An approach for normalization and quality control for NanoString RNA expression data

Arjun Bhattacharya[†1], Alina M. Hamilton[†2], Helena Furberg[3], Eugene Pietzak[4], Mark P. Purdue[5], Melissa A. Troester[2,6], Katherine A. Hoadley[7,8], Michael I. Love*[1,8]

1. Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27514, USA
2. Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27514, USA
3. Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA
4. Urology Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, New York, United States
5. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, 20850, USA
6. Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27514, USA
7. Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27514, USA
8. Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27514, USA

*To whom correspondence should be addressed.

Email: milove@email.unc.edu

[†]The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors.

ABSTRACT

The NanoString RNA counting assay for formalin-fixed paraffin embedded samples is unique in its sensitivity, technical reproducibility, and robustness for analysis of clinical and archival samples. While commercial normalization methods are provided by NanoString, they are not optimal for all settings, particularly when samples exhibit strong technical or biological variation or where housekeeping genes have variable performance across the cohort. Here, we develop and evaluate a more comprehensive normalization procedure for NanoString data with steps for quality control, selection of housekeeping targets, normalization, and iterative data visualization and biological validation. The approach was evaluated using a large cohort ($N = 1{,}649$) from the Carolina Breast Cancer Study, two cohorts of moderate sample size ($N = 359$ and 130), and a small published dataset ($N = 12$). The iterative process developed here eliminates technical variation more reliably than the NanoString commercial package, without diminishing biological variation, especially in long-term longitudinal multi-phase or multi-site cohorts. We also find that probe sets validated for nCounter, such as the PAM50 gene signature, are impervious to batch issues. This work emphasizes that preprocessing of gene expression data is an important component of study design.

INTRODUCTION

The NanoString nCounter platform offers a targeted strategy for gene expression quantification using a panel of up to 800 genes without requiring cDNA synthesis or amplification steps (1). The technology offers key advantages in sensitivity, technical reproducibility, and strong robustness for profiling formalin-fixed, paraffin-embedded (FFPE) samples (2). Given these advantages, nCounter is increasingly being used to study differential gene expression (3–6). nCounter is especially attractive for longitudinal studies involving FFPE samples carried out over several years (7) and diagnostic assays in clinical settings (8, 9).

Proper normalization and quality control of mRNA expression is necessary prior to statistical analysis to reduce unwanted variation that may be associated with technical batches or RNA degradation from sample fixation (10, 11). While some sources of variation can be measured or enumerated *a priori*, such as different research centers, batches over time, or RNA preservation metrics, not all can be captured. In

all cases, it is advisable to define a quality control and normalization pipeline to detect and account for technical variation in downstream statistical modeling. All normalization methods deal with a trade-off between bias that needs correction and bias or variance that may be introduced in normalization (12).

A variety of statistical approaches have been developed to normalize nCounter expression data. NanoString provides two forms of normalization in its nSolver Analysis Software (13): (A) a graphical user interface with optional background correction and positive control and housekeeping gene normalization and (B) the Advanced Analysis tool, an add-on that draws on the *NormqPCR* R package (14, 15) to select co-expressed housekeeping genes prior to positive control and housekeeping gene-based normalization. The *NanoStringNorm* package implements the nSolver normalization algorithms in R, combining various ways to use positive, negative, and housekeeping controls on a given nCounter chip (16). For nCounter datasets with technical replicates, Molania et al have recently suggested an unsupervised method, Remove Unwanted Variation-III (*RUV-III*), along with an iterative process to normalization, wherein several parameters (i.e. number of housekeeping genes, number of detected outliers, number of dimensions of technical noise) are tuned with relevant visual and biological checks (11). Originally developed for RNA-seq data, *RUVSeq* is a normalization method that makes use of endogenous housekeeping genes to estimate unwanted technical variation (12, 17) and has been recognized as an effective method for the normalization of RNA-seq data (18). Interestingly, results from Molania et al showed that *RUV-III* normalization frequently outperformed nSolver normalization by more efficiently removing technical sources of variation while preserving biological variation (11). However, as many nCounter gene expression cohorts do not have technical replicates, we extend Molania et al's iterative framework using *RUVSeq*, a precursor to *RUV-III*.

Here, we provide a framework for the quality control and normalization of mRNA expression count data from the NanoString nCounter platform, using a large dataset ($N = 1{,}649$) of breast tumor expression from the Carolina Breast Cancer Study (CBCS) and three other cohorts of differing sample size ($N = 12, 130,$ and 359). We illustrate some of the pitfalls in the nSolver method of background correction and positive control normalization and provide an alternative approach that uses *RUVSeq* (10, 12). We find

that, especially in longitudinal, multi-phase or multi-site cohorts, *RUVSeq* outperforms nSolver in removing differences across various technical sources of variation. Lastly, we provide quality checks for normalization and outline the impact of proper normalization on inference for biological associations and expression-based disease subtyping.

MATERIAL AND METHODS

Data collection

We considered four cohorts with nCounter gene expression data when evaluating differences between the iterative nSolver- and *RUVSeq*-normalization procedures (described below and summarized in **Figure 1**). Details regarding these cohorts and the normalization parameters we use for each cohort are given below and summarized in **Supplemental Table 1**.

CBCS gene expression data

The Carolina Breast Cancer Study (CBCS) is a multi-phase cohort of women with incident breast cancer in North Carolina. Samples were collected during three study phases: Phase 1 (1993-1996), Phase 2 (1996-2001), and Phase 3 (2008-2013). Paraffin-embedded tumor blocks were requested from participating pathology laboratories for each sample, reviewed, and assayed for gene expression using the NanoString nCounter system as discussed previously (7, 19, 20). In total, 1,649 samples from patients with invasive breast cancer from the CBCS, across all three study phases, were analyzed on a custom panel of 417 genes. All assays were performed in the Translational Genomics Laboratory at the University of North Carolina at Chapel Hill. After quality control and normalization, 1,264 samples remained in the nSolver-normalized data, and 1,219 samples remained in the *RUVSeq*-normalized data.

Bladder tumor gene expression data

This study includes 130 samples from patients with non-muscle invasive bladder cancer (NMIBC) with FFPE tissue. More specifically, biospecimens included 42 samples of NMIBC from the University of North Carolina Hospital in Chapel Hill, NC and 88 samples from a study conducted by the Memorial Sloan Kettering Cancer Center in New York, NY. Tumor slides were reviewed for each case by pathologists at

both institutions. All RNA was isolated using the RNeasy FFPE Kit (Qiagen) at UNC and NanoString assays were performed across three separate months at the Translational Genomics laboratory at UNC using a custom exploratory codeset consisting of 440 endogenous and 6 housekeeping genes. After quality control and normalization, 86 samples remained in both the nSolver-normalized and *RUVSeq*-normalized datasets.

Kidney tumor gene expression data

This study includes 359 samples from patients with clear cell renal cell carcinoma (CCRCC) with fresh-frozen tissue collected as part of a large case-control study of kidney cancer conducted in central and eastern Europe (21) . As previously described (22), slides for each case were reviewed by a pathologist to assess tumor stage and grade. Manual microdissection was performed to remove non-tumor tissue. Frozen sections were placed directly in Trizol reagent (Invitrogen, Carlsbad, CA), homogenized for 2 minutes on ice, and RNA was isolated using the manufacturer's protocol. NanoString assays were performed at UNC at the Translational Genomics Laboratory at UNC using a custom codeset consisting of 62 endogenous and 6 housekeeping genes commonly studied in kidney cancer. After quality control and normalization, 331 samples remained in both the nSolver- and *RUVSeq*-normalized data.

Sabry et al gene expression data

For the differential expression analyses using data from Sabry et al (23), we downloaded raw RCC files from the NCBI Gene Expression Omnibus (GEO) with accession number GSE130286 and imported them using functions in *NanoStringQCPro* (24). This dataset was comprised of 12 samples, all of which remained after normalization with both procedures. The dataset measures 706 endogenous genes with 40 housekeeping genes from the NanoString nCounter Human Myeloid Innate Immunity Panel (23).

Quality control

The full quality control and normalization process using nSolver and *RUVSeq* is summarized graphically in **Figure 1**, starting with technical quality control (**Figure 1.2**)**,** pre-normalization assessments of housekeeping genes (**Figure 1.3**) and data visualization to detect problematic samples (**Figure 1.4**).

Normalization is performed with either nSolver or *RUVSeq* (**Figure 1.5**), and the processed expression data is assessed for validity through relevant visualization and biological checks (**Figure 1.6**). If validation is unsatisfactory and technical variation is still present, this process is iterated.

Technical quality controls flags

The first step in quality control was an assessment of the assay quality. nSolver provides several quality control (QC) flags to assess the quality of the data for imaging, binding density, linearity of the positive controls, and limit of detection. The definition and implementation of this QC is summarized in detail in the nSolver (13) and *NanoStringNorm* (16) documentation. For this work, we mark any sample that is flagged in at least one of these four QC assessments as *technical quality control*. We use these QC flags in both nSolver normalization and *RUVSeq* normalization.

Housekeeping gene assessment

Next, we consider the genes on the nCounter assay that can potentially serve as housekeepers in nSolver- or *RUVSeq*-normalization. Housekeeping genes serve two purposes: 1) for QC purposes to remove samples with overall poor quality and 2) for assessing the amount of technical variation present in the normalization procedure. A good housekeeping gene for nCounter expression is one that has little-to-no variability across all treatment conditions and is not expressed below the limit of detection in samples that pass QC. NanoString further suggests that ideal housekeeping genes are highly expressed, have similar coefficients of variation, and have expression values that correlate well with other housekeeping genes across all samples (10, 13). Because of these definitions, these targets will ideally vary only due to the level of technical variation present. To assess the potential for housekeeping correction to introduce bias, housekeeping genes were assessed for differential expression across a primary biological covariate of interest (estrogen receptor status in CBCS, tumor stage in the kidney and bladder cancer data, and treatment groups in Sabry et al) using negative binomial regression on the raw counts from the *MASS* package (25).

