illumina | ARCHER

Technical Note

PreSeq™ RNA QC Assay

RNA derived from formalin-fixed, paraffin-embedded (FFPE) tissue can be of variable quality due to chemical modification incurred during the fixation process as well as during pre-fixation handling. While many different methods are available for measuring RNA mass and fragment length, we have developed the Archer™ PreSeq™ RNA QC Assay, a qPCR-based transcript analysis assay that is highly predictive of sequencing success using Archer FusionPlex™ Assays. PreSeq is superior to alternative RNA analytical methods because the assay is in-line with the Archer Universal RNA library production protocol and gives a quantitative readout that assesses important aspects of RNA input quality that determine suitability for NGS library generation (Table 1).

PreSeq is a SYBR®-based qPCR assay designed to assess the suitability of an RNA input for FusionPlex library generation based on the quantification of chemically available fragments greater than 100bp in length. Positive PreSeq results are predictive of successful FusionPlex library sequencing results and of passing the Archer Analysis software quality filter. While the observed PreSeq C_q value of a given sample can vary depending on the instrument and master mix used, routinely performing this assay during FusionPlex library creation can help establish a baseline pass/fail metric to predict library quality prior to sequencing and aid in determining the cause of failure after sequencing.

	PreSeq	UV Spectrometry, Flourescent Dyes	Capillary Electrophoresis	Gel Electrophoresis
Concentration	✔	✔	✔	✔
Length	✔	✖	✔	✔
Crosslinking	✔	✖	✖	✖

Table 1. PreSeq is superior to alternative RNA analytical methods. *PreSeq quantifies the concentration of fragments in an RNA sample greater than 100 base pairs that are free from crosslinking. Other analytical methods do not assess all of these key sample characteristics that influence NGS library generation.*

Introduction

Tissue preservation by formalin fixation followed by embedding in paraffin is the standard means of sample handling and storage in oncology research settings, and archives of this material represent an enormous biological library for retrospective studies in cancer genomics. However, formalin chemically modifies nucleic acid, resulting in protein-nucleic acid crosslinking, strand cleavage and base modifications. Together, these chemical modifications can inhibit downstream enzymatic manipulation and interfere with hybridization-based detection techniques such as PCR. In addition to the chemical and physical damage related to the fixation process, storage conditions and time often cause additional nucleic acid degradation. In sum, the degree of chemical damage and fragmentation in the RNA from archival FFPE samples varies greatly. While some of these forms of damage are not detected by standard methods of nucleic acid quantification and visualization, they are the major determinants of sequencing library quality. Therefore, accurate determination of RNA quality prior to library preparation is required to predict sequencing success.

Several methods to measure RNA mass and assess fragment length are available. RNA quantification using UV spectrometry or fluorescent RNA-binding dyes

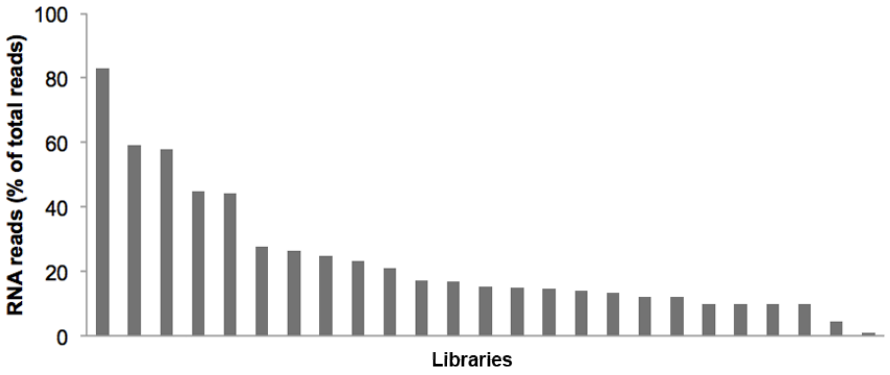

Figure 1. Total RNA input does not correlate with the fraction of library reads derived from RNA*. Libraries were prepared from 25 separate FFPE samples using the same RNA input amount from total nucleic acid for each library (200ng). The libraries were then sequenced and analyzed by Archer Analysis. The data indicate that the fraction of RNA reads in each library varies greatly, although the input RNA remained constant across all libraries. RNA mass alone is inadequate for predicting the quality of an RNA NGS library and the amount of RNA available for library generation varies per nanogram of input across samples.*

can provide accurate estimates of total RNA mass. Similarly, RNA fragment length and crude mass can be approximated by capillary or gel electrophoresis. Capillary electrophoresis can provide an RNA integrity number (RIN) in high-quality RNA samples or other metrics in more degraded samples. Nevertheless, these methods do not provide complete information about the quantity of RNA that is chemically available for library generation because they do not yield all key metrics that can significantly impact NGS library preparation.

