# Benchmarking and optimization of a Next Generation Sequencing based method for transgene Sequence Variant Analysis in Biotherapeutic Cell Line Development


## Authors:

Joost Groot[1‡]
Yizhou Zhou[2‡]
Eric Marshall[1]
Patrick Cullen[1]
Thomas Carlile[1]
Dongdong Lin[1]
Chong-Feng Xu[3]
Justin Crisafulli[2]
Chao Sun[1]
Fergal Casey[1]
Baohong Zhang[1]
Christina Alves[2]

[1]Biogen, Genome Technologies and Computational Sciences, Cambridge MA
[2]Biogen, Protein Development, Cambridge MA
[3]Biogen, Analytical Development, Cambridge MA
‡These authors contributed equally to this study

Correspondence:
**E-mail:** *To whom correspondence should be sent:
Joost Groot. joost.groot@biogen.com
Yizhou Zhou. Yizhou.zhou@biogen.com



**Keywords:** Bioprocess engineering, Next-Generation Sequencing, CHO cells, Cell Line Development, Sequence Variant Analysis


**Abbreviations:** **NGS**, Next-Generation Sequencing; **CHO**, Chinese Hamster Ovary; **GOI**, Gene of Interest; **MAb**, Monoclonal Antibody; **Lc**, Antibody Light Chain; **Hc**, Antibody Heavy Chain; **CDS**, Coding Sequence; **PCR**, Polymerase Chain Reaction; **LC-MS**, Liquid Chromatograph Mass Spectrometry; **bp**, Base Pair; **SNV**, Single Nucleotide Variant; **LOD**, Limit of Detection.


## Abstract

In recent years Next-Generation Sequencing (NGS) based methods to detect mutations in biotherapeutic transgene products have become a key quality step deployed during the development of manufacturing cell line clones. Previously we reported on a higher throughput, rapid mutation detection method based on amplicon sequencing (targeting transgene RNA) and detailed its implementation to facilitate cell line clone selection. By gaining experience with our assay in a diverse set of cell line development programs, we improved the computational analysis as well as experimental protocols. Here we report on these improvements as well as on a comprehensive benchmarking of our assay. We evaluated assay performance by mixing amplicon samples of a verified mutated antibody clone with a non-mutated antibody clone to generate spike-in mutations from ~60% down to ~0.3% frequencies. We subsequently tested the effect of 16 different sample and NGS library preparation protocols on the assay's ability to quantify mutations and on the occurrence of false-positive background error mutations (artifacts). Our evaluation confirmed assay robustness, established a high confidence limit of detection of ~0.6%, and identified protocols that reduce error levels thereby significantly reducing a source of false positives that bottlenecked the identification of low-level true mutations.

## 1 Introduction

Clinical cell line development is a multi-step process aiming to generate high-yield clones that produce the protein of interest with desired qualities. This process typically requires screening of hundreds to thousands of clones and is traditionally time consuming and labor intensive. Automation and multiplex culturing platforms have greatly streamlined the workflow and accelerated the timeline to clinical proof of concept. Along with the advancement in cell line technologies, analytical assays must co-evolve to ensure the speed and throughput to support the selection of top clones. Among these, sequence variant analysis has demonstrated applications to support early stage cell line development efforts.[1–5]

Sequence variants are unintended amino acid substitutions, deletions, or insertions that occur during protein biosynthesis.[4] It can potentially impact higher order structures, and raises concerns about potency, immunogenicity, and product heterogeneity.[6,7] To mitigate the risk of sequence variants, switch of clones or additional process control strategies have to be applied, which can greatly increase the clinical program timeline and process complexity. Therefore, it is critical to reliably detect sequence variants at the cell line screening stage to select clones with favorable quality attributes.

Sequence variants originate from amino acid misincorporations during translation or from mutations in the DNA or RNA sequences of the transgenes. While LC-MS/MS based peptide mapping has been conventionally used for sequence variant detection in purified proteins, it often consumes substantial time and skilled human resources which limits its application to only a handful of top clones. On the other hand, NGS has recently evolved to serve as an orthogonal tool to detect relatively low levels of DNA or RNA mutations that are often not detectable via traditional Sanger sequencing technology.[4,5] In addition, NGS can detect synonymous mutations which are missed by LC-MS/MS.

A variety of NGS methodologies could be applied to sequence variant detection depending on the stage of clinical cell line development.[1–5,8] Among these methods, targeted RNA amplicon-sequencing is suitable for early stage cell line screening because of the throughput, fast turnaround time, cost and scalability to a large number of clones.[1] In addition, sequencing of mRNA/cDNA can capture both genetic mutations and transcriptional errors.[2] Nevertheless, gaps remain to be addressed with the targeted RNA amplicon-sequencing method. Firstly, workflow and reagent choices are known to affect NGS assay error levels (method artifacts).[9,10] Therefore it is important to identify optimal protocols that achieve error reduction while maintaining assay speed and throughput. Secondly, a range of frequency thresholds for mutation reporting have been claimed but there is a lack of consensus on how these thresholds are established.[1–3,5] The ability to set a low-level analysis threshold, which enables a low limit of detection (LOD) above which mutations can be reliably reported, is a key parameter of assay performance. Paucity of benchmarking and error characterization currently limits insight into the noise factors that bottleneck assay performance. We found only one prior study that extensively characterized errors in a transgene targeted RNA-Seq assay.[2] Finally, there is paucity of public computational frameworks and documented analysis protocols for transgene targeted RNA-seq (amplicon) variant analysis. Based on a recent industry survey, most companies rely on commercial or in-house software with workflows specially optimized for sequence variant analysis[4], which can be costly or require skilled computational biology resources.

Here we detail our updated variant analysis protocols including an automated computational framework that improves the speed and standardization of mutated clone sample identification. We benchmarked our updated analysis with a mutation spike-in series and systematically quantified assay errors. To the best of our knowledge, we are the first to evaluate how a range of different preparation protocols (16 total) affect error levels and mutation detection in a transgene targeted RNA-Seq assay. We adopted performance criteria from NGS diagnostics such as specifying analysis thresholds by balancing finding low level mutations (true positives) with minimizing false positives from errors[11] and used the outcome to establish an assay limit of detection.[12,13] To assess error distributions we implemented a robust statistics method[14,15] and identified protocol changes that significantly reduced the upper noise levels compared to our initial protocol.[1]

## 2 Materials and Methods

Figure S1 provides an overview of our cell line development workflow, within which we have incorporated our standardized variant analysis workflow. Figure 1 provides a schema of our variant analysis workflow and

lists the specific study design and materials at each step which are the subject of this manuscript and detailed further below. Samples were sequenced in two NGS runs, and sample metadata is provided in Supporting Information (Table S1, Table S2).

