WHITEPAPER

Best Practices for Clinical Research Biomarker Studies Using the nCounter® Platform: Strategies to Control for Variability

Stacey Taylor

Camille Troup

Greg Gonye

Kirsteen Maclean

Lori Mull

Ryan Friese

Erin Piazza

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.
NanoString Technologies, Inc., Seattle, WA 98109
MK3415 | June 2021

nanoString®

Scope

To obtain the highest quality data from clinical research biomarker studies, it is crucial to design your experiments with consideration for both platform- and sample-associated sources of variability. Controlling for these types of variability is not only good laboratory practice, but it allows you to uncover relevant associations in your data more easily and accurately. In this white paper, we briefly introduce the nCounter® technology, then propose experimental designs and best practices to minimize the effects of technical variability in your assays. We also present an in-house study and summarize examples from peer reviewed publications to provide context for the application of these concepts in clinical research biomarker studies and laboratory developed tests.

It is important to define the types of variation that can occur in a biomarker study and their respective sources. A batch effect is a general term used to describe the technical variation that could come from a variety of sources, including sample collection methods, hybridization conditions, differences in system settings and performance, or operator behavior. These sources of variability are beyond the scope of this document, but careful planning and experimental design can mitigate their effects. Lot-to-lot variation is a type of batch effect that refers to the technical variability between different lots of reagents used in a study, and this topic is what will be primarily addressed in this document. There are two sources of lot-to-lot variation to consider in a NanoString experiment: consumable lots and CodeSet lots. We will discuss the use of a calibration sample to control for variability between both types of reagents.

As it is assumed that users will have some familiarity with data quality control and analysis, including statistical methods for clinical study design, we do not discuss those topics in detail here. Contact **support@nanostring.com** to learn more about data analysis with NanoString products or to ensure you have the most up-to-date guidance and documentation.

Intrinsic advantages of multiplexed, enzyme-free, single-molecule counting

The NanoString nCounter platform is a direct, enzyme-free, digital counting technology that relies on base pairing and does not require target molecule amplification (1). Due to the robustness and intrinsic technical reproducibility of nCounter technology, it has become a platform of choice for clinical research biomarker studies, especially those using clinical research samples derived from body fluids or damaged from formalin fixation (2).

NanoString's nCounter chemistry utilizes fluorescently barcoded probes that specifically hybridize to target molecules of interest. These target-probe complexes are captured and imaged for direct digital counting on the nCounter platform (3). Positive and negative control probes are built into all nCounter reagents such that every reaction can be internally monitored for performance (4). These controls can be used to calculate quality control (QC) metrics that allow investigators to quantitatively evaluate the performance of both the chemistry and instrumentation across many variables (for more details on calculating QC metrics, refer to the nSolver™ User Manual [5]). Sample processing can also be fully automated using nCounter robotics. The ability to multiplex samples and controls, combined with the simplicity of an enzyme-free reaction, effectively minimizes the need for technical replicates during test development* (2,6).

Strategies and tools for controlling lot-to-lot variability in a biomarker experiment

Lot-to-lot variability may mask changes in your data, potentially resulting in missed associations. This variability may also simulate changes in data, which can frequently be misinterpreted as having biological origin and significance. Lot-to-lot variability can originate from two main sources in a NanoString experiment: differences between lots of consumables or between lots of CodeSets. Below we will address strategies for assessing and mitigating each.

Evaluation of variance from consumables

A NanoString Master kit contains a variety of consumables, including a cartridge, reagent plates, and hybridization buffer. Each of these consumables can differ slightly in the manufacturing process from lot-to-lot, which can introduce variability into your experiment. These components usually affect all targets in a single experiment equally, and hence, the variations can be normalized through use of NanoString's intrinsic positive controls as well as the housekeeping genes in your CodeSet. We recommend normalizing your data for lot-to-lot differences in consumables across a single CodeSet lot.

nanoString®

Use of a calibration sample for normalization and mitigation of variance between CodeSet lots

Differences between lots of CodeSets can also introduce variability into your study. Due to differences in manufacturing, the hybridization efficiency for each probe pair will vary slightly between each new lot of CodeSet. For longitudinal studies across multiple CodeSet lots, it is critical to include one or more calibration samples to calibrate counts across all probes.

A calibration sample could be a pool of RNA (or lysates) of representative samples in the study, a commercially available reference RNA sample, or a pool of target oligonucleotides (6). The ideal type of calibration sample for your study depends on the size of the CodeSet as well as the purpose of your experiment.