ArcherDX has developed the Archer PreSeq RNA QC Assay, a qPCR assay to rapidly estimate the quantity of RNA in a sample that is available for information-rich library generation. PreSeq provides insight into the amount of reverse-transcribable and amplifiable RNA fragments while accounting for fragment length along with damage that inhibits downstream enzymatic manipulation and library generation. Here we demonstrate the utility and advantages of PreSeq in screening retrospective samples to establish criteria that identify samples that yield high-value targeted libraries using Archer FusionPlex panels.

Input mass does not predict library quality

Although the input RNA mass is critical for successful library preparation, this metric alone does not represent the amount of RNA in the sample that is available for amplification. To demonstrate this concept, libraries were prepared from 25 separate FFPE samples using the same input RNA amount, as measured by RNA Qubit® (200ng). After sequencing the libraries, the fraction of library reads derived from RNA was measured using the Archer Analysis bioinformatics software. This metric is an indicator of both the chemical availability as well as the fragment length of the RNA molecules in the input; it is also a critical measure of the overall quality of an RNA-derived NGS library.

The fraction of reads derived from RNA significantly varies across libraries, even though all of the libraries used the same RNA input amount *(Figure 1)*. This data demonstrates that input mass does not predict library quality.

PreSeq predicts library quality

PreSeq uses primers that amplify a universally expressed transcript to quantify the amount of amplifiable RNA of sufficient length in a sample *(Figure 2)*. PreSeq works on the principle that a quantification cycle (C_q) threshold can be identified above which insufficient amplifiable material is present for library generation. This principle also holds significant predictive power for successful library generation.

PreSeq integrates into the FusionPlex Workflow

The PreSeq protocol dovetails with the FusionPlex workflow, eliminating the need to sacrifice additional sample for quality control. As shown in Figure 3, a small aliquot of in-progress library preparation is collected after first strand synthesis. The aliquot is diluted and then used in the PreSeq qPCR assay. The remainder of the library preparation is simultaneously carried

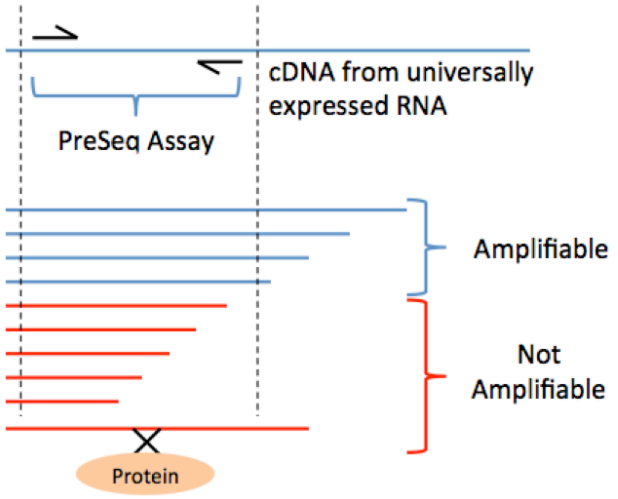

Figure 2. PreSeq measures the quantity of amplifiable RNA fragments in an input sample. *The PreSeq RNA QC Assay targets cDNAs derived from a universally expressed mRNA, which is used as an indicator of input quality. RNA molecules that fail to be reverse transcribed, potentially due to chemical damage, or which produce fragments too short for informative library generation (red fragments) will fail to amplify in the PreSeq qPCR reaction. Only RNA molecules of adequate quality and size in the sample input are quantified (blue fragments).*

