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

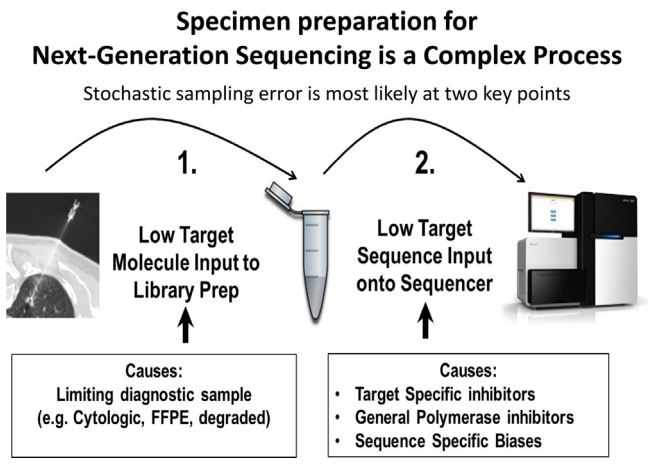

Fig. 1. Overview of specimen preparation for Next-Generation Sequencing. This schematic illustrates our hypothesis that two primary points of stochastic sampling error along the continuum of Next-Generation Sequencing (NGS) library preparation and sequencing can account for observed analytical variation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

for targets comprised by the LCRT were cloned into plasmids, and selected as pure clonal isolates, with Sanger sequencing confirmation of final sequence. This additional purification step was taken to eliminate any potential errors introduced by synthesis. We reason that these pure clonal competitive IS will have a frequency of technically acquired base substitutions, insertions and deletions that is similar to the native templates during the combined library preparation and sequencing steps.

2.2. Development of model to predict analytical variation due to stochastic sampling variation in NGS

Hypothesis 1. To test the hypothesis that analytical variation is dependent on both target analyte native template molecules added into library preparation reaction and resultant amplicon molecules added to sequencer, we developed three working models using Monte Carlo simulation and derived equations to predict expected assay coefficient of variation (CV) (Fig. 1 and Supplementary Method—Model generation). These three models and their equations were based on: target molecules in library added to sequencer (i.e., sequence read counts; Model 1), target native molecules added to library preparation (Model 2), or both (Model 3). This model is based, in part, on a model of biallelic genetic drift provided by Dr. Stephen P. DiFazio that can easily be simulated in excel <http://www.as.wvu.edu/~sdifazio/popgen_12/labs/GeneticDriftSim.pdf, last accessed 06.08.15>. We reasoned that population based founding effects that result in genetic drift of bi-allelic loci should operate statistically in the same way as stochastic sampling of a bi-allelic locus present in a test tube in the laboratory setting, and that the act of pipetting and sampling the specimen DNA is analogous to a founding effect seen in population genetics. We further reasoned that there were two primary founding (i.e., stochastic sampling) effects present in the lab test tube analogy; (1) initial pipetting of the specimen into library preparation reaction, and (2) loading of the prepared library onto the sequencer and the number of sequencing counts enumerated for each target template (Fig. 1 and Supplementary Method—Model generation). This model was varied for both the number of input molecules, as well as number of sequence reads derived (Supplementary Method—Model generation). This then produced a rich data set, from which three equations were derived by best curve fit analysis (Supplementary Method—Model generation). These derived equations were then tested against empirically derived data from cross-mixtures of cell lines to predict observed assay variance in targeted NGS (see Section 2.1).

2.3. NGS library preparation: targeted competitive multiplex-PCR

2.3.1. Cell line cross-mixture specimens

Each of four target analytes was PCR-amplified in samples derived from the cross-mixture of two cell-lines (see Mixing design in Supplementary Table 3) that had each been mixed with a known number of synthetic competitive internal standard (IS) molecules as previously described (Supplementary Table 1) [14].

2.3.2. NBEC cDNA specimens

Each of 30 target analytes (two target assays for each of 15 genes) was PCR-amplified in the presence of a known number of respective synthetic competitive IS molecules as previously described (Supplementary Table 2) [14]. Prepared libraries were then sent for Illumina HiSeq 2500 sequencing service at the University of Michigan, Genomics Core facility.

