Clin. Lab. 2020;66:XXX-XXX

ORIGINAL ARTICLE

Clinical Validation of a Novel GeneReader Next Generation Sequencing System for Tumor Specific Mutations and Bioinformatics Variant Analysis

Ibrahim Boga [1, 2], Ozge Sonmezler [1], Atil Bisgin [1, 2]

[1] *Cukurova University AGENTEM (Adana Genetic Diseases Diagnosis and Treatment Center), Adana, Turkey*
[2] *Cukurova University, Balcali Clinics and Hospital, Faculty of Medicine, Department of Medical Genetics, Adana, Turkey*

(Clin. Lab. 2020;66:xx-xx. DOI: 10.7754/Clin.Lab.2020.200123)

Correspondence:
Atil Bisgin, MD, PhD
Cukurova University
Balcali Clinics and
Hospital Faculty of Medicine
Department of Medical Genetics and
Cukurova University AGENTEM
01250, Saricam, Adana
Turkey
Phone: +90 5325652167
Email: abisgin@yahoo.com

Manuscript accepted March 23, 2020

KEY WORDS

GeneReader NGS System, bioinformatics, oncogenetics

INTRODUCTION

Next generation sequencing (NGS) has become the preferred molecular genetics testing method due to its high sensitivity and reliability. It enables multi gene sequencing of multiple samples during the same workflow. This provides copious amounts of data and information as compared to conventional methods. This technology, however, requires qualified laboratory scientists and experienced medical geneticists to process and analyze such a large quantity of data. Proficient bioinformatic analyses of this data carries great potential for the human genetics community. It was therefore inevitable that this technology would be used to pioneer a methodology in routine clinical genetic testing [1]. Recently, numerous commercialized NGS devices have been produced for clinical use. Although the various NGS systems have different approaches, the principle is still analogue. Under these circumstances, optimization of these methodologies has become essential before clinical testing takes place [2]. This is due to the fact that the user-dependent, complex workflow needs to be optimized and validated by experienced laboratories for routine clinical genetic testing [3,4]. In this study we aimed to validate, optimize, and standardize a new NGS system with a targeted multi-gene solid tumor panel starting with DNA extracted from formalin fixed paraffin embedded (FFPE) tissue samples as well as perform bioinformatic analyses of tumor specific variants.

The development of an NGS-based clinical diagnostic

application consists of several steps starting with DNA isolation from FFPE samples, followed by target enrichment, library preparation, clonal amplification, next generation sequencing, data and variant analysis, and bioinformatic interpretation. The workflow which starts with DNA extraction and the resulting clinical report is shown in Figure 1. Any application errors in the protocol may cause a repeat of the entire experiment. It is essential for all steps to be performed properly in order to achieve accurate and precise results. Therefore, it is of great importance that the manufacturers' manuals are validated and standardized by experienced laboratory scientists to establish reproducibility and reliability [5]. Precision is crucial in routine clinical laboratory service due to the many intrinsic variables. In this study, we performed the clinical and laboratory optimizations of a solid tumor panel (Qiagen Actionable Insight Solid Tumor Panel) which consists of 12 genes that are important for medical genetics and medical oncology via a novel NGS (Qiagen, GeneReader) system.

MATERIALS AND METHODS

Biological sample processing

We performed next generation sequencing via a 12-gene-solid-tumor-panel (*EGFR, ALK, KRAS, NRAS, KIT, BRAF, PDGFRA, ERBB2, PIK3CA, ERBB3, ESR1* and *RAF1*) with FFPE tissues at Cukurova University AGENTEM. Conventional methods such as pyrosequencing, real-time PCR, Sanger sequencing and even NGS are typically carried out with more homogenous samples as starting materials, such as peripheral blood. In studies with FFPE samples however, DNA can be damaged by chemical exposure. It therefore becomes obligatory to optimize the entire workflow process, starting from the deparaffinization step to acquire a high quality and quantity of DNA.

In this study, we compared two different paraffin brands to eliminate the possible quality differences of genetic material depending on the paraffinization process. We compared the isolated DNA amounts and determined which paraffinization brand results in a higher quality and quantity of DNA. We performed the workflow according to the manufacturers' instructions with modifications for laboratory optimization.

Genomic DNA was isolated from FFPE samples with an FFPE DNA Isolation Kit (Qiagen, Germany), and the DNA concentrations were measured using a Qubit 3.0 (ThermoFisher, USA). Samples that had DNA concentrations of as low as 1 ng/µL were considered acceptable to continue to the target enrichment step, which is different than the manufacturer's instructions.