Below limit of detection quality control

Lastly, as samples with high proportions of both endogenous and housekeeping genes below the limit of detection (LOD) may be indicative of reduced assay or sample quality, we define another QC flag: the number of genes measured at below the LOD per sample. Here, we define the per-sample limit of detection as the mean of the counts of negative control probes for that given sample. We further assessed the percent of counts below the LOD in the housekeeping genes per sample as an added QC step to flag both poor quality samples and poor housekeeping genes. Samples were flagged if they have more than one housekeeping gene missing and a median percent below LOD in the endogenous genes greater than the 75th percentile of the samples with all housekeeping genes present.

nSolver normalization

Background correction

NanoString guidelines suggest background correction (13, 16) by either subtraction or thresholding for an estimated background noise level for experiments in which low expressing targets are common, or when the presence or absence of a transcript has an important research implication (11, 16). Data from all four cohorts considered do not necessarily fall under this criterion, and accordingly, we did not background correct by either method. To demonstrate the effect of background correction, we illustrated differences in nSolver-normalized gene expression with and without background thresholding in CBCS using relative log expression (RLE) plots.

Positive control and housekeeping gene-based normalization

For normalization, the arithmetic mean of the geometric means of the positive controls for each lane is computed and then divided by the geometric mean of each lane to generate a lane-specific positive control normalization factor (13, 16). The counts for every gene are then multiplied by its lane-specific normalization factors. To account for any noise introduced into the nCounter assay by positive normalization, the housekeeping genes are used similarly as the positive control genes to compute housekeeping normalization factors to scale the expression values(13, 16). NanoString also flags samples with large housekeeping gene scaling factors (we call this a *housekeeping QC flag*) and large

positive control spike-in scaling factors (*positive QC flag*) but note that samples with these flags do not necessarily require removal.

***RUVSeq* normalization pipeline**

Normalization

After quality control and housekeeping assessment, we started the *RUVSeq*-based normalization process (**Figure 1.5**)**,** an alternative approach to nSolver normalization. We scaled distributional differences between lanes with upper-quartile normalization (26). *RUV-III* has been created specifically for normalization of NanoString data with technical replicates (11); however, as the datasets we discuss here do not have technical replicates, we proceeded with *RUVg*. Unwanted technical factors were estimated in the resulting gene expression data with the *RUVg* function from the *RUVSeq* Bioconductor package (12). Unwanted variation was estimated using the distribution of the endogenous housekeeping genes not associated with the outcome of interest on the NanoString gene expression panel (27, 28). In general, we chose the number of dimensions of unwanted variation removed iteratively by normalizing the data for a given number of dimensions and checking for the removal of known technical factors already identified in the raw expression data (e.g. study phase), and presence of key biological variation (e.g. bimodality of *ESR1* expression in the CBCS breast cancer data where estrogen receptor status is a known predominant feature). Further details about choosing this dimension are given by Gagnon-Bartsch et al and Risso et al (10, 12). DESeq2 was used to compute a variance-stabilized transformation of the original count data (28), and we removed the estimated unwanted variation using the removeBatchEffects function from *limma* (29). Ultimately, we removed 1, 1, 3, and 1 dimensions of unwanted variation from CBCS, kidney cancer, bladder cancer, and the Sabry et al datasets, respectively. We used RLE plots, principal component analysis and heatmaps to detect any potential outliers.

Details regarding the total sample sizes of each study, sample size after quality control and normalization, and housekeeping genes selected for each study are summarized in **Supplemental Table 1**.

Downstream analyses

We considered several data visualization and benchmarking methods for each cohort to assess the impact of the two normalization procedures on technical and biological variation.

Silhouette width analysis in CBCS

We assessed differences in expression across study phases in CBCS to search for evidence of technical batch effects. We consider here the silhouette value s_i for sample i (30), a measure used to show how similar a sample is to its own study phase as compared to other study phases. The silhouette is defined as

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)},$$

where a_i is the mean distance between sample i and all other samples in the same study phase and b_i is the smallest mean distance of sample i to all samples in any other study phase. If many samples have large silhouettes, we can interpret that the different study phases are clearly distinct, in terms of the normalized expressions of its member samples, and that a batch effect is still present in the data.

eQTL analysis in CBCS

We assessed the additive relationship between the gene expression values and germline genotypes with linear regression analysis with *MatrixeQTL* (31), using the same linear model as detailed in previous work (32). Briefly, for each gene and SNP in our data, we constructed a simple linear regression, where the dependent variable is the scaled expression of the gene with zero mean and unit variance, the predictor of interest is the dosage of the SNP, and the adjusting covariates are the top five principal components of the genotype matrix. We considered both cis- (SNP is less than 0.5 Mb from the gene) and trans-eQTLs in our analysis. We adjusted for multiple testing via the Benjamini-Hochberg procedure (33).

PAM50 subtyping in CBCS

We classified each subject into PAM50 subtypes based on both nSolver-normalized and *RUVSeq*-normalized gene expression, using the procedure summarized by Parker et al (34, 35). Briefly, for each sample, we computed the Euclidean distance of the log-scale expression values for the 50 PAM50 genes to the PAM50 centroids for each of the molecular subtypes. Each sample was classified to the subtype

with the minimal distance (34). To assess expression patterns in the PAM50 genes, we created heatmaps with hierarchical clustering on both samples and genes. The concordance of subtype classification was assessed between nSolver and *RUVseq*-normalized expression sets, excluding normal-like cases.

RNA-seq normalization and distance correlation analysis in CBCS

We followed a standard RNA-seq normalization process with *DESeq2* (28), using the median of ratios method described by Anders et al (28). We then calculated the distance correlation and conducted a multivariate permutation test of independence between the RNA-seq data set (subset to the overlapping genes on the NanoString codeset) with each of the nSolver-normalized and *RUVSeq*-normalized nCounter data using the *energy* package (36). The distance correlation and associated permutation test allow for detection of non-independence across multivariate datasets of different distribution.

Differential expression analysis with Sabry et al. dataset

We performed differential expression analysis on data from Sabry et al to illustrate differences across the two normalization methods (23). We conducted differential expression analysis using a generalized linear model where counts are assumed to follow a Negative Binomial distribution with a gene-specific dispersion parameter estimated by *DESeq2* (28), and adjusting for multiple testing with the Benjamini-Hochberg (33) procedure. We compared differential expression across IL-2–primed NK cells vs. NK cells alone and CTV-1-primed NK cells for 6 hours vs. NK cells alone.

RESULTS

We evaluated the impact of nSolver- and *RUVSeq*-normalization methods on the removal of technical variation while retaining biological patterns and associations across four cohorts of differing sample size and varying sources of technical bias (**Supplemental Table 1**). Known sources of variation included age of sample and different study sites. In addition, these cohorts varied in preservation status. Two cohorts used fresh-frozen specimens, while two used archival FFPE specimens. The number of genes targeted for both endogenous genes and housekeeping genes also varied by study. In addition, some studies used validated and optimized code sets for specific gene signatures versus a more general code set. In

cohorts with large technical biases, *RUVSeq* provided superior normalization with more robust removal of technical variation and provided stronger biological associations. In two of the datasets, we found evidence that use of nSolver-normalized data often resulted in a large number of biological associations, suggesting possible test statistic inflation from potential residual technical variation that was confounded with biological variables of interest. However, when few strong technical biases were present or if a validated and optimized code set (e.g. PAM50 genes) was used, nSolver and *RUVSeq* performed comparably. The details for each of these four studies are described below.

Case study: Carolina Breast Cancer Study

Evaluation of background correction

Background thresholding led to increased per-sample variance while per-sample medians remained relatively similar (**Supplemental Figure 1A**). In particular, the distributions of per-sample median expression values were more right-skewed (greater mean than median) when using background thresholding prior to normalization compared to not using background thresholding (**Supplemental Figure 1B**). Based on this analysis, we did not perform background correction prior to normalization for all cohorts analyzed.

Quality assessment of expression levels using LOD of housekeeping genes

We used the housekeeping genes to assess if the lack of expression of endogenous genes was due to biology or due to technical failures. We compared the level of missing endogenous genes in samples with all housekeeping genes present to those with increasing number of housekeeping genes below LOD. There was a strong positive correlation for increasing proportions of genes below the LOD in both the endogenous and housekeeping genes (**Figure 2A** and **Supplemental Figure 2)**. Furthermore, the age of the FFPE sample was associated with the number below the LOD (**Figure 2A** and **Supplemental Figure 3**). Samples with higher numbers of genes below the LOD were from earlier phases of CBCS (i.e. Phase 1 from 1993-1996 and Phase 2 from 1996-2001). Samples that had a higher proportion of endogenous genes below the LOD tended to have increased numbers of QC flags as well (**Supplemental Figure 2**).

Evaluation of normalization methods

We observed differences in the two normalization strategies, namely greater unresolved technical variation in nSolver-normalized data than *RUVSeq*-normalized data. **(Figure 2)**. We compared the deviation from the median RLE per sample across raw, nSolver-normalized, and *RUVSeq*-normalized expression (**Figure 2B**). While both methods reduced variation compared to the raw data, there were considerable differences in the deviations from the median log-expressions in the nSolver-normalized expression that are not present in the *RUVSeq*-normalized expression (**Figure 2B**). The nSolver method retained technical variation, either not fully corrected or re-introduced during the nSolver normalization process.