2.3.3. Internal standard mixture preparation

Each competitive IS was designed to contain six nucleotide differences from target analyte native template (NT) that enabled reliable differentiation between IS between IS and NT during post-sequencing data analysis (Supplementary Tables 1 and 2) [14]. For IS used in the analysis of cell line cross-mixture samples, following synthesis, each IS was PCR-amplified with specific primers to ensure full length product, isolated by gel electrophoresis, quantified using NanoDrop, and mixed with IS for other analytes at equivalent concentration to prepare an IS mixture [14]. IS used in analysis of NBEC cDNA samples were prepare by Accugenomics, Inc. (Wilmington, NC). Briefly, following synthesis IS were cloned in bacteria and purified to ensure an accurate and uniform population of sequences for each competitive IS used (see Section 2.1).

2.3.4. NGS data analysis

FASTQ data files from the University of Michigan Genomics core facility were processed as previously described [14]. FASTQ files for hypothesis 2 in this study, pertaining to the LCRT reagents, were additionally processed using Blast 2.2.26+ command line with a Practical Extraction and Reporting Language (PERL) wrapper to automate feeding of reference and query sequences to the Blast command line interface (reference sequences in Supplementary Table 2). This same PERL script then identified and stored the frequency of each Blast result for each template and for the type of base substitution, insertion or deletion that was identified across all reads in a Hash of Hashes of Hashes data table configuration (sequence error frequencies in Supplementary Tables 4 and 5). PERL wrapper for Blast 2.2.26+, and the input parameters used for Blast to enumerate base substitution, insertion and deletion frequencies, is available upon request.

Because the goal of hypothesis 2 was to identify and characterize the base by base frequency of combined sequencing and library preparation errors, and not biological variation (which was tested in hypothesis 1), we surmised that the sequencing data (NT and IS) could be aggregated into two large sub-pools of subjects (Groups 1 and 2). This is feasible and beneficial for several reasons: (1) a combined data set of normal specimens with minimal biological sequence variation (Group 1 [115 NBEC specimen library preparations] and Group 2 [98 NBEC specimen library preparations], total 213 NBEC specimen library preparations), should provide adequate sampling of very rare technically derived base substitution, deletion and insertion events (1 in 1000–100,000) across each specimen pool. (2) If these normal specimens do indeed have minimal biological variation in sequence, there should be a high degree of concordance in base substitution, insertion and deletion rates between the NTs and their respective competitive IS present in the same specimen (Supplementary Table 2). (3) By splitting the sequencing data into two pools, we can, in a surrogate way, assess the performance of external NT controls versus competitive synthetic IS controls, for accurately measuring technically derived base substitution, insertion and deletion frequencies.

All final NGS summary counts and absolute quantification of molecules (where appropriate) are provided in Supplementary Tables 3–5.

3. Results

3.1. Controlling for stochastic sampling error in NGS

For the equation derived from both sequencing coverage and input molarity (Model 3; see Supplementary Method—Model generation), expected coefficient of variation (CV) was very close to observed (average [observed CV/expected CV] = 1.01) and explained 74% of observed assay variance (Fig. 2C). In contrast,

