The Journal of Molecular Diagnostics, Vol. 19, No. 5, September 2017

ELSEVIER

the Journal of Molecular Diagnostics

jmd.amjpathol.org

TECHNICAL ADVANCE

A Method to Evaluate the Quality of Clinical Gene-Panel Sequencing Data for Single-Nucleotide Variant Detection

Chung Lee,[*†] Joon S. Bae,[*] Gyu H. Ryu,[‡] Nayoung K.D. Kim,[*] Donghyun Park,[*] Jongsuk Chung,[*§] Sungkyu Kyung,[*¶] Je-Gun Joung,[*] Hyun-Tae Shin,[*] Seung-Ho Shin,[*] Younglan Kim,[*] Byung S. Kim,[*] Hojun Lee,[*] Kyoung-Mee Kim,[‖] Jung-Sun Kim,[‖] Woong-Yang Park,[*†§] and Dae-Soon Son[*]

From the Samsung Genome Institute and the Office of R&D Strategy and Planning,[‡] Samsung Medical Center, Seoul; the Department of Health Sciences and Technology,[†] Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul; the Department of Molecular Cell Biology,[§] Sungkyunkwan University School of Medicine, Suwon; the Department of Bioinformatics and Life Science,[¶] Soongsil University, Seoul; and the Department of Pathology,[‖] Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea*

Accepted for publication June 5, 2017.

Address correspondence to Woong-Yang Park, M.D., Ph.D., or Dae-Soon Son, Ph.D., Samsung Genome Institute, Samsung Medical Center, 81 Irwon-Ro, Gangnam-Gu, Seoul 06351, Republic of Korea. E-mail: woongyang.park@samsung.com or ds3.son@samsung.com.

Customized gene-panel tests, based on next-generation sequencing, have demonstrated their usefulness in a plethora of clinical settings. As with other clinical diagnostic techniques, gene-panel sequencing for clinical purposes requires precise quality control (QC) measures to ensure its reliability. Only detected variants are currently recorded in clinical reports; however, identifying whether a nondetected variant is a true or false negative is regarded essential in a clinical setting and, thus, a comprehensive QC measure is in demand. Conventional QC metrics, such as mean coverage and uniformity, are considered inadequate for such an evaluation. As such, a more specific measure focused on clinically important variants is herein proposed. In this study, we suggest a new scoring method for assessing the quality of clinical gene-panel sequencing data, specifically for the detection of a set of single-nucleotide variants. The performance of the method was analyzed using 2295 clinical samples (1012 formalin-fixed, paraffin-embedded and 1283 fresh-frozen tissues), and was shown to provide additional information that conventional methods do not show, such as mean depth and uniformity. Customized sequencing protocols, which include QC criteria, have been optimized by each genomic laboratory. The pass rate scoring method proposed in this study provides an appropriate QC response variable for the customized panel, which strengthens the reliability of calls on clinically relevant variants implicated in clinical reports. *(J Mol Diagn 2017, 19: 651—658; http://dx.doi.org/10.1016/j.jmoldx.2017.06.001)*

Next-generation sequencing technology has been successfully extended to the field of clinical diagnostics, such as genetic testing for cancer patients and the corresponding targeted anticancer drug prescription. With the identification of new biomarkers, many clinical genetic tests, including FoundationOne,[1] perform targeted deep sequencing (panel sequencing) to detect specific regions or variants. This is a straightforward method for investigating the presence or absence of known variants.

Variant caller algorithms (eg, MuTect)[2] can detect somatic variants when there is sufficient evidence, such as the presence of supporting reads. As implied by the name, these algorithms focus on increasing detection sensitivity without specifically calling for the wild type. In other words, the chromosomal position that is not called has the possibility of producing both a true and false negative. However, if clinical reports only document detected variants, the remainder may be mistaken as being true negatives. It is, therefore, important to review hotspot confidence levels using expert judgment along with read depth.

Several conventional programs have been developed to evaluate data quality for certain levels of raw sequencing

Supported by Ministry of Food and Drug Safety grant 16173MFDS004.
C.L., J.S.B., and G.H.R. contributed equally to this work.
Disclosures: None declared.

