WHITEPAPER

Introduction to GeoMx® Cancer Transcriptome Atlas: Normalization

MK2833 | October 2020
NanoString Technologies, Inc.

nanoString®

Table of Contents

Overview . . . 3
Conceptual Background . . . 3
Dataset Background . . . 3
Data quality at a glance . . . 3
AOI QC . . . 4
Evaluating background . . . 4
Evaluating signal strength . . . 5
Gene QC . . . 5
Understanding Other Technical Variables . . . 6
Normalization . . . 6
Q3 Normalization . . . 7
Background Normalization . . . 8
Area Scaling or Nuclei Scaling . . . 8
Background Subtraction . . . 8
Evaluating Normalization Methods . . . 8

Introduction to GeoMx® Cancer Transcriptome Atlas: Normalization

Overview

This vignette demonstrates basic QC and normalization considerations for GeoMx® Cancer Transcriptome Atlas (CTA) data (1800+ genes).

We will:

- Demonstrate how to QC ROI/AOI segments and genes for sufficient signal.
- Explore the relationship between signal strength and background.
- Evaluate multiple normalization methods.
- Illustrate how to select a normalization method; Q3 normalization is typically preferred.

(Note: we present here gene count data that has been sent through the initial QC and Biological Probe QC steps in the GeoMx® DSP Data Analysis Suite)

Conceptual Background

The purpose of normalization is to adjust for technical variables, such as differing ROI/AOI surface area and tissue mRNA quality and enable meaningful biological and statistical discoveries.

We will focus on the two primary technical measures that should be understood to achieve robust normalization:

1. Signal strength: a measure of on-target counts as captured by higher expressing endogenous genes (e.g. Quartile 3).
2. Background: a measure of off-target counts as captured by unique negative probes; these probes target no known human transcript; in the CTA assay, multiple unique negative probes are condensed into one NegProbe count (geomean of the multiple negative probes) per ROI/AOI.

We study signal strength and background because they are both measured precisely and capture the impact of most other technical variables.

Dataset Background

This vignette utilizes CTA data from a small study comparing two different colorectal cancer tissue slides (CRC 13, CRC 16). 23 ROIs were sampled for each tissue. Each ROI was segmented into PanCK+ (tumor) and PanCK- (tumor microenvironment) AOIs. Thus, we have a total of 46 segments sampled per tissue.

The visualization below (Figure 1) summarizes the study design.

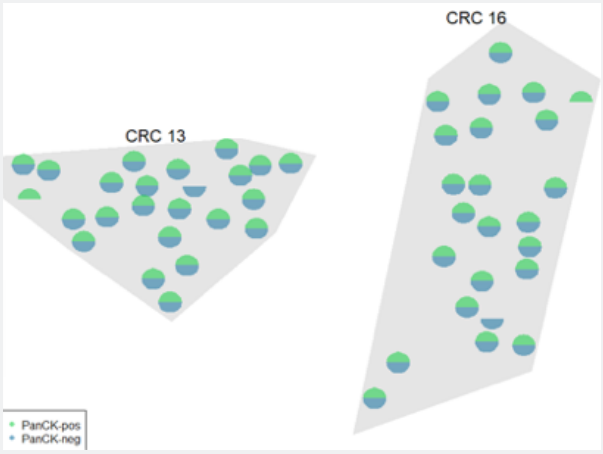

FIGURE 1

Data quality at a glance

For a high-level survey of the data, we plot a heatmap displaying the signal-to-background ratio (SNR) for every gene per AOI as well as the proportion of detected genes per AOI (Figure 2). Here we define a gene as "detected" if the SNR > 2, while an SNR of 3 is also reasonable. We will discuss background in more detail in the section Evaluating Background.

With this visualization, we are looking to:

- Develop a basic understanding of the signal levels in relation to background across the dataset.
- Visualize how AOIs vary in their proportion of detected genes.
- Approximate how many genes are above background based AOI Type and Tissue ID.

Observations for this dataset:

- We observe a rich and diverse dataset; several AOIs show large numbers of genes above the background.
- AOIs vary in the extent and distribution of their above-background expression as a function of both AOI Type and Tissue ID.
- A few AOIs have particularly low rates of above-background counts, and it might be better to remove them from the study. We will revisit excluding segments with low signal in AOI QC.
- There are small bands of genes in the heatmap with low SNR. We will discuss filtering targets below background when discussing Gene QC.

