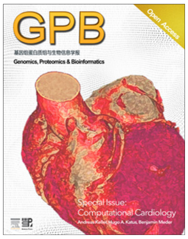
HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



ORIGINAL RESEARCH

The Role of Quality Control in Targeted Next-generation Sequencing Library Preparation

Rouven Nietsch [1,a,#], Jan Haas [1,2,b,#], Alan Lai [1,c], Daniel Oehler [1,2,d], Stefan Mester [1,2,e], Karen S. Frese [1,2,f], Farbod Sedaghat-Hamedani [1,2,g], Elham Kayvanpour [1,2,h], Andreas Keller [3,i], Benjamin Meder [1,2,*,j]

[1] *Institute for Cardiomyopathies, Department of Internal Medicine III, University of Heidelberg, 69120 Heidelberg, Germany*
[2] *German Centre for Cardiovascular Research (DZHK), Heidelberg/Mannheim, Germany*
[3] *Chair for Clinical Bioinformatics, Medical Faculty, Saarland University, 66123 Saarbrücken, Germany*

Received 29 January 2016; revised 31 March 2016; accepted 5 April 2016
Available online 28 July 2016

Handled by Songnian Hu

KEYWORDS

Next-generation sequencing;
Quality control;
Library preparation;
Target enrichment;
Sequence variants

Abstract **Next-generation sequencing** (NGS) is getting routinely used in the diagnosis of hereditary diseases, such as human cardiomyopathies. Hence, it is of utter importance to secure high quality sequencing data, enabling the identification of disease-relevant mutations or the conclusion of negative test results. During the process of sample preparation, each protocol for **target enrichment library preparation** has its own requirements for **quality control** (QC); however, there is little evidence on the actual impact of these guidelines on resulting data quality. In this study, we analyzed the impact of QC during the diverse **library preparation** steps of Agilent SureSelect XT **target enrichment** and Illumina sequencing. We quantified the parameters for a cohort of around 600 samples, which include starting amount of DNA, amount of sheared DNA, smallest and largest fragment size of the starting DNA; amount of DNA after the pre-PCR, and smallest and largest fragment size of the resulting DNA; as well as the amount of the final library, the corresponding

[*] Corresponding author.
E-mail: benjamin.meder@meduni-heidelbergde (Meder B).
[a] ORCID: 0000-0003-0751-471X.
[b] ORCID: 0000-0002-8040-8289.
[c] ORCID: 0000-0003-0916-9227.
[d] ORCID: 0000-0002-0435-4119.
[e] ORCID: 0000-0001-8799-4383.
[f] ORCID: 0000-0003-1822-8559.
[g] ORCID: 0000-0002-3266-0527.
[h] ORCID: 0000-0001-7285-2825.
[i] ORCID: 0000-0002-5361-0895.
[j] ORCID: 0000-0003-0741-2633.
[#] Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2016.04.007

smallest and largest fragment size, and the number of detected variants. Intriguingly, there is a high tolerance for variations in all QC steps, meaning that within the boundaries proposed in the current study, a considerable variance at each step of QC can be well tolerated without compromising NGS quality.

Introduction

Before the advent of next-generation sequencing (NGS), genetic testing was realized by Sanger sequencing [1], which meant analyzing a gene exon-wise or amplicon-wise in a relatively elaborate, time-consuming and costly way. This substantially limited the number of genes that could be examined in parallel. In 2005, the first commercial NGS systems were introduced, yielding up to 20 megabase (mb) output per run [2]. Genetic studies have gained enormously from NGS over the past years. There is no doubt that NGS has matured to a technique that is highly reliable if performed by following certain rules [3]. Today, it is to replace Sanger sequencing not only in research, but also in clinical applications. One major step in this path is the first marketing authorization for an NGS instrument (Illumina's MiSeqDx) by the Food and Drug Administration of the United States (US FDA) [4]. Besides such optimism, less certainty exists on the required standards for ensuring sequencing quality. It is also debated whether precise bench-work or careful data analysis is more important. For gene panel or target enrichment, a number of distinct protocols based on, *e.g.*, PCR, hybridization, or selective circularization, have been developed [5]. For each of these methods, stringent quality control (QC) steps were introduced to ensure a consistent data quality of the resulting NGS process. On the other hand, QC is expensive and requires significant hands-on time. Moreover, it is virtually unknown how QC could affect the sequencing process in case of abnormal results obtained.

