Biomolecular Detection and Quantification 5 (2015) 30–37

Contents lists available at ScienceDirect

Biomolecular Detection and Quantification

journal homepage: www.elsevier.com/locate/bdq

Research paper

Control for stochastic sampling variation and qualitative sequencing error in next generation sequencing

Thomas Blomquist [a], Erin L. Crawford [b], Jiyoun Yeo [b], Xiaolu Zhang [b], James C. Willey [a,b,*]

[a] Department of Pathology, University of Toledo Health Sciences Campus, Toledo, OH 43614, USA
[b] Department of Medicine, University of Toledo Health Sciences Campus, Toledo, OH 43614, USA

ARTICLE INFO

Article history:
Received 1 June 2015
Received in revised form 12 August 2015
Accepted 18 August 2015
Available online 28 August 2015

Keywords:
NGS
Targeted
Sequencing
Diagnostics
Internal standards

ABSTRACT

Background: Clinical implementation of Next-Generation Sequencing (NGS) is challenged by poor control for stochastic sampling, library preparation biases and qualitative sequencing error. To address these challenges we developed and tested two hypotheses.
Methods: Hypothesis 1: Analytical variation in quantification is predicted by stochastic sampling effects at input of (a) amplifiable nucleic acid target molecules into the library preparation, (b) amplicons from library into sequencer, or (c) both. We derived equations using Monte Carlo simulation to predict assay coefficient of variation (CV) based on these three working models and tested them against NGS data from specimens with well characterized molecule inputs and sequence counts prepared using competitive multiplex-PCR amplicon-based NGS library preparation method comprising synthetic internal standards (IS). Hypothesis 2: Frequencies of technically-derived qualitative sequencing errors (i.e., base substitution, insertion and deletion) observed at each base position in each target native template (NT) are concordant with those observed in respective competitive synthetic IS present in the same reaction. We measured error frequencies at each base position within amplicons from each of 30 target NT, then tested whether they correspond to those within the 30 respective IS.
Results: For hypothesis 1, the Monte Carlo model derived from both sampling events best predicted CV and explained 74% of observed assay variance. For hypothesis 2, observed frequency and type of sequence variation at each base position within each IS was concordant with that observed in respective NTs ($R^2 = 0.93$).
Conclusion: In targeted NGS, synthetic competitive IS control for stochastic sampling at input of both target into library preparation and of target library product into sequencer, and control for qualitative errors generated during library preparation and sequencing. These controls enable accurate clinical diagnostic reporting of confidence limits and limit of detection for copy number measurement, and of frequency for each actionable mutation.

1. Introduction

Quantitative analysis of transcript abundance and/or sequence variant frequency are common applications of next generation sequencing (NGS) [1,2]. One important diagnostic NGS application includes accurate identification of clinically actionable sequence variation in tumors and the estimation of tumor cell fraction with the actionable mutation [2,3]. However, lack of appropriate quality control limits wider clinical diagnostic application of NGS in this context. For example, under-loading of target analyte into library preparation and/or library product into sequencer will result in analytical variation due to stochastic sampling [4]. At the same time, over-loading of prepared library onto sequencer will result in re-sampling of library amplicons from the same target analyte molecule, and without proper controls will give false assurance of adequate sampling. Moreover, qualitative errors in sequence generated by polymerase during library preparation and/or sequencing steps can confound accurate estimation of the true cellular fraction containing clinically actionable sequence mutations [4,5].

Thus, for diagnostic NGS applications, it is important to control for several sources of analytical variation, including sample loading into library preparation, efficiency of target amplification in library preparation, loading of prepared NGS library onto a sequencing platform, and the combined polymerase error rates throughout library preparation and sequencing [6–8]. Currently, the most prevalent practice is to rely on sequence count data alone to

* Corresponding author.
E-mail address: james.willey2@utoledo.edu (J.C. Willey).

http://dx.doi.org/10.1016/j.bdq.2015.08.003

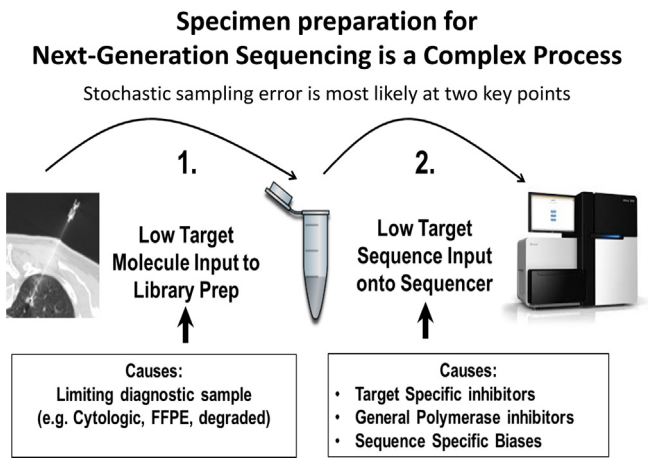

Fig. 1. Overview of specimen preparation for Next-Generation Sequencing. This schematic illustrates our hypothesis that two primary points of stochastic sampling error along the continuum of Next-Generation Sequencing (NGS) library preparation and sequencing can account for observed analytical variation in targeted PCR based NGS assays.