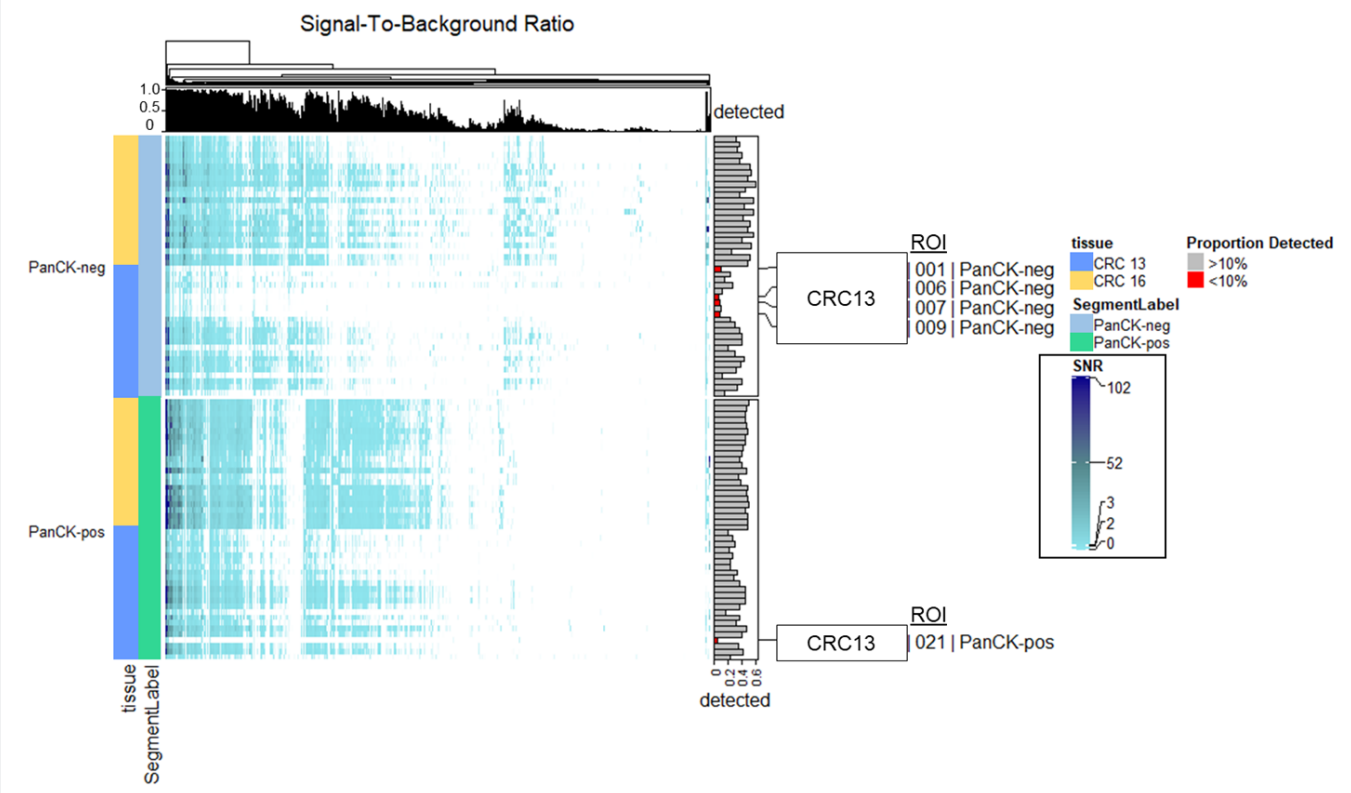

FIGURE 2

AOI QC

The purpose of AOI-level QC is to identify low-performing AOIs that should be removed. We begin by investigating the two primary technical variables: signal strength and background. Typically, the two are highly correlated, but they are not redundant with each other. In other words, they measure overlapping but not identical technical effects.

Evaluating background

Background arises in CTA data when probes bind non-specifically to nucleotides, proteins, and ECM material. To measure background, the CTA panel has 80+ negative control probes targeting sequences not present in the human genome. Together, these probes provide an accurate estimate of each segment's background level. When CTA data is processed into gene-level measurements, these probes are averaged with a geometric mean to create a single measurement, "NegProbe".

Let's create a histogram (Figure 3) plotting NegProbe counts for the study to investigate further (note the use of a Log2 x-axis).