Next generation sequencing process

Target enrichment was performed with the AIT panel primers and reagents (4.4 µL GeneRead DNAseq Panel 5 x PCR Buffer, 1.5 µL GeneRead HotStar Taq DNA Polymerase, 11 µL Primer Mix Pool A1-A4 (2x), and 0.7 µL DNase free water) to amplify the specific regions of interests on the selected 12 genes. PCR amplicons were purified twice with magnetic beads to eliminate shorter or longer amplicons than the desired ~150 base pairs.

Quality controls with the real time PCR step were not performed due to the non-concordance of the quality of the results based on real time PCR and the next generation sequencing results of the same samples. Although the real time PCR assays suggested inefficacious results, the sequencing results were successful. Therefore, we determined that the manufacturer's suggested quality control process in this step was time consuming and ineffectual.

Capillary electrophoresis (QIAxcel) was used to measure base pair length of the amplimers against a size marker panel of 50 - 800 bp and an alignment marker of 15 bp/3 kb. Samples within the optimal base pair length range, with a minimum of 1 ng of DNA concentrations were taken to the library preparation step.

To differentiate the samples in the library preparation step, adapter ligation was performed with 12 specific molecular barcodes (8 base long oligonucleotides). At the end of the library preparation step, concentrations and qualities of samples were measured via capillary electrophoresis. Concentrations of each of the barcoded samples were normalized and pooled according to quality control results.

For the clonal amplification step, each specific barcoded amplicon was amplified clonally via droplet making, emulsion PCR, and break/enrichment steps. This was achieved by attaching the amplicons to a primer loaded PCR bead which was amplified in an oil drop.

Next generation sequencing was performed for ~40 hours by the GeneReader NGS System (Qiagen, Germany).

Data interpretation

Even when the laboratory workflow is done accurately, next generation sequencing data should be analyzed precisely to make the large amounts of data generated useful for clinical applications [6]. With this aim in mind, the data quality assessment was performed after the next generation sequencing steps. At this point, we checked the sizes of the FASTQ files (for example ~1 Gbyte or higher) as one of the QC parameters. The other parameters stated in the analysis report of the GeneReader NGS device which were investigated, included the Sequence Quality Score, the Cycle Quality Score, the Yield and Reads Past Filtering. Yields of every sample were checked individually and compared with other samples in the same run to determine sample-based errors during the NGS procedure.

After data monitoring, we proceeded to variant analysis via the QCI Analyze interface (version 1.5.0), starting with the quality control parameters which were determined by analyze interface. Next, eligible variants were selected and transmitted for bioinformatic analyses for clinical interpretations. Variant selection was made for

each detected genetic change by depth of coverage, quality of variant reads, forward/reverse read balance, variant frequencies, coverage of the region of interest, and whether the variants were in the region of interest. Bioinformatic analyses of the selected variants were performed using the QCI Interpret Interface. Variant pathogenicities and their effects on diagnosis and treatment were determined according to the patients' clinical findings in terms of the ACMG (American College of Medical Genetics and Genomics) guideline classification [6,7].

We used CADD and Mutation Taster to estimate the general effects while protein level effects were examined using PolyPhen, SIFT, and B-SIFT. Conservation was investigated with BLOSUM, QCI Inferred Activation, and Phylop, while splice site effects were checked by MaxEntScan and Gene Splicer in-silico analysis tools.

We evaluated the selected somatic variants utilizing two different approaches. First, we assessed variants independently of the patient's clinical findings and diagnosis. Secondly, we interpreted variants taking into consideration the patients' clinical and other laboratory findings.

Once bioinformatic analyses were completed, the results were reported according to international criteria and the ACMG guidelines.

RESULTS

Here we share our experiences and results of a clinical diagnostic pipeline we developed using a novel NGS system that integrated a multi-gene panel. Our system was based on a commercial kit that we further optimized and modified as shown Figure 1.

First, we optimized DNA isolation from the FFPE samples to obtain the maximum quality and quantity. This was achieved by comparing the two different paraffin brands. [According to our DNA measurements we could obtain sufficient DNA to do the further molecular studies from the FFPE samples, from only one of the two paraffin brands (MERCK KGaA, Germany) ($p < 0.05$)]. Another parameter that had an effect on DNA isolation was tissue section thickness cut from the FFPE blocks. We compared four different section sizes: 5 µm, 10 µm, 15 µm, and 20 µm. The sections that gave the best DNA concentrations with an optimal one-step of deparaffinization was 10 µm.