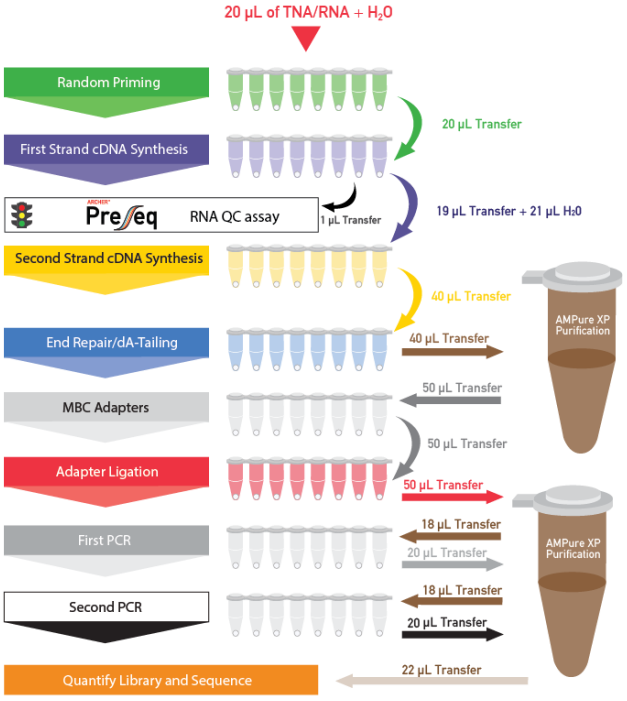

Figure 3. PreSeq integrates into the FusionPlex workflow. *The PreSeq QC protocol is carried out as an in-line and parallel step in the Archer FusionPlex workflow. Note: The workflow has been updated to include an additional magnetic purification step. Please refer to the workflow in your Universal RNA Reagent Kit v2 documentation when preparing a FusionPlex library.*

into second strand synthesis and either halted at that point or continued to library completion.

If PreSeq results indicate that the quality of the input is poor, library preparation can be arrested early in the process, which saves time and sequencing costs. Additionally, if a poor-quality sample is taken to sequencing and subsequently fails the quality filters in Archer Analysis, PreSeq results also serve as confirmation that the input contained insufficient amplifiable material for library preparation.

PreSeq measures RNA quantity and quality

Two experiments were performed to demonstrate the robustness of the PreSeq RNA QC Assay for reporting RNA quantity and quality. In the first, samples with pre-defined RNA amounts were prepared from lung and thyroid tissues, and each sample was then carried through the first two steps of the FusionPlex workflow. An aliquot from each was then analyzed using the PreSeq RNA QC Assay. Figure 4A shows that the C_q values are inversely correlated with input amount, demonstrating that the PreSeq assay can effectively report RNA quantity based on measured C_q values.

To demonstrate that the PreSeq assay provides accurate RNA quality metrics, zinc-mediated fragmentation was performed with Ambion® Fragmentation Reagent separately on equal amounts of RNA purified from fresh colon and prostate samples, along with an FFPE sample. Aliquots were fragmented for various lengths of time to achieve varied fragment length distributions. Reactions were stopped using the provided EDTA solution. Aliquots from each sample were removed, and the DV_{200} was measured using an Agilent Bioanalyzer. DV_{200} is a metric that indicates the percentage of RNA fragments in the sample that are longer than 200 nucleotides. In this experiment, it was used as an orthogonal approach to measure the RNA fragment size. PreSeq C_q value measurements were also collected at each timepoint. Figure 4B shows that DV_{200} values decreased with increased fragmentation time, confirming that the RNA fragment size, and thus the quality of RNA in the sample, continually decreased throughout the timecourse. The PreSeq assay reveals this increased fragmentation via increased C_q values *(Figure 4C)*. Additionally, we found a moderately high inverse correlation between the DV_{200} and C_q values, demonstrating that the PreSeq assay effectively reports RNA quality *(Figure 4D)*.