## 2.1 Cell line generation

Biogen's proprietary CHO-K1 host was transfected with vectors encoding the proteins of interest. Transfected pools were selected and recovered in Biogen's proprietary media. The selected pools of cells were further amplified in MTX-containing media, enriched by ClonePixFL[16] , single cell cloned using limiting dilution in combination with brightfield and fluorescence imaging (CellaVista, SynenTec, Munster, Germany) to ensure clonality, expanded into 96 well plates, and evaluated in a 14-day fed-batch process. Growth and product titers were evaluated in fed batch cultures (Supporting Information for Section 2) which identified thirty-six top candidate clones that were subjected to NGS-based variant analysis out of which one mutated and one clean candidate clone were selected to create a benchmark spike-in series (see Results Section 3.1 for more details).

## 2.2 Peptide mapping

MAb protein was were purified using Protein A chromatograph. The Protein A purified samples were analyzed together with a reference sample using an in-house Lys-C peptide mapping method. Briefly, 100 micrograms of Protein A purified samples were denatured and reduced by 6M Guanidine HCl /4mM DTT, diluted 1:4 using a 50mM sodium phosphate, 10mM EDTA, pH 7.2 buffer, and then digested by adding 10 micrograms of endoproteinase Lys-C (Wako, Richmond VA) and incubating overnight at 25°C. Five micrograms of the resulting peptides were separated with a HSS T3 2.1mm x 15cm C18 column (Waters, Milford MA) heated at 55°C, using H2O/ACN gradient with 0.03% TFA as the additive. The peptide elutes were online analyzed by an Orbitrap Fusion mass spectrometer (Thermo Scientific, Waltham MA). The LC-MS data was processed with Pinpoint 1.4 software (Thermo Scientific, Waltham MA) to generate peptide ion lists for all samples. Peptide ions unique to the clone or detected in the clone with at least 100% increase in ion counts (compared to those in the reference), were considered as up-regulated ions, whose tandem mass spectra were further examined for post-translational modifications and sequence variants.

## 2.3 Transgene amplicon generation

Total RNA was isolated using Qiagen's RNeasy kit (Qiagen, Hilden, Germany) following manufacturer's protocol. cDNA was synthesized using Superscript III (IIIScpt) first strand synthesis kit, Superscript IV (IVScpt) first strand synthesis kit (Thermo Fisher Scientific Baltics UAB, Vilnius, Lithuania), or Accuscript (AcScpt) High Fidelity first strand cDNA synthesis kit (Agilent, La Jolla, CA) with an input of 25 ng/uL ng total RNA and 2.5 uM oligo dT per reaction. For amplicon generation, one set of PCR primers flanking the gene of interest at 0.5 uM and 2.5 ng/uL of the cDNA synthesis product were used per reaction. PCR reactions were performed using Phusion high-fidelity (Phu) pcr master mix or Q5® High-Fidelity (Q) 2X Master Mix (New England Biolabs, Ipswich, MA) at the following 3 amplification conditions. For PCR reactions with lower cycle numbers, the elongation step was extended to generate comparable amplicon yield for library preparation.[9]
Condition 25 cycles: 98°C 30 s., [98°C 10 s, 62°C 30 s, 72°C for 1 min] x 25 cycles, 72°C 5 min, 4°C hold
Condition 20 cycles: 98°C 30 s., [98°C 10 s, 62°C 30 s, 72°C for 3 min] x 20 cycles, 72°C 5 min, 4°C hold
Condition 15 cycles: 98°C 30 s., [98°C 10 s, 62°C 30 s, 72°C for 3 min] x 15 cycles, 72°C 5 min, 4°C hold
PCR products were cleaned using Qiagen's PCR purification kit (Qiagen, Hilden, Germany). Concentrations of the purified PCR products were measured with Qubit dsDNA BR Assay (Life Technologies, Grand Island, NY) and normalized to 20 ng/uL. Qualities of the PCR products were analyzed by 1.2% agarose gels and D1000 DNA ScreenTape analysis (Agilent, La Jolla, CA).

## 2.4 NGS library preparation and sequencing

Libraries were prepared with the following DNA library kits: TruSeq (Tru) DNA PCR Free (Illumina, San Diego, CA), Nextera XT (NXT) DNA Library Prep Kit (Illumina, San Diego, CA), QIAseq FX (FX) DNA Library Kit

(QIAgen, Hilden, Germany), KAPA (Kap) HyperPlus PCR-Free Kit (Roche Sequencing, Pleasanton, CA). All libraries were prepared per the manufacturer's instructions detailed below. Nextera, QIAseq, and KAPA kits use enzymatic fragmentation while TruSeq uses mechanical fragmentation. Nextera XT Library prep kit (Illumina, San Diego, CA) was used to prepare libraries from 1ng of DNA and were barcoded using 12 PCR cycles after ligation with Nextera XT Index kit v2 set-C. For TruSeq libraries, 1 μg (50uL) of DNA was fragmented with a Covaris E220 instrument with the parameters of 120sec, 175 peak power, Duty Factor 10.0, and cycles/Burst of 200, for a 250 bp average fragment size. The libraries were indexed with TruSeq DNA CD Adapters (Illumina, San Diego, CA) without PCR (according to the manufacturer's instructions). For QIAseq FX libraries 100 ng of input DNA was used as input. DNA was fragmented for 8 min at 32°C (450 bp fragment size targeted), and libraries were indexed with the QIAseq Adapters included in the kit without PCR. For KAPA HyperPlus libraries 100 ng of input DNA was used as input. DNA was fragmented for 12.5 min at 37°C (~300 bp fragment size), A-tailing was carried out with the ER&AT Plus Enzyme, and libraries were indexed with the KAPA Dual-Indexed Adapter Kit (Roche Sequencing, Pleasanton, CA) without PCR. Libraries were quantified with a Lab Chip GX (Perkin Elmer, Waltham, MA), and KAPA Library Quantification Kit (Roche Sequencing, Pleasanton, CA). Equimolar amounts of libraries were pooled, denatured, and sequenced on a MiSeq with a 300-cycle v2 Reagent Kit (Illumina, San Diego, CA) with 2x150bp paired-end reads, yielding >121,000 (with minimum sample read count of 114,000) fragments per sample. Each run generated > 4.9 Gb of data (5.15 Gb for the Nextera/TruSeq sequencing run; run one) with over 89 % ≥Q30 for NGS run one and over 94% ≥Q30 for NGS run two.

### 2.5 Computational Variant Analysis

Raw read quality was checked with FASTQC. Reads were demultiplexed per sample and processed using a pipeline of python scripts that starts with mapping reads to the transgene vector reference using the BWA aligner v0.7.12.[17] The aligned reads are further filtered by soft-clipping the ends of partially aligned reads and by marking of duplicate reads with Picard Tools v2.6.0 (http://Broadinstitute.Github.Io/Picard/). Alignment statistics were checked for total number of reads per sample (0.1-1e6 reads per sample), mapping rate (~80-100%), and insert size distribution. Subsequent variant calling is done with samtools v1.3.1 mpileup[18] ignoring duplicate reads (potential PCR duplicates) and using thresholds on mapping quality of >q20 and on base quality of >Q20 (both indicating ≤1% probability of being in error). The resulting mpileup files are parsed into mutation tables and further processed as described in Results. All mutations, including the less common small insertions and deletions[19] are logged but the analysis focuses on the more common Single Nucleotide Variants (SNVs). Pipeline scripts are available at https://github.com/GrootJ/TransgeneSeq.