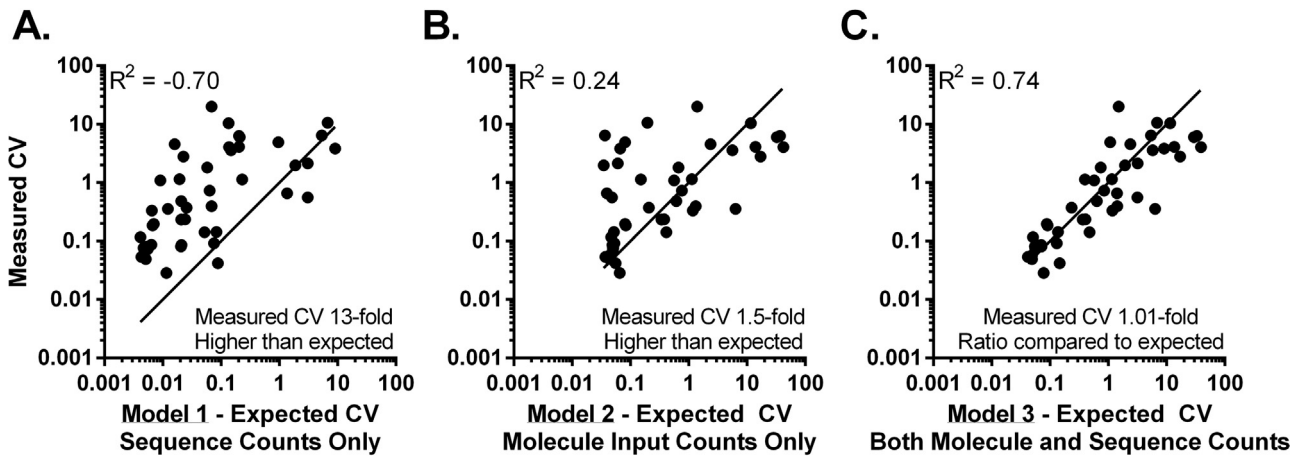

Fig. 2. Performance of Monte Carlo simulation models to predict observed assay variance.
Equations used to plot expected coefficient of variance (CV) are presented in Supplementary Methods—Model design. Measured CV was obtained by 46-quadruplicate technical measurements; 46 measurements of CV and calculated CV based on Models 1, 2 and 3 are available in Supplementary Table 3.

