METHOD

Background modeling, Quality Control and Normalization for GeoMx RNA data with *GeoDiff*

Lei Yang, Zhi Yang, Patrick Danaher, Stephanie Zimmerman, Tyler Hether, David Henderson* and Joseph Beechem

*Correspondence:
dhenderson@nanostring.com
NanoString Technologies, Seattle, US
Full list of author information is available at the end of the article

Abstract

Background: NanoString's GeoMx Digital Spatial Profiler (DSP) RNA assay can measure mRNA from hundreds of regions of customizable shape and size, yet it gives unique challenge in Quality Control(QC) and normalizating due to the omnipresent background noise incurred by the non-specific probe binding, which could not be addressed by conventional methods.

Results and discussion: Using Poisson Background model, Background Score Test, Negative Binomial threshold model and Poisson threshold model for normalization from the R package **GeoDiff**, we perform tasks including size factor estimation, QC and normalization on GoeMx RNA assay data. They are shown to outperform conventional methods like Limit of Quantification for QC as to consistency/false positive rate and 75% quantile normalization as to eliminating technical variability and recovering true signal.

Conclusions: We present a statistical model based workflow for QC and normalizing GeoMx RNA data using **GeoDiff**, justified by statistical theory and validated by real/simulated data.

Keywords: GeoMx RNA assay; **GeoDiff package**; threshold model

Background

Spatial gene expression platforms, which measure gene expression within tiny regions from across the span of a tissue sample, have opened a frontier in biology. These platforms harness idiosyncratic chemistries that introduce distinct technical effects into their data; however, spatial data is often analyzed with tools developed for bulk or single RNA-seq data, leading to bias and statistical inefficiency. Here we derive data analysis methods addressing the important technical effects of the GeoMx platform for spatial gene expression.

Given a tissue sample, the GeoMx RNA assay can measure mRNA from hundreds of regions of customizable shape and size. Its ability to assay flexibly-defined regions comes from a unique chemistry. Probes are constructed of two parts: an oligonucleotide that hybridizes to a target mRNA sequence in the tissue, and a second oligonucleotide barcode specifying the gene identity. These two halves are connected by a photocleavable linker. Probes are flowed across a tissue, binding their targets wherever they lie; then, precisely focused UV light cleaves the probes within user-defined regions and releases the barcodes for counting by short-read sequencing.

This chemistry leads to three complications for data analysis, each creating pitfalls for analyses designed for other platforms. First, probes stick at low rates to

biological material other than their intended mRNA target, leading to background counts. Background impacts the expression of genes, especially genes with very low expression, introducing bias if ignored. Background also complicates the task of calling a gene as definitively detected. Second, in very small regions, many genes will have expression levels dropping to zero or background-level counts. These near-zero counts become statistically unstable after log-transformation, which has been commonly favored by gene expression analysts since the early days of microarrays. Third, the size of regions sampled in a single experiment can vary from one to thousands of cells, leading to a much wider range of technical effects in the raw data.

Normalization of GeoMx data has thus far required great care, often forcing a choice between multiple unsatisfactory methods. To facilitate future analyses, we have established data-generating models for GeoMx, and we have models to derive methods for the fundamental operations of GeoMx analysis. Most importantly, we define a normalization procedure that removes all major technical effects and allows for log-scale analyses. We also define a suitable model for the background and a test for whether genes are detected above background, as well as a model for size factor estimation. This suite of tools is implemented in the **GeoDiff** R package.

The background model is the starting point of the **GeoDiff** analysis. While there is some previous attempt to model negative control probes using Poisson based model for data from NanoString nCounter [1] and DSP [2] platforms, none of them allow parameter estimation for each negative probe and each sample as in **GeoDiff**. With full parameterization, model diagnostics can be carried out and outliers and batch effect can be identified, which provides more insight to the data than previously possible.

The Quality Control(QC) is a prerequisite before gene expression analysis. In QC for DSP, one commonly asked question is: "whether a feature/Region of Interest(ROI) has good signal". One traditional way to answer this question is to define a per ROI Limit of Quantitation(LoQ) cutoff, to determine whether a single target is above the background per ROI by the cutoff, then to calculate the overall pass rate for each target or ROI as the metric for flagging them[3]. This approach is very heuristic. In this paper we formally clarify the meaning of "a target/ROI has good signal" in the context of GeoMx RNA dataset, and provide a statistical test for features and a metric for ROIs for signal QC purpose, with necessary justification and validation.

Normalization, the endeavor of removing technical variation from the gene expression data by finite step operations, is usually performed before other gene expression analysis[4]. This tradition dates back to early days of microarray gene expression analysis[5]. Among those, scaling method[6, 7, 8], i.e., dividing gene expression by a sample specific factor is commonly used on RNA-seq data for its simplicity. The output of normalization procedure is called "normalized expression". Currently, the most used scaling normalization method for DSP RNA data is 75% quantile normalization(or Q3 normalization as 75% quantile is also called 3rd quartile)[3, 9]. Another popular scaling normalization method is from [7], which tried to find the best scaling factor using is trimmed mean of M values (TMM) compared to simply using quantiles. Log transformation with base 2 or 10 is often further applied on the

normalized data to stabalize heavy tail distribution[10] of gene expression and facilitate interpretation using ratios. To avoid $-\infty$ in log transformation, 0 is usually replaced by a small positive number. Although some non-scaling method exists, the interpretation is usually more tricky. For example, [11] takes the Pearson residuals from "regularized negative binomial regression" as the normalized expression.

This scaling normalization and log transformation approach has caused two major problems in GeoMx DSP data analysis. First, the relation between observed counts and real expression often is not(purely) scaling. Instead, genes are subject to influence of the background as well as the scaling(size) factor. The background has more of a "shift" effect on the data, which can not be corrected by scaling alone. Second, there is no base to the choice of that small positive number to replace 0, and it also causes artificial bimodality in the results. Here, we address both problems by reframing the normalization problem as a parameter estimation problem using appropriate saturated model[12] with regularization on the normalization parameters. Empirical Bayes(EB) procedure[6] could be implemented to select a data driven regularization term. We show that this method corrects for the background as well as fixes the artificial spikes. We also show the empirical bayes shrinkage increase precision of normalized expression by leverage information across genes.

Results and discussion

Overview

Our workflow begins by modeling the negative probes and performing corresponding model diagnostics. After that, we use the background model as a reference to perform statistical test for features and estimate signal size factor. Finally, we use information gathered from these steps to perform normalization.

We explain our approach using a dataset of GeoMx Human Whole Transcriptome Atlas (WTA) RNA assay on diabetic kidney disease (DKD) vs. healthy kidney tissue, a dataset of GeoMx Human Whole Transcriptome Atlas RNA assay on cell pellet array containing 11 human cell lines, and a dataset of GeoMx Cancer Transcriptome Atlas (CTA) RNA assay on an FFPE cell pellet array of mixed HEK293T and CCRF-CEM cell lines.

The WTA diabetic kidney dataset is used for NanoString Spatial Omics Hackathon, This dataset consists of three normal tissue samples and four samples with diabetic kidney disease. The details of this dataset are in [13].

For the WTA cell pallet array data, we profiled a formalin fixed paraffin embedded (FFPE) cell pellet array containing 11 human cell lines with human WTA. We placed a range of sizes of circular ROIs from 50-360 μm diameter on each cell line, with 2-4 replicate ROIs per cell line and size[14].

In the CTA cell mixture data, HEK293T and CCRF-CEM (Acepix Biosciences, Inc.) were mixed in varying proportions, and aliquoted into a FFPE cell pellet array. Expression of 1414 genes in 700 μm diameter circular regions from the cell pellets were measured with the GeoMx platform, and each gene has 3 probes[15].

For all datasets, ROIs were illuminated on the GeoMx Digital Spatial Profiler and tags were collected and sequenced as previously described. FASTQ files were processed using the Nanostring GeoMx NGS Pipeline v2.0 as previously described. Reads were deduplicated by UMI and deduplicated counts for each ROI and probe

were used for analysis[16, 14]. The methods described in this paper all require raw count data without any preprocessing.

Poisson Background model

The Poisson background model takes a form of

$$X_{ij} \sim \text{Poisson}\left(\gamma_i \alpha_{0j}\right), \tag{1}$$

$$\sum_{j=1}^{J} \alpha_{0j} = 1, \tag{2}$$

where $\gamma_i, i = 1, \cdots, I$ are feature specific factors, and $\boldsymbol{\alpha}_0 = (\alpha_{0j}), j = 1, \cdots, J$ are sample size factors for each ROI. This model assumes, aside from random noise, variation in levels of ith feature in different ROIs is explained by the technical variation in the form of multiplicative size factor $\boldsymbol{\alpha}_0$. Therefore, this model is most suitable to features without biological variation and are subject to non-specific probe binding, i.e. negative control probes designed against synthetic sequences from the External RNA Controls Consortium(ERCC), which mimic the properties of mammalian sequences but have no homology to any known transcript. The constraint $\sum_{j=1}^{J} \alpha_{0j} = 1$ is imposed for identifiability. See Methods for further details.

Model diagnostics for Poisson Background model is available in **GeoDiff**. An empirical dispersion is calculated and a probability–probability(P-P) plot is generated in the diagnostics. The expected dispersion for Poisson distribution is 1. It is called overdispersion when the empirical dispersion is larger than 1, which is the most direct consequence of deviation from model assumption, likely caused by outliers or batch effect. The recommended remedy for outliers in negative probes is setting them to be missing and refiting the model. For batch effect we recommend people to find the root cause, as it impacts all aspects of data analysis. Fitting a Poisson Background model with a grouping variable is a workaround when batch effect is present(see Methods for details).