We assessed the extent to which key sources of technical variation were reflected in the raw, nSolver-normalized and *RUVseq*-normalized datasets. We found that the median overall silhouette value was highest in the raw expression (median = 0.08), lower in the nSolver-normalized data (median = 0.05), and lowest in the *RUVSeq*-normalized data (median = 0.00). The higher, positive silhouette values indicated that some samples had expression profiles that were closest to expression profiles of other samples in the same study phase even after normalization. This showed that technical bias from study phase was dominant in the raw data, still present in the nSolver data, and minimal in the RUVSeq data (**Figure 2C**). The distributions of per-sample median log-expression in the raw and nSolver-normalized expression were multimodal with respect to CBCS study phase, reflective of the technical variation due to age of the sample (**Figure 2D**). With *RUVSeq*, the per-sample median expression was more consistent across all three study phases with less separation by study phase in the top two principal components (PC) compared to nSolver (**Supplemental Figures 5A**).

Lastly, we examined the ability to retain biological sources of variation known to be important in breast cancer after applying the two normalization procedures. Estrogen Receptor (ER) status is one of the most important clinical and biological features in breast cancer. In the raw expression, variation due to ER status was captured in PC2 rather than PC1, further highlighting the need to perform appropriate normalization (**Figure 2E**). In the nSolver-normalized data, ER status was shared between PC1 and

PC2, suggesting that unresolved technical variation was still present. ER status was reflected predominantly in PC1 of the *RUVSeq*-normalized expression, demonstrating effective removal of technical variation and boosting of the true biological signal. The PAM50 molecular subtypes (34) were also strongly separated by PC1 only in the *RUVSeq*-normalized expression (**Supplemental Figure 5B**).

Genomic analyses and expression profiles across normalization methods

We evaluated the impact of normalization choices on downstream eQTL analysis, PAM50 molecular subtyping, known expression patterns, and similarity to RNA-seq data. We conducted a full cis-trans eQTL analysis accounting for race and genetic-based ancestry using both normalization methods. Thresholding at nominal $P < 10^{-3}$, we found considerably more eQTLs using the nSolver-normalized expression as opposed to the *RUVSeq*-normalized expression (2,050 vs. 1,143). We identified a strong cis-eQTL signal using both normalized expression sets, as indicated by the densely populated cis-line in **Figure 3A**. We also observed considerably more trans-eQTLs at FDR-adjusted P-value less than 0.25 using the nSolver-normalized data (**Figure 3B**). Further, the proportion of trans eQTLs detected in the total was substantially higher using the nSolver-normalized data than the *RUVSeq*-normalized data, across various FDR-adjusted significance levels (**Figure 3B**) with histograms of raw *P*-values shown in **Supplemental Figure 6**. We further investigated cross-chromosomal trans-eQTLs with FDR-adjusted $P < 0.25$. The distribution of absolute differences in minor allele frequency (MAF) across women of African and European ancestry were wide across nSolver-specific and RUV-specific cross-chromosomal trans-eQTLs (**Supplemental Figure 7)**. However, we observed larger absolute MAF differences across phase in cross-chromosomal trans-eQTLs in nSolver-normalized data than in *RUVSeq*-normalized data.

We compared each normalization method for the ability to classify samples into PAM50 intrinsic molecular subtype using the classification scheme outlined by Parker et al (34). Our PAM50 subtyping calls were robust across normalization methods with approximately 91% agreement and a Kappa of 0.87 (95% CI (0.85, 0.90)) (**Supplemental Table 2**). Among discordant calls, approximately half had low confidence values, and half had differences in correlations to centroids less than 0.1 between the

discordant calls (data not shown). Most of these discordant calls were among HER2-enriched, luminal B and luminal A subtypes, which are molecularly similar.

We observed noticeable differences between the *RUVSeq*- and nSolver-normalized gene expression when visualized via heatmaps, emphasizing differences seen in the principal components across technical and biological variables. These heatmaps allowed us to identify samples without any QC flags that had significant deviations in expression patterns. In nSolver-normalized data, 14 samples were flagged in **Supplemental Figure 8,** emphasizing the need for post-normalization data visualization. ER status is the most distinguishing biological factor in breast tumors, which characteristically subdivides basal-like and HER2-enriched molecular subtypes (predominantly ER-negative) and luminal A and B molecular subtypes (predominantly ER-positive) into two major clusters (34). However, this characteristic biological clustering pattern was only partially preserved after nSolver normalization (**Supplemental Figure 9A**)**.** Hierarchical clustering of nSolver data revealed four major dendrogram branches driven by a combination of ER status and study phase. While the first major branch was enriched for ER-negative tumors and contained an even distribution of study phases, two of the remaining three branches were formed by Phase 1 and Phase 3 samples, respectively. Conversely, we did not observe a similar influence of sample age on sample clustering in the *RUVSeq*-normalized data and ER status drove the primary division of samples (**Supplemental Figure 9B**).

Lastly, we compared normalization choices for NanoString data to RNA-seq data performed on the same samples. CBCS collected RNA-seq measurements on 70 samples that have nCounter expression measured on a different codeset (162 genes instead of 417). These samples were not included in the cohort we present in **Figure 2**. We used distance correlation to measure whether datasets appeared independent from another in terms of sample-sample relationships within each dataset. On this RNA-seq dataset, we performed astandard RNA-seq normalization procedure as described in **Methods**. A permutation-based test of independence using the distance correlation (36, 37) revealed that the distance correlation between the RNA-seq and nSolver-normalized distance matrices was small and near 0 (distance correlation = 0.051, $P = 0.24$) while the distance correlation between the RNA-seq and

RUVSeq-normalized distance matrices was larger (distance correlation = 0.36, $P = 0.02$). The permutation-based test rejected the null hypothesis of independence (distance correlation of zero for unrelated datasets) between *RUVSeq*-normalized nCounter data and RNA-seq data but fails to reject the null hypothesis for nSolver-normalization nCounter and RNA-seq data. We conclude that *RUVSeq* produced normalized data with closer relation to the RNA-seq, in terms of distance correlation and test of independence, compared to nSolver.

Case Study: Sabry Cohort: differential expression analysis in natural killer cells

We looked at the impact of the two normalization methods in a small cohort ($N = 12$) on DE analysis across natural killer (NK) cells primed for tumor-specific cells and cytokines from Sabry et al (23). RLE plots before and after normalization showed minor differences between the two normalization methods (**Supplemental Figure 10**).

Using *DESeq2* (28), we identified genes differentially expressed in NK cells primed by CTV-1 or IL-2 cytokines compared to unprimed NK cells at FDR-adjusted $P < 0.05$. The two normalization methods led to a different number of differentially expressed genes with a limited overlap of significant genes by both methods (**Figure 4A).** The raw P-value histograms from differential expression analysis using nSolver-normalized expression exhibited a slope toward low P-values, whereas the distributions of P-values using the *RUVSeq*-normalized data were closer to uniform, with few significant tests of differential expression (**Figure 4B**). While the $\log_2$-fold changes were correlated between the two normalization procedures, there was a shift toward larger fold changes in the nSolver data. Genes found to be differentially expressed only in the nSolver-normalized data had effect sizes with large standard errors in the *RUVSeq* analysis, while using the *RUVSeq*-normalized data identified more low effect size genes with smaller standard errors (**Figure 4C**).

Case study: bladder cancer gene expression

The *RUVSeq*-normalization procedure reduced technical variation (e.g. variation associated with study site) while maintaining the biological variation observed by tumor grade. *RUVSeq*-normalized data

showed the most homogeneity in per-sample median deviation of log-expressions compared to raw and nSolver-normalized data (**Figure 5A**). We observed significant differences in the first principal component of nSolver-normalized expression between tumors from different study sites, which was not present in the *RUVseq*-normalized data (**Figure 5B**). In addition, there was a stronger biological association in the first principal component of expression with tumor grade using *RUVSeq*-normalization (**Figure 5C**).

Case study: kidney cancer gene expression

We only found subtle differences in the deviations from the median expression between the normalization procedures for the kidney cancer dataset (**Figure 6A**). This cohort did not have the same known technical variables observed in the other cohorts such as study site or sample age, and the RNA came from fresh-frozen material (**Supplemental Table 1**). We evaluated normalization methods on a source of technical variation, DV300, the proportion of RNA fragments detected at greater than 300 base pairs as a source of technical variation, and tumor stage as a biological variable of interest. A scatterplot of the first two principal components marked by level of DV300 (**Figure 6B**) and tumor stage (**Figure 6C**) shows little difference across the two normalization methods. When there were limited sources of technical variation and a robust, high quality dataset, we found both normalization methods performed equally well.

DISCUSSION

Proper normalization is imperative in performing correct statistical inference from complex gene expression data. Here, we outline a sequential framework for NanoString nCounter RNA expression data that provides both quality control checks, considerations for choosing housekeeping genes, and iterative normalization with biological validation using both NanoString's nSolver software (13, 16) and *RUVSeq* (10, 12). We show that *RUVSeq* provided a superior normalization on three out of four datasets by more efficiently removing sources of technical variation, while retaining robust biological associations.

Our major focus here is a large dataset of breast tumor samples from the Carolina Breast Cancer Study with a custom NanoString nCounter assay of 417 endogenous genes and 11 housekeeping genes. We observed that both normalization methods were sensitive to the quality and the set of housekeeping

genes. It has become increasingly appreciated over the years that several genes thought to behave exclusively in a “housekeeping” fashion in fact associate with biological variables under certain conditions (38) or across different tissue types (39). A careful validation of housekeeping gene stability on a case-by-case basis, considering both technical and biological sources of variation in each dataset is therefore imperative for an optimized normalization procedure.