## 3 Results

### 3.1 Cell line generation and initial sequence variant analyses

To evaluate the assay threshold of the NGS-based sequence variant analysis, two clones expressing the same MAb were selected from a routine cell line development campaign (Figure S1). Clone 9 (C9) and clone 29 (C29) were single cell cloned from two independent transfection pools and were among the top 36 performers based on growth, product titers, and productivities in fed batch cell culture processes (Figure S2, Supporting Information). Initial sequence variant analysis of the top 36 clones using amplicon based NGS identified three variants in clone 29 MAb heavy chain at ~60% frequencies. The three variants C1775A, A1983C, C2608G were re-analyzed by NGS and confirmed to occur at 60%, 58%, 59% frequencies respectively (Figure 2B, Table S3, Supporting Information). Moreover, mutations A1983C, C2608G would result in non-synonymous mutations T219P and A427G in the protein sequence. To orthogonally validate the mutations in the protein sequence, MAb samples purified from fed-batch cultures of clone 9 and clone 29 were analyzed by peptide mapping using LC-MS (Figure 2C, D). As expected, variants T219P and A427G were detected in protein sample from clone 29 and measured to be at 58% and 56% frequency, respectively (Table S4, Supporting Information). These mutation frequencies suggest there are multiple genomic copies of the transgene. No mutations over 1% frequency were found in clone 9 which was confirmed by both an NGS re-analysis and a lack of protein sequence variants from peptide mapping. After these validations we mixed

mutated clone 29 (C29) with non-mutated clone 9 (C9) mRNA in various ratios to create a spike-in series of true mutations at different frequencies.

### 3.2 Computational Variant Analysis pipeline

We upgraded our previous computational analysis by developing an automated pipeline of scripts (Figure 1).[1] This pipeline includes alignment and variant calling (updated to BWA and Samtools – see Methods) as well as alignment Quality Control providing insert sizes between forward and reverse reads and alignment statistics per sample (Figure S4, Table S5, Supporting Information). The variant calling output of samtools is parsed into a mutation table that lists per position the sequencing depth, the basecalls (A/C/T/G), insertions/deletions, and the frequency of mismatches (%) from the reference base (Table S6, Supporting Information). Mutation tables of all samples are then processed in a custom R programming workflow that plots mutation (mismatch) frequencies (Figure 2A,B,E,F, File S1, Supporting Information) and read coverage (Figure S3, Supporting Information) across the regions of interest (here the coding sequence/CDS regions). If read coverage falls below 500 reads at any CDS position, the sample is automatically flagged. This workflow summarizes mutation frequency data (base-calls) for all positions that have mutations at or above an inputted set of thresholds (here set to ≥0.2%, ≥0.5%, ≥1%, ≥2%, ≥5%) and lists samples that harbor mutations at or above each of these thresholds (File S2, Supporting Information).
Our protocol inspects mutation summary data and CDS frequency plots by first focusing on higher frequency mutations before evaluating lower frequency mutations; this is done to distinguish true mutations from potential errors. Errors are method artifacts that often constitute a low frequency background noise signal which can be enriched in specific sequence regions. The protocol inspection looks for patterns; when multiple lower frequency mutations occur in very close proximity across all samples, based on historical experience (data not shown) and reports of amplicon-seq noise[19], they are more likely to be errors, and they are used to define an analysis threshold frequency below which distinguishing true mutations from errors becomes uncertain. For example, in Figure 2E and F the MAb Heavy chains of the C9 and C29 samples show multiple ~0.4% apparent mutations (i.e. errors) near the CDS 3'end (below the maximum error dashed line) as well as more region-wide background noise ~0.1% (below the 95$^{th}$ quantile error red line). Similar patterns can be found in all samples (Figure S5, File S1, Supporting Information). By subjecting low-level mutations to a systematic curation, we can adapt analysis thresholds per transgene/program, per sample, and, depending on the error profile, per site. By standardizing and summarizing outputs our upgraded pipeline improved the identification of mutations and error patterns thereby greatly reducing the analysis time down to one workday.

### 3.3 Variant assay accurately captures high to low-frequency mutations of the spike-in series

The variant analysis workflow was benchmarked by evaluating the true mutation and the error calls of the spike-in series. The impact of amplicon preparation and NGS library preparation was investigated by processing the spike-in series using different combinations of preparation methods including cDNA reagent choices, PCR conditions, and NGS library methods (Figure 1). Figure 3 shows our variant analysis accurately captured true spike-in mutations from ~60% down to ~0.3% frequency across all preparation methods (high correlation with spike-ins, adjusted $R^2$ of 0.99). Samples processed with the Nextera library kit (NXT) properly identified mutation presence but quantified mutation frequencies with greater variability than the other methods. For example, the two Nextera kit samples of C29 estimated the frequency of the A1983C mutation to be 46% whereas the ten other methods that all used the TruSeq PCR free library kit (Tru) indicated 58.3% (±1.3%) (Tables S3 and S4, Supporting Information). This frequency difference might be related to a C>A substitution bias in the Nextera kit (see section 3.4). We created consensus "benchmark truth" spike-in frequencies by taking the mean of the three C29 mutations measured across the prep methods but left out the more variable Nextera samples. Testing our assay at a 0.2% analysis threshold we detected 36 out of the 36 ~0.3% spike-ins (100% sensitivity) across all prep methods (12 samples) but found high false positives; on average 9.6 per TruSeq sample and 428 for the two Nextera samples. With errors above 0.2% in all samples, a 0.2% threshold would incorrectly identify all samples as mutated, even if no true mutations

were present. Assay performance improved when taking a 0.475% analysis threshold for the ~0.6% spike-ins; we detected 34 out of 36 true mutations across all prep methods (12 samples) with only two false positives in ten TruSeq samples and 173 false positives on average for the two Nextera samples. For all TruSeq samples we considered 0.6% a limit of detection (LOD) with reasonable confidence (i.e. reasonably low false positive and false negative rates[12], see definitions in Table S7, Supporting Information). Differences in false positive rates across prep methods (incl. much higher error with Nextera) were subsequently investigated through an extensive characterization of error profiles.