We fit the Poisson Background model without and with grouping variable indicating slide ID to the negative probes of WTA kidney dataset. The estimated dispersion are close to 1, and probability–probability(P-P) plot(Figure 1A, 1C) align with the diagonal line, showing both models are good fit for the data. This is further confirmed by the consistency of the feature factors across slides(Figure 1B, 1D).

Background Score Test

After applying the Poisson Background Model to the negative probes, it is imperative to test whether a target is above the background using the fitted model as a reference. It is called Background Score test for that it is derived using the score test paradigm(also known as the Lagrange multiplier test in econometrics[17]).

There are two versions of Background Score Test derived without and with a (Gamma distribution) prior assumption for feature factors.

$$\gamma_k \sim \text{Gamma}(.,.).$$

Background Score Test with prior takes into account the variation between feature factors thus is recommended and set as default. The "with prior" version is also implemented for all Background Score Test in this manuscript. For details of Background Score Test without and with prior, refer to Methods for details.

We compare the Background Score Test to the current default feature QC method LoQ, defined as the geometric mean times exponential of two standard deviations of the logarithm negative probes[3], in both the case study and simulation study.

0.0.1 Case study

In this analysis, we apply both Background Score test and LoQ method to the WTA cell pallet array data described above by five different ROI sizes, including 50, 80, 110, 250, 360 μm. The Background Score test with prior is applied to raw count data to extract the p values. In contrast, we flag the gene above the background if it exceeds LOQ in at least one ROI. The heatmap displays how -log(p value) from background score test and indicators on whether above background (1: Yes; 0:No) from LOQ change over increasing ROI sizes (Figure 4).

It is notable that Background Score test yield p values consistently increase with ROI sizes, meaning that it is more likely to be detected above background in larger ROI size. In contrast for LoQ, some genes are present in smaller ROI sizes but not larger ROI sizes, shows the LoQ method is less stable and more prone to be influenced by random noise.

Transcripts per million (TPM) from RNA-seq for different cell lines was used as a reference. We have shown that score statistics have better correlation with TPM values, measured in Spearman correlation, compared to the proportion of ROIs with gene expression above the LOQ (Figure 5).

0.0.2 Simulation study

The simulation study evaluates the false positive rates (FPRs) based on the 136 negative probes from the WTA cell pallet array data. For each ROI size, we estimate the size factors and feature factors and randomly sample 10,000 negative probes for each ROI. Then, the raw counts of genes are drawn from a Poisson distribution with mean of the size factor multiplied by the feature factor. Similar to the case study, we applied both Background Score test and LOQ method on the simulated population of negative probes (n = 10,000). False positive gene is defined as a gene where the null hypothesis is falsely rejected and the proportion of the total false rejected genes as the FPR. By setting the nominal level to be 0.001, we observe that LOQ method yields much higher FPRs in smaller sized ROIs, i.e., 50 μm, 80 μm, and 110 μm. It is consistent with the finding where LOQ only detects a subset of genes in smaller-sized ROIs (Figure 6), which could be false positives.

Negative Binomial threshold model: The signal size factor

In GeoMx RNA data, the size factor of targets are usually not the same as the size factor of the background, as is demonstrated in Figure 2, the correlation of features with background size factor decreases as the abundance increases. To estimate the signal size factor with background taken into consideration, we fit the following Negative Binomial threshold model to a set of "high genes/features" or "above the

background genes/features", defined as features above the background determined by the Background Score Test, and assuming the real expression level of these gene is constant across all ROIs.

$$Y_{kj} \sim \text{NB}(\mu = \max(\gamma_k - \gamma_t, 0)\alpha_j + \max(\gamma_i, \gamma_t)\alpha_{0j}, r = r_k),$$
$$\sum_{j=1}^{J} \alpha_j = 1$$

We use Negative Binomial distribution due to the overdispersion brought by the biology heteogenity as well as the "oversimplication" of the constant assumption, which is necessary for deriving size factors and is widely used for scaling normalization method[6, 7, 8]. For example, the commonly used 75% quantile normalization is implicitly assuming genes are roughly constant after dividing by the 75%. In [6], a median-of-ratios method is used to derive the size factor, which implicitly assumes the ratios are roughly constant for all genes. Here we explictly make this assumption using a representative set of features above the background.

The model is fitted by taking the estimated background size factor $\hat{\boldsymbol{\alpha}}_0$ as input. The threshold γ_t can be either set to be the mean of background feature factor $\gamma_0 = \frac{1}{I}\sum_{i=1}^{I}\hat{\gamma}_i$ or unspecified and estimated from the model. Since $\gamma_t \approx \gamma_0$, when γ_t is estimated, it should be within 20% of γ_0, otherwise the model might not be a good fit for the particular dataset. Other parameters $\gamma_k, r_k, \boldsymbol{\alpha}$ are also estimated from the model.

An important implication of this model is that the size factor of each feature is defined as a linear combination of background size factor and signal size factor depending on its abundance as

$$\max(\gamma_k - \gamma_t, 0)\boldsymbol{\alpha} + \max(\gamma_i, \gamma_t)\boldsymbol{\alpha}_0$$

For low abundance features, it basically behaves the same as the background size factor; as the abundance gets higher, the linear combination is weighted heavier towards the signal size factor.

By the constraint $\sum_{j=1}^{J}\alpha_{0j} = 1$ and $\sum_{j=1}^{J}\alpha_j = 1$, the γ_k is approximately the total count of kth feature, since if we use Poisson distribution instead of Negative Binomial distribution in the model, the total count of kth feature is the maximum likelihood estimator of γ_k. With this knowledge, the linear combination is easy to calculate for each feature without applying the model on all of them.

Figure 2 shows this linear combination has the best correlation with features from different abundance ranges among all size actors to be compared, including background/signal size factor alone and 75% quantile. The performance of 75% quantile is comparable for features with moderate abundance, but is considerably worse for low/high adundance genes. The background sizefactor has the same correlation with low abundance features as the linear combination, but the correlation deteriorate as the abundance increases; while the signal sizefactor has the overall worst correlation but the correlation increases as the abundance increases.

0.1 ROI signal metric

The signal size factor $\boldsymbol{\alpha}$ is an estimator for ROI signal levels, which can be naturally used to flag ROI with low signal. However this should not done directly on $\boldsymbol{\alpha}$ since they are scaled to sum up to 1, so on average the more samples the lower the value of each element of $\boldsymbol{\alpha}$. We could define signal level of ROI by

$$\boldsymbol{\alpha}(\gamma_k - \gamma_t),$$

for γ_k representing a certain quantile of genes. One can just use 75% quantile. To encourage keeping ROIs, one can also use a higher quantile, such as 90% quantile. Then this metric represents the expected above-the-background counts of certain quantile of features.

However, $\boldsymbol{\alpha}$ is defined and estimated via a model, which is not intuitive and straightforward for some users. We propose using the following metric named "quantile range" as a proxy. Denote $\tilde{\boldsymbol{\alpha}}_0 = \boldsymbol{\alpha}_0\gamma_0$ (see Methods for rationale), the 75% quantile range is defined as

$$\text{Q75r} = \text{Q75} - \tilde{\boldsymbol{\alpha}}_0.$$

Q75r is just the difference between 75% quantile and the average background level. Quantile ranges are shown to be better correlated with $\tilde{\boldsymbol{\alpha}}$ than quantiles (Figure 3) so that it can be used as an substitute for ROI signal metric for QC purpose without fitting the model.

Poisson threshold Normalization

From a statistical perspective, normalization is to fit a saturated model in which there is one parameter($\log_2$ normalized expression) for each observation. This is called saturated since the number of parameters is identical or larger than the number of obervations[12].

Assume $\boldsymbol{\beta}_k = (\beta_{k1}, \cdots, \beta_{kJ})$ be the vector of $\log_2$ real expression of kth feature for J samples. The goal of normalization is to optimally estimate $\boldsymbol{\beta}_k$ with all available information. There is no overdispersion caused by lack of information in saturated model, so Poisson distribution is applied.

For traditional scaling-normalization, we fit the following Poisson model

$$Y_{kj} \sim \text{Poisson}(\rho_j 2^{\beta_{kj}}).$$

With observed Y_{kj} and any size factor $\boldsymbol{\rho}$, the normalized expression is equivalent to

$$\hat{\beta}_{kj} = \log_2\left(\frac{Y_{kj}}{\rho_j}\right).$$

Following the same argument for Negative Binomial threshold model, a Poisson threshold model is more appropriate for GeoMx RNA data. The size factors calculated by Poisson Background model and Negative Binomial threshold model need to be rescaled to stay invariant to constraints (see Methods for more details).

Let $\tilde{\boldsymbol{\alpha}}_0 = \gamma_0 \boldsymbol{\alpha}_0 (\text{or } \gamma_t \boldsymbol{\alpha}_0)$ and $\tilde{\boldsymbol{\alpha}} = \gamma_0 \boldsymbol{\alpha}$,

$$Y_{kj} \sim \text{Poisson}(\tilde{\alpha}_j 2^{\beta_{kj}} + \tilde{\alpha}_{0j})$$

One direct consequence of such model is the observations will be fitted perfectly without constraint or priors, and when $Y_{kj} = 0$, the MLE of β_{kj} is $-\infty$. But perfect fitting is usually suboptimal due to overfitting. Parameter regulation is desired for getting parameters generalize better as well as restrict infinite parameter estimation, so we specify a prior for the in addition to the model

$$\boldsymbol{\beta}_k \sim \text{N}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}}).$$

The precision matrix $\Sigma_{\boldsymbol{\beta}}^{-1}$ is estimated by a 2-step Empirical Bayes procedure: in Step 1, the model is applied to high abundance features with a default weak prior; in Step 2, a precision matrix is calculated using the estimated parameters, then the model is applied to all features of interest, using the estimated precision matrix.