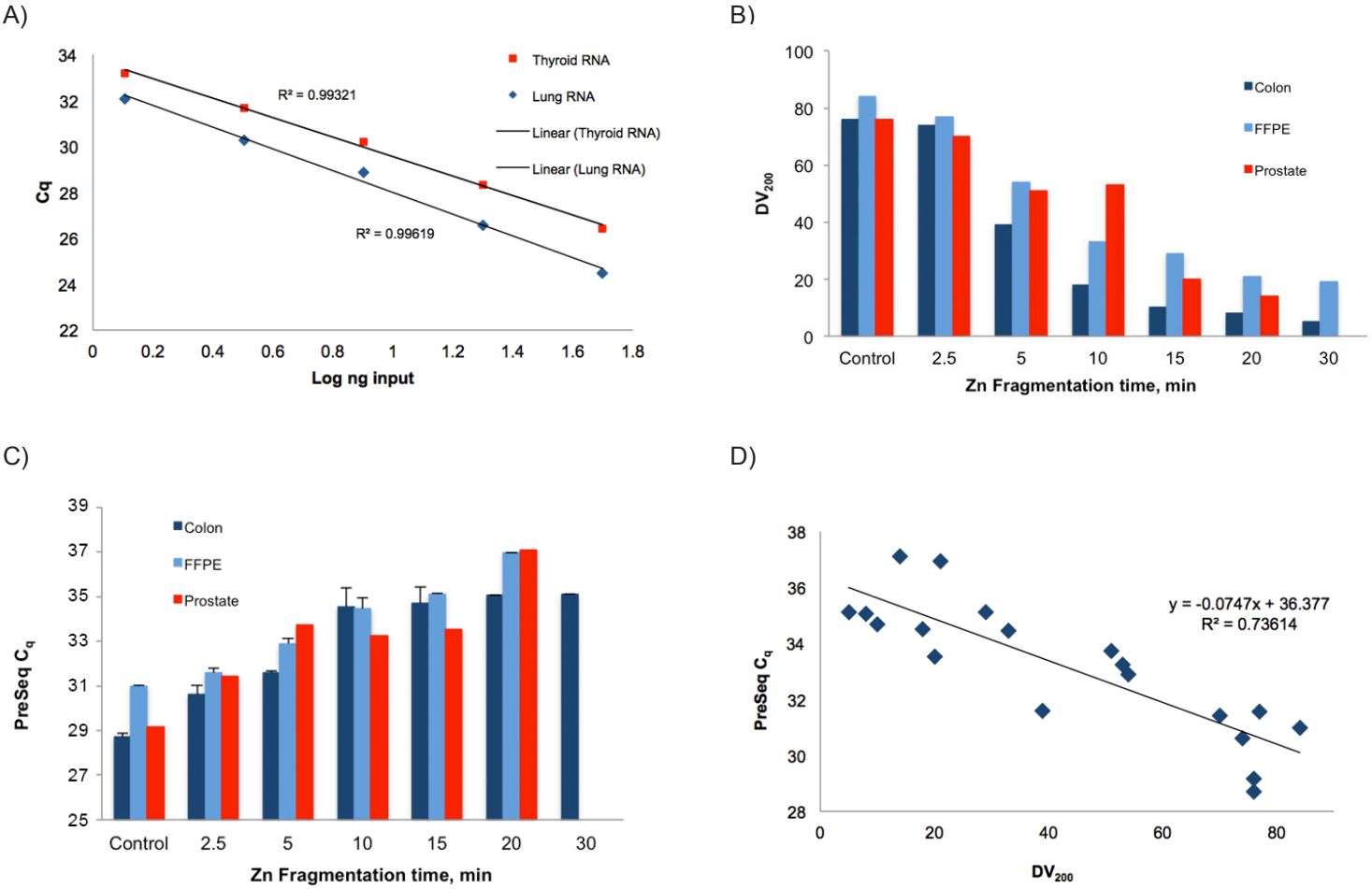

Figure 4. PreSeq is sensitive to RNA quantity and quality. A) *Varying quantities of lung (blue) or thyroid (red) RNA were carried through the first two steps of the FusionPlex workflow. After first strand synthesis, an aliquot of each sample was analyzed using the PreSeq RNA QC Assay. The C_q values were plotted against the log of the input mass, and for both inputs, the expected linear relationship was observed.* ***B)*** *RNA from three sources were individually fragmented in a time-dependent manner in the presence of Zn^{2+}. The DV_{200} was measured at each time point using a Bioanalyzer. As expected, longer incubation times resulted in higher fragmentation and lower DV_{200} values.* ***C)*** *Similarly, PreSeq measurements made at each time point revealed increasing C_q values, indicating that although the input mass was kept constant, PreSeq detects the increase in fragmentation.* ***D)*** *PreSeq C_q and DV_{200} values were plotted against each other, demonstrating the expected inverse correlation.*