We introduced a quality metric to assess sample quality: samples with high proportions of genes detected below the LOD in both endogenous genes and housekeepers were indicative of either low-quality samples or reduced assay efficiency. Sample age was correlated with higher proportions of genes below the LOD in both endogenous and housekeeping genes, which was likely due to RNA degradation over time. Missing counts in endogenous genes alone does not suggest technical variation, but rather, represent genes not expressed and therefore not detected under certain biological conditions or cell types. An example includes using an immuno-oncology gene panel in a tumor sample with little to no immune cell infiltration. Many of these samples had additional quality control flags including those derived from nSolver’s assessment of data quality. We excluded these samples for analysis in both the nSolver- and *RUVSeq*-based procedures.

nSolver-normalized data was prone to residual unwanted technical variation when there were known technical biases, such as in CBCS and the bladder example. We checked for known biological associations that are intrinsic to the sample, as in eQTL analysis, to judge the performance of the normalization process (40, 41). A full cis-trans eQTL analysis using nSolver- and *RUVSeq*-normalized expression data showed a strong cis-eQTL signal in data from both normalization methods. We find significantly more trans-eQTLs with the nSolver-normalized data, though at smaller P-value ranges (Figure 3). We suspect that these trans-eQTL associations may be driven by residual technical variation in gene expression and ancestry-based variation in genotypes. nSolver-normalized data retained some technical bias by study phase, which could serve as a confounder that is conflated with race in the eQTL analysis. Such spurious associations from population stratification have been described in previous studies of eQTL analysis (42, 43). Our observation of minor allele frequency differences across both race

and phase in nSolver-specific cross-chromosome trans-eQTLs supports this hypothesis. Follow-up eQTL studies could possibly disentangle the relationship between study phase and race in CBCS.

The choice of normalization procedure is less of a concern in cohorts with minimal sources of technical variation or in nCounter targeted gene panels that have been optimized for robust measurement across preservation methods. In the CBCS breast cancer cohort, we identified significant differences in gene expression between normalization methods across the entire gene set (417 total genes). However, PAM50 subtyping was robust across the two normalization procedures, showing little difference in subtyping calls. The genes in the PAM50 classifier were selected due to their consistently robust measurement in both FFPE and fresh frozen breast tissues (34). Robustly measured genes may be less affected by different normalization procedures. Furthermore, we see minimal differences in residual technical variation in the kidney cancer dataset and the Sabry et al dataset, both of which were measured on either robustly validated genes or nCounter panels. The kidney cancer example had newer, fresh frozen specimens that were profiled using a small and well-validated set of genes important in that cancer type. This dataset gives an opportunity to stress the importance of the general principles of normalization: as Gagnon-Bartsch et al and Molania et al recommend (10, 11), normalization should be a part of scientific process and should be approached iteratively with visual inspection and biological validation to tune the process. One normalization procedure is not necessarily applicable to all datasets and must be re-evaluated on each dataset.

We outline a systematic and iterative framework for the normalization of NanoString nCounter expression data. Even without background correction, a technique which has been shown to impair normalization of microarray expression data (44, 45), we believe that relying solely on positive control and housekeeping gene-based normalization may result in residual technical variation after normalization. Here, we show the merits of a comprehensive procedure that includes sample quality control checks, assessments of housekeeping genes, normalization with *RUVSeq* (17) and data analysis with popular count-based R/Bioconductor packages, as well as iterative data visualization and biological validation to assess

normalization. Researchers should pay close attention to the normalization process and systematically assess pipelines that best suit each dataset.

AVAILABILITY

Relevant R code for these analyses are freely available at https://github.com/bhattacharya-a-bt/CBCS_normalization. Summary statistics for eQTL analysis are available at https://github.com/bhattacharya-a-bt/CBCS_TWAS_Paper/, as a part of Bhattacharya et al (32).

CBCS genotype datasets analyzed in this study are not publicly available as many CBCS patients are still being followed and accordingly CBCS data is considered sensitive; the data is available from M.A.T upon reasonable request. Raw and normalized expression data from CBCS will be available on GEO upon publication. For replication prior to publication, this data can be requested from the authors. Data from the bladder and kidney cancer datasets will be provided by the authors upon reasonable request.

ACCESSION NUMBERS

Raw RCC files for nCounter expression from Sabry et al (23) are available NCBI Gene Expression Omnibus (GEO) with the accession numbers GSE130286. Raw and normalized expression data from CBCS will be available on GEO upon publication. For replication prior to publication, this data can be requested from the authors.

SUPPLEMENTARY DATA

Supplementary Data and Figures are available at NAR online.

ACKNOWLEDGEMENT

First and foremost, we thank the Carolina Breast Cancer Study participants and volunteers. We thank Halei Benefield, Xiaohua Gao, Erin Kirk, Linnea Olsson, and Jessica Tse for their invaluable support during the research process.

FUNDING

Susan G. Komen® provided financial support for CBCS study infrastructure. Funding was provided by the National Institutes of Health, National Cancer Institute P01-CA151135, P50-CA05822, and U01-CA179715 to M.A.T. M.I.L. is supported by P01-CA142538 and P30-ES010126. KAH is supported by a Komen Career Catalyst Grant (CCR16376756). AMH is supported by 1T32GM12274. The Translational Genomics Laboratory is supported in part by grants from the National Cancer Institute (3P30CA016086) and the University of North Carolina at Chapel Hill University Cancer Research Fund. The kidney cancer study and gene expression analysis were supported by the Intramural Research Program of the National Institutes of Health and the National Cancer Institute.

The funders had no role in the design of the study, the collection, analysis, or interpretation of the data, the writing of the manuscript, or the decision to submit the manuscript for publication.

CONFLICT OF INTEREST

The authors have no conflicts of interest to disclose.

REFERENCES

1. Geiss,G.K., Bumgarner,R.E., Birditt,B., Dahl,T., Dowidar,N., Dunaway,D.L., Fell,H.P., Ferree,S., George,R.D., Grogan,T., *et al.* (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat. Biotechnol.*, **26**, 317–325.
2. Veldman-Jones,M.H., Brant,R., Rooney,C., Geh,C., Emery,H., Harbron,C.G., Wappett,M., Sharpe,A., Dymond,M., Barrett,J.C., *et al.* (2015) Evaluating Robustness and Sensitivity of the NanoString Technologies nCounter Platform to Enable Multiplexed Gene Expression Analysis of Clinical Samples. *Cancer Res.*, **75**, 2587–2593.
3. Brasó-Maristany,F., Filosto,S., Catchpole,S., Marlow,R., Quist,J., Francesch-Domenech,E., Plumb,D.A., Zakka,L., Gazinska,P., Liccardi,G., *et al.* (2016) PIM1 kinase regulates cell death, tumor growth and chemotherapy response in triple-negative breast cancer. *Nat. Med.*, **22**, 1303–1313.
4. Urrutia,A., Duffy,D., Rouilly,V., Posseme,C., Djebali,R., Illanes,G., Libri,V., Albaud,B., Gentien,D., Piasecka,B., *et al.* (2016) Standardized Whole-Blood Transcriptional Profiling Enables the Deconvolution of Complex Induced Immune Responses. *Cell Rep.*, **16**, 2777–2791.
5. Scott,D.W., Wright,G.W., Williams,P.M., Lih,C.-J., Walsh,W., Jaffe,E.S., Rosenwald,A., Campo,E., Chan,W.C., Connors,J.M., *et al.* (2014) Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood*, **123**, 1214–1217.
6. Ng,S.W.K., Mitchell,A., Kennedy,J.A., Chen,W.C., McLeod,J., Ibrahimova,N., Arruda,A., Popescu,A., Gupta,V., Schimmer,A.D., *et al.* (2016) A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*, **540**, 433–437.
7. Troester,M.A., Sun,X., Allott,E.H., Geradts,J., Cohen,S.M., Tse,C.-K., Kirk,E.L., Thorne,L.B., Mathews,M., Li,Y., *et al.* (2018) Racial Differences in PAM50 Subtypes in the Carolina Breast Cancer Study. *JNCI J. Natl. Cancer Inst.*, **110**, 176–182.
8. Wallden,B., Storhoff,J., Nielsen,T., Dowidar,N., Schaper,C., Ferree,S., Liu,S., Leung,S., Geiss,G., Snider,J., *et al.* (2015) Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med. Genomics*, **8**, 54.

9. Vieira,A.F. and Schmitt,F. (2018) An Update on Breast Cancer Multigene Prognostic Tests-Emergent Clinical Biomarkers. *Front. Med.*, **5**, 248.

10. Gagnon-Bartsch,J.A. and Speed,T.P. (2012) Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**, 539–552.

11. Molania,R., Gagnon-Bartsch,J.A., Dobrovic,A. and Speed,T.P. (2019) A new normalization for Nanostring nCounter gene expression data. *Nucleic Acids Res.*, **47**, 6073–6083.

12. Risso,D., Ngai,J., Speed,T.P. and Dudoit,S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, **32**, 896–902.

13. nSolver™ 4.0 Analysis Software User Manual (2018).

14. Vandesompele,J., De Preter,K., Pattyn,F., Poppe,B., Van Roy,N., De Paepe,A. and Speleman,F. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.*, **3**, research0034.1.

15. Perkins,J.R., Dawes,J.M., McMahon,S.B., Bennett,D.L.H., Orengo,C. and Kohl,M. (2012) ReadqPCR and NormqPCR: R packages for the reading, quality checking and normalisation of RT-qPCR quantification cycle (Cq) data. *BMC Genomics*, **13**, 296.