There are two options for calculating the precision matrix in **GeoDiff**. One option is "equal", which simply calculates the empirical precision matrix of estimated parameters; one option is "contrast", which simply calculates the precision matrix with only contrast information from the estimated parameters. Precision matrix estimated by the "contrast" option avoids the abundance information of selected high abundance features influence model fitting other features by using the contrast information only, thus is recommended and set as default in **GeoDiff**. Both "contrast" and "equal" way of estimating precision matrix is parameterization invariant (see Methods for details).

The 2-step Empirical Bayes procedure can be applied for whole dataset or separately for each level of a grouping variable. The usual choice of grouping variable is the slide ID. When fitted without grouping variable, ROIs in different slides could influence each other's normalized expression, thus the option of a grouping variable is recommended for normalization in multiple slides data.

0.1.1 Case study: WTA diabetic kidney dataset

We evaluate the performance of this approach to other normalization methods by WTA diabetic kidney dataset and CTA cell mixture dataset.

For WTA diabetic kidney dataset, we apply different normalization methods and the distribution of their $\log_2$ normalized expression in slide "disease4" are compared in Figure 7. The Poisson threshold normalization(Figure 7E, Figure 7F for using slide ID as grouping variable) can align gene expression of different ROI much better than other methods. Specifically, the scaling normalization methods (Figure 7B, Figure 7C) do not correct for shift and the wiggles on the lower end while the scaling method with background subtraction (Figure 7D) incur a big spike on the lower end due the negative numbers it induces, both Poisson threshold normalization (Figure 7E, Figure 7F) can align the distribution of gene expression in a unimodal fashion.

Furthermore, PCA plot Figure 8 shows Poisson threshold normalization better removes technical variability and reveals better clustering pattern in terms of grouping

variable and biological variable than 75% quantile normalization for "middle" abundance genes, defined as genes between the 40% and 60% quantiles of scores from the Background Score Test. "middle" abundance genes are above the background as indicated by the Background Score Test with $p < 1e-3$, but they are still heavily influenced by the background. 75% quantile normalization yield results with its PC1 dominated by technical variation defined as the $\log_2$ ratio of 75% quantile and background size factor, showing it is not able to remove this technical variability, while PCs in both Poisson threshold normalization are less dictated by this variable, showing they both remove this technical variability better. Poisson threshold normalization results both cluster better to the grouping variable and the main biological variable: region. Interestingly, the Poisson threshold model without grouping variable yield better overall clustering in terms of ROI regions with apparent and not distinct clustering in terms of slides, while the Poisson threshold model with slide ID as grouping variable yield better overall clustering in terms of slides, with ROI regions clustered in the same direction within the same slide.

0.1.2 Case study: CTA cell mixture dataset

For CTA dataset of mixture of HEK293T and CCRF-CEM cell lines with varying proportions, we has "partial truth" we can leverage. Specifically, while we do not know the "true $\log_2$ expression" of every gene of HEK293T and CCRF-CEM, if we assume τ_{k1}, τ_{k2} are the "true $\log_2$ expression" of kth gene of HEK293T and CCRF-CEM, then the "true $\log_2$ expression" of any mixture with the proportion of HEK293T as x is

$$\log_2\left(2^{\tau_{k1}}x + 2^{\tau_{k2}}(1-x)\right).$$

By the following steps, we evaluate how normalized expression from different methods align with this unique functional relationship.

1 Perform normalization

2 For kth gene, estimate τ_{k1}, τ_{k2} using the $\log_2$ normalized expression $\boldsymbol{\beta}_k$ by minimizing mean squared error (MSE)

$$\frac{1}{J}\sum_{j=1}^{J}\left(\log_2\left(2^{\tau_{k1}}x_j + 2^{\tau_{k2}}(1-x_j)\right) - \beta_{kj}\right)^2$$

3 For kth gene and jth ROI, where x_j is the proportion of HEK293T, calculate the expected $\log_2$ normalized expression by

$$z_{kj} = \log_2\left(2^{\hat{\tau}_{k1}}x_j + 2^{\hat{\tau}_{k2}}(1-x_j)\right)$$

4 Calculate expected raw count by the model assumption of Poisson threshold normalization and 75% quantile normalization by the Poisson threshold model and Poisson scaling model assumptions.

5 Calculate the correlation of expected raw count vs the raw count

Figure 9A showcases this unique functional relationship between true $\log_2$ expression and percentage of HEK293T in the mixture using gene ABCB1 as example,

in which $\hat{\tau}_{k1} = 3.37, \hat{\tau}_{k2} = -0.33$ estimated by Poisson threshold normalization results. Figure 9B shows how $\hat{\tau}_{k1}, \hat{\tau}_{k2}$ are fitted by the steps 1,2 and Poisson threshold normalization with and without EB prior gives very close estimates but the MSE is smaller in Poisson normalization with EB prior, which is further confirmed by Figure 9C where MSE of all genes from the two models are compared. Thus we can say the EB prior of Poisson threshold normalization imposed information implied by high features to all features and yield more precise results.

The results of Step 4 of the workflow is shown in Figure 10, in which correlation of expected raw count from different normalization methods vs raw count are compared. In Figure 10A, Poisson threshold normalization with and withour EB prior are compared; Figure 10B, Poisson threshold normalization with EB prior and 75% quantile normalization are compared, and Poisson threshold normalization with EB prior has better performance in both comparison.

Targets with multiple probes

There are products on GeoMx platforms with multiple probes for one target. When the number of probes are big, some screening can be conducted to ensure the probes correlate well with each other and are in similar dynamic range. After that, the sum of multiple probes for that target is used for analysis, as Poisson and Negative Binomial distributions have the following property.

For

$$Y_m \sim \text{Poisson}(\lambda),$$
$$m = 1, \cdots, M,$$

then

$$\sum_{m=1}^{M} Y_m \sim \text{Poisson}(M\lambda);$$

and

$$Y_m \sim \text{NB}(\mu, r),$$
$$m = 1, \cdots, M,$$

then

$$\sum_{m=1}^{M} Y_m \sim \text{NB}(M\mu, Mr).$$

Furthermore, the sum of counts are sufficient statistics[18] for both models so no information is lost in the aggregation counts for parameter estimation. As long as the number of probe involved in the probe sum is correctly specified in the models, the estimated parameter will be scaled properly. The CTA cell line mixture dataset has 3 probes for each target and has been aggregated as described. The previous section shows the **GeoDiff** methods work well in such case.

Conclusions

This paper describes statistical models/tests from **GeoDiff** including Poisson Background model, Background Score Test, Negative Binomial threshold model and Poisson threshold model for normalization, and recommends a GeoMx RNA data analysis workflow starting with modeling the negative probes for background size factor, followed by modeling the high features for estimating signal size factor, testing whether a target is above the background, assessing the signal level of an ROI, and perform normalization using outputs from these models/tests. We have shown the proposed methods outperform conventional methods in various aspects.

Methods

0.2 Identifiability

The Poisson Background model is intrinsic non-identifiable, i.e. different sets of parameters $c\gamma_i, \alpha_{0j}/c$ for any positive constant c yield the same model. To impose identifiability, a constraint like $\sum_{j=1}^{J} \alpha_{0j} = 1$ must be enfored. Constraints are arbitrary. However, $\sum_{j=1}^{J} \alpha_{0j} = 1$ is equivalent to $\sum_{j=1}^{J} X_{ij} \sim \text{Poisson}\,(\gamma_i)$, making the total count $\sum_{j=1}^{J} X_{ij}$ the MLE of γ_i. This interpretation is convenient in a lot of applications.

The Negative Binomial threshold model suffers from identifiability problems just like Poisson Background model. Specifically, $\gamma_k, k = 1, \cdots, K, \gamma_t, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}$ and $\gamma_k/c, k = 1, \cdots, K, \gamma_t/c, c\boldsymbol{\alpha}_0, c\boldsymbol{\alpha}$ give the same model for any positive constant c. So similarly, we impose a constraint $\sum_{j=1}^{J} \alpha_j = 1$ to the model.

Applying different constraints can change the parameters by a factor, while some functions of parameters, including $\gamma_0\boldsymbol{\alpha}_0$ and $\gamma_0\boldsymbol{\alpha}$ are invariant to choice of constraints. Let $\tilde{\boldsymbol{\alpha}}_0 = \gamma_0\boldsymbol{\alpha}_0$ and $\tilde{\boldsymbol{\alpha}} = \gamma_0\boldsymbol{\alpha}$, which are parameterization(constraint) invariant size factors and we use them as size factor in normalization.

0.3 Diagnostics and Remedy for Poisson Background model

One important metric of any Poisson model is the dispersion, i.e. the ratio of variance vs mean, which equals to 1 theoretically. In Poisson Background model, dispersion of ith feature and jth sample is

$$d_{ij} = \frac{E(x_{ij} - E(x_{ij}))^2}{E(x_{ij})} \equiv 1,$$

which can be estimated by squared Pearson residual at x_{ij}

$$\widehat{d_{ij}} = \frac{(x_{ij} - \widehat{\gamma_i}\widehat{\alpha_{0j}})^2}{\widehat{\gamma_i}\widehat{\alpha_{0j}}}.$$

To reduce variation caused by randomness of a single point, we calculate the empirical (mean) dispersion as a diagnostics metric for this model.

$$\widehat{d} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(x_{ij} - \widehat{\gamma_i}\widehat{\alpha_{0j}})^2}{\widehat{\gamma_i}\widehat{\alpha_{0j}}}}{IJ}$$

We call the data overdispersed when $\widehat{d}$ is much larger than 1, a rule of thumb cutoff can be 2. Overdispersion is a sign of extra variation not captured by model (1).