We see a wide range of background levels across the study segments. This is typical, but it speaks to the importance of accounting for this technical difference in the context of both

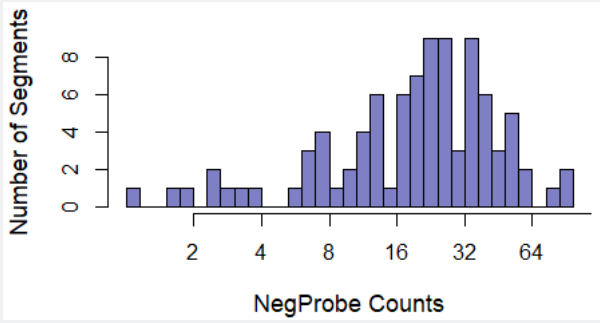

FIGURE 3

Gene QC and Normalization, especially when analyzing low-expression genes. We do not typically exclude AOIs from a study based on background alone.

We can, however, use the background to determine which genes are poorly detected and may not be suitable to include within the study (discussed in Gene QC).

To determine AOIs that may need to be removed from a study, we study background in comparison to on-target signal strength.

Evaluating signal strength

AOI segments differ in their tendency to return on-target counts. Variables such as surface area, amount of targetable mRNA, and in-situ hybridization binding may differ between segments; we use the term "signal strength" to refer to their collective effects on endogenous gene count levels.

To define and quantify signal strength, we take a representative quartile of the data from each segment. Taking the median count has been a popular approach when measuring signal strength in microarray and RNAseq datasets. In GeoMx data, the median count is applicable, but we typically take the Quartile 3 count (Q3) as a higher expressing measure of signal strength. Q3 is simply the 75th percentile gene count value for an AOI; in other words, Q3 is a single gene count (or a mean of two gene counts based on the odd/even number of total genes).

To confirm that the Q3 count is a reflective measure of signal for an AOI, we would like to confirm it is above background. The below plots (Figure 4) address this consideration (note the use of a log-scale). Solid lines track background, and dashed lines show 2 times the background level.

We see that chosen measure of signal is safely above background; most AOIs have a Q3 signal that is 2-fold or higher above background. We do not see any segments that drop to the background level and need to be removed.

Regardless of background, segments with substantially lower signal strength relative to the rest of the data can be unreliable for analysis, and we may consider removing them from the study. Let's create a histogram (Figure 5) plotting Q3 counts for the study to investigate further (note the use of a Log2 x-axis).

With this visualization, we're looking to:

- Understand the range of signal strength in the dataset.
- Look for segments with outlier low signal values that should be removed.

Based on the histogram, we see a wide range of signal strength values, which is common in studies with segmented ROIs. There is one segment with substantially lower signal than the others (red line). This segment is best removed. Removing the low signal segments on the left tail of the histogram could also be reasonable (blue line).

Background considerations with CTA spike-ins

In datasets with multiple probe pools, for example the CTA panel plus a spike-in, each probe pool will have a distinctive background level. Analyses must be performed separately for the data from each probe pool. This will result in two expected background levels per segment: e.g. 1800 CTA genes with one background level, and 30 spike-in genes with a different background level.

Gene QC

Before performing normalization, it is valuable to remove genes with insufficient signal. The general approach is to test each gene for whether its study-wide signal is above background. Based on the definition for background and AOI representation, the number of genes we choose to exclude can differ.

As a basic heuristic to illustrate this concept of Gene QC, let's define a limit of quantitation (LOQ), below which we exclude genes. For example, we can define the LOQ as 2 standard deviations above the NegProbe background value; we could choose to be more or less conservative. At this point, we need to define what percentage of AOIs a gene should be above the LOQ for, before it is included. The smaller the percentage, the more genes we keep (and the fewer genes we filter out).

