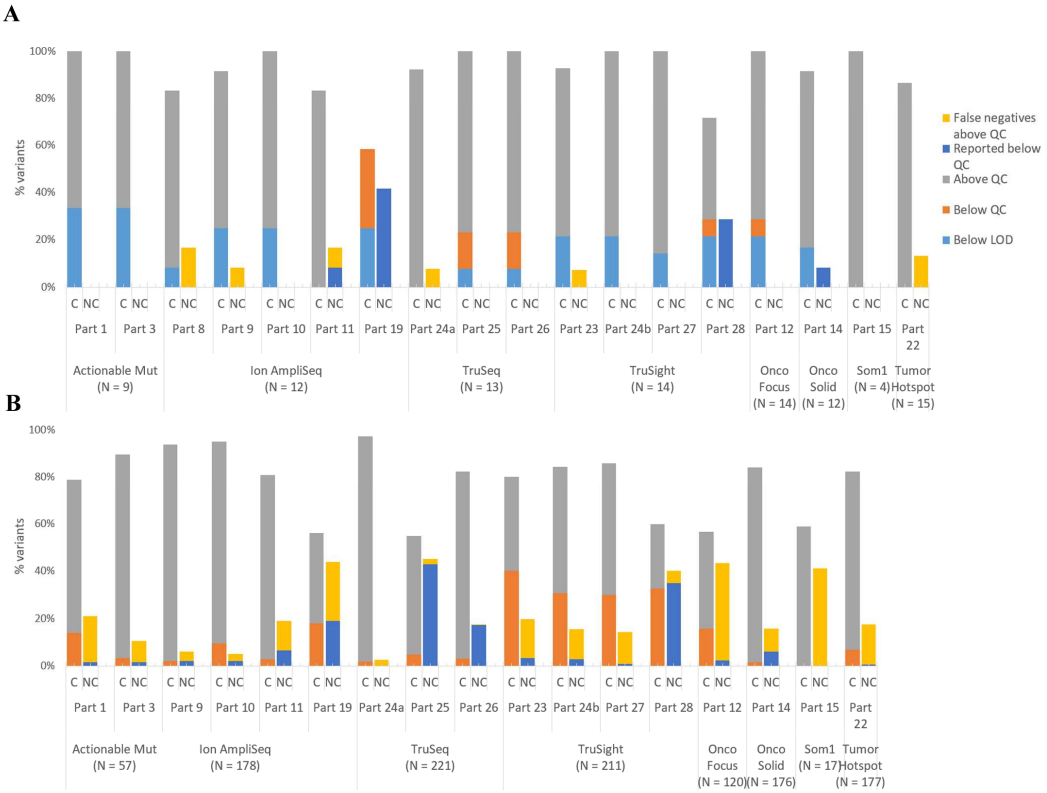

8%
7%
6%
3%
2%
1%
0%
St Dev of observed VAF (%)
0%
10%
20%
30%
Expected VAF (%)
TruSight tumor
TruSeq
Ion AmpliSeq
Actionable mut