According to manufacturer's manual, a minimum starting DNA input concentration of 2.5 ng/µL is required. However, we were able to successfully perform the entire workflow with lower DNA concentrations (1.51 ng/µL, 1.62 ng/µL, and 1.05 ng/µL, respectively, from three different specimens). These produced similar next generation sequencing yields when compared to the samples that had concentrations of 2.5 ng/µL DNA concentration. Therefore, the minimum DNA input threshold was determined as 1 ng/µL.

The quality control step of target enrichment by capillary electrophoresis has been used to detect a target which indicates larger fragments besides our target. The first bead incubation time was modified from 5 minutes to 7 minutes during purification to create optimal results. Therefore, by increasing our first incubation time, we avoided larger fragments and obtained only targeted fragments without any loss.

Gel electrophoresis results from the library preparation step need to be evaluated carefully. Even at very low concentrations, we were able to achieve optimum sequencing results as long as the fragment sizes were the correct bp length. When fragment sizes were not optimal, even in samples with high DNA concentrations, getting good results was not feasible.

According to the variant analysis interface, samples that have a QC score under 80% should be avoided. However, we found that samples with QC scores as low as 60% gave excellent results.

Variant analysis is another one of the most important steps because of the difficulties in differentiating between background dirt and real variants. In these kinds of situations, users must be highly focused to eliminate and discriminate artifacts. One of the key parameters for variant analysis is the forward/reverse variant reads ratio. Normally, ratios should be around one. Especially for somatic variant analysis, allele fraction is a critical parameter due to tissue heterogeneity. Germline analysis however, does not have such heterogeneity and an expected allele fraction ratio is 1 or 0.5. It is also very important to have the tumor percentage and the necrosis percentage of the section. This information gives insight to the users for identifying the real variants.

During bioinformatic analyses, it is imperative that an experienced medical geneticist be involved. It is also crucial to have the patient's clinical information and indications such as diagnosis, primary tumor site, previous treatments, and whether they were metastatic or not. Related to this information, variant pathogenicity and/or clinical significances may vary. Even when bioinformatic tools classify a variant as benign, a user still needs to carry out interpretations due to the possible clinical significance according to the patients' clinical findings or vice versa. In some cases, multiple benign variants might be detected but with a cumulative effect that might have clinical importance.

Population frequency is a meaningful factor for bioinformatic analyses because variant frequencies may vary according to ethnicity. Rare variant interpretations require a more detailed literature search of related clinical case reports or studies in consideration with patients' diagnoses. The source of these reports should be checked to eliminate redundant reports of the same cases in different databases.

Variants that are classified as variants of uncertain significance (VUS) or novel variants are the most challenging part of the bioinformatic analyses. The literature source of VUS might be insufficient, inconsistent and/or reported without clinical information. At this stage

