Manuscript title: Application of targeted-Next-Generation Sequencing, TruSeq Custom Amplicon assay for molecular pathology diagnostics on formalin-fixed and paraffin embedded samples.

Running title: Targeted- Next-Generation sequencing for diagnostic use

Authors: Erzsébet Csernák 1, János Molnár 2, Gábor E. Tusnády 2, Erika Tóth 1

1 National Institute of Oncology, Department of Surgical and Molecular Pathology, H-1122 Budapest, Hungary

2"Momentum" Membrane Protein Bioinformatics Research Group, Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, H-1117 Budapest, Hungary

Correspondence: Erika Tóth, e-mail: erika66toth@gmail.com

Abstract

The implementation of targeted therapies revolutionized oncology. As the number of new oncogenic driver mutations, which provide molecular targets for prediction of effective and selective therapies, is increasing, the implementation of fast and reliable methods by molecular pathology labs is very important. Here we report our results with TruSeq Custom Amplicon (TSCA) assay performed on formalin-fixed and paraffin embedded (FFPE) material. The oligo capture probes targeted the hotspot regions of ten well known oncogenes linked to clinical diagnosis and treatment of lung and colorectal adenocarcinomas, melanomas and gastrointestinal stromal tumors. Fifteen previously genotyped FFPE DNA samples from different tumor types were selected for massively parallel sequencing. A bioinformatics pipeline was developed to identify high quality variants and remove sequence artifacts. With the exception of one sample, which was of lower quality than the others, relevant mutations corresponding to tumor types could be reliable detected by the developed bioinformatical pipeline. This study indicates that the application of TSCA assay is a promising tool in molecular pathology diagnostics, but it is important to standardize sample processing (including fixation, isolation procedure, sample selection based on quality assessment and rigorous variant calling) in order to achieve the highest success rate and avoid false results.

Introduction

NGS has a fundamental importance in the genetic profiling of various diseases and in the translation of the obtained data back into medical applications. In the field of oncology, molecular characterization of tumors is growing rapidly and the number of new biomarkers implicated in tumor progression and therapeutic sensitivity is expanding.

As the number of new oncogenic driver lesions, which provide molecular targets for prediction of effective and selective therapies is increasing, the implementation of fast and reliable methods by molecular pathology departments is very important. The aim of molecular pathology labs is giving the most accurate and precise results within the shortest period of time. „Turnaround time"(TAT) in molecular pathology is as much important as in surgical pathology.

This raises the need for deployment of high performance and rapid molecular tests that can screen a large panel of genes for tumor actionable mutations both specifically and accurately. In this respect, standard PCR based approaches can only detect a limited number of genomic alterations owing to low-level multiplicity. The genetic analysis should also be appropriate for working with DNA extracted from formalin fixed tissues, because this is the type of material that is generally available at routine pathological laboratories. Although numerous studies have been successfully used FFPE material for NGS, several caveats and considerations should be taken into account when we use FFPE samples for diagnostic purposes.[1–3] The main issue with FFPE samples is DNA degradation caused by formalin fixation resulting incorporation of incorrect bases during PCR.[4] Since artificial mutations are present at low rate, there could be difficulties when we use sensitive techniques to detect low frequency variants or to improve analysis from tissue samples with very low tumor content on a normal background. As NGS method implements clonal sequencing and enables low detection limit, the probability of erroneously identifying mutations is higher in such cases, resulting in reduced specificity. Nevertheless, the use of efficient bioinformatics tools has been thought to overcome such problems.[5]

In light of this, our aim was to investigate what conditions are necessary to achieve reliable mutation detection from FFPE material by targeted next generation sequencing. To do this, formalin-fixed tumor tissues with different tumor ratio, mutational status and varying degrees of quality were selected. The sequencing reaction was performed on Illumina MiSeq instrument using TruSeq Custom Amplicon (TSCA) technology. For data evaluation we developed our own bioinformatics pipeline to adapt to damaged samples with particular

attention to removal of sequencing artifacts. Finally, we also assessed FFPE sample criteria concerning quality and tumor cell content required to obtain a successful sequencing result.

Materials and methods

Sample and testing selection

We used TruSeq Custom Amplicon assay with the Illumina Miseq instrument. We expanded our currently used gene panels with new marker genes using Illumina DesignStudio which contains 3,493 Kb of cumulative sequence and 36 amplicons (Supplemental Table).