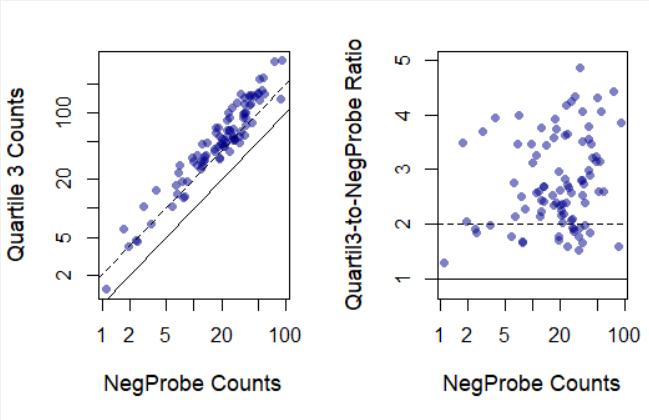

FIGURE 4

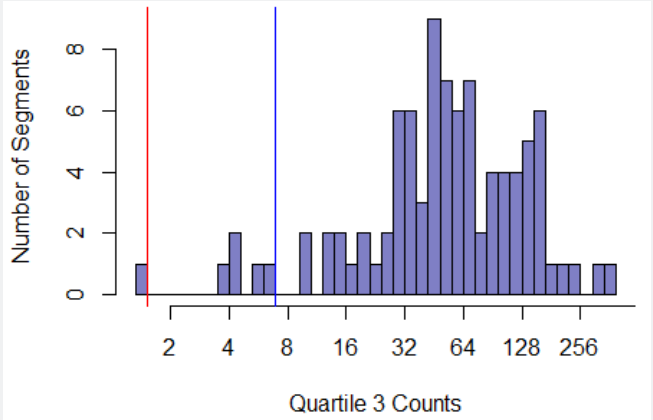

FIGURE 5

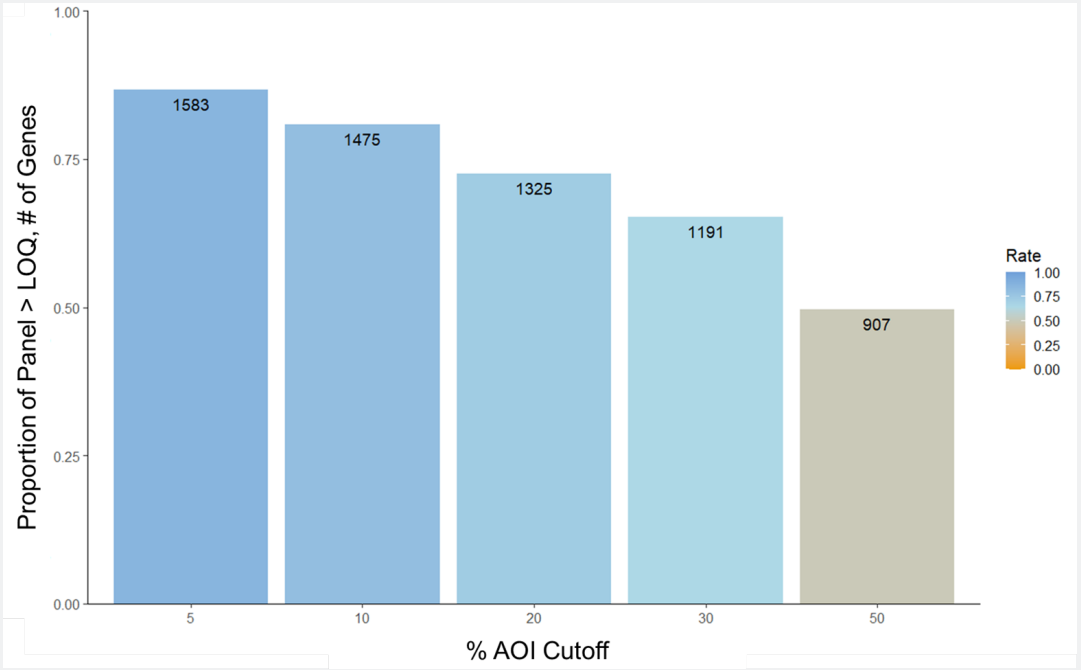

FIGURE 6

Here is a visualization (Figure 6) of what the number of genes to be analyzed would look like based on different percentage AOI cutoffs.

Based on the % AOI cutoff we choose, we will filter out the low-signal genes from most of our analyses.

Note: we do not drop filtered genes from the data entirely; some analyses may make use of the information that a gene is not detectable; alternatively, some genes may simply be low expressors that are biologically relevant in a study.

Understanding Other Technical Variables

Now that we have understood signal strength and background and their relationship, we would like to understand how other technical variables, such as cell populations, tissue slide ID, and clinical annotations, are behaving in a study.

The below plots (Figure 7) are a useful way to understand the signal/background relationships as a function of these different technical variables.

(Note the log-scale axes on the plots)

Our goal here is to judge whether normalization using only one technical effect (signal or background) will adequately address considerations for the data.

Specifically, we ask the following questions:

- Is the ratio between signal and background constant or highly variable?
- Does signal/background correlate preferentially with some study variable like tissue ID or segment type?

Observations from the plots for this study:

- Signal-to-background ratio is fairly consistent, with the highest ratio being ~2-fold above the lowest ratio; this variability is reasonable, but we will consider this further during when discussing normalization.
- Signal-to-background ratio is higher in CRC 16 than CRC 13; this introduces some bias; for example, if we normalize to signal strength, the background counts could be inflated in the CRC 13 data; as a result, near-background genes could appear higher in that tissue. So, if our objective is to compare these tissues, we may need to address this bias with the appropriate normalization strategy.