Besides empirical dispersion, a ppplot is generated as well. Let $F(x|\lambda)$ be the cumulative distribution function of the Poisson distribution, by model assumption

$$F(x_{ij}|\gamma_i\alpha_{0j}) \overset{iid}{\sim} U(0,1)$$

We calculate

$$F(x_{ij}|\widehat{\gamma_i}\widehat{\alpha_{0j}}),$$

the empirical cumulative probability function values, then sort them from the smallest to the largest, and plot the sorted vector against a even grid over $[0,1]$ with IJ points, the theoretical cumulative probability function values. In practice, we have found the count data for negative probes could be very low, a decent amount could just be 0, and the ppplot would end up with zigzag shape on the lower end due to that. A way to mitigate that is to simulate data from the fitted model, generate empirical cumulative probability function values, sort them and use the sorted vector as the theoretical cumulative probability function values.

For a good model fit, the ppplot is almost a diagonal line from 0 to 1. The most common aberration of model assumption results in an "S" shape in ppplot, indicating overdispersion.

For real targets with biological variability, fitting Poisson Background model is inadequate and will lead to overdispersion. More complicated models explained in the following sections are needed to account for their additional variation.

Even the features are only negative probes, overdispersion could still occur. For data from multiple sources, mean expression level of each probe could be different. For example, if there is some dramatic difference in experimental setting among different slides of samples, the mean expression level of each probe could be different from slide to slide. Assuming there are $s = 1, \cdots, S$ different sources of data, a quick check is to fit the alternative model below and recalculate $\hat{d}$ or generate heatmap using γ_{is}.

$$X_{ij} \sim \text{Poisson}\left(\gamma_{is}\alpha_{0j}\right)$$
$$\sum_{j=1}^{J} \alpha_{0j} = 1$$

where γ_{is} are the mean expression level for ith feature from sth source (batch, slide,... etc.). A much smaller $\hat{d}$ and a clear clustering in the heatmap of feature factors indicate batch effect.

Overdispersion could also occur due to outliers. We detect outliers by the empirical probability of each point using the estimated parameters. Detected outliers are mostly large outliers for negative probes. On the lower end, 0 is unlikely to be determined as outliers in a negative probe count matrix, of which counts are usually low. After outliers are identified, they can be directly set as missing and Poisson background model could be refitted with missing values. Furthermore, if ith feature

or the jth ROI have a lot of counts are outliers, the ith feature or the jth sample should be considered as an outlier, and be removed from the Poisson Background model. Furthermore, the jth sample outlier should be removed universally from all downstream analysis since the underlying mechanism impact the negative probes could impact other features too.

0.4 Background Score Test

0.4.1 Without prior

After applying the Poisson Background Model to the negative probes, it is imperative to test whether a target is above the background use the fitted model as a reference. Such test is called background score test.

For kth feature, assuming it has only one probe, and it still follows the Poisson background model

$$X_{kj} \sim \text{Poisson}\left(\gamma_k \alpha_{0j}\right)$$

where the α_{0j} are background size factors estimated from Poisson background model on negative probes

$$X_{ij} \sim \text{Poisson}\left(\gamma_i \alpha_{0j}\right).$$

Let $\gamma_0 = \dfrac{\sum_{i=1}^{I} \gamma_i}{I}$, we are interested in testing

$$H_0 : \ \gamma_k \leq \gamma_0,$$
$$H_a : \ \gamma_k > \gamma_0.$$

Let $\boldsymbol{x}_k$ be the vector of x_{ik}, the observed count for kth feature in ith sample.

$$\log f\left(x_{ik}|\gamma_k\right) = x_{ik} \log\left(\alpha_{0i}\gamma_k\right) - \alpha_{0i}\gamma_k - \log\left(x_{ik}!\right)$$

and the likelihood function

$$L(\gamma_k|\boldsymbol{x}) = \sum_{i=1}^{I} \log f\left(x_{ik}|\gamma_k\right)$$

The score, i.e, the gradient of the log-likelihood function:

$$\frac{dL(\gamma_k|\boldsymbol{x})}{d\gamma_k} = \sum_{i=1}^{I}\left(\frac{x_{ik}}{\gamma_k} - \alpha_{0i}\right) = \frac{\sum_{i=1}^{I} x_i}{\gamma_k} - \sum_{i=1}^{I} \alpha_{0i}$$

The Fisher information

$$\begin{aligned} I(\gamma_k) &= \mathrm{var}\left(\frac{dL(\gamma_k|\boldsymbol{x})}{d\gamma_k}\right) \\ &= \mathrm{var}\left(\frac{\sum_{i=1}^{I} x_i}{\gamma_k} - \sum_{i=1}^{I}\alpha_{0i}\right) \\ &= \mathrm{var}\left(\frac{\sum_{i=1}^{I} x_i}{\gamma_k}\right) \\ &= \frac{\sum_{i=1}^{I}\mathrm{var}(x_i)}{\gamma_k^2} = \frac{\sum_{i=1}^{I}\alpha_{0i}\gamma_k}{\gamma_k^2} = \frac{\sum_{i=1}^{I}\alpha_{0i}}{\gamma_k} \end{aligned}$$

the score statistic

$$\begin{aligned} \frac{\left(\frac{dL(\gamma_k|\boldsymbol{x})}{d\gamma_k}\right)^T \frac{dL(\gamma_k|\boldsymbol{x})}{d\gamma_k}}{I(\gamma_k)} &= \frac{\left(\frac{\sum_{i=1}^{I} x_i}{\gamma_k} - \sum_{i=1}^{I}\alpha_{0i}\right)^2}{\frac{\sum_{i=1}^{I}\alpha_{0i}}{\gamma_k}} \\ &= \left(\frac{\sum_{i=1}^{I} x_i - \sum_{i=1}^{I}\alpha_{0i}\gamma_k}{\sqrt{\sum_{i=1}^{I}\alpha_{0i}\gamma_k}}\right)^2 \end{aligned}$$

Under null hypothesis

$$\left(\frac{\sum_{i=1}^{I} x_i - \sum_{i=1}^{I}\alpha_{0i}\gamma_k}{\sqrt{\sum_{i=1}^{I}\alpha_{0i}\gamma_k}}\right)^2\Bigg|_{\gamma_k=\gamma_0} \sim \chi_1^2$$

Or

$$\frac{\sum_{i=1}^{I} x_i - \sum_{i=1}^{I}\alpha_{0i}\gamma_0}{\sqrt{\sum_{i=1}^{I}\alpha_{0i}\gamma_0}} \sim \mathrm{N}(0,1)$$

Since this is a one-sided test, we reject the null when

$$\frac{\sum_{i=1}^{I} x_i - \sum_{i=1}^{I}\alpha_{0i}\gamma_k}{\sqrt{\sum_{i=1}^{I}\alpha_{0i}\gamma_0}} > Z_\alpha$$

where α is the significance level, default is $\alpha = 0.001$.

0.4.2 With prior

For kth feature, assuming it has only one probe, and it still follows the Poisson background model with Gamma prior

$$\begin{aligned} X_{kj}|\gamma_k &\sim \mathrm{Poisson}(\gamma_k\alpha_{0j}), \\ \gamma_k &\sim \mathrm{Gamma}\left(\frac{1}{\sigma}, \sigma\mu_k\right). \end{aligned}$$

With α_{0j}, σ and μ_0 estimated from Poisson background model using negative probes

$$X_{ij}|\gamma_i \sim \text{Poisson}\left(\gamma_i \alpha_{0j}\right),$$
$$\gamma_i \sim \text{Gamma}\left(\frac{1}{\sigma}, \sigma\mu_0\right),$$

we are interested in testing

$$H_0: \ \mu_k \leq \mu_0$$
$$H_a: \ \mu_k > \mu_0$$

$$\begin{aligned}
f\left(x_{ik}|\mu_k,\sigma\right) &= \int_0^\infty f\left(x_{ik}|\gamma_k\right) f\left(\gamma_k|\sigma,\mu_k\right) d\gamma_k \\
&= \int_0^\infty \frac{\left(\alpha_{0i}\gamma_k\right)^{x_{ik}} \ e^{-\alpha_{0i}\gamma_k}}{x_{ik}!} \ \cdot \frac{1}{\Gamma\left(\frac{1}{\sigma}\right)\left(\sigma\mu\right)^{\frac{1}{\sigma}}} {\gamma_k}^{\frac{1}{\sigma}-1} e^{-\frac{\gamma_k}{\sigma\mu_k}} d\gamma_k \\
&= \int_0^\infty \frac{\left(\sigma\mu_k\right)^{-\frac{1}{\sigma}}}{x_{ik}!\ \Gamma\left(\frac{1}{\sigma}\right)} \alpha_{0i}^{x_{ik}} {\gamma_k}^{x_{ik}+\frac{1}{\sigma}-1} e^{-(\alpha_{0i}+\frac{1}{\sigma\mu_k})\gamma_k} d\gamma_k \\
&= \frac{\left(\sigma\mu_k\right)^{-\frac{1}{\sigma}}}{x_{ik}!\ \Gamma\left(\frac{1}{\sigma}\right)} a_{0i}^{x_{ik}} \int_0^\infty {\gamma_k}^{x_{ik}+\frac{1}{\sigma}-1} e^{-(\alpha_{0i}+\frac{1}{\sigma\mu_k})\gamma_k} d\gamma_k \\
&= \frac{\left(\sigma\mu_k\right)^{-\frac{1}{\sigma}}}{x_{ik}!\ \Gamma\left(\frac{1}{\sigma}\right)} a_{0i}^{x_{ik}} \frac{\Gamma\left(x_{ik}+\frac{1}{\sigma}\right)}{\left(\alpha_{0i}+\frac{1}{\sigma\mu_k}\right)^{x_{ik}+\frac{1}{\sigma}}} \\
&= \frac{\Gamma\left(x_{ik}+\frac{1}{\sigma}\right)}{x_{ik}!\ \Gamma\left(\frac{1}{\sigma}\right)} \left(\frac{\frac{1}{\sigma\mu}}{\alpha_{0i}+\frac{1}{\sigma\mu_k}}\right)^{\frac{1}{\sigma}} \left(\frac{\alpha_{0i}}{\alpha_{0i}+\frac{1}{\sigma\mu_k}}\right)^{x_{ik}}
\end{aligned}$$