The oligo probes targeted the hotspot regions of clinically relevant oncogenes including KRAS, NRAS, BRAF, EGFR, KIT, PDGFRA, MAP2K1, AKT1, PIK3CA, and FGFR2 (Supplemental Table). Fifteen FFPE tissue samples of colorectal cancer (CRC), melanoma, gastrointestinal stromal tumor (GIST), and non-small cell lung cancer (NSCLC) patients were selected for targeted-NGS, harboring known mutations in oncogenes previously detected by in-house fluorescence probe based real-time PCR assay, Sanger sequencing and Cobas KRAS and EGFR mutation tests (Roche Diagnostics).

Sample preparation and DNA quality assessment

Prior to isolation, the tumor / normal cell ratio was estimated by our local pathologist on hematoxylin and eosin (H&E) stained slides and tumor tissues were macro dissected from a single representative block. Depending on the diameter of the tumor area, three or five 5µm sections were deparaffinized and subjected to cell lysis with proteinase K treatment at 56 ºC for 24 hours. After this, DNA extraction was carried out with the Cobas DNA Sample Preparation kit (Roche Diagnostics) according to the manufacturer's instruction. The purified DNA concentration was determined by fluorescent method using the Quant-iT High-Sensitivity DNA Assay Kit (Life Technologies). We used two PCR based quality control methods namely the Illumina FFPE QC assay (Illumina) and KAPA hgDNA Quantification and QC Kits (Kapa Biosystems) in order to assess whether there is sufficient number of amplifiable template in the DNA sample for library construction.

Library preparation and sequencing analysis

Library preparation was performed according to the TSCA protocol. The successfully produced samples were pooled and sequenced on the Illumina MiSeq instrument using the Miseq reagent nano kit v2 generating 2x150 paired-end reads at a level of 300Mb output. The reads were aligned to the human (*Homo sapiens*) reference sequence version *GRCh37.7*,

which was obtained from the Ensembl database.[6] The alignments of paired-end reads were generated with the *bwa-mem* algorithm. Duplicated reads were removed from the alignment by the Picard tool *MarkDuplicates*. The SAMtools[7] pipeline was used in the variant-calling process**.**

Filtering out sequencing artefacts

Generating the final variants from the deduplicated alignments, we used strict filtering options. In case of single nucleotide variants, the variant must be covered with high quality reads from each sequencing strand at least one time and the Phred variant quality must be equal or higher than 45. Additionally, when the number of high quality reads is less than 5 the alternative variant containing high quality reads must be equal or greater than 2, else the ratio of alternative variant containing reads must be greater than 0.25. In case of short indels, the variants must be covered with high quality reads containing the variant from each sequencing strand at least one time and the Phred variant quality must be equal or higher than 45. Additionally, if the coverage of high quality reads is 4 than the alternative variant must be covered with high quality reads at least two times. Else the ratio of alternative variant containing reads must be greater than 0.25. These criteria were used to distinguish sequencing artefacts (e.g. formalin induced mutations) from real heterozygous variants. Finally the detected variants were annotated by the Ensembl Variant Effect Predictor[8].

Results

Library construction from FFPE samples

Table 1 shows the results of DNA quality assessment. We determined those parameters that mainly affect the success of library preparation. The DNA concentration had minimal influence, because as little as 5 ng/µl worked properly (case 5, Table 1) which is favorably lower than 25 ng/µl specified in the protocol. This flexibility is likely because of the fact that the FFPE samples contain certain amount of single stranded DNA, which is also suitable for amplicon based assays unlike other types of library preparation protocols, where double stranded DNA is required. However, it should be noted that if the concentration was too low as in case 3 (0.4ng/µl), the sample didn't function for TSCA even if it was of good quality (Table 1).