Normalization

Normalization is a prerequisite to interpreting spatial data. The goal is to account for technical effects (e.g. segment area, tissue quality, etc.) to the greatest extent. Different analyses and study designs may warrant different normalization approaches. Here we introduce, demonstrate, and evaluate potential normalization strategies for our colorectal cancer dataset.

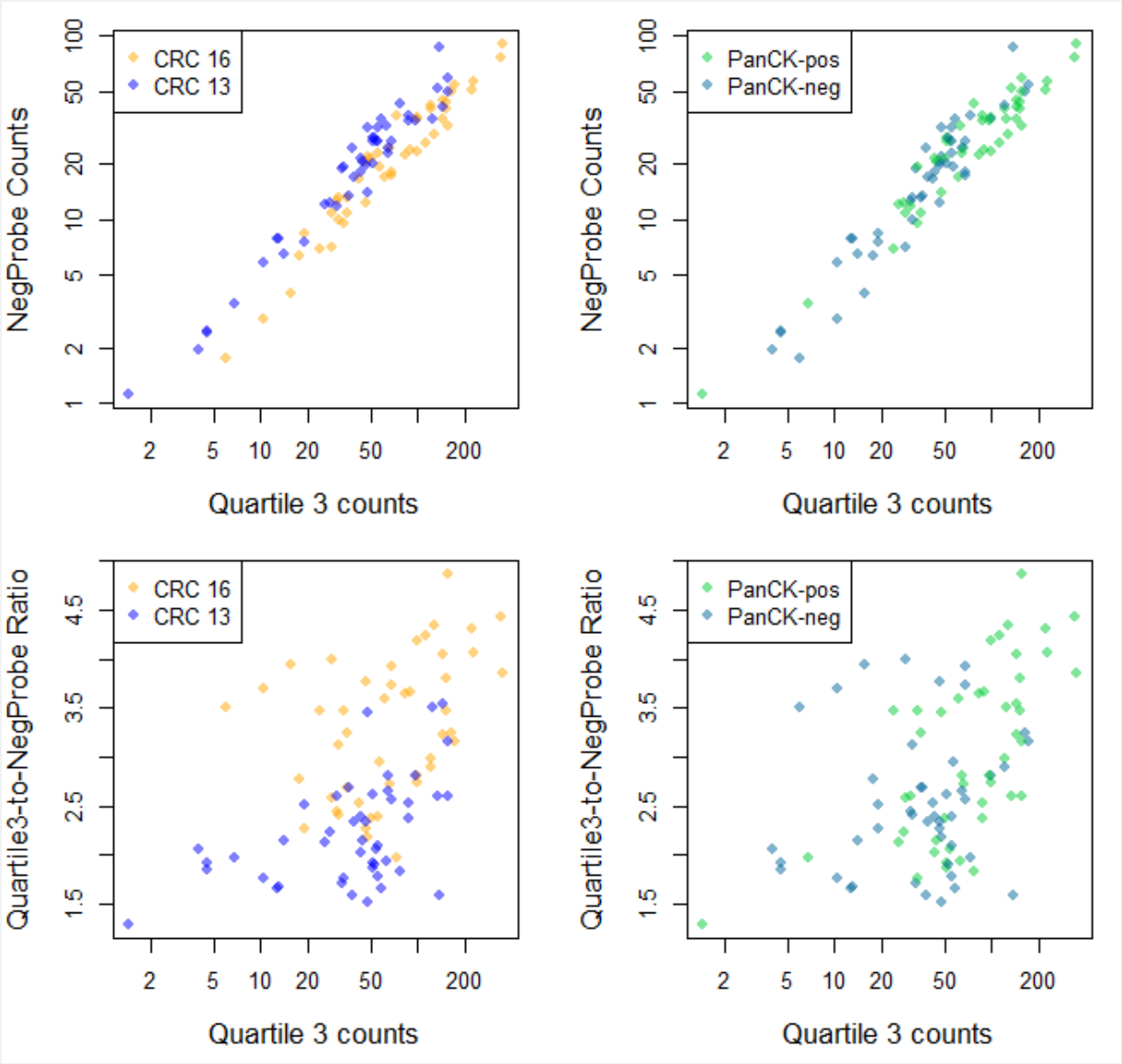

FIGURE 7

Normalization methods include:

- Q3 normalization
- Background Normalization
- Area Scaling or Nuclei Scaling (not preferred)
- Q3 normalization with Background Subtraction
- Background Normalization with Background Subtraction

Note: a given dataset should not be consecutively normalized multiple times (e.g. Q3 normalization followed by Area scaling) as this could lead to "double" normalization of your data. We can certainly, however, compare different normalization methods for a given dataset (e.g. Q3 normalization vs. Background Normalization).