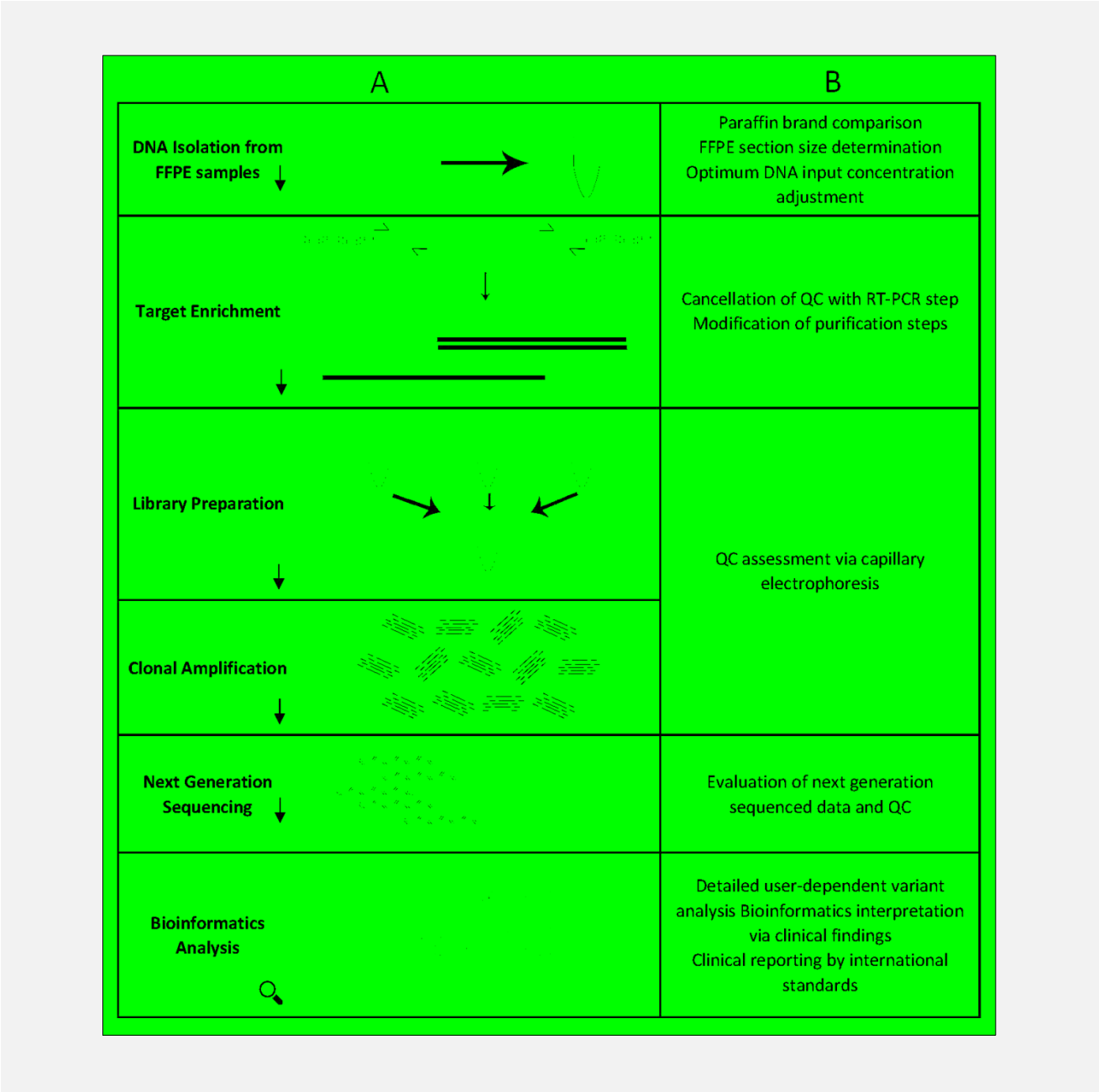

Figure 1.

Column A - Workflow of GeneReader next generation sequencing system.
Column B - Modifications processed in our study.

in-silico tools may provide useful information to estimate the functional impact of the variants. It should, however, not be forgotten that these tools only make predictions; therefore, to obtain more accurate estimations and higher quality data, multiple in-silico tools should be used.

From an overall point of view, a modification of the DNA extraction method had been adapted to monitor the variations in the initial protocol of the manufacturer. Modifications we brought to the initial protocol are detailed in below.

Quality control step:

According to manufacturer's protocol, real time PCR (QantiMIZE, Qiagen, Germany) quality control step was used following the procedure. From our results, it appears that these collected QC results have no real effect on the sequencing data, while we could have the

optimal data from the samples with lower QC. Thus, the quality control step was improved by using capillary gel electrophoresis.

DNA purification procedure:
According to the commercial purification protocol, the purification was not efficient enough. Hence, incubation periods were increased to increase the final elution quality when compared to the initial protocol.

Library quality control step:
Manufacturer`s protocol does not include any specific optimal library QC parameters that limits to have additional information about the variations in the study. However, the protocol we have modified in the present work used fragment size as the indicator, since the optimal results had been collected from the samples with proper fragment size but quite low concentrations after the library preparation step.

Bioinformatic analysis:
The sequence QC scores should also be deeply investigated during bioinformatic analysis to evaluate and differentiate which sequencing data can or cannot be analyzed. In this context, even though the bioinformatics tools (QCI Analyze - 1.5.0, Qiagen, Germany) present lower QC scores, the sequencing still can provide enough and qualitative data to analyze but with definitions and cautions.