We used two quantitative real-time PCR (qPCR) assays to assess the amplification efficiency of the input DNA. The Illumina FFPE QC Kit calculates a ΔCt value, which is a subtraction of the quantification threshold cycle (Ct) of a control template included in the kit, from the Ct

value of a given sample, both used in equal concentration. According to the recommendations, all samples with values below or equal to 2 can be selected for TSCA. Interestingly, case 1 with 1.7 ΔCt didn't work; by contrast, case 14 with 2.4 ΔCt gave good result (Table 1). These minimal discrepancies can be attributed to the fact that the ΔCt value only characterizes the amplification capacity of a defined target length and does not give any information if there is sufficient number of amplifiable template in wider size range, that is the extent of fragmentation in damaged FFPE DNA. Using KAPA hgDNA Quantification and QC assay allows us to get this kind of quality information by measuring Q129/Q41 ratio. For calculating this value, standard curves were generated with amplifying targets of 41bp and 129bp in the human genome. The relative quality can be derived by normalizing the concentration obtained with the 129bp assay against the concentration from the 41bp assay. The value can vary between 0 and 1, depending on the sample quality. We found that a Q129/Q41 ratio below 0.2 was a negative predictor of library success. These values could also explain the deviations mentioned above at cases 1 and 14, by modifying the expected outcome based on ΔCt. For sample 4, despite its poor quality (Q=0,104) the library worked well probably due to its good amplification efficiency (Table 2). In spite of the low quality, we carried out the subsequent sequencing analysis to see its influence on final data interpretation. Based on the 129bp assay, we also calculated a ΔCt value (by subtracting the expected Ct value of a sample calculated from its concentration and standard curve, from the measured Ct) that gave relatively small differences to Illumina QC ΔCt, which was likely caused by the various lengths of targets used in qPCRs. The inconsistency of samples 1 and 14 was also observed at the Q129 ΔCt value, which confirmed the necessity of using the Q129/Q41 ratio or a similar quality value for better selection of samples suitable for sequencing (Table 1).

TSCA assay results

Sequencing metrics

Full coverage was relatively high (6929 in average, range between 4764-8342) due to limited samples were running, but this could be corrected, and it did not affect the recovery of data. The TSCA sequencing run gave an excellent specificity, (defined as the percent of filtered reads mapping to target regions) and was very similar among formalin-fixed samples at an average value of 96.6%. We assessed the uniformity of target read distribution between amplicons and samples after duplicate removal (Figure 1). The coverage values deviated from what we expected, with a range from 8.45 +/- 3.77 to 21.34 +/- 7.94 within the samples,

which may reflect the uneven quality of FFPE samples available for sequencing. Furthermore, this variation can be seen within the amplified regions as well (Figure 1). It should be noted that due to the deduplication of redundant sequences (PCR duplicates) our relative coverage seems quite small (Figure 1), but this step was important to improve variant calling and eliminate bias influencing the real allelic representation of the sample. In the cases of NRASex3, AKT1ex2, and EGFRex20 target regions, a high GC sequence content (above 80%) was found near the TSCA oligo probes that likely caused this observation. However, these discrepancies did not affect the reference mapping and the subsequent variant calling.

Detecting high quality variants

Using our variant detection pipeline, we were able to identify high quality variants (Table 2). As our experience showed, it is crucial to remove the duplicated sequences, which can interfere the variant calling by the biased read coverage (Figure 2). This was presented in other publication as well, i.e. the FFPE samples contain high number of PCR duplicates within the 60% - 85% range[9]. We made a test about the reliability of variant calling. The selected loci (KIT_ex11) from sample 11 was downsampled using the Picard tool *DownsampleSam* within the range of 100%-10% (the coverage was in the range from 16 to 2), and we were able to identify the alternative allele when the covarege was as low as two. Using this knowledge, we achieved a high concordance in the identification of clinically relevant mutations that were previously characterized by our currently used laboratory tests. In case of sample 4 we were not able to apply our strict automated variant detection pipeline to identify mutations, as it generated a proportionately high number of low-quality reads making it unsuitable for reliable analysis.

Discussion

The advents of next generation sequencing technologies have opened the possibility to get enormous amount of genetic information from a given sample and have a deeper knowledge of the association between the genomic structure and function, and various biological processes or diseases. Applying this information in clinical practice has many benefits including refining diagnosis, individualizing treatments, predicting drug effects or developing new therapies.[10] In light of these advantages, we considered the possibilities to replace our currently used low scale diagnostic tests with a large scale NGS method. Using fifteen pre-selected FFPE tumor tissue samples as controls harboring clinically important mutations, we investigated the main criteria required to the successful adoption of NGS in clinical

diagnostics. Basically, two target enrichment techniques can be applicable for this purpose: amplicon-based and sequence capture approaches. We chose an amplicon based method, because we focused on a small set of clinically important genes with defined hotspots. We preferred TruSeq custom amplicon (TSCA) assay as it has simple workflow, the entire procedure takes only two days, thus allowing short turnaround time of reports and rapid and efficient patient management. In addition, during work with million copies of PCR amplicons, care must be taken to avoid cross-contamination between templates. In this regard, TSCA seemed to us more optimal than other amplicon based sequencing strategies, because only a reduced cycling time PCR occurs when sample-specific indices are added to each library, thereby minimizing the possible presence of unwanted constituents.