### 3.4 Sample prep and NGS library prep affect background error rates

Background errors are low-level NGS method artifacts that originate from sample prep, NGS library prep, and sequencing itself. We investigated error levels in the spike-in series across prep methods, comparing groups of samples that differed in reverse-transcriptase and/or PCR and/or NGS library prep (Figure 1). Errors were defined as all mutations (mismatches) that were not spiked-in. Error rates were quantified by summarizing error frequencies per position across the CDS region of each sample (see definitions in Table S7, Supporting Information). We subsequently pooled the error rates of all samples processed with the same prep method. The error rates presented as highly skewed distributions (Figure 4A&B). Comparing higher level errors across prep methods is of most interest since these errors force one to set a higher analysis threshold as to limit the incorrect removal of mutation free clones. The error rate maxima are not stable statistical estimates which limited their use in comparing across NGS methods but the 95$^{th}$ quantiles of error rate are a more robust measure, representative of the region-wide noise in each prep method (Figures 2E&F, S6, Supporting Information). Samples prepared with the Nextera library kit (NXT) of our original protocol[1] harbored the largest errors and use of the TruSeq PCR free library kit (Tru) reduced errors substantially (~0.5% at 95$^{th}$ quantile, Figure 4A). Error rate differences were smaller, more subtle between all methods that used TruSeq (Figure 4B) and we implemented a statistically robust estimator[14,15] to quantify these subtle differences in error rate upper quantiles. Using the IIIScpt_Ph20_Tru method as an error rate reference distribution, four prep methods significantly reduced higher level errors (IIIScpt_Q20_Tru, IVScpt_Ph15_Tru, IVScpt_Q20_Tru, IVScpt_Ph20_Tru, Figure 4C) and all did so at similar levels (~0.031% at error rate 95$^{th}$ quantile, Table 1).
These four methods used either high fidelity Q5 PCR polymerase (IIIScpt_Q20, IVScpt_Q20) or Phusion PCR polymerase at 15 or 20 cycles with IVScpt (IVScpt_Ph15, IVScpt _Ph20). We also pooled error rates by PCR method which confirmed a modest error reduction with high fidelity Q5 polymerase (Figure S7, Supporting Information). The effect of Phusion PCR cycles depended on the reverse transcriptase: for IIIScpt lowering PCR cycles to 15 (IIIScpt_Ph15) slightly increased error but with IVScpt (IVScpt_Ph15) this slightly lowered error (Figure 4B,C). IIIScpt reverse transcriptase has lower processivity than IVScpt and perhaps effects around template finishing could explain why an intermediate number of PCR cycles (beyond 15 but below 25) provided a slight reduction in error. Interestingly, from Figure 4B error rate maxima of Phusion PCR samples (IVScpt_Ph15, IVScpt _Ph20) seem reduced compared to Q5 PCR samples (IIIScpt_Q20, IVScpt_Q20) but given the variability in error rate maxima the significance of this difference could not be determined (Figure 4C). Taken together, our evaluation established an absolute reduction in error rate of ~0.5% between IIIScpt_Ph25_NXT and IVScpt_Ph15_Tru (Table 1) which equated to a relative reduction of 6 fold (at the 95$^{th}$ quantile).
We subsequently investigated whether certain sequence features influenced (local) error rates and found that increased error patterns matched high GC-content and/or lower sequence complexity (Figure S8 and Figure S9 respectively, Supporting Information), both well-established sources of NGS error.[19,20] Additionally we characterized substitution patterns amongst the errors and found Nextera kit errors to be enriched in C>A/G>T substitutions (Figure S10, Supporting Information) which also matches prior reports.[19]

### 3.5 A second NGS run found minor error differences between additional NGS library prep kits

The first sequencing run demonstrated substantial error reduction from the Nextera to the PCR-free TruSeq library kit and promising additional error reductions with IVScpt_Ph15_Tru and IVScpt_Ph20_Tru methods for amplicon generation (Table 1). Yet the TruSeq PCR free library prep (Tru) is more time-consuming than

Nextera which prompted us to additionally evaluate PCR free library kits QIAseq FX DNA (FX) and KAPA HyperPlus (Kap) which are 2-4X faster than TruSeq. In a second NGS run we evaluated these kits with both the IVScpt_Ph15_Tru and IVScpt_Ph20_Tru methods using a duplicate spike-in series (Figure S11A, Table S2, Supporting Information). The IVScpt_Ph15_Tru and IVScpt_Ph20_Tru samples showed similar error profiles as in NGS run one. As in NGS run one, error rate maxima were not stable statistical estimates (had large confidence intervals) and prep method differences were evaluated at the 95$^{th}$ quantile of error rate (Figure S11B, Table S8, Supporting Information). We pooled the FX and Kap library kit samples across PCR methods (Ph15, Ph20) and did not find a significant difference in higher level errors compared to TruSeq (Figure S11C, Supporting Information). Lowering PCR cycles from 20 (Ph20) to 15 (Ph15) did result in a significant, albeit minor reduction in higher level error (~0.007% 95$^{th}$ quantile error reduction across library kits, Figure S11D, Supporting Information). Here again FX and Kap provide a practical advantage over TruSeq as they require 10-fold less input DNA which makes them more amenable to lowering cycles in the preceding PCR step. With the FX kit with 15 PCR cycles (IVScpt_Ph15_FX) we found that a 0.475% analysis threshold yielded zero false positives (across duplicate spike-ins, 10 samples) while detecting all six ~0.6% true mutations (2 samples), establishing a further improvement in statistical confidence for the ~0.6% limit of detection.

## 4 Discussion

We created an updated variant analysis pipeline that further automated and standardized outputs for NGS QC, mutation data visualization, and threshold-based identification of mutated samples. This improved the speed, efficiency, and accuracy of our variant analysis workflow. Using a mutation spike-in series we benchmarked our updated workflow and verified it accurately quantified spiked-in mutations from ~60% down to ~0.3%. Our spike-in analysis also established a higher confidence limit of detection (LOD) of ~0.6% that allows for zero false-positive errors for preferred prep method IVScpt_Ph15_FX (Figure S11A). By contrast, the samples prepared with the Nextera library kit that was part of our initial protocol [1] contained errors up to ~2%, well above a 0.6% threshold (Figure 4A). Our improved LOD of ~0.6% is similar to the ~0.5% reported for similar RNA variant analysis assays[2,3,5,21], where LOD is sometimes referred to as sensitivity[5] and assay statistical performance[12] is not always detailed[1,3,5].

Errors are low-frequency background artifacts introduced at different steps; Reverse Transcriptase[21,22], PCR[19,21–23], NGS library prep[19,22,24], as well as in the sequencing process.[19,21,25] The origins of NGS assay error remain the subject of ongoing debate and investigation.[19,26] Background errors are widely studied in DNA-sequencing applications such as somatic cancer variant analysis but are less comprehensively characterized in RNA-seq variant analysis, a less common application which is more prone to errors.[27,28] Despite targeted RNA-Seq variant analysis becoming the method of choice in cell line mutation screening[4], we found just two prior studies that comprehensively characterized errors in this assay[2,21], only one of which was targeting transgenes.[2] Our study comprehensively characterized errors and is, to the best of our knowledge, the first to evaluate how different combinations of prep methods (16 total) affect error levels in transgene targeted RNA-Seq.