Q3 Normalization

With this method, we scale our AOIs so that they all have the same value for their Quartile 3 value. In particular, we i) first, divide all the genes per AOI by their respective Q3 count, and ii) second, multiply all the genes in all AOIs by a constant, defined as the geometric mean of Q3 counts for all AOIs.

This approach generally performs well for CTA data and is often the preferred normalization.

When approaching Q3 normalization, we keep in mind that this approach does not necessarily capture differences in background; in datasets where signal strength and background

are not in a consistent ratio and correlation with each other, background effects will persist in the normalized data. Another consideration is that segments may differ in true overall expression levels. Q3 normalization does normalize this difference.

Background Normalization

The NegProbe count for CTA data captures many of the sources of variability underlying signal strength. Thus, under the appropriate conditions, the NegProbe readout is a useful normalization factor that captures technical effects without being skewed by changing gene expression levels.

When approaching background normalization, we keep in mind that this approach does not necessarily capture differences in signal strength; in datasets where signal strength and background are not in a consistent ratio and correlation with each other, effects from differing signal strength will persist in the normalized data; as shown in Figure 7, we observe this mild effect when comparing CRC 13 to CRC 16; in such instances, we may consider background subtraction before normalization. Lastly, when considering background for normalization, we typically look for NegProbe values > 10. Single digit NegProbe values are usually statistically unstable to serve as a denominator for normalization.

In short, normalization tackles the following technical variables:

- Segment area: this impacts signal strength and background in nearly an identical way.
- Binding efficiency: this may differ slightly between on-target and off-target counts, though in most experiments there is little discernable difference.
- Amount of RNA: this is likely not well-captured by the negative probes, since much of their binding is not necessarily to RNA but biological material generally.

Area Scaling or Nuclei Scaling

Segment surface area or nuclei count provide a partial measurement of signal strength. But because these variables do not capture ROI/AOI differences in tissue binding efficiency, they may not capture important technical variability in the data, particularly for inter-tissue comparisons. Neither of these approaches are typically used or recommended.

Background Subtraction

Background subtraction is not a normalization method but rather a correction technique that is used prior to normalization. A key reason to perform background subtraction is that AOIs have widely varying ratios of signal strength to background, making it difficult for normalization alone to remove both technical effects from the data. For example, in this dataset, we see higher background levels in CRC 13 than in CRC 16. Note that this concern is lessened in the context of differential expression testing, where technical effects can be explicitly accounted for in well-chosen statistical models (e.g. linear mixed model).

The use of background subtraction does require care for downstream data analysis when the log-scale is used. Background-subtracted data is rich in 0s, and the log of 0 is -Infinity. A simple correction avoids this problem: select a small number, e.g. the largest non-zero value in a dataset, and round all data points below that number up to it (e.g. thresholding).

(Note: perform thresholding prior to log-scale analyses, but continue to use the unthresholded data for linear-scale analyses and plots)

Evaluating Normalization Methods

Let's compare our 4 reasonable normalization methods:

1. Q3 Normalization
2. Background normalization
3. Background subtraction then Q3 Normalization
4. Background subtraction then Background normalization

We use two sets of plots to compare normalization methods: i) heatmaps and ii) principal component plots.

First, let's draw heatmaps (Figure 8) to create an organized visualization of the data under each normalization schema. Note: for the heatmap visualizations, we scale every gene separately to have at most a maximum value of 1.

We make the following key observations:

- Segments cluster by cell population first (e.g. Tumor, TME) and Tissue ID second (e.g. CRC 16, CRC 13); this clustering is especially clear for Q3 normalized data.
- We do not see genes cluster by expression level (green color bar displayed on left of each heatmap).
- No normalization method has created data with an imbalance, such that, for example, one tissue or study variable has uniformly higher expression levels than another.

Our second QC step to compare normalization strategies is principal component analysis (PCA). We plot the first principal components of the data after each normalization (Figure 9). Before running PCA, we log-transform the normalized data. For the background-subtracted data, we first lower threshold the data at its smallest non-zero value.