DISCUSSION

NGS systems have become an integral tool in clinical genetic testing because of their pace, sensitivity, reliability, and accuracy. When compared with traditional methods, NGS's long and complicated protocols that generate vast amounts of data which need clinical interpretations. Due to complexity of the procedures, internal optimizations are required. In addition to the manufacturers' protocols, laboratory scientists need detailed written instructions that they can use as a laboratory handbook. This is often quite difficult to find in literature. In this study we share our optimization and validation process of a novel NGS system using FFPE tissue samples.
In this study, as the first step, we clarified material handling and sampling problems, which is fundamentally paramount for somatic studies. For this purpose, we refer to some basic parameters, such as tumor heterogeneity, tumor content, necrosis score and primary tumor site, which may be crucial for both the procedure application and the data analysis.
During the workflow, minor errors or modifications may affect assay quality and the results. To address this issue, we specified all possible problems and obstacles we encountered in every step and pointed to possible solutions and necessary modifications for all workflow.
In the data analysis step, we indicate the importance of evaluating the next generation sequencing results on a sample basis as well as general quality checks of the overall study. The fact that the samples with low quality may affect the general quality score and that the yields of other samples may be sufficient should be considered during evaluation. This one-sample-inefficiency may occur either way. Our sample might be degraded at the beginning or an error can be made at any point during the process.
Last but not the least, the challenging part of NGS workflows is the interpretation of the high volume of data. Each parameter that affects this output quality must be evaluated carefully. Here we suggest a precise data interpretation which covers all the mentioned parameters directly affecting the clinical report. Accurate and precise data handling is a must for gathering valuable information for proper patient and disease management.
The utilization of all these data still requires experienced geneticists. In addition to the qualified laboratory scientists and geneticists needed, all the optimizations should be made by each laboratory to achieve reproducibility and reliability in NGS workflows, especially when working with a brand-new system such as the GeneReader System in this study.
Although we used the QCI Interpret interface, there are some other commercially available bioinformatic tools to help with variant interpretation. It is, however, important to note that these analysis tools are only predictive. This means they cannot accomplish bioinformatic analyses as an autonomous drive. In fact, they only provide quick access to relevant information by data mining for users. The evaluation of this information always relies on the end user. In this study, we supported our analyses with HGMD (Human Gene Mutation Database), COSMIC (Catalogue Of Somatic Mutations In Cancer), ClinVar, NCBI (National Center for Biotechnology Information), VarSome (The Human Genomic Variant Search Engine), ExAC (The Exome Aggregation Consortium), 1000 Genome Frequency, ESP (Exome Sequencing Project), Ancestry, Ingenuity Knowledge Base, OMIM (Online Mendelian Inheritance in Man) databases and showed the significance of using different databases and in-silico tools comparatively.
In this study, we validated a novel NGS system for analysis of a multi-gene panel through development of complete NGS workflow to achieve optimal studies starting with FFPE samples. The difficulties in interpreting the vast amount of data produced mandate the need for an integrated clinical and laboratory team.

Acknowledgment:
We wish to acknowledge the help provided by Kaan Ihsan Eskut during the creation of NGS workflow and laboratory modifications figure.

Source of Support:
This study was supported by grants from Cukurova University Scientific Research Project (Project No. FYL-2017-7434).

Declaration of Interest:
The authors declare that they have no competing interests.

References:

1. Reinecke F, Satya RV, DiCarlo J. Quantitative analysis of differences in copy numbers using read depth obtained from PCR-enriched samples and controls. BMC Bioinformatics 2015;16:17 (PMID: 25626454).

2. Bisgin A, Boga I, Yilmaz M, Bingol G, Altintas D. The Utility of Next-Generation Sequencing for Primary Immunodeficiency Disorders: Experience from a Clinical Diagnostic Laboratory. Biomed Res Int 2018;2018:9647253 (PMID: 29888287).

3. Choudhary A, Mambo E, Sanford T. Evaluation of an integrated clinical workflow for targeted next-generation sequencing of low-quality tumor DNA using a 51-gene enrichment panel. BMC Med Genomics 2014;7:62 (PMID: 25395014).

4. Smith A, Boycott KM, Jarinova O. Lake Louise mutation detection meeting 2013: clinical translation of next-generation sequencing requires optimization of workflows and interpretation of variants. Hum Mutat 2014;35(2):265-9 (PMID: 24282140).

5. Jia H, Guo Y, Zhao W, Wang K. Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer. Sci Rep 2014;4:5737 (PMID: 25034901).

6. Sonmezler O., Boga I., Bisgin A.: 2020 Integration of Liquid Biopsies into Clinical Laboratory Applications via NGS in Cancer Diagnostics. Clin Lab 66: 5, (PMID: (32390404).

7. D'Antonio M, D'Onorio De Meo P, Paoletti D. WEP: a high-performance analysis pipeline for whole-exome data. BMC Bioinformatics 2013;14 Suppl 7(Suppl 7):S11 (PMID: 23815231).

8. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015;17(5):405-24 (PMID: 25741868).