We tested DNA samples obtained from formalin-fixed, paraffin-embedded tissue blocks. Although this type of material has great potential in clinical medicine, due to the fixation process DNA can be degraded at various degrees, which can impact the reliability and quality of genetic analyses. For NGS, high quality starting material is essential for reliable and accurate sequencing results, therefore qualitative monitoring of FFPE DNA specimens is an important step. Using different qPCR methods we assessed the main parameters that best characterize the quality of the samples and predict the success of assay performance. The samples used in our experiments were obtained from a variety of tissue types and hospitals using different fixation and embedding protocols. As it was expected, the level of degradation and quality between samples varied over a wide range (Q: 0.052-0.91, Table 2). The best approach to determine the quality of DNA is proved to be a qPCR assay targeting different sizes of genomic regions. This method gives the most precise information about DNA degradation level. Our results show that at the moment the quality requirements are higher for NGS compared to conventional PCR techniques (sequencing was successful in 67% of our FFPE samples), but it can be improved in the future by using more standardized protocols for tissue fixation, and DNA isolation tailoring to the particularities of FFPE material.

For accurate variant detection it is essential to reach consistent uniformity and minimum required depth of coverage of target regions. The amplicon-based technique has the advantage of generating more uniform coverage across all target bases than sequence capture approach, however it is possible that the poor amplification of some targets mainly caused by high GC content as presented above. Low or no coverage regions could produce false negative results, which is not acceptable in diagnostic applications. In our case, 8% of the target amplicons were underrepresented. With our high coverage depth, we achieved a few hundred fold coverage in these regions, which proved to be sufficient for accurate variant calling. In the

future more emphasis should be placed on optimal probe design to obtain more consistent coverage and avoid false results. On the basis of our findings, for accurate variant calling needs at least one high quality read from both the reverse and forward strand after duplicate removal. But the minimal required raw read number/coverage is hard to tell: this depends highly on sample quality, and more data will be needed to predict precisely

In case of FFPE samples, it should be taken into account that the samples do not amplify in the same quality, resulting in greater variability in the distribution of sequencing reads and piling up high number of PCR duplicates within the 60% - 85% range. Developing an alternative PCR amplification protocol could solve this problem.[11]

A properly configured bioinformatics pipeline is essential for high confidence variant detection. Aligned to the properties of FFPE samples, we introduced rigorous filtering parameters including duplicate removal, minimal coverage of high quality reads representing the alternate allele, coverage from the forward and reverse sequencing strand, and higher variant quality as well. Using our strict filtering pipeline, we were able to detect short genetic variants and filter out sequencing errors and artifacts caused by the formalin fixation. We identified safely the previously detected genetic variants of samples with tumor cell content between 40-90%. Due to the very useful nature of sequencing, we were able to detect other genetic variants in the selected regions as well, which enables even finer determination of the different mutations.

It is also important to consider when applying diagnostics tests that the quality of the samples determines data quality; bioinformatics could not manage every input data optimally, as it was seen in the case of our samples. Therefore it is essential to set up quality criteria for efficient sequencing.

In summary, we described the feasibility of successful validation of targeted NGS for molecular pathology diagnosis of different hot spot mutations in FFPE samples. As the number of targetable genes increasing in oncology, it becomes more important to implement NGS workflows into the routine laboratory work.

Disclosure/Conflict of interest

The authors declare no conflict of interest.