16. Waggott,D., Chu,K., Yin,S., Wouters,B.G., Liu,F.-F. and Boutros,P.C. (2012) Gene expression NanoStringNorm: an extensible R package for the pre-processing of NanoString mRNA and miRNA data. *Bioinforma. Appl. NOTE*, **28**, 1546–1548.

17. Gagnon-Bartsch,J.A. and Speed,T.P. (2012) Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**, 539–552.

18. Li,S., Labaj,P.P., Zumbo,P., Sykacek,P., Shi,W., Shi,L., Phan,J., Wu,P.Y., Wang,M., Wang,C., *et al.* (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.*, **32**, 888–895.

19. D'Arcy,M., Fleming,J., Robinson,W.R., Kirk,E.L., Perou,C.M. and Troester,M.A. (2015) Race-associated biological differences among Luminal A breast tumors. *Breast Cancer Res. Treat.*, **152**, 437–448.

20. Hall,I.J., Moorman,P.G., Millikan,R.C. and Newman,B. (2005) Comparative Analysis of Breast Cancer Risk Factors among African-American Women and White Women. *Am. J. Epidemiol.*, **161**, 40–51.

21. Brennan,P., Van Der Hel,O., Moore,L.E., Zaridze,D., Matveev,V., Holcatova,I., Janout,V., Kollarova,H., Foretova,L., Szeszenia-Dabrowska,N., *et al.* (2008) Tobacco smoking, body mass index, hypertension, and kidney cancer risk in central and eastern Europe. *Br. J. Cancer*, **99**, 1912–1915.

22. Moore,L.E., Nickerson,M.L., Brennan,P., Toro,J.R., Jaeger,E., Rinsky,J., Han,S.S., Zaridze,D., Matveev,V., Janout,V., *et al.* (2011) Von Hippel-Lindau (VHL) inactivation in sporadic clear cell renal cancer: Associations with germline VHL polymorphisms and etiologic risk factors. *PLoS Genet.*, **7**.

23. Sabry,M., Zubiak,A., Hood,S.P., Simmonds,P., Arellano-Ballestero,H., Cournoyer,E., Mashar,M., Pockley,A.G. and Lowdell,M.W. (2019) Tumor- and cytokine-primed human natural killer cells exhibit distinct phenotypic and transcriptional signatures. *PLoS One*, **14**, e0218674.

24. Nickles,D., Sandmann,T., Ziman,R. and Bourgon,R. NacoStringQCPro.

25. Venables,W.N. and Ripley,B.D. (2002) Modern Applied Statistics with S 4th ed. Springer, New York.

26. Bullard,J.H., Purdom,E., Hansen,K.D. and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

27. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

28. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

29. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47–e47.

30. Rousseeuw,P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

31. Shabalin,A.A. (2012) Gene expression Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.

32. Bhattacharya,A., García-Closas,M., Olshan,A.F., Perou,C.M., Troester,M.A. and Love,M.I. (2019) A Framework for Transcriptome-Wide Association Studies in Breast Cancer in Diverse Study Populations. *bioRxiv*, 10.1101/769570.

33. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple.

34. Parker,J.S., Mullins,M., Cheang,M.C.U., Leung,S., Voduc,D., Vickery,T., Davies,S., Fauron,C., He,X., Hu,Z., *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, **27**, 1160–1167.

35. Gendoo,D.M.A., Ratanasirigulchai,N., Schröder,M., Pare,L., Parker,J.S., Prat,A. and Haibe-Kains,B. (2018) genefu: a package for breast cancer gene expression analysis.

36. Székely,G.J. and Rizzo,M.L. (2017) The Energy of Data. *Annu. Rev. Stat. Its Appl.*, **4**, 447–479.

37. Mantel,N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res.*, **27**, 209–220.

38. Sikand,K., Singh,J., Ebron,J.S. and Shukla,G.C. (2012) Housekeeping gene selection advisory: glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and β-actin are targets of miR-644a. *PLoS One*, **7**, e47510.

39. Barber,R.D., Harmer,D.W., Coleman,R.A. and Clark,B.J. (2005) GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol. Genomics*, **21**, 389–395.

40. Raulerson,C.K., Ko,A., Kidd,J.C., Currin,K.W., Brotman,S.M., Cannon,M.E., Wu,Y., Spracklen,C.N., Jackson,A.U., Stringham,H.M., *et al.* (2019) Adipose Tissue Gene Expression Associations Reveal Hundreds of Candidate Genes for Cardiometabolic Traits. 10.1016/j.ajhg.2019.09.001.

41. Aguet,F., Brown,A.A., Castel,S.E., Davis,J.R., He,Y., Jo,B., Mohammadi,P., Park,Y.S., Parsana,P., Segrè,A. V., *et al.* (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.

42. Lee,C. (2018) Genome-wide expression quantitative trait loci analysis using mixed models. *Front. Genet.*, **9**.

43. Jiang,N., Wang,M., Jia,T., Wang,L., Leach,L., Hackett,C., Marshall,D. and Luo,Z. (2011) A robust statistical method for association-based eQTL analysis. *PLoS One*, **6**.

44. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level

data. *Biostatistics*, **4**, 249–264.

45. Freytag,S., Gagnon-Bartsch,J., Speed,T.P. and Bahlo,M. (2015) Systematic noise degrades gene co-expression signals but can be corrected. *BMC Bioinformatics*, **16**, 309.

FIGURE CAPTIONS

Figure 1: **Graphical summary of both nSolver and RUVSeq normalization pipelines.** Steps in purple are shared across both pipelines, steps in orange are specific to the nSolver pipeline, and steps in blue are specific to the RUVSeq pipeline, with summaries for each step indicated. Though we do not recommend background correction, we include it as a consideration. See more details in **Methods**. The iterative process of any normalization method is reflected via the red arrows and box: if visualization across technical variables or known biological associations are unsatisfactory, the normalization process can be tuned and iterated. Note that here an RCC file is a raw nCounter expression file for a single sample and an RLE plot is a relative log expression plot.

Figure 2: **Quality control and normalization validation in CBCS. (A)** Boxplot of percent of endogenous genes below the limit of detection (LOD) (Y-axis) over varying numbers of the 11 housekeeping genes below LOD (X-axis), colored by CBCS study phase. Note that the X-axis scale is decreasing. **(B)** Relative log expression plots colored by estrogen receptor subtype. Samples are arranged in relative temporal order on the X-axis and deviation from the median $\log_2$-expression is plotted on the Y-axis. **(C)** Boxplots of per-sample silhouette values by phase across raw (left), nSolver-normalized (center), and RUVSeq-normalized (right). **(D)** Kernel density plots of median per-sample $\log_2$-expression from the raw (top), nSolver-normalized (center), and RUVSeq-normalized (bottom) expression matrices, colored by CBCS study phase. **(E)** Plots of the first principal component (X-axis) vs. second principal component (Y-axis) colored by estrogen receptor subtype of the raw (top), nSolver-normalized (center), and RUVSeq-normalized (bottom) expression matrices.

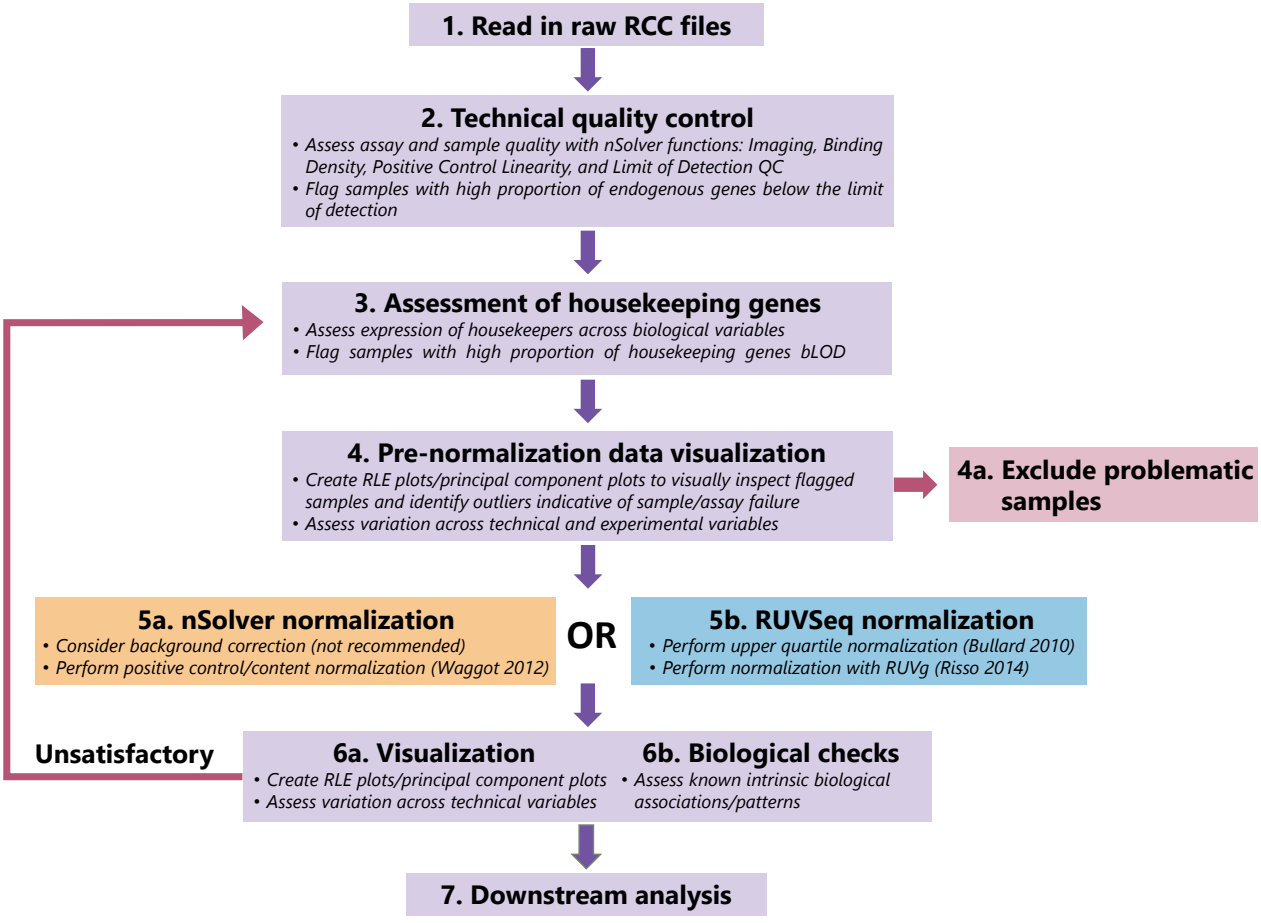
Figure 3: **eQTL analysis in CBCS. (A)** Cis-trans plots of eQTL results from nSolver-normalized (left) and RUVSeq-normalized data with chromosomal position of eSNP on the X-axis and the transcription start