Thus

$$x_{ik} \sim \text{NB}(\frac{\alpha_{0i}}{\alpha_{0i}+\frac{1}{\sigma\mu_k}}, \frac{1}{\sigma})$$

$$\begin{aligned}
\log f\left(x_{ik}|\mu_k,\sigma\right) &= -\log x_{ik}! + \log\left(\frac{\Gamma\left(x_{ik}+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)}\right) \\
&\quad - \left(x_{ik}+\frac{1}{\sigma}\right)\log\left(\alpha_{0i}+\frac{1}{\sigma\mu_k}\right) + x_{ik}\log\left(\alpha_{0i}\right) - \frac{1}{\sigma}\log\left(\sigma\mu_k\right) \\
&= -\log x_{ik}! + \log\left(\frac{\Gamma\left(x_{ik}+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)}\right) \\
&\quad - \left(x_{ik}+\frac{1}{\sigma}\right)\log\left(\alpha_{0i}\sigma\mu_k+1\right) + \left(x_{ik}+\frac{1}{\sigma}\right)\log\left(\sigma\mu_k\right) + x_{ik}\log\left(\alpha_{0i}\right) - \frac{1}{\sigma}\log\left(\sigma\mu_k\right) \\
&= -\log x_{ik}! + \log\left(\frac{\Gamma\left(x_{ik}+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)}\right) + x_{ik}\log(\alpha_{0i}\sigma\mu_k) \\
&\quad - \left(x_{ik}+\frac{1}{\sigma}\right)\log\left(\alpha_{0i}\sigma\mu_k+1\right)
\end{aligned}$$

Given σ, the likelihood function is

$$L(\mu_k|\boldsymbol{x}) = \sum_{i=1}^{I}\log f\left(x_k|\mu_k,\sigma\right)$$

$$\frac{dL(\gamma_k|\boldsymbol{x})}{d\mu_k} = \sum_{i=1}^{I}\frac{x_{ik}}{\mu_k} - \frac{x_{ik}\sigma\alpha_{0i}+\alpha_{0i}}{\alpha_{0i}\sigma\mu_k+1} = \sum_{i=1}^{I}\frac{x_{ik}-\alpha_{0i}\mu_k}{(\alpha_{0i}\sigma\mu_k+1)\mu_k}$$

The Fisher information

$$\begin{aligned}
I\left(\gamma_k\right) &= \mathrm{var}\left(\frac{dL(\gamma_k|\boldsymbol{x})}{d\gamma_k}\right) \\
&= \mathrm{var}\left(\sum_{i=1}^{I}\frac{x_{ik}-\alpha_{0i}\mu_k}{(\alpha_{0i}\sigma\mu_k+1)\mu_k}\right) \\
&= \mathrm{var}\left(\sum_{i=1}^{I}\frac{x_{ik}}{(\alpha_{0i}\sigma\mu_k+1)\mu_k}\right) \\
&= \sum_{i=1}^{I}\frac{\mathrm{var}\left(x_i\right)}{((\alpha_{0i}\sigma\mu_k+1)\mu_k)^2} \\
&= \sum_{i=1}^{I}\frac{\frac{\frac{\alpha_{0i}}{\alpha_{0i}+\frac{1}{\sigma\mu_k}}\frac{1}{\sigma}}{(\frac{\frac{1}{\sigma\mu_k}}{\alpha_{0i}+\frac{1}{\sigma\mu_k}})^2}}{((\alpha_{0i}\sigma\mu_k+1)\mu_k)^2} \\
&= \sum_{i=1}^{I}\frac{\alpha_{0i}}{(\alpha_{0i}\sigma\mu_k+1)\mu_k}
\end{aligned}$$

Under the null hypothesis

$$\left.\frac{\left(\frac{dL(\gamma_k|\boldsymbol{x})}{d\mu_k}\right)^T\frac{dL(\gamma_k|\boldsymbol{x})}{d\mu_k}}{I(\mu_k)}\right|_{\mu_k=\mu_0} = \left.\left(\frac{\sum_{i=1}^{I}\frac{x_{ik}-\alpha_{0i}\mu_k}{(\alpha_{0i}\sigma\mu_k+1)\mu_k}}{\sqrt{\sum_{i=1}^{I}\frac{\alpha_{0i}}{(\alpha_{0i}\sigma\mu_k+1)\mu_k}}}\right)^2\right|_{\mu_k=\mu_0} \sim \chi_1^2$$

Or

$$\frac{\sum_{i=1}^{I} \frac{x_{ik}-\alpha_{0i}\mu_0}{(\alpha_{0i}\sigma\mu_0+1)\mu_0}}{\sqrt{\sum_{i=1}^{I} \frac{\alpha_{0i}}{(\alpha_{0i}\sigma\mu_0+1)\mu_0}}} \sim \mathrm{N}(0,1)$$

and we reject the null when

$$\frac{\sum_{i=1}^{I} \frac{x_{ik}-\alpha_{0i}\mu_0}{(\alpha_{0i}\sigma\mu_0+1)\mu_0}}{\sqrt{\sum_{i=1}^{I} \frac{\alpha_{0i}}{(\alpha_{0i}\sigma\mu_0+1)\mu_0}}} > Z_\alpha$$

where α is the significance level, default is $\alpha = 0.001$.

0.5 Deriving $\Sigma_{\boldsymbol{\beta}}^{-1}$ in Poisson threshold model

The covariance matrix $\Sigma_{\boldsymbol{\beta}}^{-1}$ is determined by Empirical Bayes approach in 2 steps.

1. Solving this model using the set of high features using a default prior $\Sigma_{\boldsymbol{\beta}}^{-1} = \frac{1}{\sigma^2}B^TB$, where $B = \left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)$. This means we are only adding a penalty of $\frac{1}{\sigma^2}$ to the mean of each of these high abundance feature, the default is $\sigma = 5$. This prior amounts to a belief of mean $\log_2$ expression of each feature follows distribution $N(0, \frac{1}{25})$. It is a weak penalty helps with the numerical stability especially with 0 counts.
2. Using the $\hat{\boldsymbol{\beta}}$ estimated from high features, calculate $\Sigma^{-1} = kA^T\hat{\Sigma}_c^{-1}A + \frac{1}{\sigma^2}B^TB$, where $B = \left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)$, and $A_{(n-1)\times n}$ is any full rank matrix satisfying $AB^T = \mathbf{0}$. By definition, $A_{(n-1)\times n}$ consists of vectors of contrasts. $\hat{\Sigma}_c$ is the empirical covariance matrix of $A\boldsymbol{\beta}_k$, $A^T\hat{\Sigma}_c^{-1}A$ is invariant with respect to different choices of A. The contrast factor $k \in (0, 1)$ can adjust the penality level of contrast. This form of EB prior is based on the idea of decomposing the precision matrix into the orthogonal space of precision matrix of mean and precision matrix of contrasts, rooted from the belief the contrast information of the high features should be passed to the parameter estimation of other features, but they are in different dynamic range so the behavior of mean expression of high features should not impact the behavior of other features.

It is more natural to assume only the ROIs in the same slide are correlated, so in the case of multiple slides data, it is advisable to apply this normalization function on each slide separately, which is implemented in **GeoDiff** as default for normalization of multiple slides data.

Acknowledgements
Text for this section...

Funding
Text for this section...

Abbreviations
Text for this section...

Availability of data and materials
Text for this section...

Ethics approval and consent to participate
Text for this section...

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Text for this section. . .

Authors' contributions
Text for this section . . .

Authors' information
Text for this section. . .

Author details
NanoString Technologies, Seattle, US.