References

1. Zhang, W., Cui, H. & Wong, L. J. C. Application of next generation sequencing to molecular diagnosis of inherited diseases. *Top. Curr. Chem.* **336,** 19–46 (2014).
2. Kerick, M. *et al.* Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med. Genomics* **4,** 68 (2011).
3. Wong, S. Q. *et al.* Targeted-capture massively-parallel sequencing enables robust detection of clinically informative mutations from formalin-fixed tumours. *Sci. Rep.* **3,** 3494 (2013).
4. Williams, C. *et al.* A high frequency of sequence alterations is due to formalin fixation of archival specimens. *Am. J. Pathol.* **155,** 1467–1471 (1999).
5. Yost, S. E. *et al.* Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res.* **40,** e107 (2012).
6. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** (2014).
7. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
8. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–70 (2010).
9. Hedegaard, J. *et al.* Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. *PLoS One* **9,** e98187 (2014).
10. Meldrum, C., Doyle, M. a & Tothill, R. W. Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin. Biochem. Rev.* **32,** 177–95 (2011).
11. Hoeijmakers, W. a M., Bártfai, R., Françoijs, K.-J. & Stunnenberg, H. G. Linear amplification for deep sequencing. *Nat. Protoc.* **6,** 1026–36 (2011).

Titles and legends to figures

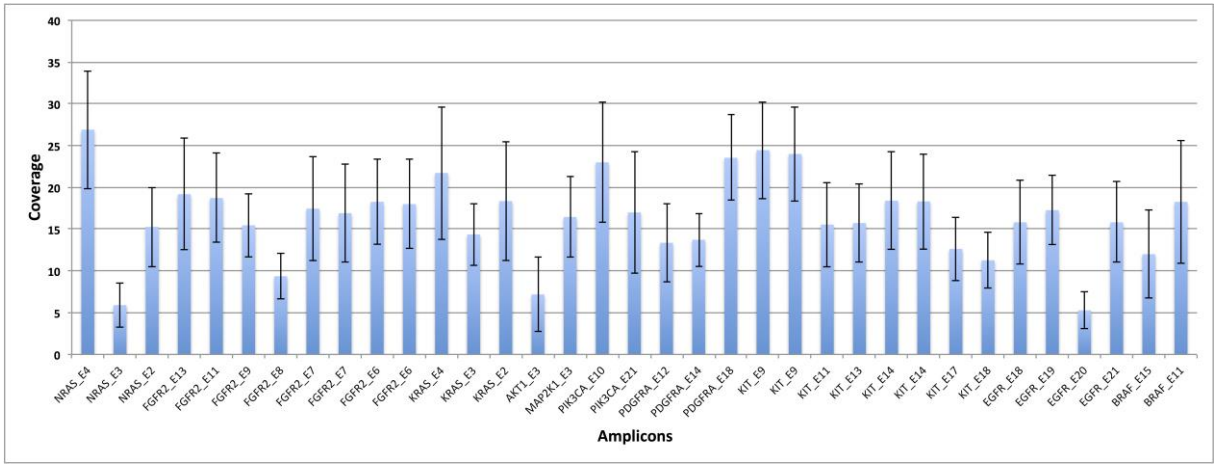
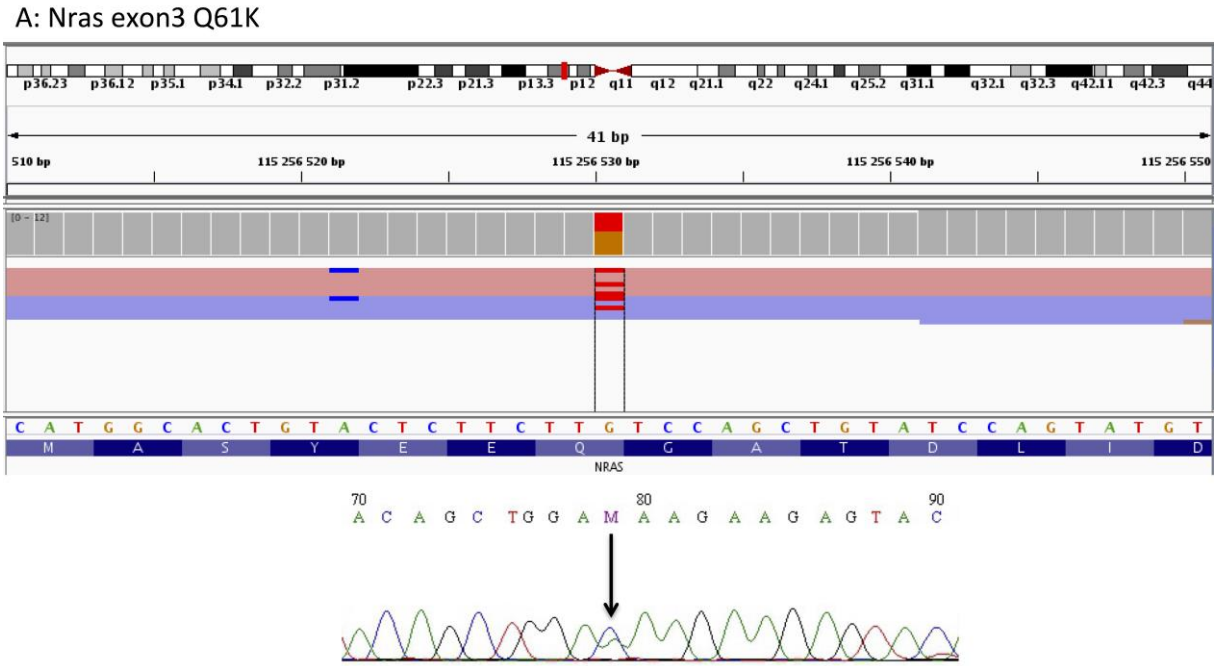