provide quality control for each potential source of analytical variation. For example, many recently developed programs seek to quantify the fractional representation of actionable tumor mutations, and enumeration of sequence read counts are the only source of data for assay variance analysis [3–5,9,10]. While these approaches address many issues, they provide false assurance regarding control for stochastic sampling variation due to low input of sample into the library preparation, and do not provide frequency limit of detection for each type of base substitution, insertion and deletion at each base position, in each target analyte [2,4]. Recent barcoding methods combined with bait-capture targeted sequencing provide better control for low sample input while, again, using only sequence count data to estimate analytical variance [1,4,5,11–13]. However, these methods do not provide a way to assess limit of detection for observed biological sequence variation [12], and the bait-capture method is associated with 100–1000-fold loss in signal [4]. Signal loss is a particular liability for analysis of small or degraded specimens, such as those routinely encountered in the clinical setting [3]. Furthermore, sequencing read counts are not always concordant with number of molecules "captured" during library preparation, resulting in false negative results [9]. In addition, it is less well recognized that if the number of target analyte molecules loaded into the library preparation is low the analyte may be poorly quantified due to over-amplification of a stochastically sampled specimen, regardless of the number of analyte amplicons loaded into the sequencer. In order to address these challenges, we developed and tested two hypotheses.

Hypothesis 1. We hypothesized that analytical variation in target analyte quantification can be predicted by Poisson (i.e. stochastic) sampling effects at two primary points: (a) input of intact nucleic acid target molecules loaded into the library preparation reaction, and (b) input of derived amplicons from library preparation into the sequencer (i.e. sequence counts) (Fig. 1). Using Monte Carlo simulation we derived equations to predict assay coefficient of variation (CV) based on three working models: number of target molecules added to library preparation, number of target amplicons in library added to sequencer (i.e., sequence read count), or both (Fig. 1). We then tested these working models using cell lines with known allelic composition. Cell lines were mixed and prepared for NGS such that a broad range of limiting allelic molar proportions and/or sequence read counts were observed. Each target allele was measured relative to a known number of synthetic internal standard molecules using a competitive multiplex-PCR amplicon-based NGS library preparation method [14].

Hypothesis 2. The accuracy of frequency measurement of acquired mutations in specimens (e.g., circulating plasma DNA, tumors, etc.) is confounded by both sampling error (described above and tested in hypothesis 1), and nucleotide substitution, insertion and deletion errors encountered during both library preparation steps and sequencing [3,9]. This latter, technically derived, sequence variation may to some extent be systematic for certain types of sequence variations, but may also vary largely on local sequence context. We hypothesized that technically derived base substitution, insertion and deletion frequencies observed at each base position in each target analyte is concordant with frequencies observed in respective synthetic internal standards present in the same reaction. In order to characterize the contribution of technically derived nucleotide sequence error rate, we measured the frequency of base substitution, insertion and deletion errors in a NGS data set derived from 213 normal airway brushing derived cDNA specimens with both ample intact nucleic acid loading and sequence counts. Each normal airway brushing derived cDNA specimen was mixed with a known number of synthetic internal standard (IS) molecules for each target analyte prior to competitive multiplex PCR amplicon NGS library preparation to determine if frequency of observed base substitution, insertion and deletions in each native target was concordant with frequency observed in each respective synthetic IS. If concordant, synthetic IS could provide control for both stochastic sampling in quantitative NGS, as well as control for technically derived sequencing error in qualitative NGS of low frequency alleles.

2. Methods

2.1. Sample preparation

Hypothesis 1. To test the effect of stochastic sampling on variance in allelic frequency measurements, genomic DNA (gDNA) was extracted by FlexiGene DNA kit (Qiagen) and quantified by NanoDrop (ThermoScientific, Wilmington, DE) spectrophotometry for two cell lines (H23 [ATCC CRL-5800] and H520 [ATCC HTB-182]). The cell lines were previously characterized as homozygous for opposite alleles at four polymorphic sites (rs769217, rs1042522, rs735482 and rs2298881) [14]. Cross-mixtures of these two cell-lines were performed so as to create a well characterized extreme limiting dilution of each of the four bi-allelic loci (see Mixing design in Supplementary Table 3). These limiting dilutions of alleles were then loaded into the library preparation (see Section 2.3), then limiting dilutions of NGS libraries were added to the Illumina HiSeq 2500 flow cell (see Section 2.3).

Hypothesis 2. In order to characterize the base-specific substitution, insertion and deletion rates imparted by combined library preparation and sequencing error, we used 213 normal human bronchial epithelial cell (NBEC) cDNA specimens. These specimens were obtained as part of the ongoing Lung Cancer Risk Test (LCRT) study at the University of Toledo Medical Center [15]. Approval for specimen acquisition for this study was obtained by the institutional review board at the University of Toledo Medical Center. These samples were chosen based on several key features: (1) they represent a source of normal nucleic acid templates with presumably low, or absent, acquired somatic mutations. (2) They were previously confirmed to have high copy numbers of intact template for each native target, which minimized chance that stochastic sampling of templates would confound assessment of combined library preparation and sequencing error on base-specific substitution, insertion and deletion rates. (3) Competitive synthetic IS