References

1. H, W., C, H., H, W., Y, L., S, S., J, L., H, W., AJ, S., C, W.: NanoStringDiff: a novel statistical method for differential expression analysis based on NanoString nCounter data. Nucleic acids research **44**(20), 151 (2016). doi:10.1093/NAR/GKW677
2. Roberts, K., Aivazidis, A., Kleshchevnikov, V., Li, T., Fropf, R., Rhodes, M., Beechem, J., Hemberg, M., Bayraktar, O.: Transcriptome-wide spatial RNA profiling maps the cellular architecture of the developing human neocortex. biorxiv.org. doi:10.1101/2021.03.20.436265
3. Brady, L., Kriner, M., Coleman, I., Morrissey, C., Roudier, M., True, L.D., Gulati, R., Plymate, S.R., Zhou, Z., Birditt, B., Meredith, R., Geiss, G., Hoang, M., Beechem, J., Nelson, P.S.: Inter- and intra-tumor heterogeneity of metastatic prostate cancer determined by digital spatial gene expression profiling. Nature Communications 2021 12:1 **12**(1), 1–16 (2021). doi:10.1038/s41467-021-21615-4
4. Brumbaugh, C.D., Kim, H.J., Giovacchini, M., Pourmand, N.: NanoStriDE: Normalization and differential expression analysis of NanoString nCounter data. BMC Bioinformatics **12** (2011). doi:10.1186/1471-2105-12-479
5. Park, T., Yi, S.-G., Kang, S.-H., Lee, S., Lee, Y.-S., Simon, R.: Evaluation of normalization methods for microarray data. BMC Bioinformatics 2003 4:1 **4**(1), 1–13 (2003). doi:10.1186/1471-2105-4-33
6. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 2014 15:12 **15**(12), 1–21 (2014). doi:10.1186/S13059-014-0550-8
7. Robinson, M.D., Oshlack, A.: A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 2010 11:3 **11**(3), 1–9 (2010). doi:10.1186/GB-2010-11-3-R25
8. Li, X., Cooper, N.G.F., O'Toole, T.E., Rouchka, E.C.: Choice of library size normalization and statistical methods for differential gene expression analysis in balanced two-group comparisons for RNA-seq studies. BMC Genomics **21**(1) (2020). doi:10.1186/S12864-020-6502-7
9. Desai, N., Neyaz, A., Szabolcs, A., Shih, A.R., Chen, J.H., Thapar, V., Nieman, L.T., Solovyov, A., Mehta, A., Lieb, D.J., Kulkarni, A.S., Jaicks, C., Pinto, C.J., Juric, D., Chebib, I., Stone, J.R., Ting, D.T.: Temporal and Spatial Heterogeneity of Host Response to SARS-CoV-2 Pulmonary Infection. doi:10.1101/2020.07.30.20165241
10. Friedman, N., Cai, L., Xie, X.S.: Linking stochastic dynamics to population distribution: an analytical framework of gene expression. Physical review letters **97**(16) (2006). doi:10.1103/PHYSREVLETT.97.168302
11. Hafemeister, C., Satija, R.: Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biology 2019 20:1 **20**(1), 1–15 (2019). doi:10.1186/S13059-019-1874-1
12. Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Linde, A.V.D.: Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**(4), 583–639 (2002). doi:10.1111/1467-9868.00353
13. NanoString: nanostring-spatial-omics-hackathon (2021). http://emarketing.nanostring.com/nanostring-spatial-omics-hackathon Accessed 2021-08-11
14. Merritt, C.R., Ong, G.T., Church, S.E., Barker, K., Danaher, P., Geiss, G., Hoang, M., Jung, J., Liang, Y., McKay-Fleisch, J., Nguyen, K., Norgaard, Z., Sorg, K., Sprague, I., Warren, C., Warren, S., Webster, P.J., Zhou, Z., Zollinger, D.R., Dunaway, D.L., Mills, G.B., Beechem, J.M.: Multiplex digital spatial profiling of proteins and RNA in fixed tissue. Nature Biotechnology 2020 38:5 **38**(5), 586–599 (2020). doi:10.1038/s41587-020-0472-9
15. Danaher, P., Kim, Y., Nelson, B., Griswold, M., Yang, Z., Piazza, E., Beechem, J.M.: Advances in mixed cell deconvolution enable quantification of cell types in spatially-resolved gene expression data. doi:10.1101/2020.08.04.235168
16. Zimmerman, S.M., Fropf, R., Kulasekara, B.R., Griswold, M., Appelbe, O., Bahrami, A., Boykin, R., Buhr, D.L., Fuhrman, K., Hoang, M.L., Huynh, Q., Isgur, L., Klock, A., Kutchma, A., Lasley, A.E., Liang, Y., McKay-Fleisch, J., Nelson, J.S., Nguyen, K., Piazza, E., Rininger, A., Zollinger, D.R., Rhodes, M., Beechem, J.M.: Spatially resolved whole transcriptome profiling in human and mouse tissue using Digital Spatial Profiling. bioRxiv, 2021–0929462442 (2021). doi:10.1101/2021.09.29.462442
17. Engle, R.F.: Chapter 13 Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. Handbook of Econometrics **2**, 775–826 (1984). doi:10.1016/S1573-4412(84)02005-5
18. Kleven, H.J.: Sufficient Statistics Revisited. https://doi.org/10.1146/annurev-economics-060220-023547 **13**, 515–538 (2021). doi:10.1146/ANNUREV-ECONOMICS-060220-023547

Additional Files
Additional file 1 — numerics.pdf
Description of numerical methods for solving the models included in the paper

Figures

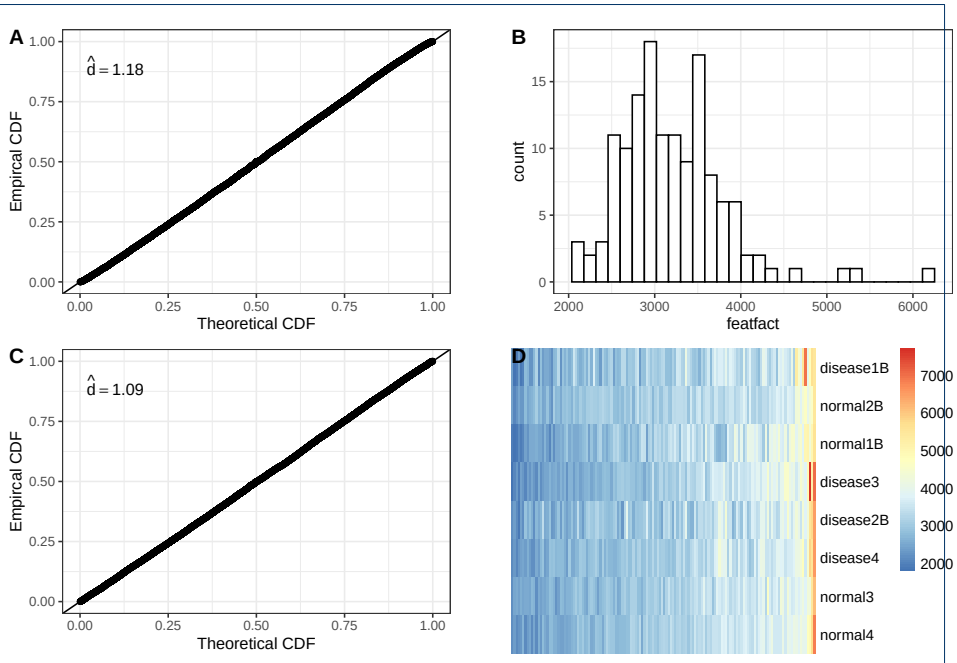


Figure 1: Diagnostics and distribution of feature factors from Poisson Background Model

Poisson Background model, treating the feature factors constant in the whole dataset(A, B), or within each slide(C, D). The P-P plot (A, C) show both models have a good fit with estimated dispersion close to 1. This can be explained by homogeneous distribution of the slide specific feature factors (D).

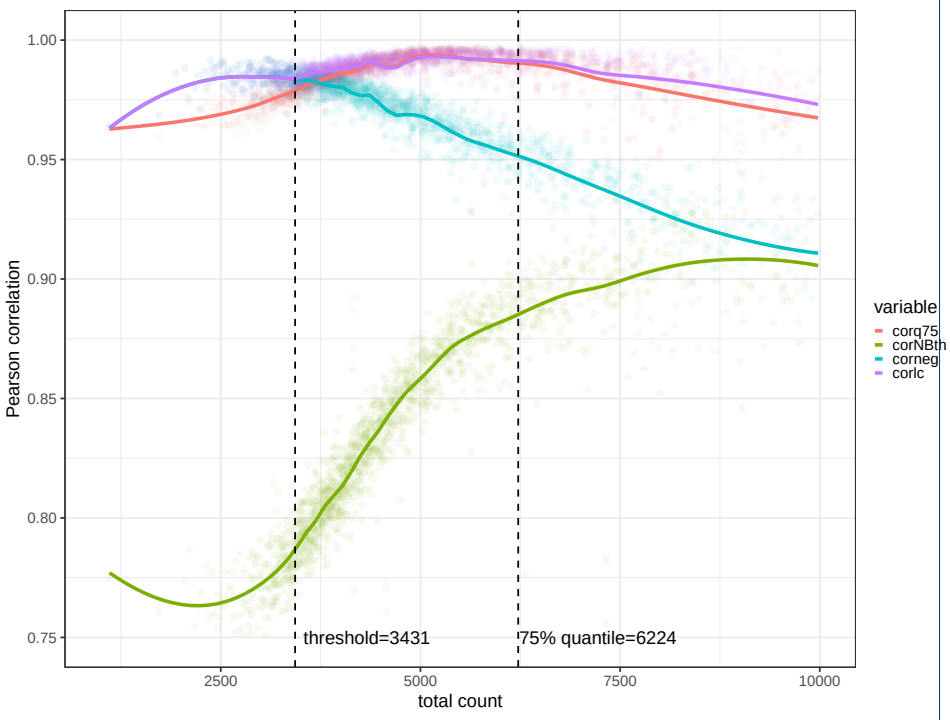


Figure 2: Correlation of counts and size factors

This plot shows the correlation of size factors: background sizefactor estimated from Poisson Background Model(corneg), signal sizefactor estimated from Negative Binomial threshold model(corNBth), linear combination of both based on the Negative Binomial threshold model(corlc), and 75% quantile(corq75) with average counts of binned features(bin size=10) sorted by scores of their Background Score Test.