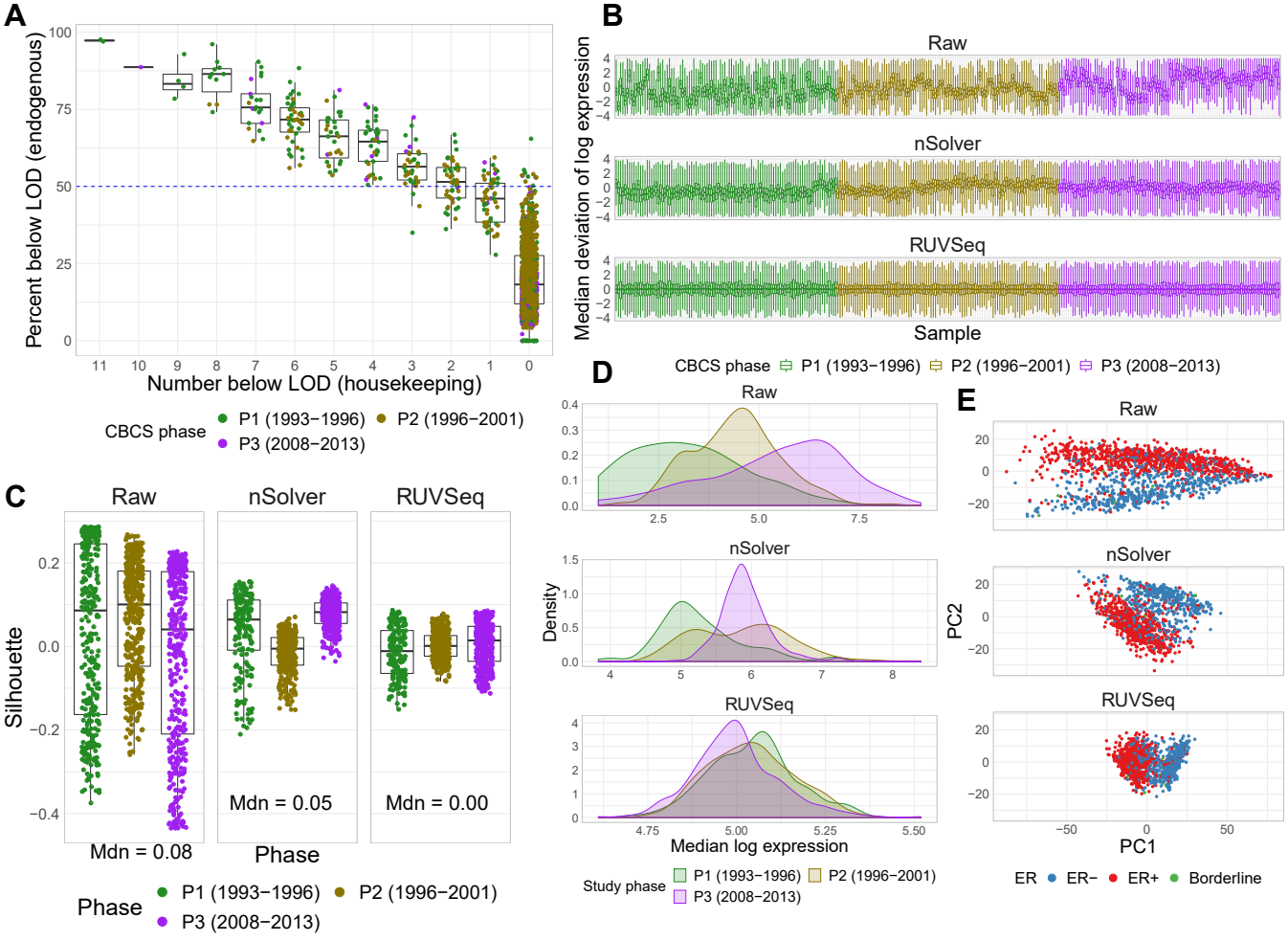
site of associated gene in the eQTL (eGene) on the Y-axis. Points for eQTLs are colored by FDR-adjusted P-value of the association. The dotted line provides a 45-degree reference line for cis-eQTLs. **(B)** Number of cis- (left) and trans-eQTLs (right) across various FDR-adjusted significance levels. The number of eQTLs identified in nSolver-normalized data is shown in red and the number of eQTLs identified in RUVSeq-normalized data is shown in blue.

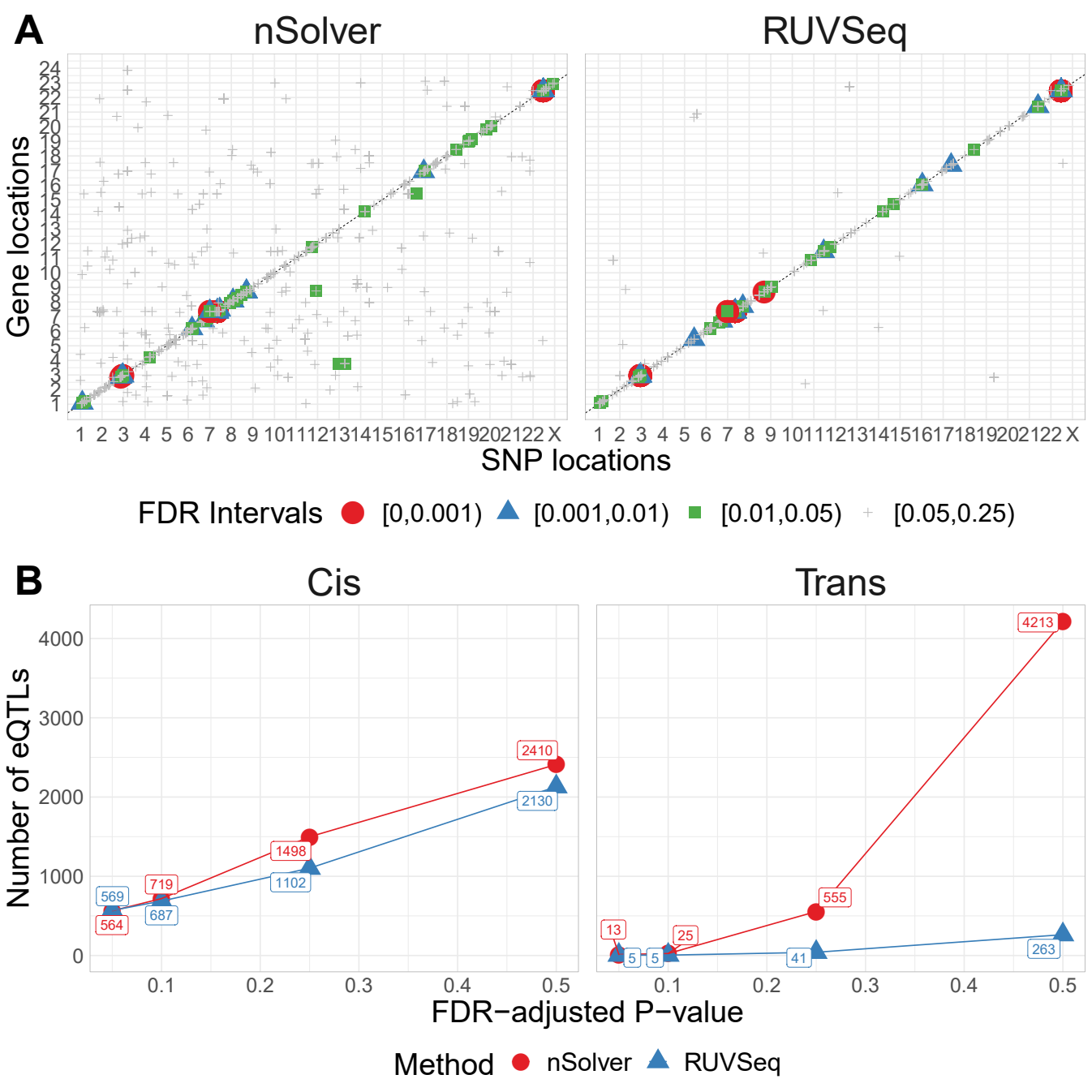
Figure 4: **Differential expression analysis from Sabry et al** (23)**. (A)** Venn diagram of the number of differentially expressed genes using nSolver-normalized (blue) and RUVSeq-normalized data (red) across comparisons for IL-2-primed (top) and CTV-1-primed NK cells (bottom). **(B)** Raw P-value histograms for differential expression analysis using nSolver-normalized (blue) and RUVSeq-normalized (red) data across the two comparisons. **(C)** Scatterplots of $\log_2$-fold changes from differential expression analysis using RUVSeq-normalized data (X-axis) and nSolver-normalized data (Y-axis) for any gene identified as differentially expressed in either one of the two datasets. Points are colored by the datasets in which that given gene was classified as differentially expressed. The size of point reflects the standard error of the effect size as estimated in the RUVSeq-normalized data. $X = 0, Y = 0$, and the 45-degree lines are provided for reference.