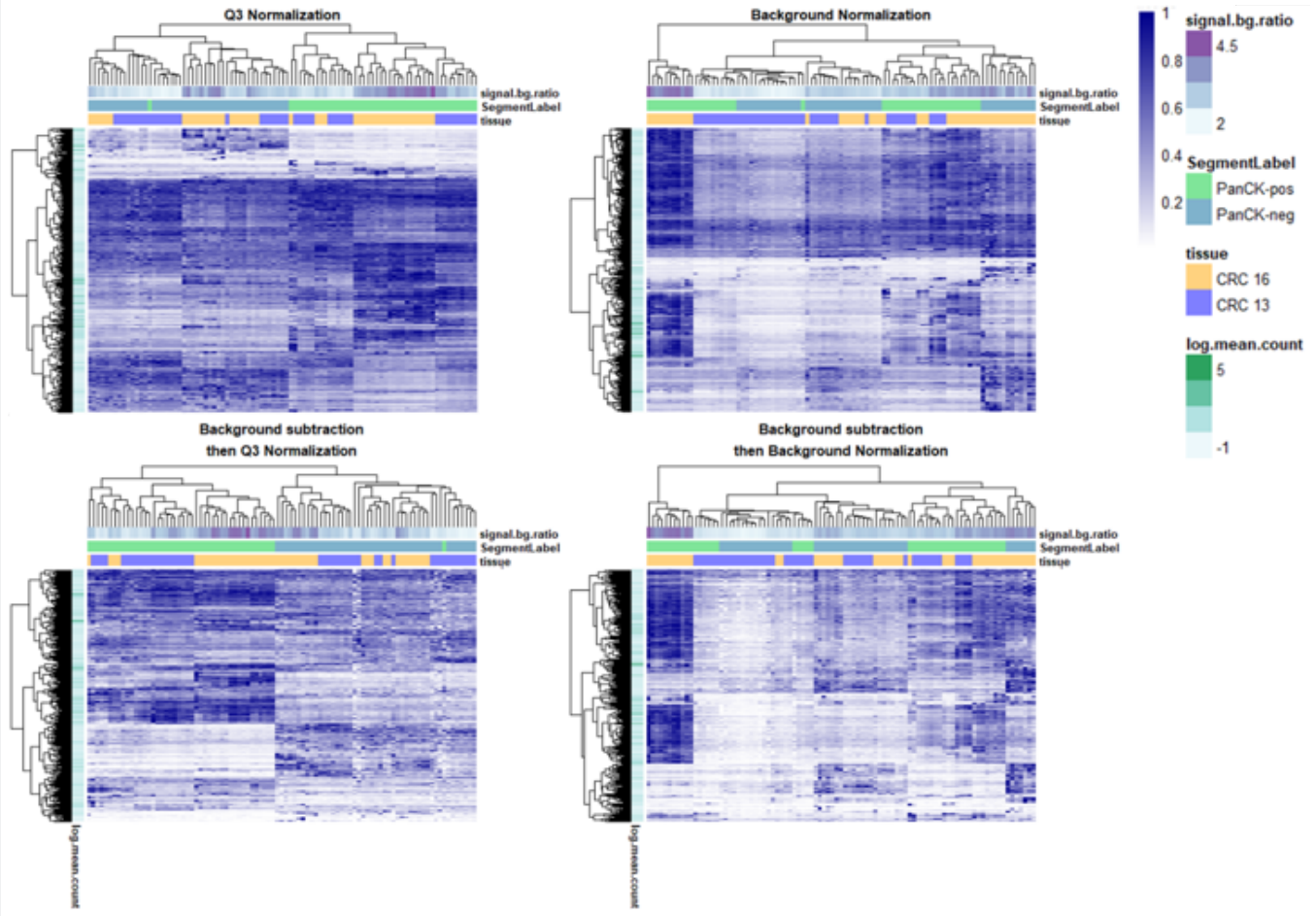

FIGURE 8

Here is what we generally look for in PCA plots when comparing normalization methods:

1. If different segment types are present, they are a major driver of the PCs.
2. Technical effects like the NegProbe value and the signal strength are not major drivers of PCs.
3. Tissue ID is an appropriately strong/weak signal in the PCs (what is appropriate depends on the study).

For our case study, we observe the following:

- In all normalizations, the two segment types (Tumor, TME) are well-separated in the PC space; this is what we would expect.
- In all normalizations, the signal-to-background ratio affects clustering; this is intuitive as we would expect higher SNR AOIs to display more biological differences and more pronounced cluster separation; we should ensure that when presented with an interesting finding, this finding cannot be entirely explained by this technical effect.
- Tissue ID (CRC 16 vs. CRC 13) appears to be a major contributor for the first two PC's; this is largely expected given that these two tissue are two different colorectal cancer types (MSS vs. MSI); given that CRC 13 has lower signal relative to background this may partly explain the PC clustering by signal-to-background ratio.
- In the normalization-only (i.e. no background subtraction) approaches, the two clusters approach each other when signal-to-background is low, while when using background subtraction, they remain well-separated along their lengths. This suggests that background subtraction is correcting some of the technical effects arising from SNR differences between AOIs.