Variant analyses need to balance the finding of true mutations (sensitivity) with avoiding false positive/errors (specificity); this requires setting a proper analysis threshold frequency above which mutations are reported. [12,13] Higher level errors force the analysis threshold to be raised (Figure 2E, F) as to avoid incorrectly reporting errors and discarding candidate clones that are mutation free. Errors presented as highly skewed, non-normal distributions which complicated comparing across prep methods. We addressed this by implementing a statistically robust quantile comparator[14] and found the 95$^{th}$ quantile of error rate to be an estimate that is both stable and representative of the prep method. Our original assay used IIIScript reverse transcriptase, 20 cycles of Phusion PCR and the Nextera library prep kit; this commonly created ~1-2% errors thereby forcing the analysis threshold to ≥2% frequency to avoid reporting errors.[1] We identified a substantial reduction in higher level (95$^{th}$ quantile) errors by using PCR free library kit(s) as opposed to the Nextera kit (~0.5%, Table 1), and modest additional reductions by using IVScript reverse transcriptase in combination with lowering PCR cycles from 20 to 15 (~0.03%, Table 1). These findings agreed with previous reports of higher errors with Nextera in targeted DNA-Seq[19] and with higher error with higher PCR cycles (note Nextera also entails PCR).[19,21,29] Finally, we evaluated additional PCR-free library kits including the fast QIAseq FX kit (2.5 hour workflow) that can handle low DNA input. QIAseq FX proved to be a pragmatic, low noise NGS library kit to facilitate lowering PCR cycles to 15 during the preceding amplicon

generation step. Combined these protocol improvements lowered error levels significantly, thereby enabling variant analysis at lower thresholds.

Errors in our assay (median ~0.04%, maximum ~0.4-0.8% - excluding Nextera samples) seem to be at similar levels as reported for other targeted RNA-Seq assays (median ~0.05%, maximum ~0.66%[21]) and (mean ~0.02%, maximum 0.30-1.0%[2]). That said, error levels and analysis thresholds should be interpreted with caution and within context. NGS errors are highly sequence specific and like prior reports[19,20] we found increased regional error with high GC-content and lower sequence complexity (Figure S8 and Figure S9 respectively, Supporting Information). In addition to being sequence specific, errors can vary per sequencing run/machine[12,19,20] which is why our protocol carefully inspects mutation profiles for each transgene program. In general NGS error rate reports vary from ~0.1% up to ~1% and calling mutations below 1% abundance requires specific attention as it cannot be done universally with high confidence.[26] To further aid with distinguishing errors from true genetic variation, we started implementing an additional control: adding a mutation-free vector control sequence to the sample set (subjected to PCR, library prep). For future error mitigation efforts molecular barcoding strategies are of interest as they are reported to be a promising way to denoise artifacts introduced after cDNA synthesis.[26,30] Using bioinformatic tools that overlap paired-end reads is also reported to reduce error[19] but requires a very careful evaluation as recent reports indicate a detrimental impact on the calling of true variants.[31]

Finally, the performance of sequence variant analysis needs to be evaluated in the context of its use in the time-constrained process of cell line development. In addition to the ability to detect true mutations and avoid errors at lower thresholds, assay speed, throughput, and efficiency need to be appropriate for each specific process development stage.[2] Sequence variant analysis using novel technologies harbors promising new capabilities. TLA-NGS (Target Locus Amplification-NGS) is a new, specialized DNA pull-down technology that can identify structural variants in addition to SNVs.[32] Single molecule real-time circular consensus sequencing demonstrated sensitive DNA-level mutation detection across a vector region.[8] Coupling NGS with DNA-level digital PCR allowed for an orthogonal verification of mutations.[11] However, these technologies work or have only been demonstrated on the DNA-level which would miss potential transcriptional errors[33] and could bias mutation frequency estimates relative to the protein level.[11] Moreover, these assays typically demand more resources and have a longer turnaround time, which is better suited for sequence confirmation of a handful of lead cell lines at a later stage of process development. The targeted RNA-sequencing method with optimized preparation conditions described here demonstrated competitive assay performance and suitable efficiency to facilitate early clone screening of cell line development.

## Acknowledgement

We want to thank Spring Liu for consult on pipeline construction.

## Conflict-of-Interest

All authors are employees of Biogen.

## 5 References

[1] C. Wright, J. Groot, S. Swahn, H. McLaughlin, M. Liu, C. Xu, C. Sun, E. Zheng, S. Estes, Biotechnol. Prog. **2016,** 32, 813.

[2] S. Zhang, J.D. Hughes, N. Murgolo, D. Levitan, J. Chen, Z. Liu, S. Shi, S. Zhang, J.D. Hughes, N. Murgolo, D. Levitan, J. Chen, Z. Liu, S. Shi, BioMed Res. Int. BioMed Res. Int. **2016,** 8.

[3] S. Zhang, L. Bartkowiak, B. Nabiswa, P. Mishra, J. Fann, D. Ouellette, I. Correia, D. Regier, J. Liu, Biotechnol. Prog. **2015,** 31, 1077.

[4] J. Valliere-Douglass, L. Marzilli, A. Deora, Z. Du, L. He, S. Kumar, Y.-H. Liu, H. Martin-Mueller, C. Nwosu, J. Stults, Y. Wang, S. Yaghmour, Y. Zhou, PDA J. Pharm. Sci. Technol. **2019,** pdajpst.2019.010009.

[5] T.J. Lin, K.M. Beal, P.W. Brown, H.S. DeGruttola, M. Ly, W. Wang, C.H. Chu, R.L. Dufield, G.F. Casperson, J.A. Carroll, O.V. Friese, B.F. Jr, L.A. Marzilli, K. Anderson, J.C. Rouse, MAbs **2019,** 11, 1.

[6] R. Jefferis, M.-P. Lefranc, MAbs **2009,** 1, 332.