PreSeq predicts Archer Analysis QC Result

The Archer Analysis bioinformatics software reports successful library generation using data obtained from 8 internal control primers in the FusionPlex panels. Sequencing reads originating from the control primers are mapped to the reference sequence, and the mean number of unique RNA start sites per control for these reads is indicative of the overall molecular complexity of the library in which the reads were created. Archer Analysis reports a QC pass if a sample generates 10 or more mean control RNA start sites.

The value of PreSeq is its ability to predict if a given sample will pass Archer Analysis QC. To demonstrate this, FusionPlex libraries were generated from RNA purified from 158 FFPE samples using FusionPlex Sarcoma, Heme, Solid Tumor and ALK, RET, ROS1 v2 Panels. PreSeq C_q data and mean control RNA start site data obtained from Archer Analysis for these 158 samples were subjected to receiver operating characteristic analysis to identify the C_q value that best predicts whether or not a library with greater than 10 mean control RNA start sites can be generated from a given input. A C_q value of 28.303 was determined to predict this outcome with 70.6% specificity and 91.8% sensitivity *(Figure 5)*.

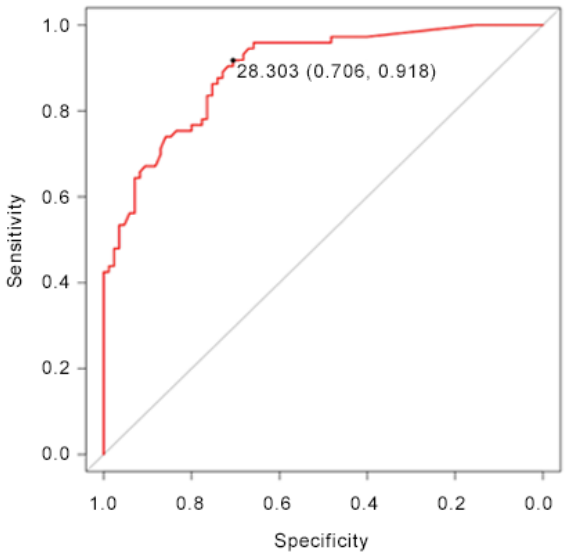

Figure 5. PreSeq values correlate with Archer Analysis QC values. *Libraries were generated from 158 samples using four FusionPlex Assays, with PreSeq assays performed for each. The libraries were then sequenced and analyzed using Archer Analysis. The data was subjected to receiver operating characteristic analysis to identify a PreSeq C_q value that predicts software QC pass. A PreSeq C_q value of 28.3 was found to be most predictive of Archer Analysis QC pass in our lab.*

Fusions are robustly detected in libraries that pass Archer Analysis QC

Library complexity is primarily a function of the quality and quantity of RNA input used in the FusionPlex assay. Insufficient quality or quantity of the starting material will produce low-complexity libraries in which fusion calling capability is compromised, though low complexity can also result from insufficient sequencing depth.

In order to examine the relationship between library complexity and fusion calling sensitivity, the PreSeq C_q data from the aforementioned 158 FusionPlex libraries was plotted against the respective Archer Analysis QC metric (mean number of control RNA start sites). Rearrangements were previously identified either by RT-PCR or FISH in 55 of the 158 samples, and the expected fusions were detected using FusionPlex assays and Archer Analysis in 51 of the 55 confirmed samples *(Figure 6)*. All of the expected fusions were detected in samples with greater than 10 mean control RNA start sites. The four libraries in which fusions were not called had less than 10 mean control RNA start sites. This data indicates that low-complexity libraries can lead to false-negative fusion calls, while libraries with greater than 10 mean control RNA start sites have sufficient complexity to call fusions robustly. This data substantiates the Archer Analysis quality control pass/