In general, the calibration sample should contain targets that produce a robust signal for all probes in the CodeSet, and as much as possible, these targets should mimic the intended natural targets in your experimental samples. Furthermore, the calibration sample should be affordable to generate and relatively easy to reproduce or change when necessary. Table 1 describes the types of calibration samples that can be utilized to normalize data across multiple CodeSet lots.

The calibration sample you choose may be either a biological or synthetic template but should be run and analyzed as outlined in tech note TN_MK3415_Panel Standard.

Calibration Sample	Nucleic Acid	Cost	Pros	Cons
Biological sample(s)	Total RNA (for gene expression) or Genomic DNA (for CNV)	$	• Cost-effective • Easily prepared • Best representation of size and secondary structure of naturally occuring target molecules	• Some targets may be expressed below background levels • Signal may vary considerably • Difficult to reproduce consistently
Unpurified oligonucleotides	Single-stranded DNA (approx. 100 bases)	$	• Cost-effective • Easily prepared • Adequate for most applications	• Variable size profile • Signal varies even nominally equimolar targets • Difficult to reproduce consistently • DNA does not fully represent RNA targets, including size and secondary structure
PAGE-purified oligonucleotides	Single-stranded DNA (approx. 100 bases)	$$	• Full strength for all targets • Easily reproducable	• DNA does not fully represent RNA targets, including size and secondary structure
In vitro-transcribed RNA	RNA (approx. 250 bases or greater)	$$$	• Representative of RNA targets • Can be made to mimic the size of naturally occuring targets • Reproducible	• Expensive, especially for studies involving secondary targets • Labor-intensive
Panel Standard	Single-Stranded DNA (100 bases)	$	• Cost-effective • No preparation needed • Specific to the nCounter Panel you are using	• DNA does not fully represent RNA targets • Only for some nCounter panels

TABLE 1: Overview of different types of calibration samples

Use of a calibration sample for normalization and mitigation of variance between CodeSet lots

The calibration sample can also be used to determine assay linearity for each probe in a CodeSet prior to the commencement of a study. The concentration of the calibration sample should be high enough to produce a robust signal for all target nucleic acids but not so high as to cause saturation of the imaging surface. Low concentrations of sample will produce low and inherently more variable counts, which are less reliable for normalizing experimental data.

Figure 2 illustrates the dynamic range in a NanoString hybridization reaction using an equimolar probe pool of DNA oligonucleotides. A titration curve was generated for a calibration sample at final target concentrations of 0.0625 fM to 16 fM. The accompanying data in Table 2 suggest that 4 fM is a sufficient input concentration to produce high counts and a low coefficient of variance (CV). If using synthetic targets as a calibration sample, NanoString recommends a concentration between 4 and 16 fM, however empirical testing and validation of the optimal concentration for your study may be required.

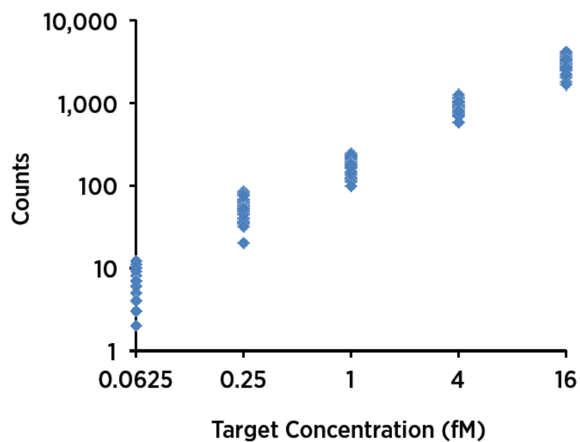

FIGURE 2: Titration of a PAGE-purified DNA calibration sample from 0.0625 - 16 fM of each synthetic target.

Case study comparison of four CodeSet lots

To illustrate the high correlation of expression data across multiple CodeSet lots, NanoString constructed four separate lots of the nCounter Human Inflammation panel, a 189-gene CodeSet. Although the target sequences and barcode assignments remained unchanged, each lot was manufactured separately. Each of the four separate lots was hybridized to a commercially available human RNA calibration sample (referred to as Human Reference RNA). Each set was run in quadruplicate using standard nCounter procedures (3). The data sets were normalized separately to the internal positive controls as described in our Gene Expression Data Analysis Guidelines (8). Normalized counts were averaged across the four replicates for each sample per CodeSet. The number of genes detected in the Human Reference RNA for each CodeSet was determined based on two standard deviations above the average background; genes below background levels were omitted from further analysis.