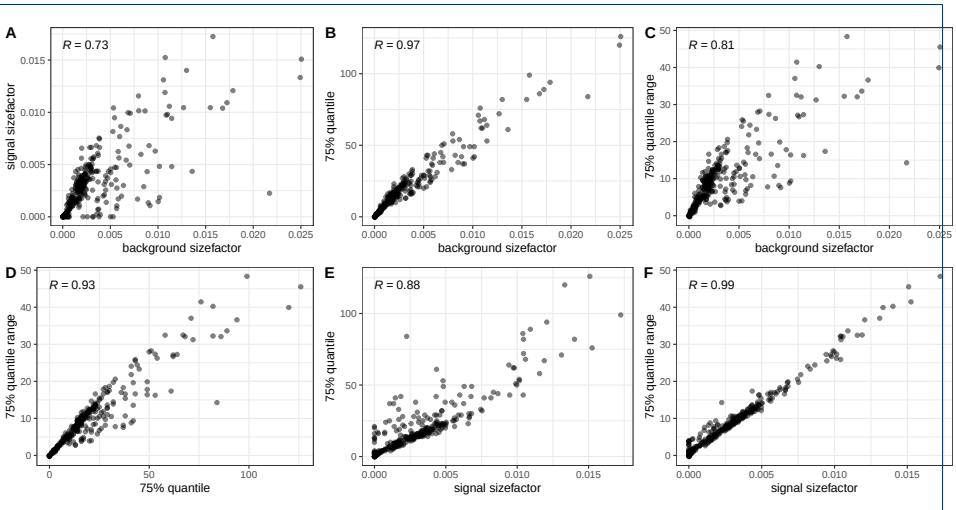


Figure 3: Scatterplots between size factors, 75%quantile and 75% quantile range

Scatterplot between background size factor vs signal size factor(A), 75% quantile(B) and 75% quantile range(C); between 75% quantile(B) and 75% quantile range(D), and signal size factor vs 75% quantile(E) and 75% quantile range(F).

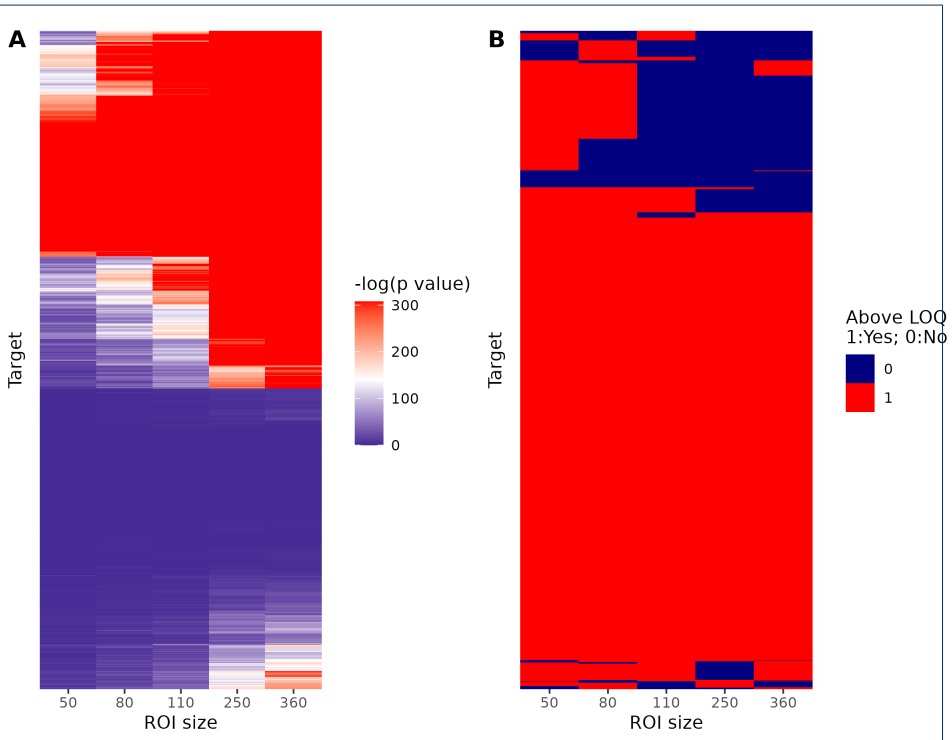


Figure 4: Heatmaps of p-value and indicators of targets above Limit of Quantitation (LOQ) across different ROI size in WTA cell pellet array data

A:Each row of the heatmap represents a target's -log(p-value) from background score test across different ROI sizes. B:Each row of the heatmap represents the indicator of whether a target is above LOQ (1: above; 0: below) across different ROI sizes.

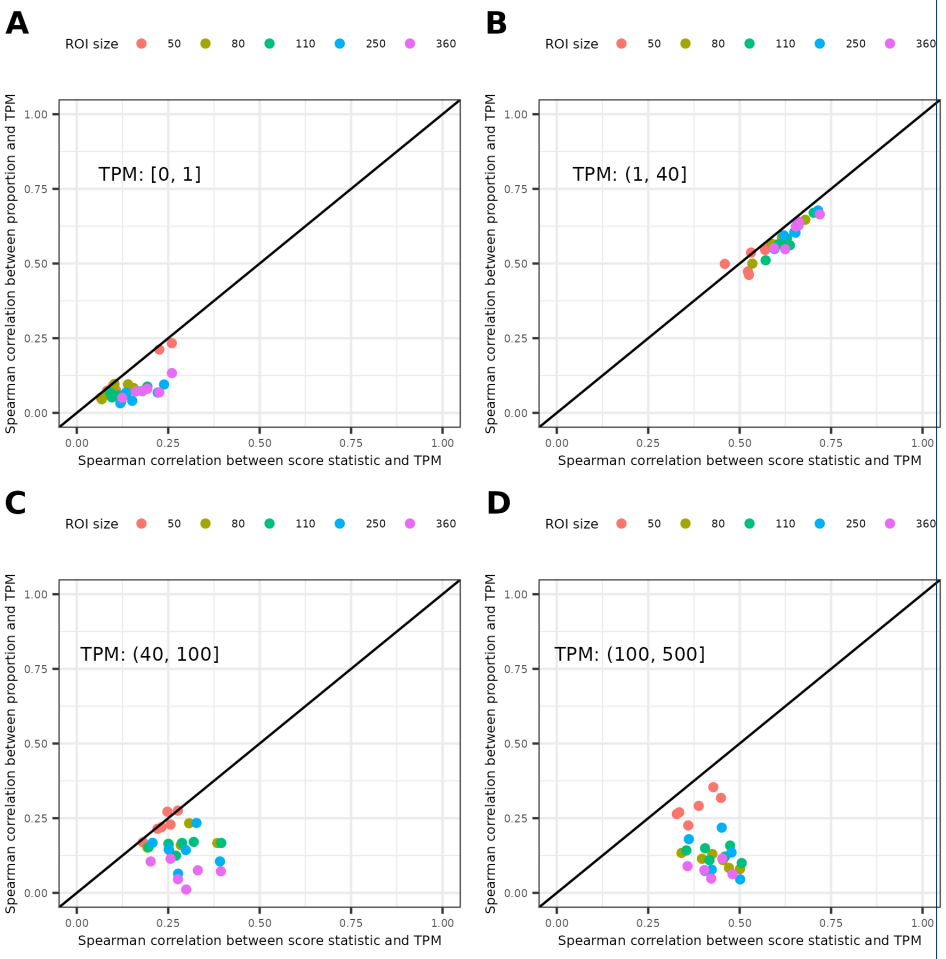


Figure 5: A comparison of Spearman's correlations between Transcripts Per million (TPM) and score statistics vs. TPM and the proportions above LOQ

Comparison is split at varying TPM intervals, including (A) [0, 1], (B) (1, 40], (C) (40, 100], and (D) (100, 500], in WTA cell pellet array data.

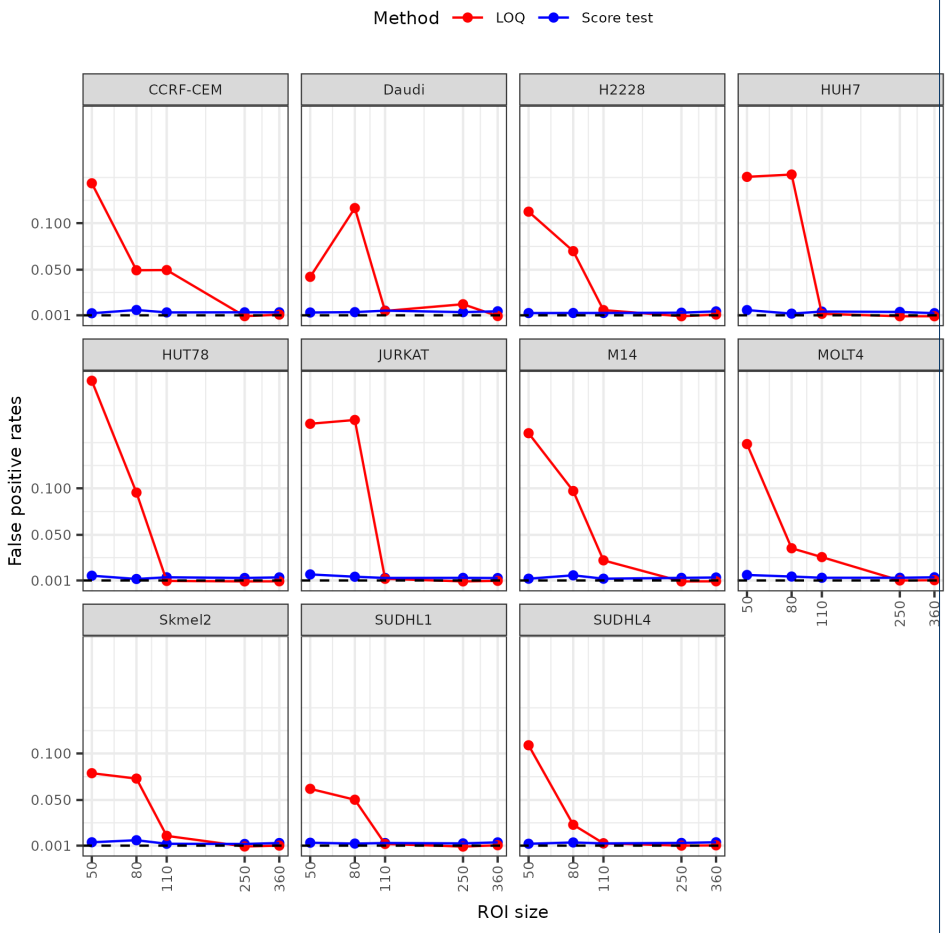


Figure 6: False positive rates (FPRs) of Background Score test and LOQ method for varying ROI sizes and different cell lines in simulated negative probes

The two solid lines represents the FPRs of background score test (blue) and LOQ method (blue), respectively. The dashed line represents the threshold of nominal alpha level, 0.001.