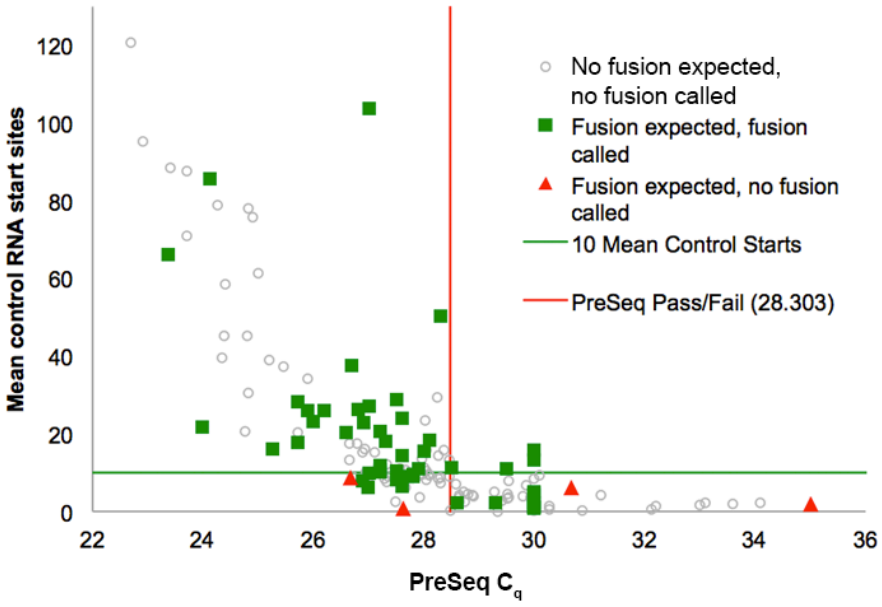

Figure 6. Fusions are called robustly in libraries that pass Archer Analysis QC*. Libraries were generated from 158 samples using four FusionPlex Assays, with PreSeq assays performed for each. The libraries were then sequenced and analyzed using Archer Analysis. The Archer Analysis QC value (mean control RNA start sites) for each library was plotted against the respective PreSeq C_q value. Data points were color-coded on the basis of whether or not a fusion was expected and/or detected in the sample.*

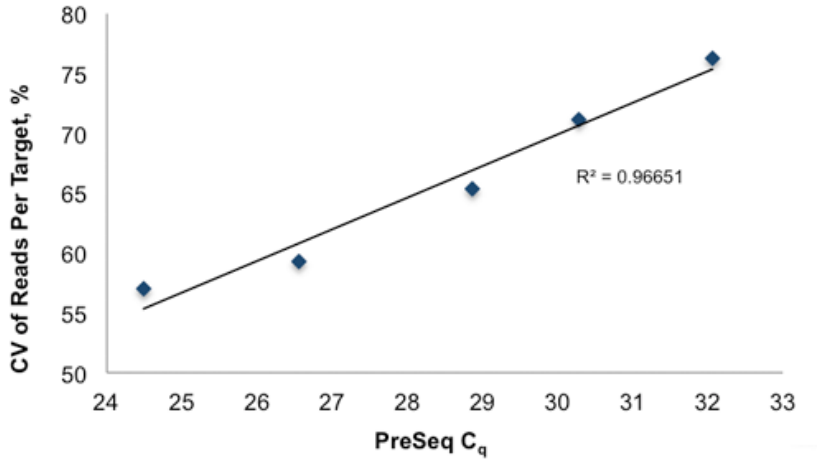

Figure 7. PreSeq values predict the reproducibility of read distribution in FusionPlex libraries*. FusionPlex Solid Tumor libraries were prepared in triplicate from samples with varying PreSeq C_q values. Variability in read distribution was defined as the coefficient of variation (CV) for each probe in each triplicate, averaged across the panel of probes. This value was then plotted against the PreSeq C_q value. As PreSeq C_q increases, the variability in read distribution also increases, indicating that higher quality inputs yield more reproducible FusionPlex libraries.*

fail threshold of 10 mean control RNA start sites.

While expected fusions might be detected in samples that do not pass software QC, sensitivity in this population is reduced, as the 4 samples in which fusions were expected but not called all failed the QC metric. Samples that showed evidence of a genomic rearrangement by FISH were considered to have an expected fusion; however, a positive signal by FISH does not guarantee expression of a fusion transcript. Note that as the PreSeq C_q value increases, the software QC value generally decreases, indicating that PreSeq can identify samples that produce high-complexity RNA libraries.