Precision

We first examined the precision of the quadruplicate measurements as a function of overall expression levels for all four CodeSets. Figure 3 shows a scatter plot of the percent coefficient of variation (CV) versus the average expression for the Human Reference RNA compared across all four CodeSet lots. We can conclude from this that the precision of the assay does not vary between lots and is clearly a function of the number of counts.

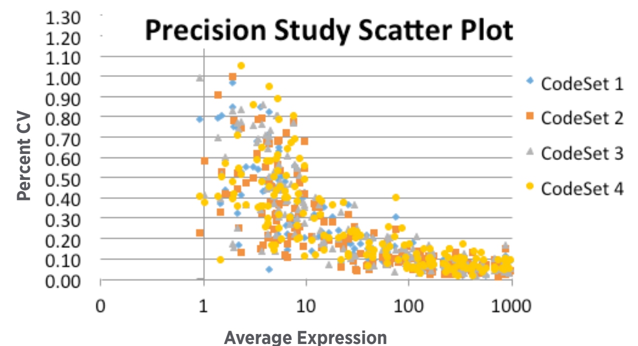

FIGURE 3: Percent CV vs average expression over four CodeSets.

Concentration (fM)	Average Counts	Average SD	Average CV
0.0625	11	3.4	32.3%
0.25	41	5.1	13.2%
1	165	11.9	7.3%
16	2,849	81.7	2.9%

TABLE 2: Average counts, standard deviation (SD), and coefficient of variation (CV) for a titrated calibration sample. CV values were calculated separately for each individual probe in the CodeSet and then averaged.

Correlation in Absolute Expression

The correlation plot in Figure 4 illustrates how raw counts vary across different CodeSet lots. The difference between individual probe counts across lots may vary by as much as 5-fold (data not shown). While these raw counts do have some variance, normalization to controls and housekeeping genes frequently removes much of this. When fold change ratios are calculated for each experiment (see next section), most of the variation in raw counts between lots is eliminated (Figure 5).

Raw Counts Study Correlation Plot

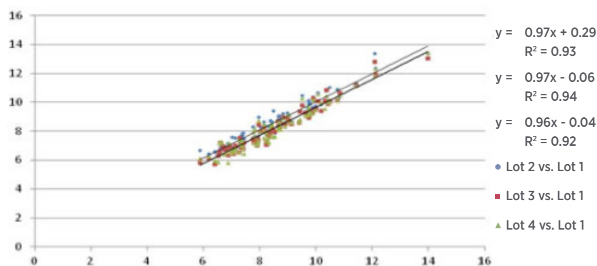

FIGURE 4: Raw counts of CodeSet Lots: 2 vs. 1, 3 vs. 1, & 4 vs. 1.

Fold Change Correlations

We explored how fold changes between Human Reference RNA and a second calibration sample, Brain Reference RNA, differed across the different CodeSet lots. To reduce the impact of low-expressing genes, we removed those with expression levels close to background (lower than 50 counts in the calibration sample for all four CodeSets). The data in Figure 5 is the log2 ratio of Brain Reference RNA to Human Reference RNA.

Using a conservative approach to background thresholding (8), the R2 value for log2 ratios between CodeSet1 and the other three CodeSets is 0.99. Thus, variation in measuring relative changes in gene expression levels between samples using different CodeSet lots is low.

Fold Change Study Correlation Plot

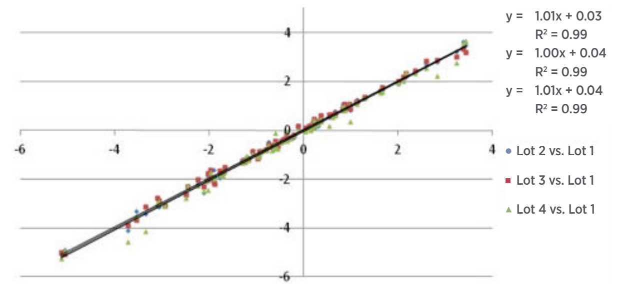

FIGURE 5: log2 ratio of Brain Reference RNA to Human Reference RNA (>50 counts) in CodeSet Lot 2 vs. Lot 1, 3 vs. 1, & 4 vs. 1

Observations from peer reviewed publications

Numerous publications have evaluated the NanoString nCounter platform for use in clinical research biomarker studies. Collectively, these pieces analyze the major technical and biological sources of variability inherent in clinical research biomarker studies. We recommend careful review of the publications referenced in the next section for strategies and recommendations on best practices to successfully implement experiments using the nCounter platform.