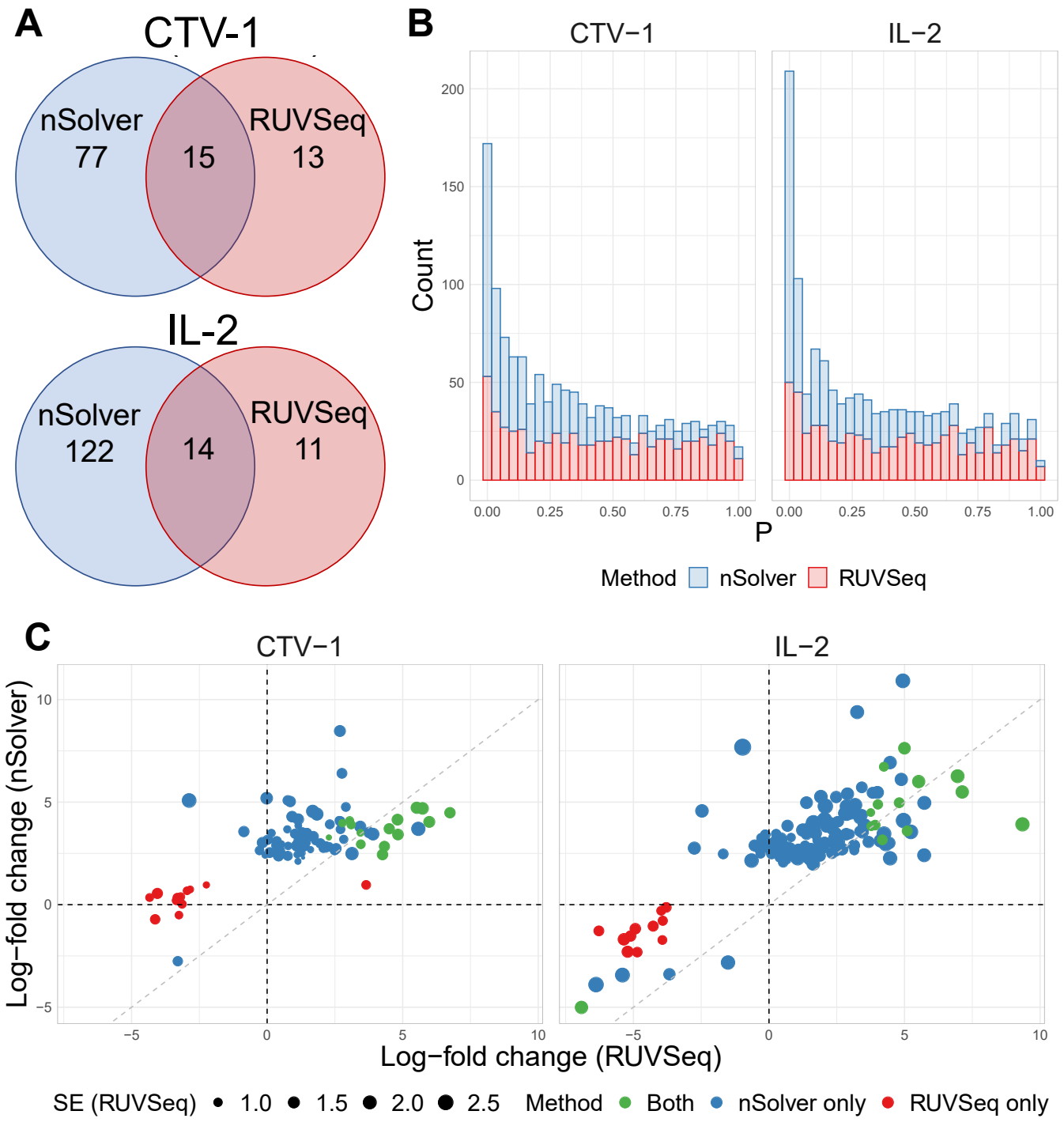
Figure 5: **Normalization differences in bladder cancer dataset. (A)** RLE plot from bladder cancer dataset, colored by assay month. **(B)** Boxplot of first principal component of expression by tumor collection site (location) across nSolver- (top) and RUVSeq-normalized (bottom) data. **(C)** Boxplot of first principal component of expression by tumor grade across nSolver- (top) and RUVSeq-normalized (bottom) data.

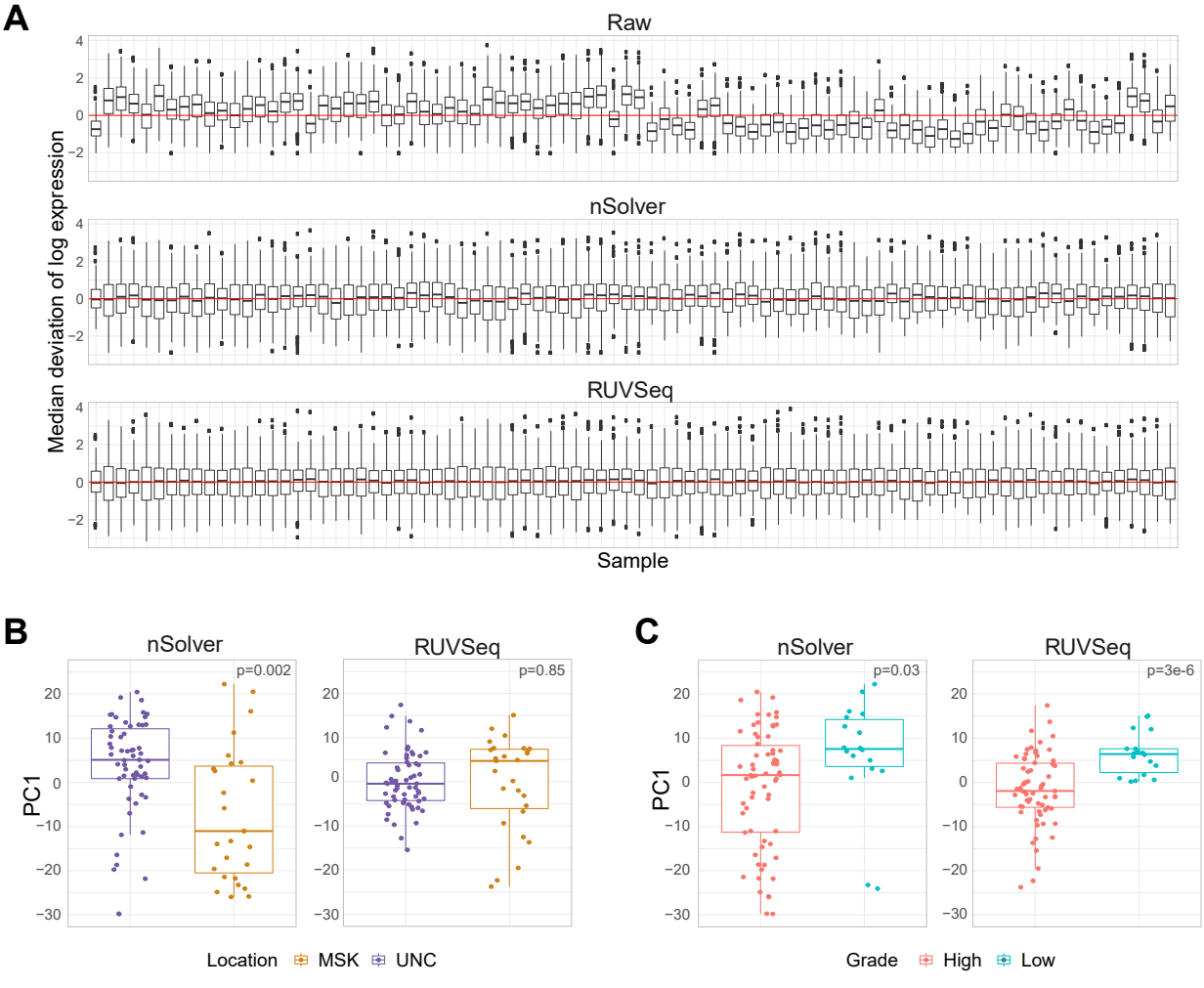
Figure 6: **Equal performance of normalization procedures in kidney cancer dataset. (A)** RLE plot of per-sample deviations from the median for raw, nSolver-, and RUVSeq-normalized data. **(B)** Scatter plot of the first and second principal component of nSolver- (left) and RUVSeq-normalized (right) expression, colored by high and low DV300. **(C)** Scatter plot of the first and second principal component of nSolver- (left) and RUVSeq-normalized (right) expression, colored by tumor stage.