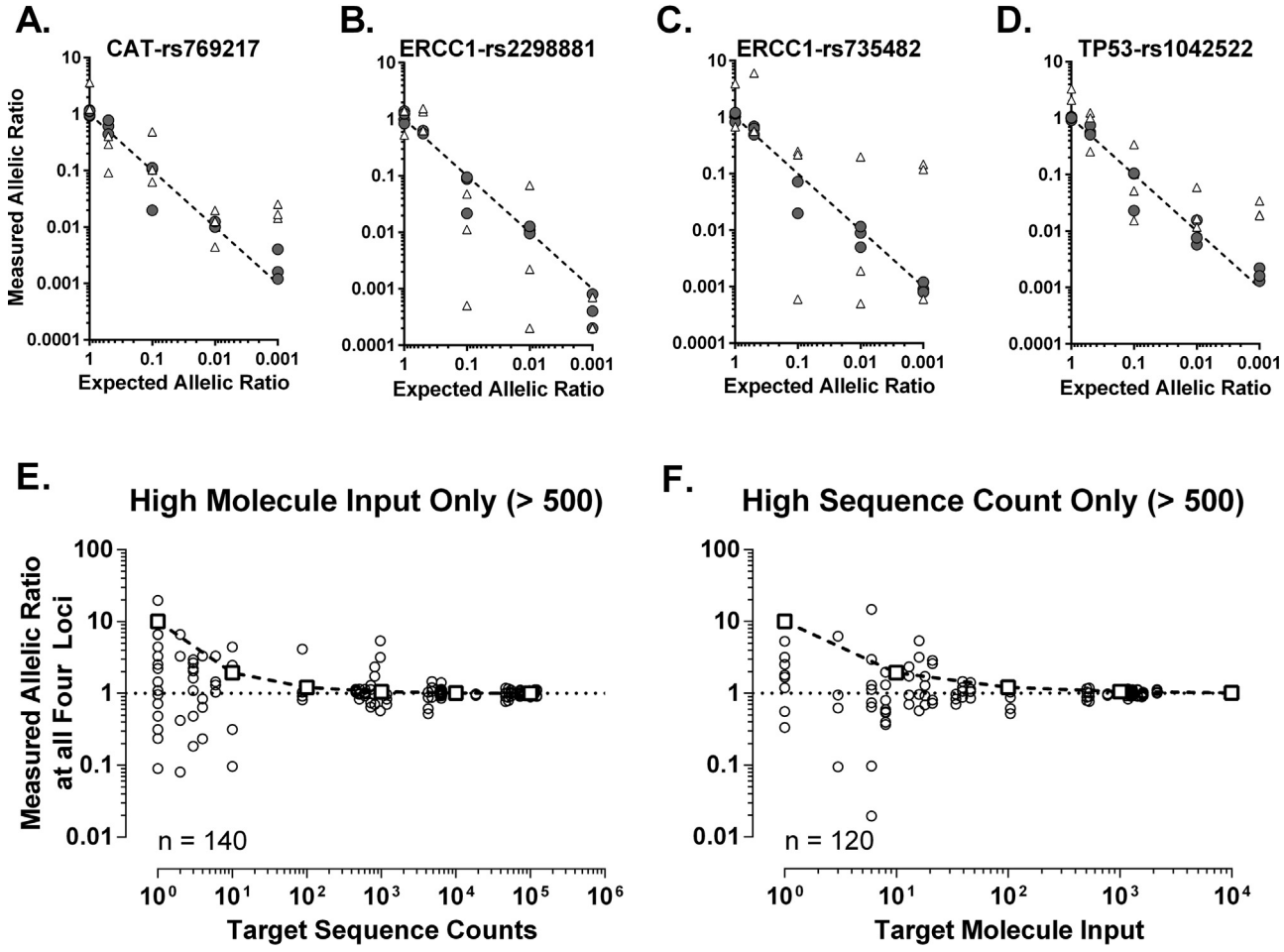

Fig. 3. Independent effects of sequence counts and sample molecule loading on measured allelic ratios.
(A–D) Effect of low molecule input into library preparation on measured allelic-ratio relative to expected. To eliminate effect of low sequence counts, only values based on at least 500 sequence counts were included. Closed circles = high molecule input (median = 3313 molecules each replicate; Supplementary Table 3, rows 33–58). Open triangles = low molecule input (median = 15 molecules each replicate; Supplementary Table 3, rows 60–85). Each data point is a single technical replicate. (E) Serially diluted PCR amplicon library samples from the undiluted 1:1 cell line mixture were loaded into sequencer. Effect of sequences counted (*X*-axis) on allelic-ratio (*Y*-axis) for each target with high molecule input (>500 molecules in each replicate). Combined results from all four loci are presented (Supplementary Table 3, rows 88–112). (F) Undiluted PCR amplicon library samples from serially diluted 1:1 cell line mixture were loaded into sequencer. Effect of target molecule number (*X*-axis) on allelic-ratio (*Y*-axis) for each target with high sequence count (>500 sequence read counts in each replicate) for each target (Supplementary Table 3, rows 115–139). Dashed line with open squares represents an expected frequency of error based on a Poisson distribution (Model 1 and 2). Mixing design of cell line DNA and titration of sequencing counts, and all measurements derived from these specimens are available as full and individual subset analysis tables in Supplementary Table 3.