[7] D. Wen, M.M. Vecchi, S. Gu, L. Su, J. Dolnikova, Y.-M. Huang, S.F. Foley, E. Garber, N. Pederson, W. Meier, J. Biol. Chem. **2009,** 284 , 32686.
[8] J.F. Cartwright, K. Anderson, J. Longworth, P. Lobb, D.C. James, Biotechnol. Bioeng. **2018,** 115, 1485.
[9] C. Brandariz-Fontes, M. Camacho-Sanchez, C. Vilà, J.L. Vega-Pla, C. Rico, J.A. Leonard, Sci. Rep. **2015,** 5, 8056.
[10] J.D. Ring, K. Sturk-Andreaggi, M.A. Peck, C. Marshall, Forensic Sci. Int. Genet. **2017,** 29, 174.
[11] T.J. Lin, K.M. Beal, H.S. DeGruttola, S. Brennan, L.A. Marzilli, K. Anderson, Biotechnol. Bioeng. **2017,** 114 1744.
[12] L.J. Jennings, M.E. Arcila, C. Corless, S. Kamel-Reid, I.M. Lubin, J. Pfeifer, R.L. Temple-Smolkin, K.V. Voelkerding, M.N. Nikiforova, J. Mol. Diagn. **2017,** 19, 341.
[13] S.A. Hardwick, I.W. Deveson, T.R. Mercer, Nat. Rev. Genet. **2017,** 18, 473.
[14] F.E. Harrell, C.E. Davis, Biometrika **1982,** 69, 635.
[15] P. Mair, R. Wilcox, Behav. Res. Methods **2019,** 52, 464.
[16] C. Wright, C. Alves, R. Kshirsagar, J. Pieracci, S. Estes, Biotechnol. Prog. **2017,** 33, 1468.
[17] H. Li, R. Durbin, Bioinformatics **2010,** 26, 589.
[18] H. Li, Bioinformatics **2011,** 27, 2987.
[19] M. Schirmer, U.Z. Ijaz, R. D'Amore, N. Hall, W.T. Sloan, C. Quince, Nucleic Acids Res. **2015,** 43, e37.
[20] M.G. Ross, C. Russ, M. Costello, A. Hollinger, N.J. Lennon, R. Hegarty, C. Nusbaum, D.B. Jaffe, Genome Biol. **2013,** 14, R51.
[21] R.J. Orton, C.F. Wright, M.J. Morelli, D.J. King, D.J. Paton, D.P. King, D.T. Haydon, BMC Genomics **2015,** 16 229.
[22] D.A. Shagin, I.A. Shagina, A.R. Zaretsky, E.V. Barsova, I.V. Kelmanson, S. Lukyanov, D.M. Chudakov, M. Shugay, Sci. Rep. **2017,** 7, 2718.
[23] S. Filges, E. Yamada, A. Ståhlberg, T.E. Godfrey, Sci. Rep. **2019,** 9, 3503.
[24] E.L. van Dijk, Y. Jaszczyszyn, C. Thermes, Exp. Cell Res. **2014,** 322, 12.
[25] D. Laehnemann, A. Borkhardt, A.C. McHardy, Brief. Bioinform. **2016,** 17, 154.
[26] J.J. Salk, M.W. Schmitt, L.A. Loeb, Nat. Rev. Genet. **2018,** 19, 269.
[27] C. Xu, Comput. Struct. Biotechnol. J. **2018,** 16, 15.
[28] Y. Guo, S. Zhao, Q. Sheng, D.C. Samuels, Y. Shyr, BMC Genomics **2017,** 18, 690.
[29] E.N. Smith, K. Jepsen, M. Khosroheidari, L.Z. Rassenti, M. D'Antonio, E.M. Ghia, D.A. Carson, C.H.M. Jamieson, T.J. Kipps, K.A. Frazer, Genome Biol. **2014,** 15, 420.
[30] A.M. Newman, A.F. Lovejoy, D.M. Klass, D.M. Kurtz, J.J. Chabon, F. Scherer, H. Stehr, C.L. Liu, S.V. Bratman, C. Say, L. Zhou, J.N. Carter, R.B. West, G.W. Sledge Jr, J.B. Shrager, B.W. Loo Jr, J.W. Neal, H.A. Wakelee, M. Diehn, A.A. Alizadeh, Nat. Biotechnol. **2016,** 34, 547.
[31] D.L. Cameron, L. Di Stefano, A.T. Papenfuss, Nat. Commun. **2019,** 10, 3240.
[32] S.H. Aeschlimann, C. Graf, M. Dmytro, H. Lindecker, L. Urda, N. Kappes, A.L. Burr, M. Simonis, E. Splinter, M. van Min, H. Laux, Biotechnol. J. **2019,** 14, 201800371.
[33] P. Cui, F. Ding, Q. Lin, L. Zhang, A. Li, Z. Zhang, S. Hu, J. Yu, Genomics Proteomics Bioinformatics **2012,** 10, 4.

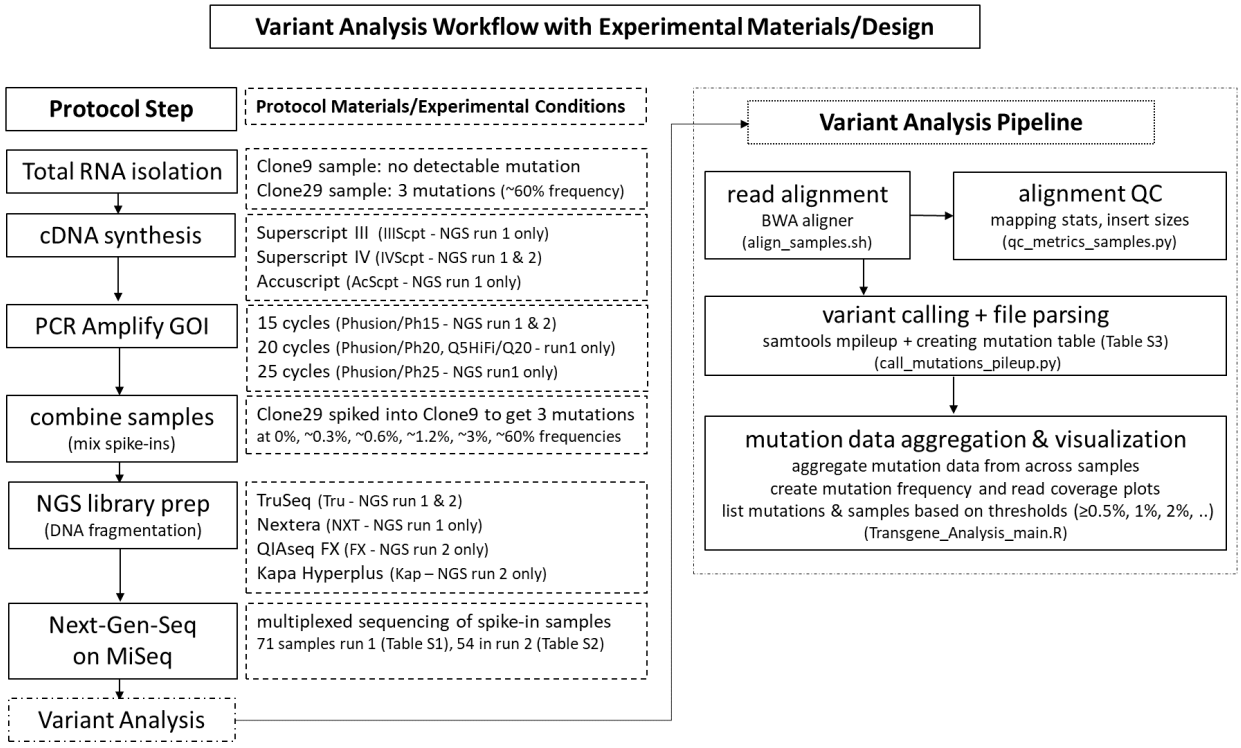


**Figure 1.** Schema of NGS transgene variant analysis workflow with experimental materials/design listed per step (dashed boxes, samples detailed in Table S1, Table S2, Supporting Information) and the computational variant analysis pipeline detailed separately (scripts listed in brackets, see **Methods**). The NGS transgene variant analysis is a key cell line development screen to select lead and back-up manufacturing clones out of 30-40 candidate clones (Figure S1, Supporting Information).