With respect to the influence of data analysis on sequence quality, numerous studies and recommendations provide a guideline toward reproducible and comparable NGS results [6]. This so-called post-sequencing QC typically starts with raw-data processing covering measures of base quality, nucleotide distribution, GC content distribution, and read duplication rate. Then post-alignment QC is mostly based on the BAM-files, which provides QC parameters like the number of mappable reads, mapping quality, depth of coverage, and the number of reads mapped to the target region. Finally, on the variant level, data quality can be analyzed by the transition/transversion (Ti/Tv) ratio, heterozygosity rate, or occurrence in variant databases [3].

In this study, we investigated, using a large-scale dataset from nearly 600 patients, the impact of the many different QC phases during library preparation on the resulting sequencing data, and provided a recommendation on library quality requirement.

Results and discussion

Impact of library preparation on NGS quality

While it is broadly appreciated that post-processing of sequencing data is inevitable, less certainty exists on the influence of wet-lab steps during library preparation on the final quality of variant calls. Hence, we collected data from stringent QC during a larger-scale diagnostic target-enrichment study, which has underlined the high analytical quality and feasibility of NGS in a clinical genetic diagnostic setting [3].

Our aim in the current study was to investigate whether sequencing results are affected by quality differences during the library preparation. We thus tested if QC during the diverse library preparation protocol can foresee any impact on the quality of the resulting sequencing library. To do so, we first examined the statistical distributions of all assessed QC parameters over a set of 581 patient samples undergoing SureSelect target enrichment (referred as "main cohort" hereafter). The QC steps examined include initial DNA-shearing and cleanup (QC1), pre-PCR and clean-up (QC2), as well as post-PCR and clean-up (QC3). **Figure 1** depicts Violin plots of the distributions of the following parameters at each QC step: DNA concentration, largest fragment size, and smallest fragment size, which are all approximately normal.

Then we tested by Pearson correlation, as well as Spearman and Kendall, whether the aforementioned parameters measured at different steps of the library preparation protocol exert significant impact on the quality of the resulting sequencing library. Surprisingly, we did not find any obvious correlation between the different QC steps and library QC measures, all correlation coefficients were below 0.4 (**Figure 2**). Next, we calculated linear correlations of the different QC steps with the total number of detected sequence variants as an indicator of final sequence quality. The lower triangular part of the matrix in Figure 2A shows the absolute values of the Pearson correlation coefficients between every possible pair of parameters, whereas the upper triangular part shows the scatter plots. Figure 2B shows the absolute values of the Spearman correlation coefficients below the diagonal and the absolute values of the Kendall correlation coefficients above the diagonal. Again, we did not detect obvious correlations.

Robustness of library preparation for NGS

To further underline these findings, we applied Mann–Whitney *U*-test and Székely's distance correlation on the total number of variant calls to rule out the possibility of undetected correlation in outliers and dependency of variant calls. As shown in Figure 1, the horizontal red bars indicate the total number of variant calls on the top 10% of the study population for each parameter and the blue bars on the bottom 10%, respectively. At QC1, the numbers of variant calls differ significantly between the bottom 10% population and top 10% population in terms of the smallest fragment size obtained ($P = 0.04$; *U*-test). According to Székely's distance correlation coefficients (**Table 1**), there is a weak dependency between the smallest fragment size and number of variant calls at QC1 ($R = 0.25$), whereas the distance correlation coefficients are consistently less than 0.2 for the remaining parameters. This