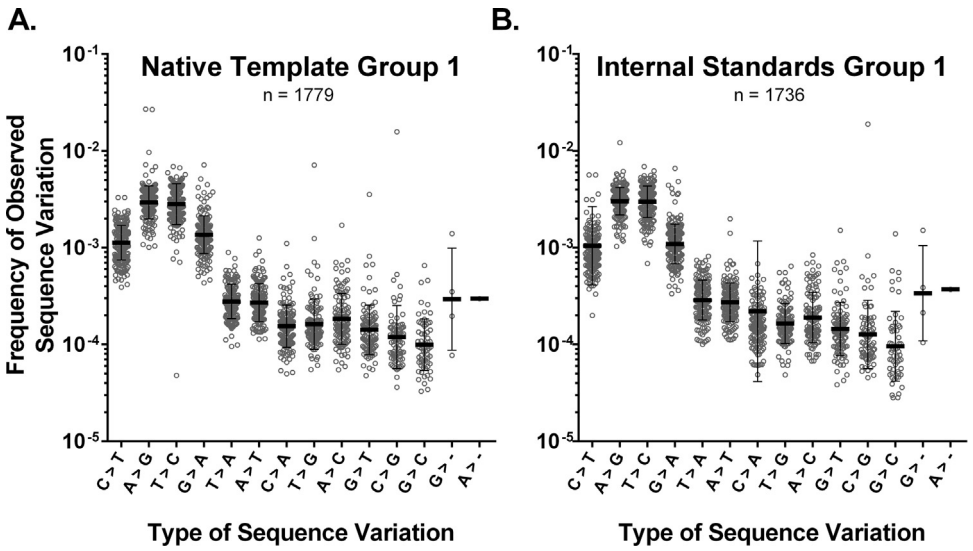

Fig. 4. Frequency plot of observed technically derived sequencing variation.
(A,B) Type of base substitution is plotted on *X*-axis. For example, "C > T" represents a transition from a cytosine to thymine base, and "G > –" represents a deletion of a guanine. The first base listed is the expected consensus base at that position based on sequences listed in Supplementary Tables 1 and 2. Each base position, for each template, and the frequency of that type of sequence variation is plotted as an individual data point along the *Y*-axis. In this figure, only Group 1 data are presented. Means and standard deviation error bars are plotted for each type of sequence variation. Group 2 data plotted essentially identically, and was moved as raw data to Supplementary Table 5.

observed CV was on average 13-fold, or 1.5-fold, higher than expected CV based on sequencing coverage (Model 1), or input molarity (Model 2), prediction models alone (Fig. 2A and B). For each assay, when input of target allele copies into library preparation was low (median of 15 molecules; open triangles) assay variance for measured allelic ratio was much higher, compared to high molecule input (median of 3313 molecules; closed circles) (Fig. 3A–D). Although there was an approximately 200-fold difference in median molecules loaded into library preparation for low and high loading conditions, sequence counts were high for both conditions (see Mixing design and raw data in Supplementary Table 3). When only specimens with high molecule input (>500 molecules) were assessed, variance in measured allelic ratio followed a Poisson distribution (plotted boxes and dashed line) for target sequence counts (Fig. 3E). Similarly, when only specimens with high sequence counts (>500 sequence counts) were assessed, variance in measured allelic ratio followed a Poisson distribution for target molecule input (Fig. 3F). All data presented in this section are available in Supplementary Table 3.

3.2. Controlling for qualitative sequencing error in NGS

Varying frequency of base substitutions were observed for all nucleotides, and rare frequency deletion events were detected for guanine and adenine bases (Fig. 4). In general, most observed base substitution rates were lower than 1 in 100 for each base location. Adenine to guanine and cytosine to thymine base transitions (purine–purine or pyrimidine–pyrimidine) were the most common type of sequence variation observed, followed by base tranversions (purine–pyrimidine or pyrimidine–purine) by a factor of approximately 10-fold lower frequency (Fig. 4). Furthermore, the type of sequence base substitution and its average frequency was concordant between NT and IS for Group 1 (Fig. 4). The coefficient of variation (CV) around the mean frequency of each type of base substitution was on average 0.28. This roughly translates to a standard deviation of 1.9-fold on either side of the population measurement mean for each type of sequence variation (2.8-fold detection limit with 95% confidence limits for detection of fold change). Data for Group 2 are available in Supplementary Table 5, and are nearly identical to those presented in Fig. 4. Bivariate plots of the frequency of technically derived sequence variation for NT and corresponding type of sequence variation for each base position in competitive IS for Groups 1 and 2 (see Section 2.3.4) are presented in Fig. 5A,C,E and G. Frequency of observed sequence variation in IS explained 93–94% of observed sequence variation in NT (Fig. 5A and C). Importantly, the vast majority of deviation from the regression line is explainable by the minimum sequence counts observed for the technically derived sequence variation (Fig. 5B and D). Concordance was slightly higher between NT and NT, or IS and IS comparisons between groups 1 and 2 respectively, with each explaining 96–97% of the frequency of base-specific sequence variation observed between the two groups (Fig. 5E and G). Again, deviation from the regression line in Fig. 5E and G was largely explainable by the minimum sequence counts observed for the rare technically derived sequence variation (Fig. 5F and H).

4. Discussion

Next-Generation Sequencing (NGS) technologies have the potential to disrupt a large number of technologies presently used in clinical diagnostics. However, NGS implementation in the clinical setting is impeded by a complex specimen and data analysis process (Fig. 1), and this is compounded by an equally complex goal of analyzing large multi-target panels. Because of the profound clinical implications on treatment decision management based on NGS methods, they should be held to the same analytical performance standards applied to other methods used in the clinical chemistry laboratory. In an effort to achieve this goal we developed a competitive multiplex PCR-based amplicon library preparation method that utilizes competitive IS (also known as internal amplification controls) [14]. The method enables control for sample overloading, excessive amplification cycles, other signal saturation effects and technical biases that can lead to inter-assay and inter-specimen variation in signal measurement. Data also suggested that this method controls for sub-optimal loading of sample into library preparation, suboptimal loading of library preparation into sequencer, and sequencer errors generated during library preparation and sequencing. We decided to address these important challenges by formulating and experimentally testing Hypotheses 1 and 2.