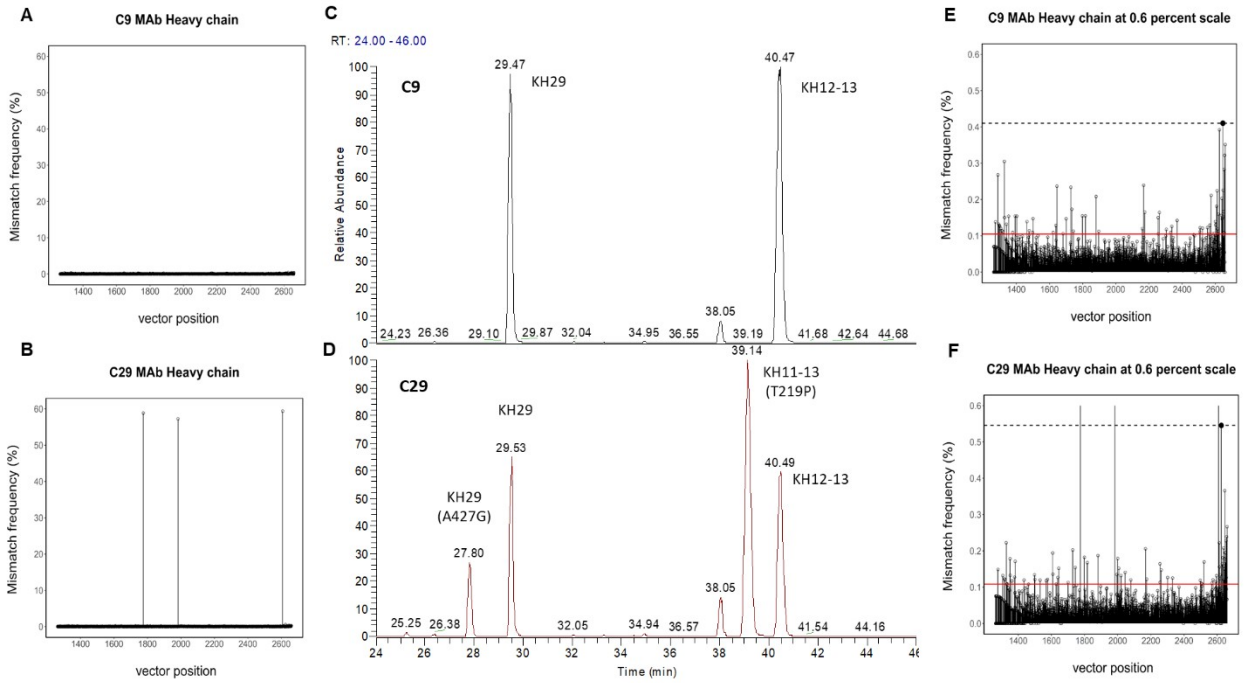


**Figure 2.** NGS derived mutation frequency plots for (A) non-mutated MAb Heavy chain of clone 9 (C9) and (B) mutated MAb Heavy chain of clone 29 (C29), and corresponding peptide mapping total ion chromatograms (C, D respectively) that indicate corresponding amino acid changes. Extracted ion chromatogram (EIC) of C29 (D) shows A427G (A1983C) substitution (KH29) at 58.3%, and T219P (C2608G) substitution (KH11-13) at 55.9% (Tables S3 and S4, Supporting Information). NGS mutation frequency plots (panel E) and (F) are for the same C9 (A) and C29 (B) samples but with the frequency scale lowered to 0.6% as to show the low-level background errors; a dashed line threshold is drawn through the maximum error (black dot) and a red line indicates the upper 95$^{th}$ quantile of error rates (see explanation in main text). Both C9 and C29 NGS samples were prepared with reverse transcriptase IIIScpt, Phusion PCR 20 cycles and the TruSeq library kit (prep method IIIScpt _Ph20_Tru).

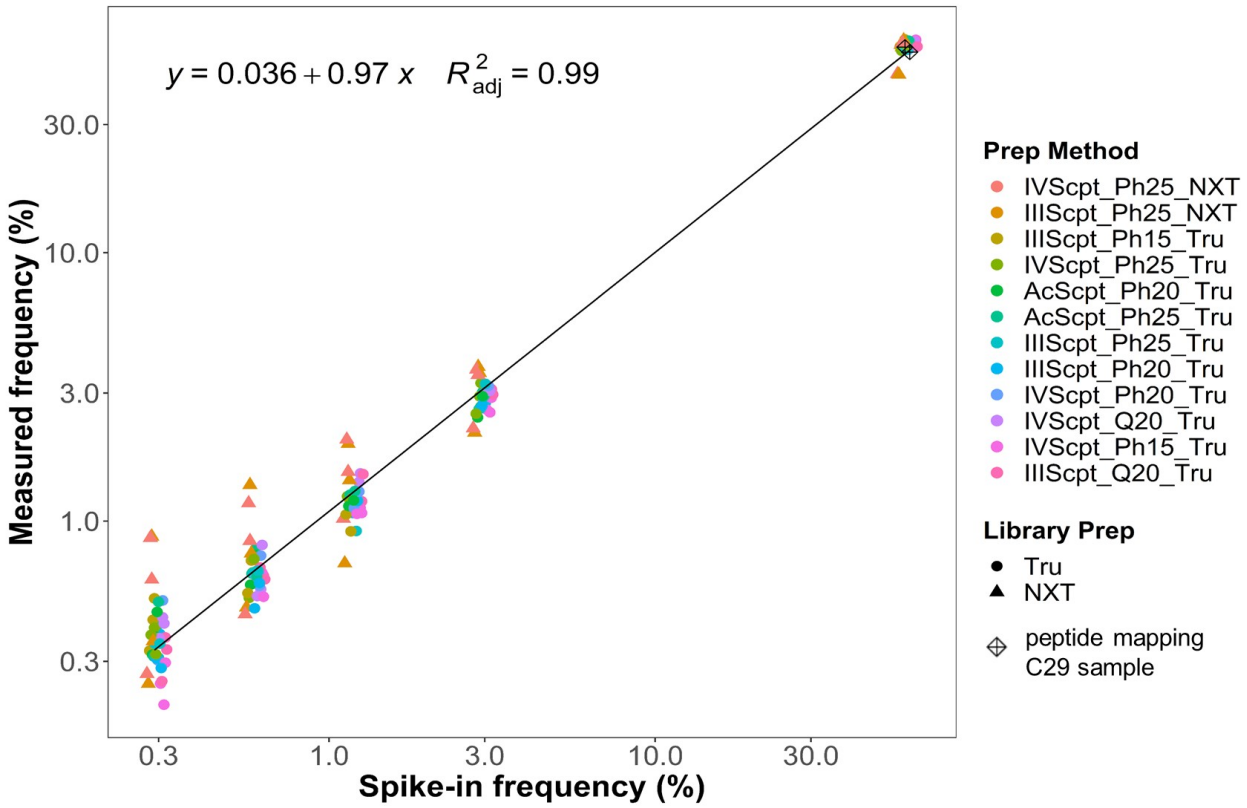


**Figure 3.** Measured mutation frequencies versus spiked-in frequencies at 60%, 3%, 1.2%, 0.6%, 0.3% (with 3 mutations per sample, x-axis points jittered to limit over plotting). The spike-in series were prepared using 12 different prep conditions (from NGS batch 1) as indicated by colors, and by shape for library prep kit: TruSeq (dots) and Nextera (triangles). Peptide mapping frequencies for 2 non-synonymous mutations in C29 are also plotted (⌖points). A correlation of measurements versus spike-ins is computed with the fit and coefficient (adjusted Pearson $R^2$) displayed on the plot.