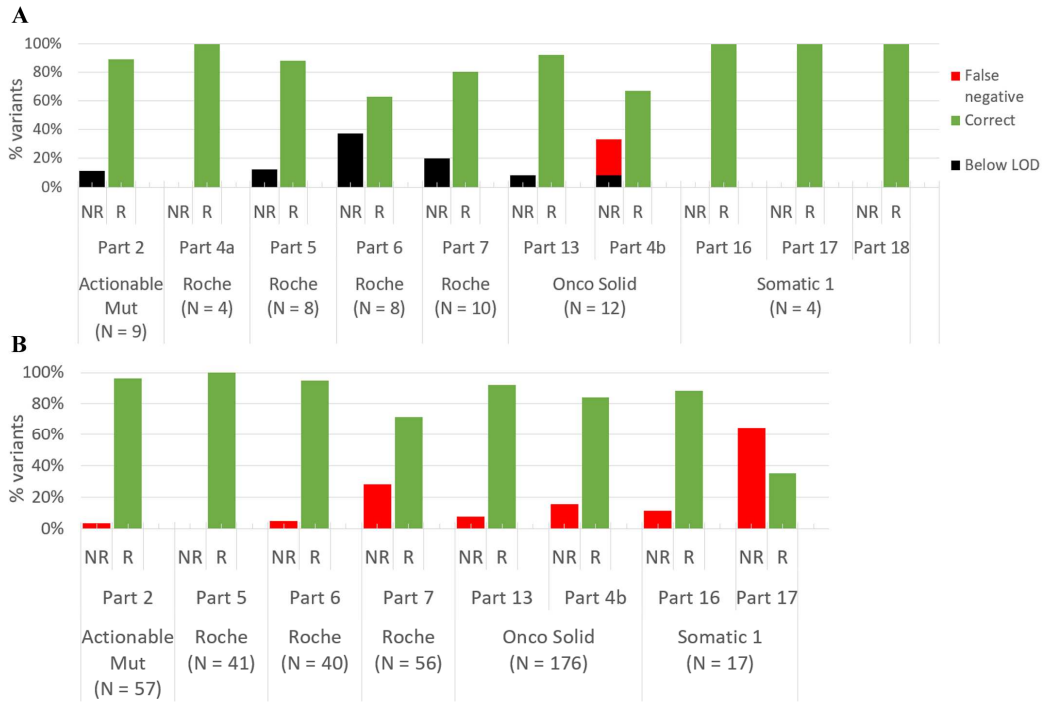
A


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
% of variants
Targeted variants without coverage in the BAM file
Targeted variants below accepted quality criteria
Targeted variants above accepted quality criteria
Part 1
Part 3
Part 8
Part 9
Part 10
Part 11
Part 19
Part 24a
Part 25
Part 26
Part 23
Part 24b
Part 27
Part 28
Part 12
Part 14
Part 15
Part 22
Actionable Mut (N = 9)
Ion AmpliSeq (N = 12)
TruSeq (N = 13)
TruSight (N = 14)
Onco Focus (N = 14)
Onco Solid (N = 12)
Som1 (N = 4)
Tumor Hotspot (N = 15)

B


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
% of variants
Part 1
Part 3
Part 9
Part 10
Part 11
Part 19
Part 24a
Part 25
Part 26
Part 23
Part 24b
Part 27
Part 28
Part 12
Part 14
Part 15
Part 22
Actionable Mut (N = 57)
Ion AmpliSeq (N = 178)
TruSeq (N = 221)
TruSight (N = 211)
Onco Focus (N = 120)
Onco Solid (N = 176)
Som1 (N = 17)
Tumor Hotspot (N = 177)


A
100%
80%
60%
40%
20%
0%
% variants
False negatives above QC
Reported below QC
Above QC
Below QC
Below LOD
C NC
Part 1
Part 3
Part 8
Part 9
Part 10
Part 11
Part 19
Part 24a
Part 25
Part 26
Part 23
Part 24b
Part 27
Part 28
Part 12
Part 14
Part 15
Part 22
Actionable Mut (N = 9)
Ion AmpliSeq (N = 12)
TruSeq (N = 13)
TruSight (N = 14)
Onco Focus (N = 14)
Onco Solid (N = 12)
Som1 (N = 4)
Tumor Hotspot (N = 15)
B
100%
80%
60%
40%
20%
0%
% variants
C NC
Part 1
Part 3
Part 9
Part 10
Part 11
Part 19
Part 24a
Part 25
Part 26
Part 23
Part 24b
Part 27
Part 28
Part 12
Part 14
Part 15
Part 22
Actionable Mut (N = 57)
Ion AmpliSeq (N = 178)
TruSeq (N = 221)
TruSight (N = 211)
Onco Focus (N = 120)
Onco Solid (N = 176)
Som1 (N = 17)
Tumor Hotspot (N = 177)


A
100%
80%
60%
40%
20%
0%
% variants
False negative
Correct
Below LOD
NR R
Part 2
Actionable Mut (N = 9)
Part 4a
Roche (N = 4)
Part 5
Roche (N = 8)
Part 6
Roche (N = 8)
Part 7
Roche (N = 10)
Part 13
Part 4b
Onco Solid (N = 12)
Part 16
Part 17
Part 18
Somatic 1 (N = 4)
B
100%
80%
60%
40%
20%
0%
% variants
NR R
Part 2
Actionable Mut (N = 57)
Part 5
Roche (N = 41)
Part 6
Roche (N = 40)
Part 7
Roche (N = 56)
Part 13
Part 4b
Onco Solid (N = 176)
Part 16
Part 17
Somatic 1 (N = 17)