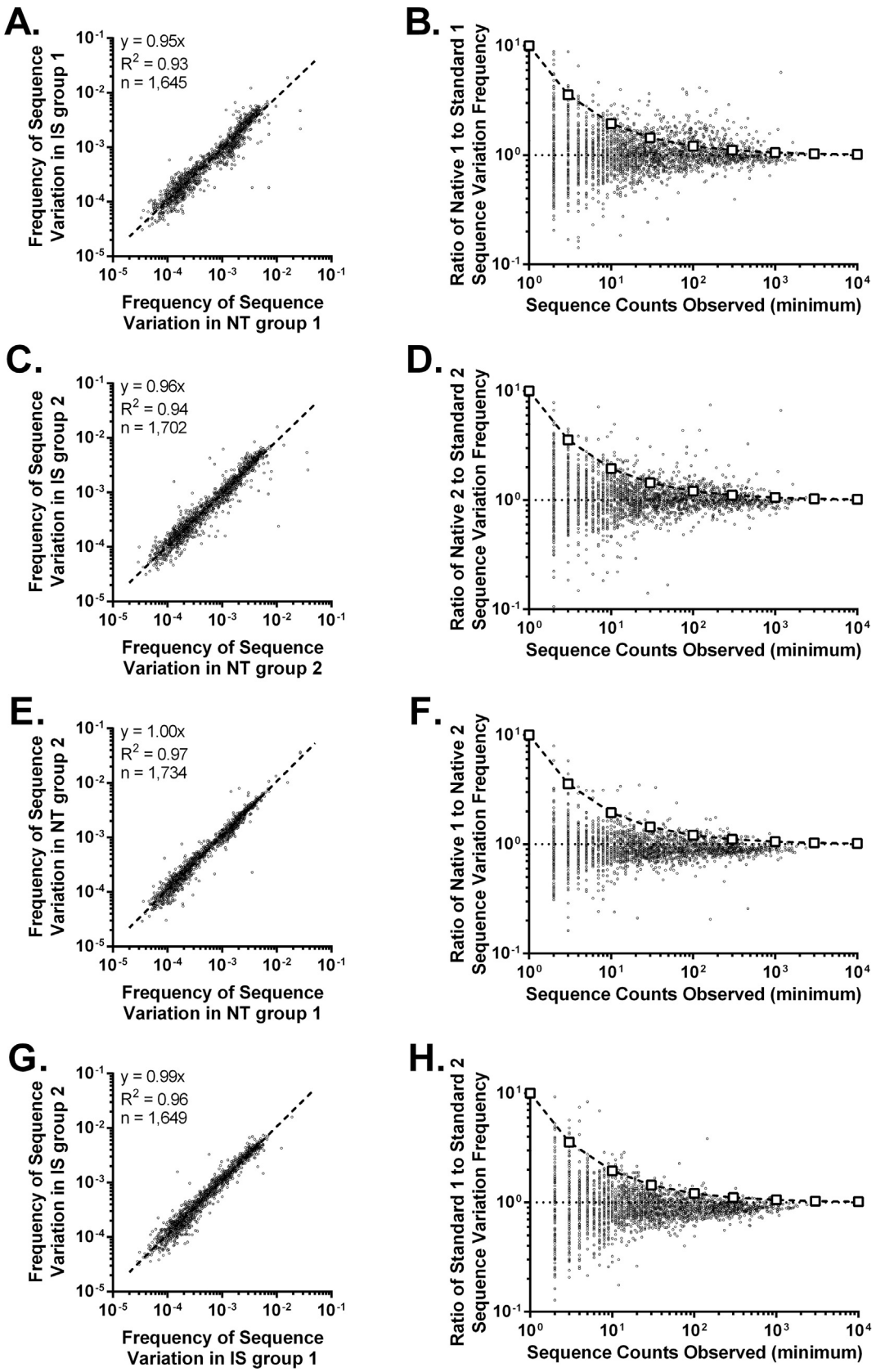

Fig. 5. Performance of competitive internal standards to measure frequency of technically derived sequence variation.
(A,C,E and G) Bivariate plots of measured sequence variation frequency, for each base position along the length of each native template (NT) and internal standard (IS) for Groups 1 and 2 (see Section 2: NGS data analysis). (B,D,F and H) Plots representing fold-deviation of NT:IS ratio away from regression line in respective plots A,C,E and G. Sequence counts observed (minimum) on the *X*-axis is the number of sequence counts for the observed type of sequencing error, and not the total number of sequence counts for that assay. Dashed line with open squares represents an expected frequency of error based on a Poisson distribution (Model 1).

Hypothesis 1 is supported by the data reported here. Specifically, the mathematical equation based on both NT loading into NGS library preparation and sequence read counts from NGS instrument (Monte Carlo simulation Model 3) predicted observed assay coefficient of variation in four targeted NGS assays (Figs. 2 and 3 and Supplementary Methods—Model design). While it remains to confirm the predictive value of this equation across other types of NGS library preparation methods and sequencing platforms, generalizability is likely based on the similarity of biochemical reactions involved.

Implementation of the Model 3 equation in the clinical setting may be particularly helpful when only one technical or biological replicate measurement is feasible. This is common in the clinical setting due to the limited size of biopsy, blood, plasma, and other specimens. In this context, the laboratory clinician will be asked to comment on the confidence in the measurement of target analyte, or frequency of a clinically actionable mutation present in a tumor specimen. Using this equation, the laboratory information system will be able to easily derive confidence intervals for reporting. As an example, this would simplify a decision regarding whether to direct treatment to an actionable mutation. Importantly, as is clear from Fig. 3F, large analytical variation from stochastic sampling will be observed if an insufficient concentration of target molecules is sampled, regardless of the concentration of amplification products sampled for loading into sequencer. This is why it is important to use quality control thresholds that address each of these sources of variation.