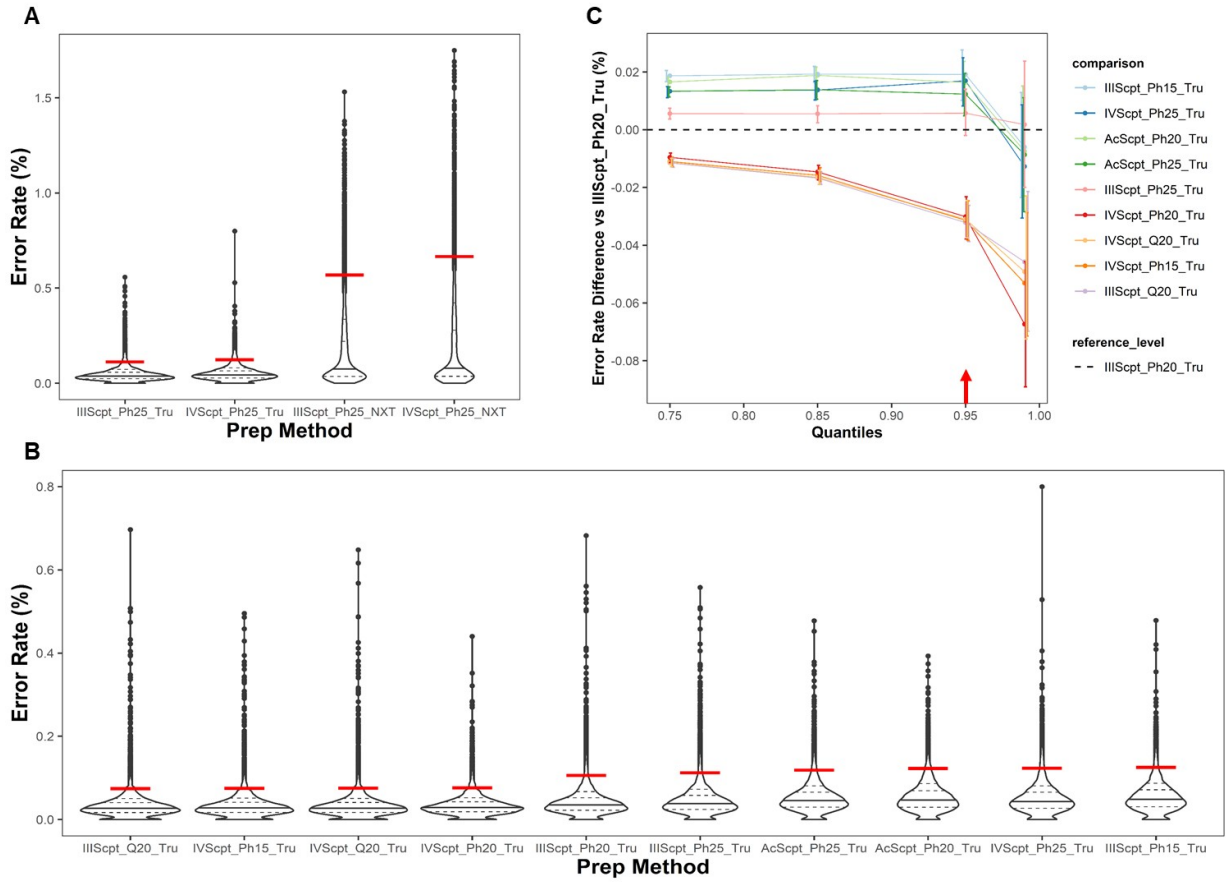


**Figure 4.** A) Mutation error rate distributions of four prep methods that used either the Nextera (NXT) or Truseq (Tru) NGS library kits. Error rates were pooled from all spike-in samples per prep method with spike-in mutations removed. Violin plots of distributions were ordered left to right according to increasing 95$^{th}$ quantiles (red line), with 25$^{th}$, 75$^{th}$, 85$^{th}$ quantiles indicated by dashed lines, the 50$^{th}$ quantile (median) by a solid line, and outlier mutations by dots.

B) Mutation error rate distributions of prep methods that used different combinations of reverse transcriptases (IIIScpt/IVScpt/AcScpt) and PCR methods (Ph15/Ph20/Ph25/Q20). Error rates were pooled from the spike-in samples per prep method with spike-in mutations removed. Distributions were ordered left to right according to increasing 95$^{th}$ quantiles (red line), with 25$^{th}$, 75$^{th}$, 85$^{th}$ quantiles indicated by dashed lines, the 50$^{th}$ quantile (median) by a solid line, and outliers by dots.

C) Differences in error rate upper quantiles of prep methods in (B) relative to IIIScpt_Ph20_Tru (reference method), statistics computed using a robust estimator.[14,15] Differences at the 95$^{th}$ quantile (red arrow) are given in Table 1.

**Table 1.** Statistical comparisons of the 95$^{th}$ quantile of the error rates for each prep method relative to reference IIIScpt_Ph20_Tru (prep method in the current protocol). These quantile statistics correspond to the confidence intervals on the 95$^{th}$ quantile in Figure 4C, are computed with a robust quantile estimator.[14,15]

| Comparison | Error rate difference (%) at 95$^{th}$ quantile | confidence int. low | confidence int. up | Significance[a] |
|---|---|---|---|---|
| IIIScpt_Q20_Tru_v_IIIScpt_Ph20_Tru | -0.032 | -0.039 | -0.026 | **** |
| IVScpt_Ph15_Tru_v_IIIScpt_Ph20_Tru | -0.031 | -0.038 | -0.025 | **** |
| IVScpt_Q20_Tru_v_IIIScpt_Ph20_Tru | -0.031 | -0.037 | -0.025 | **** |
| IVScpt_Ph20_Tru_v_IIIScpt_Ph20_Tru | -0.030 | -0.038 | -0.023 | **** |
| IIIScpt_Ph25_Tru_v_IIIScpt_Ph20_Tru | 0.006 | -0.002 | 0.014 | n.s. |
| AcScpt_Ph25_Tru_v_IIIScpt_Ph20_Tru | 0.012 | 0.005 | 0.019 | **** |
| AcScpt_Ph20_Tru_v_IIIScpt_Ph20_Tru | 0.016 | 0.008 | 0.024 | **** |
| IVScpt_Ph25_Tru_v_IIIScpt_Ph20_Tru | 0.017 | 0.008 | 0.025 | **** |
| IIIScpt_Ph15_Tru_v_IIIScpt_Ph20_Tru | 0.019 | 0.010 | 0.028 | **** |
| IIIScpt_Ph25_NXT_v_IIIScpt_Ph20_Tru | 0.464 | 0.436 | 0.496 | **** |
| IVScpt_Ph25_NXT_v_IIIScpt_Ph20_Tru | 0.559 | 0.535 | 0.579 | **** |

a) indicated by p-value magnitude: $<10^{-4}$ (****), $<10^{-3}$ (***), $<10^{-2}$ (**), $<10^{-1}$ (*), or not significant (n.s.)