Effects of lot-to-lot CodeSet variance

Talhouk et al., 2016 (9) describe a strategy for conducting both prognostic and diagnostic clinical tests using the NanoString platform with a calibration sample to correct for batch effects. In this study, the authors used a DNA oligonucleotide probe pool as a calibration sample, which contained all the targets of interest for cohorts of two cancer samples, ovarian and Hodgkin's lymphoma. Two different CodeSet lots were used to assess ovarian targets, and three different lots were used to assess lymphoma targets. After quality control metrics were assessed, the authors calculated the fold-change in expression levels across all targets relative to the housekeeping genes. Next, they assessed the variability associated with the different CodeSet lots used in the study. They computed a gene-wise percentage change in log-expression between batches of CodeSets, and noted a median percent change of about 10% for both ovarian and lymphoma clinical samples, with some genes more stable than others across the CodeSets. The authors concluded that this relatively small amount of variability can be easily mitigated by good experimental design and appropriate planning.

nanoString®

To further address the potential for variability across multiple barcoded-probe lots, a similar approach was adopted by Adam et al., 2016 (10) using renal allograft biopsies from FFPE blocks. Oligonucleotide probes representing a 34-gene signature were designed and manufactured as an nCounter Elements assay. The authors first performed an RNA titration experiment with four different input RNA quantities: 50, 100, 200, and 400 ng, with replicates run of each. The correlation between the recommended 100ng input and replicates with lower or higher amounts was excellent, with a mean correlation coefficient of 0.998 ($p<0.001$). To compare results across users, three different operators of varying technical experience performed replicate analysis using the same three samples. Again, results showed excellent correlation between results from each operator, with a correlation coefficient of 0.998 ($p<0.001$). To simulate the factors introduced during a multi-site study, a third reproducibility experiment tested six replicates of twelve samples; each sample was run twice with three different barcoded-probe lots. Each pair of replicates showed high correlation, with a mean correlation coefficient of 0.983 ($p<0.001$). Analysis of lot-to-lot variance resulted in high correlation, as well, with an F-test statistic of 0.0001 ($p=0.9998$). The authors conclude from these experiments that NanoString's platform is reproducible across a wide range of input RNA quantities, that it is technically easy to use, and that its methodology is robust.

Effects of sample variance

Acquisition, purification, and quantitation of samples introduce the largest and most important sources of variability in a clinical research biomarker study. A recent publication by Veldman-Jones et. al., 2015 (6) investigated these as well as other sources of variability in multiplexed gene expression analysis using the nCounter platform.

The authors first evaluated the effects of platform-associated variables. They used two different Prep Stations to prepare cartridges, comparing the effects of different FOV scan settings, and examined the impact of using different scanning slots in the Digital Analyzer. In each case, no significant variability or bias was found in the data, underscoring the inherent robustness of the nCounter platform.

Next, sample and tissue type were evaluated for their effects on data reproducibility and platform sensitivity. Various sample types—fresh, frozen, and FFPE—were tested, as were multiple tissue types, including diffuse large B-cell lymphoma (DLBCL), melanoma, gastric, lung, breast, pancreas, and prostate tissues. To assess technical variability, data from replicate DLBCL samples processed in a single day were compared to data from a similar set-up processed on two different days. Low- and high-quality samples were both tested under this scenario to determine the impact of sample quality on nCounter data. No difference in technical reproducibility was observed, as the R2 value was greater than 0.98 for all samples tested.

The authors also assessed technical variability within biological sample replicates by using a set of gastric tumor samples. RNA was extracted from adjacent tissue sections of several gastric tumor blocks and the gene expression profiles across sample replicates were compared for both intra- and inter-sample reproducibility. A good correlation of samples from within the same tumor was observed; the mean intra-sample CV was 6.47% and the R2 values ranged between 0.59-0.93. Comparatively, data from samples taken from different tumor blocks showed a CV of 12.09%, and R2 values between 0.33-0.7. This data suggests that the nCounter platform is robust enough to discern differences in gene expression between different tumor samples.

In addition, this group examined the effect of sample quality and quantity on nCounter performance. To assess sample quality, gene expression levels between matching fresh, frozen, and FFPE sample types were compared. The correlations were found to be robust, as only small differences between matching fresh and frozen replicates were observed, and these values were comparable with technical variation. To assess the effects of sample quantity, fresh versus FFPE samples were titrated, showing that equivalent data could be generated with 100 ng and 400 ng, respectively. Four low-quality FFPE lung samples were evaluated across a 2-fold dilution series from 50 ng to 3.13 ng. Even at a low input of 6.25 ng, the highly-expressed genes retained a good correlation relative to those in 100 ng samples, illustrating that even small inputs of low-quality samples can generate meaningful data with the nCounter platform.