We now routinely implement the Model 3 equation in our NGS pipeline to determine the confidence limits for each value. By this approach, each value is associated with a confidence limit based on loading of sample into the library preparation and library preparation into sequencer. This is particularly important for transcriptome analysis, or assessment for tumor fraction containing actionable mutation because in each sequencing run, the representation of a particular transcript or actionable mutation among hundreds of samples included in the library may range over six $\log_{10}$. As such, it is not possible to ensure that sampling of each target transcript or actionable mutation and respective library product minimizes stochastic variation.

Hypothesis 2 also is supported by data from these studies. Specifically, the frequency of technically derived sequence variation for each NT was largely explained by that observed in the respective IS template (Fig. 5A and C). Furthermore, any deviation from the regression line observed (in Fig. 5A and C), was largely explained by stochastic sampling of low sequence counts for the technically derived sequence variation (Fig. 5B and D). Thus, with sufficient molecules loaded into the library preparation, and sequence counts obtained, the limit of detection of rare biological single nucleotide variations in native material can be easily determined using a competitive internal standard (Fig. 5), and is more accurate than the 2.8-fold change limit of detection estimated by the type of sequence variation only (Fig. 4). Importantly, base transitions were observed in approximately 10-fold excess compared to base transversion events (Fig. 4). The error rates observed here are specific to the chosen combination of specimen preparation, sequencing and data analysis pipeline methods, and should not be blindly applied to other NGS pipelines.

In summary, we present data that synthetic IS, in the context of a targeted competitive PCR amplicon library preparation method [14], control for both stochastic sampling in quantitative NGS and technically derived sequencing error in qualitative NGS detection of low frequency alleles. By applying quality-control parameters based on these experimentally validated models that predict key sources of NGS analytical variation, we can now accurately report confidence limits for NGS measurement of clinically important analytical targets, as well as provide an accurate limit of detection for observed base substitution, insertion and deletion rates at each base position within each native target. We are implementing quality control measures described here in analysis of promising diagnostic tests, including a lung cancer diagnostic test [16] and a lung cancer risk test [14]. Incorporation of these quality controls provides an analysis pathway consistent with previously reported College of American Pathologists (CAP) and Nex-StoCT guidelines for NGS diagnostics in the clinical setting [6–8].