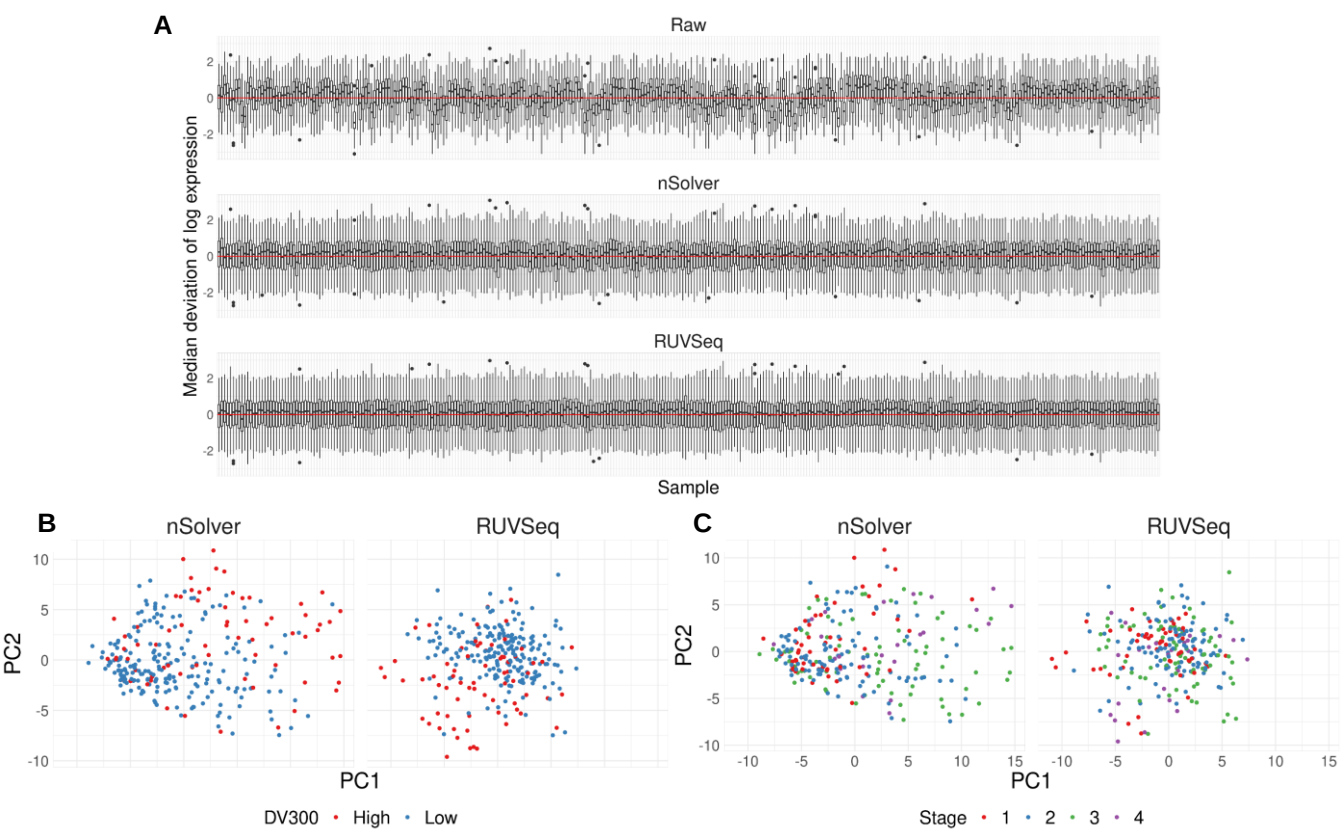

1. Read in raw RCC files
2. Technical quality control
• Assess assay and sample quality with nSolver functions: Imaging, Binding Density, Positive Control Linearity, and Limit of Detection QC
• Flag samples with high proportion of endogenous genes below the limit of detection
3. Assessment of housekeeping genes
• Assess expression of housekeepers across biological variables
• Flag samples with high proportion of housekeeping genes bLOD
4. Pre-normalization data visualization
• Create RLE plots/principal component plots to visually inspect flagged samples and identify outliers indicative of sample/assay failure
• Assess variation across technical and experimental variables
4a. Exclude problematic samples
5a. nSolver normalization
• Consider background correction (not recommended)
• Perform positive control/content normalization (Waggot 2012)
OR
5b. RUVSeq normalization
• Perform upper quartile normalization (Bullard 2010)
• Perform normalization with RUVg (Risso 2014)
Unsatisfactory
6a. Visualization
• Create RLE plots/principal component plots
• Assess variation across technical variables
6b. Biological checks
• Assess known intrinsic biological associations/patterns
7. Downstream analysis

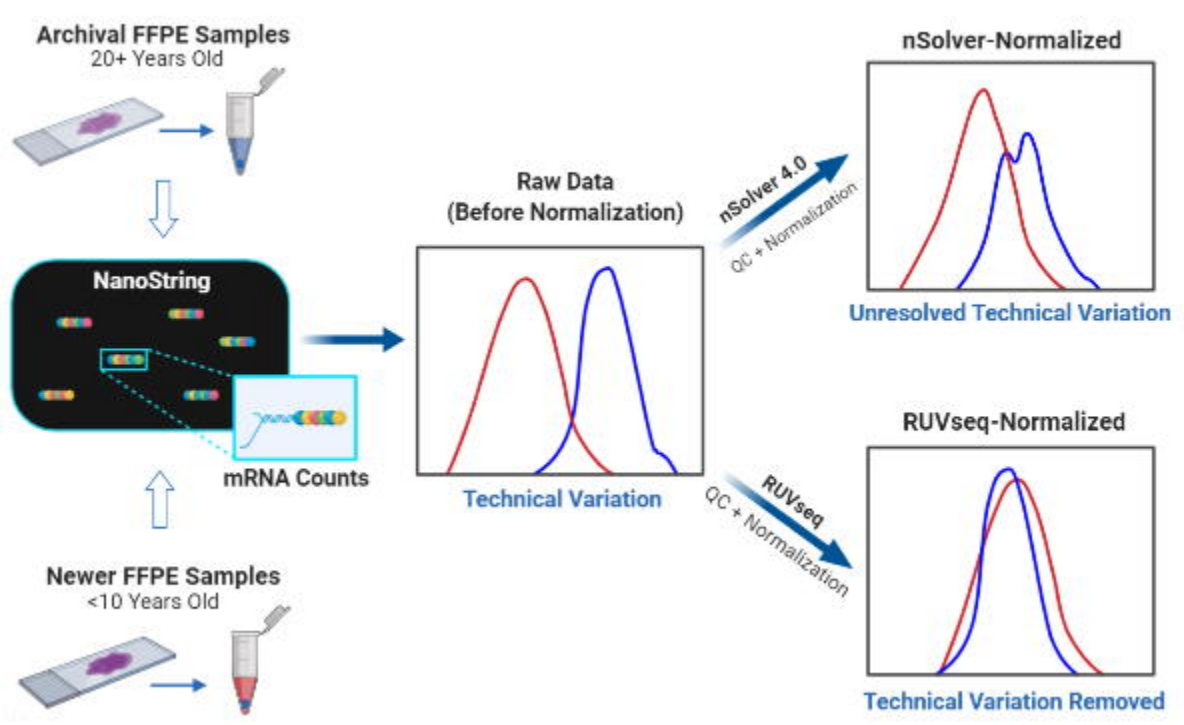

A
Percent below LOD (endogenous)
Number below LOD (housekeeping)
CBCS phase
P1 (1993−1996)
P2 (1996−2001)
P3 (2008−2013)
B
Raw
nSolver
RUVSeq
Median deviation of log expression
Sample
CBCS phase
P1 (1993−1996)
P2 (1996−2001)
P3 (2008−2013)
C
Raw
nSolver
RUVSeq
Silhouette
Mdn = 0.08
Mdn = 0.05
Mdn = 0.00
Phase
Phase
P1 (1993−1996)
P2 (1996−2001)
P3 (2008−2013)
D
Raw
nSolver
RUVSeq
Density
Median log expression
Study phase
P1 (1993−1996)
P2 (1996−2001)
P3 (2008−2013)
E
Raw
nSolver
RUVSeq
PC2
PC1
ER
ER−
ER+
Borderline


A
nSolver
RUVSeq
Gene locations
SNP locations
FDR Intervals
[0,0.001)
[0.001,0.01)
[0.01,0.05)
[0.05,0.25)
B
Cis
Trans
Number of eQTLs
FDR−adjusted P−value
2410
2130
1498
1102
719
687
569
564
4213
555
13
25
5
5
41
263
Method
nSolver
RUVSeq


A
CTV-1
nSolver
77
15
RUVSeq
13
IL-2
nSolver
122
14
RUVSeq
11
B
CTV−1
IL−2
Count
P
Method
nSolver
RUVSeq
C
CTV−1
IL−2
Log−fold change (nSolver)
Log−fold change (RUVSeq)
SE (RUVSeq)
1.0
1.5
2.0
2.5
Method
Both
nSolver only
RUVSeq only


A
Raw
nSolver
RUVSeq
Median deviation of log expression
Sample
B
nSolver
RUVSeq
p=0.002
p=0.85
PC1
Location MSK UNC
C
nSolver
RUVSeq
p=0.03
p=3e-6
PC1
Grade High Low


A
Raw
nSolver
RUVSeq
Median deviation of log expression
Sample
B
nSolver
RUVSeq
PC2
PC1
DV300 High Low
C
nSolver
RUVSeq
PC2
PC1
Stage 1 2 3 4


Archival FFPE Samples
20+ Years Old
NanoString
mRNA Counts
Newer FFPE Samples
<10 Years Old
Raw Data
(Before Normalization)
Technical Variation
nSolver 4.0
QC + Normalization
nSolver-Normalized
Unresolved Technical Variation
RUVseq
QC + Normalization
RUVseq-Normalized
Technical Variation Removed