http://dx.doi.org/10.1016/j.jmoldx.2017.06.001

data. For example, the FASTX-Toolkit (*http://hannonlab.cshl.edu/fastx_toolkit*, last accessed April 19, 2017) provides various quality control (QC) metrics from FASTQ files. FastQC (*http://www.bioinformatics.babraham.ac.uk/projects/fastqc*, last accessed April 19, 2017), operated with the Picard tool, also offers various QC metrics, including basic statistics, per base sequence quality (score), GC content distribution, identification of the most duplicated reads, distribution of sequence length, and kmer content.[3] Recently, new programs, such as PRINSEQ,[4] NGS QC Toolkit,[5] and QC-Chain,[6] have been released. Among the various QC factors, mean depth and uniformity of coverage have been widely used to represent overall quality of sequencing data.[7–9] Mean depth refers to the average depth of targeted regions, whereas uniformity (>*p*%) refers to the rate of sites that exceed the *p*% of the mean depth. Other measures, such as the percentage of all target bases achieving >100× coverage (on target rate at 100×) or the average insert size for a paired-end library, have also been used to estimate data quality.[10]

The purpose of QC measurements is to effectively summarize the entire sequencing data. However, there are limitations to panel sequencing that require detection of specific variants with high accuracy, because QC measures are only representative of the general quality of the data. As quality varies across the regions targeted for sequence analysis, current QC tools cannot provide a clear answer regarding the reliability of results, which state the presence or absence of mutations of interest. For example, it is difficult to use FastQC for QC measurement in targeted panel sequencing because the depth of the variants of interest cannot be explained by high mean depth or uniformity. Therefore, a new QC measurement to estimate each targeted variant, which also satisfies various clinical purposes, is required. This study provides a new method for determining overall panel sequencing quality. The proposed method primarily focuses on the presence or absence of targeted variants for which detection reliability, based on the depth of each site, can be precisely estimated.

Materials and Methods

Samples

Tumor specimens included formalin-fixed, paraffin-embedded (FFPE) and fresh-frozen (FF) samples obtained from the tissue bank, Pathology Department, and individual researchers at the Samsung Medical Center (Seoul, Republic of Korea) (Table 1). This study was approved by the Institutional Review Board of the Samsung Medical Center (2015-01-112).

DNA Extraction, Library Preparation, Sequencing, and Variant Calling

Genomic DNA was extracted from FF samples using QIAamp DNA Mini kits (Qiagen, Valencia, CA), and from FFPE samples using a Promega Maxwell 16 CSC DNA FFPE kit (Promega, Madison, WI) and QIAamp DNA FFPE Tissue kit. DNA concentration and purity were checked using a Qubit 2.0 fluorometer (Life Technologies, Grand Island, NY) and Nanodrop 8000 UV-Vis spectrometer (Thermo Scientific, Waltham, MA). The degree of DNA degradation was measured using a 2200 TapeStation Instrument and by real-time PCR (Agilent Technologies, Santa Clara, CA), according to the manufacturer's instructions. Our criterion of DNA input amount for sequencing was 200 and 300 ng for the FF and FFPE samples, respectively. However, a lower amount of input DNA was used when it was difficult to obtain additional samples. To generate a sequencing library of target genes with the SureSelect XT Reagent Kit, HSQ (Agilent Technologies), DNA was sheared using a Covaris S220 instrument (Covaris, Woburn, MA). The paired-end sequencing library was purified and amplified using a barcode tag, and library quality and quantity were subsequently determined. Sequencing, using the 100-bp paired-end mode of the TruSeq Rapid PE Cluster kit (Illumina, San Diego, CA) and TruSeq Rapid SBS kit (Illumina), was performed on a HiSeq 2500 sequencing platform (Illumina).

Raw sequencing reads were aligned to the human reference genome (the Genome Reference Consortium Human genome build 37/human genome build 19; *http://genome.ucsc.edu*, last accessed June 1, 2017), using BWA software version 0.7.5a (*http://bio-bwa.sourceforge.net*, last accessed June 5, 2017)[11] to generate SAM files. SAMtools version 0.1.18 (*http://samtools.sourceforge.net*, last accessed June 5, 2017),[12] GATK version 3.1 (*https://software.broadinstitute.org/gatk*, last accessed June 5, 2017),[13] and Picard version 1.93 (*https://broadinstitute.github.io/picard*, last accessed June 25, 2017) were used for sorting the SAM/BAM files, local realignment, and duplicate markings, respectively. Reads were filtered to remove duplicates, improper pairs, and off-target reads.

Table 1 Quality Control Measurement Averages for FFPE and FF Tissues

Sample type	Samples, *n*	Depth, ×	Uniformity (>50%)	PR score, %
FFPE	1012	727.8 ± 279.2	0.876 ± 0.081	92.9 ± 21.9 (PR_{200})
FF	1283	954.3 ± 181.5	0.849 ± 0.034	97.3 ± 8.9 (PR_{500})
Total	2295	854 ± 255.8	0.86 ± 0.061	95.3 ± 16.1

Data are expressed as means ± SD unless otherwise indicated.

FF, fresh-frozen; FFPE, formalin-fixed, paraffin-embedded; PR, pass rate.