Conflict of interest

Authors T.B., E.L.C. and J.C.W. are inventors of competitive internal standard mixtures for use in next generation sequencing reported in this manuscript. In addition, J.C.W. serves as a consultant for Accugenomics, Inc. which licenses the technology and has 5–10% equity in Accugenomics, Inc. These relationships do not alter the authors' adherence to all Biomolecular Detection and Quantification policies on sharing data and materials.

Funding

Significant portions of this study were paid for with funding provided by National Institutes of Health, National Cancer Institute (RC2-CA148572 and IMAT R21-CA138397) and National Heart Lung and Blood Institute (RO1-HL108016); and the University of Toledo Medical Center George Isaac Research Fund. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bdq.2015.08.003.

References

[1] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nat. Methods 5 (2008) 621–628.
[2] D.H. Spencer, M. Tyagi, F. Vallania, A.J. Bredemeyer, J.D. Pfeifer, et al., Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data, J. Mol. Diagn. 16 (2014) 75–88.
[3] K. Cibulskis, M.S. Lawrence, S.L. Carter, A. Sivachenko, D. Jaffe, et al., Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples, Nat. Biotechnol. 31 (2013) 213–219.
[4] G.K. Fu, W. Xu, J. Wilhelmy, M.N. Mindrinos, R.W. Davis, et al., Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 1891–1896.
[5] M.W. Schmitt, S.R. Kennedy, J.J. Salk, E.J. Fox, J.B. Hiatt, et al., Detection of ultra-rare mutations by next-generation sequencing, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 14508–14513.
[6] A.S. Gargis, L. Kalman, M.W. Berry, D.P. Bick, D.P. Dimmock, et al., Assuring the quality of next-generation sequencing in clinical laboratory practice, Nat. Biotechnol. 30 (2012) 1033–1036.
[7] N. Aziz, Q. Zhao, L. Bry, D.K. Driscoll, B. Funke, et al., College of American Pathologists' laboratory standards for next-generation sequencing clinical tests, Arch. Pathol. Lab. Med. 139 (2015) 481–493.
[8] A.S. Gargis, L. Kalman, D.P. Bick, S. da, C. ilva, D.P. Dimmock, et al., Good laboratory practice for clinical next-generation sequencing informatics pipelines, Nat. Biotechnol. 33 (2015) 689–693.
[9] G.M. Frampton, A. Fichtenholtz, G.A. Otto, K. Wang, S.R. Downing, et al., Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing, Nat. Biotechnol. 31 (2013) 1023–1031.
[10] H. Xu, J. DiCarlo, R.V. Satya, Q. Peng, Y. Wang, Comparison of somatic mutation calling methods in amplicon and whole exome sequence data, BMC Genomics 15 (2014) 244.
[11] J.A. Casbon, R.J. Osborne, S. Brenner, C.P. Lichtenstein, A method for counting PCR template molecules with application to next-generation sequencing, Nucleic Acids Res. 39 (2011) e81.
[12] C.B. Jabara, C.D. Jones, J. Roach, J.A. Anderson, R. Swanstrom, Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 20166–20171.

[13] I. Kinde, J. Wu, N. Papadopoulos, K.W. Kinzler, B. Vogelstein, Detection and quantification of rare mutations with massively parallel sequencing, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 9530–9535.
[14] T.M. Blomquist, E.L. Crawford, J.L. Lovett, J. Yeo, L.M. Stanoszek, et al., Targeted RNA-sequencing with competitive multiplex-PCR amplicon libraries, PLoS One 8 (2013) e79120.
[15] T. Blomquist, E.L. Crawford, D. Mullins, Y. Yoon, D.A. Hernandez, et al., Pattern of antioxidant and DNA repair gene expression in normal airway epithelium associated with lung cancer diagnosis, Cancer Res. 69 (2009) 8629–8635.
[16] J. Yeo, E.L. Crawford, T.M. Blomquist, L.M. Stanoszek, R.E. Dannemiller, et al., A multiplex two-color real-time PCR method for quality-controlled molecular diagnostic testing of FFPE samples, PLoS One 9 (2014) e89395.