Conclusion

Based on the heatmaps and principal component analyses for this study, we could comfortably use any of our four normalization strategies. However, we would lean towards one of our two Q3 normalization approaches (e.g. with or without background subtraction) due to the clear heatmap and PC clustering by cell compartment and colorectal cancer type.

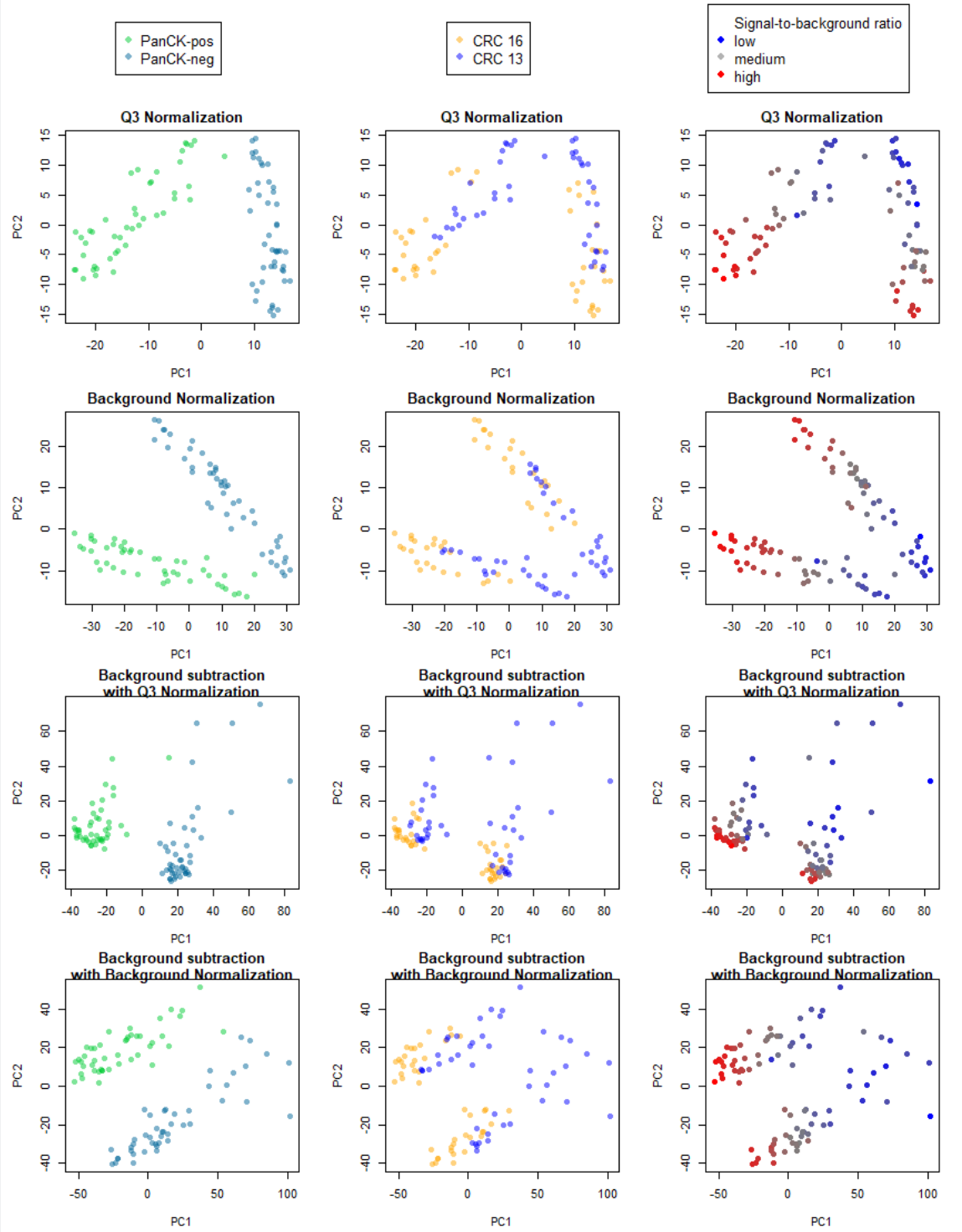

FIGURE 9

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

OCT 2020 MK2833