PreSeq predicts high-confidence libraries

Reproducibility of sequencing results is positively correlated to input quality and quantity. One way to assess reproducibility is to examine the variance of reads per target across replicate library preparations. PreSeq measures quantity and quality of input and hence closely correlates with variance in reads. *(Figure 7).*

PreSeq C_q values vary between labs

While PreSeq performs well in several commercially available qPCR master mixes, we have observed that different qPCR master mixes yield different C_q values, even though the same input amount and instrument were used for data collection *(Table 2).* The C_q value might also vary based on qPCR instrument used. As such, the C_q threshold of 28.3 identified in Figure 6 might need to be adjusted empirically for use in another lab.

Discussion

After validation, most routine sequencing is performed on specimens that are less than weeks old. However, we have observed that many Archer users perform their initial evaluation of the FusionPlex system using archived FFPE samples, some of which are five to ten years old. We therefore designed an inexpensive QC assay that is integrated into the FusionPlex library preparation workflow to identify problematic samples upfront, reduce the number of failed samples and lower the cost of sequencing.

The SYBR-based PreSeq RNA QC Assay is a qPCR assay that selectively amplifies fragments of a specified length to evaluate the likelihood of success of FusionPlex library sequencing. PreSeq C_q values exhibit a strong concordance with the number of mean control RNA start sites in a library, and correspondingly, a strong agreement with Archer Analysis QC pass. Indeed, the value of this assay is that it can determine if a sample can produce a library good enough to detect fusions should they be present in a given sample.
The integration of this quick step into the FusionPlex workflow is inexpensive and simple; by taking a single

Master Mix	Cat Number	R-squared	Efficiency	C_q 1pM	Auto Threshold
BioRad® iTaq™	172-5122	1	89.1	12.51	0.023
BioRad SsoAdvanced™	172-5270	0.999	96.7	16.81	0.90
Thermo Scientific™ Maxima®	K0221	0.999	95.9	15.44	1.06
Applied Biosystems® *Power*	4368577	0.999	94.6	15.80	0.60
KAPA® SYBR® Fast	KK4600	0.999	96.9	16.04	1.59
Qiagen® QuantiTect®	204141	0.999	98.4	17.19	1.17

Table 2. PreSeq performs well in various master mixes, but C_q values vary between master mixes. *A dilution series of DNA template was prepared and various master mixes were used for PreSeq analysis using a single instrument. Reproducibility (r-squared) and efficiency (%) of the assay in each master mix is reported, along with the C_q values at one dilution point on the curve (1pM template). The threshold reported is the default threshold selected by the instrument for each master mix.*

1-µl aliqout of first-strand cDNA product and performing the PreSeq QC assay in parallel with second-strand synthesis, one can effectively eliminate the cost of sequencing a library that would likely fail the Archer Analysis QC.

We have seen two modes of PreSeq utilization emerge:

- Users in low- to mid-throughput laboratories utilize the PreSeq reagents supplied in the Archer Universal RNA Reagent Kit v2 to effectively screen up to 16 samples for every 8 libraries that they choose to sequence.
- High-throughput labs often sequence an entire batch of FusionPlex libraries upfront but reference the PreSeq results as confirmational data when troubleshooting failed libraries.

While C_q values can vary by instrument and SYBR master mix, the implementation of PreSeq into the standard FusionPlex library preparation workflow will build a dataset to establish the predictive value of this assay. The C_q value shown here should be considered a starting point for evaluation but will need to evolve as users build data from multiple samples and sequencing runs. As different laboratories use different SYBR master mixes, qPCR machines, etc., it is more effective for individual labs to develop an individual C_q threshold metric rather than using a generalized cutoff. While the C_q cutoff will ultimately need to be developed based on individual user's definition of high-quality libraries, this self-developed metric will ultimately take on additional meaning, based on the operating procedures within each user group.

For more information, visit archerdx.com/preseq-rna

Limitations of Use
:
For Research Use Only. Not for use in diagnostic procedures.

This product was developed, manufactured, and sold for in vitro use only. This product is not suitable for administration to humans or animals. Safety data sheets (SDS) relevant to this product are available upon request.