In light of their results, Veldman-Jones et. al. concluded that "the nCounter platform [is] favorable over other techniques based upon sensitivity, technical reproducibility, robustness, ease of use, hands-on analysis time, and utility for clinical application".

Effects of sample quality

Omolo, et. al., 2016 (11) describe the development and validation of a RAS pathway signature score for colorectal cancer samples derived from fresh frozen and FFPE tissue. This cross-platform study evaluated Next Generation Sequencing, microarray platforms, and NanoString nCounter technology. The initial cohort included 54 samples, 27% of which were of poor quality. The NanoString platform alone was able to replicate the signature on FFPE when all samples were included, with only 39 samples in the experiment successfully analyzed by the other platforms. The authors concluded that the NanoString platform is best suited to analyze poor quality samples.

Summary

Given the regulatory requirements and significant financial investment of undertaking a clinical research study, it is important to understand the sensitivity of any diagnostic test to the possible intrinsic sources of variance. The inherent robustness of nCounter technology, coupled with over ten years of continuous improvements in our manufacturing processes, has resulted in a “best-in-class” technical reproducibility when using clinical research samples. The nCounter platform is ideal for longitudinal biomarker studies. It generates reproducible and robust measurements from diverse clinical samples, which are notoriously variable in both quality and quantity. Successful execution of an oncology biomarker study relies upon the use of a carefully selected calibration sample and a sample population which represents biological sources of variability across the disease state

Best practices for clinical research studies using the nCounter platform

- **Consider appropriate study design as early as possible in study development; to optimize your experimental design, diagnostic development expertise is available from NanoString in a collaboration setting.**
- **Select appropriate synthetic or biological calibration samples to evaluate and normalize for the inherent variance in your study.**
- **Consider regulatory requirements and the final format of your actual test (multi-analyte signature, multiplex pathogen detection, companion diagnostics, required controls, etc.) as early as possible in the study design and development.**

1. Geiss et al. (2008) “Direct multiplexed measurement of gene expression with color-coded probe pairs.” Nat Biotechnol .

2. Northcott et al. (2012) “Rapid, reliable, and reproducible molecular sub-grouping of clinical medulloblastoma samples.” Acta Neuropathol

3. nCounter XT Assay User Manual (MAN-10023). https://www.nanostring.com/download_file/view/602

4. Kulkarni (2011) “Digital Multiplexed Gene Expression Analysis Using the NanoString nCounter System”. Current Protocols in Mol Bio.

5. nSolver Analysis Software User Manual (MAN-C0019) https://www.nanostring.com/download_file/view/1168

6. Veldman-Jones et al. (2015) “Evaluating Robustness and Sensitivity of the NanoString Technologies nCounter Platform to Enable Multiplexed Gene Expression Analysis of Clinical Samples.” Cancer Res.

7. Urrutia et al. (2016) “Standardized Whole-Blood Transcriptional Profiling Enables the Deconvolution of Complex Induced Immune Responses.” Cell Reports.

8. Gene Expression Data Analysis Guidelines (MAN-C0011). https://www.nanostring.com/download_file/view/251

9. Talhouk et al. (2016). “Single-Patient Molecular Testing with NanoString nCounter Data Using a Reference-Based Strategy for Batch Effect Correction.” PLoS One.

10. Adam et al. (2016) “Multiplexed color-coded probe-based gene expression assessment for clinical molecular diagnostics in formalin-fixed paraffin-embedded human renal allograft tissue.” Clin Transplant.

11. Omolo et al. (2016) “Adaptation of a RAS pathway activation signature from FF to FFPE tissues in colorectal cancer.“ BMC Med Genomics.

For more information, please visit nanostring.com

NanoString Technologies, Inc.
530 Fairview Avenue North
Seattle, Washington 98109
T (888) 358-6266
F (206) 378-6288
nanostring.com
info@nanostring.com

Sales Contacts
United States **us.sales@nanostring.com**
EMEA: **europe.sales@nanostring.com**
Asia Pacific & Japan **apac.sales@nanostring.com**
Other Regions **info@nanostring.com**

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.
© 2021 NanoString Technologies, Inc. All rights reserved. NanoString, NanoString Technologies, nCounter, nSolver, IO360, 3D Flow, Vantage 3D, and 3D Biology are registered trademarks of NanoString Technologies, Inc., in the United States and/or other countries.

JUL 2021 MK3415