Figure 1 Coverage distribution among the eleven samples after duplicate removal

Median coverage values were calculated for every amplified region within one sample. For every amplified region the average value was determined from these coverage data among the samples, and the standard deviation was calculated as well. The blue bars represent the average values for the amplified region, and the standard deviation values are visualized as well



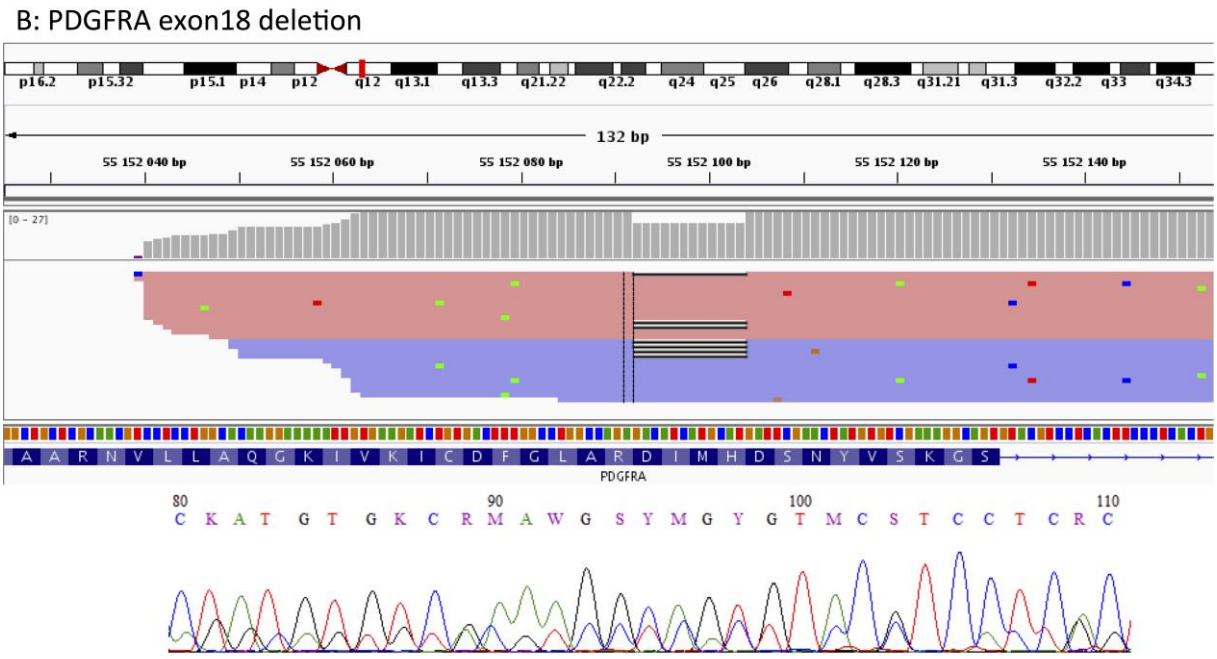


Figure 2 Example of sequence results

Manual inspection by the IGV program (Robinson et al. 2011) of the NGS reads (a) and the sequencing chromatograms from Sanger method (b) shows the same genotypes. In the case of NRAS the Sanger sequencing data is from the sense strand, which is reversed to the human genomic reference sequence