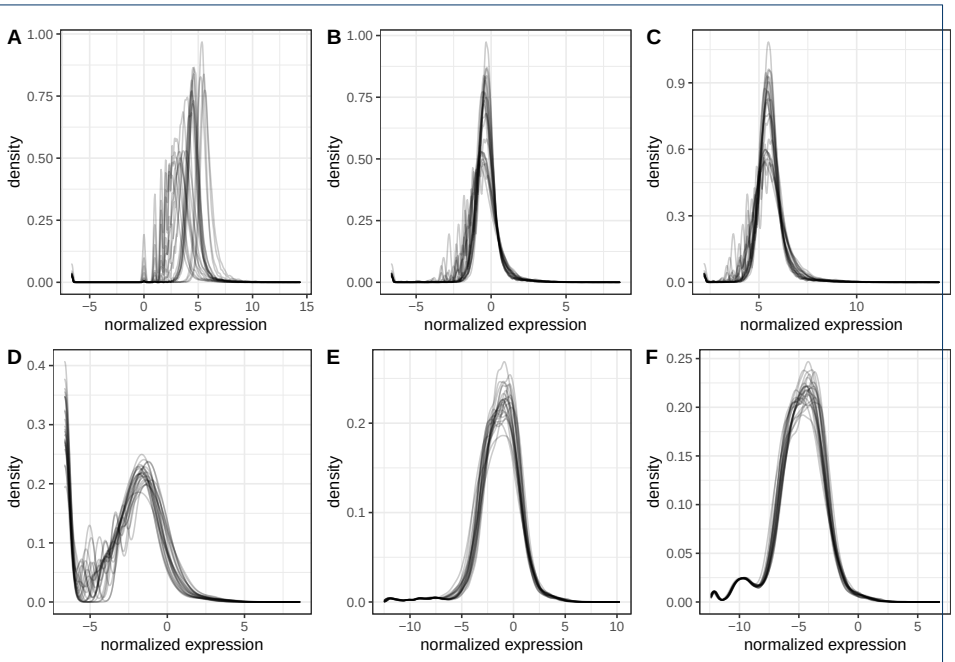


Figure 7: Density plot for gene expression for slide "disease4" in kidney data

the density of log2 expression of raw counts(A), 75% quantile normaliztion(B), TMM[7](C), 75% quantile normaliztion with background subtraction(D), Poisson threshold normalization(E) and Poisson threshold normalization by grouping variable (F) for the slide "disease4".

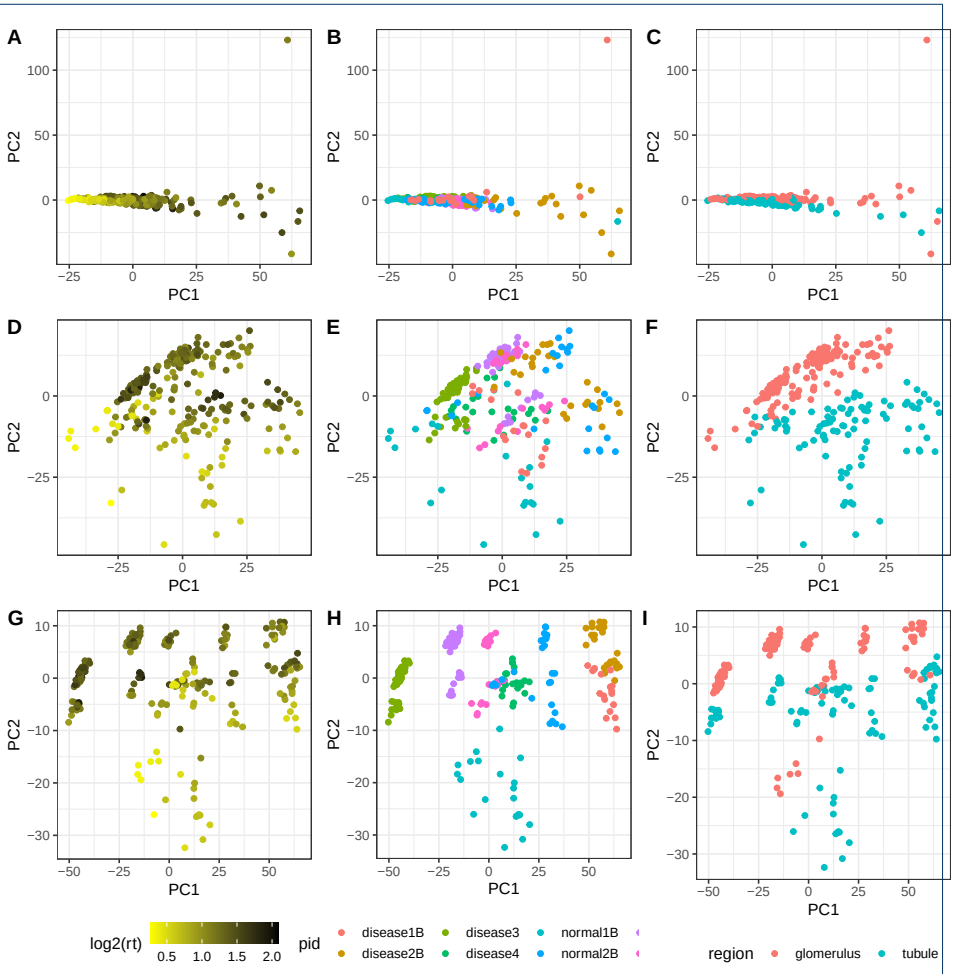


Figure 8: PCA plot for normalized expression of middle abundance genes

Figures of first two principle components of normalized expression of 75% quantile normalization(A, B, C), Poisson threshold normalization(D, E, F) and Poisson threshold normalization by grouping variable(G, H, I) on "middle" abundance genes, colored by the $\log_2$ ratio of 75% quantile and $\hat{\alpha}$ representing a key technical variation (A, D, G), slides ID (B, E, H) and ROI regions (C, F, I).

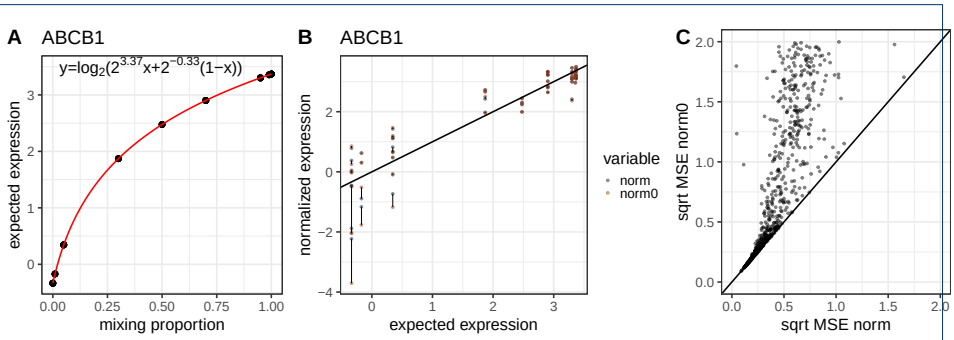


Figure 9: The relation between true $\log_2$ expression and HEK293T percentage

Using gene ABCB1 as an example(A, B), taking its estimated $\log_2$ expression of HEK293T and CCRF-CEM as true, the expected $\log_2$ expression for any mixture of the two with the percentage of HEK293T x follows a functional form as in A, and B shows the distance of $\log_2$ expression from Poisson threshold normalization without and with EB prior and the calculated expected $\log_2$ expression, in which Poisson threshold normalization with EB prior gives results closer to the expected values for ABCB1; C shows for all genes, the square root of MSE of estimating the $\log_2$ expression of HEK293T and CCRF-CEM using the Poisson threshold normalization without and with EB prior.

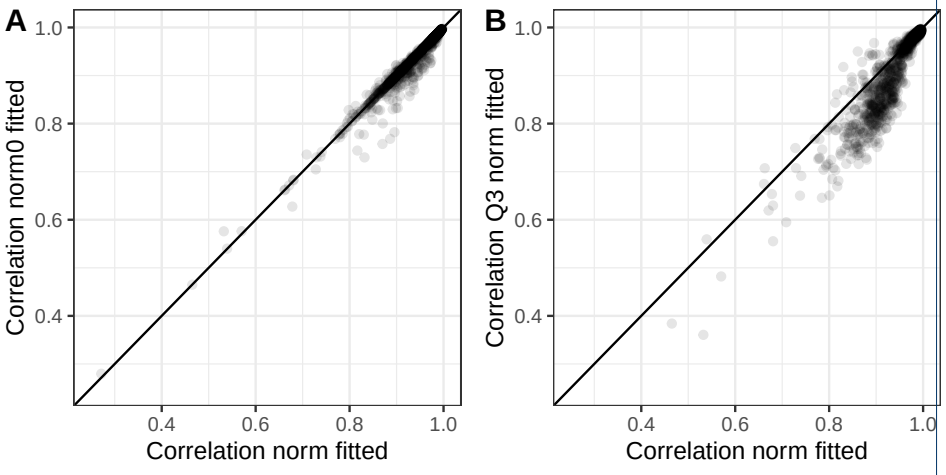
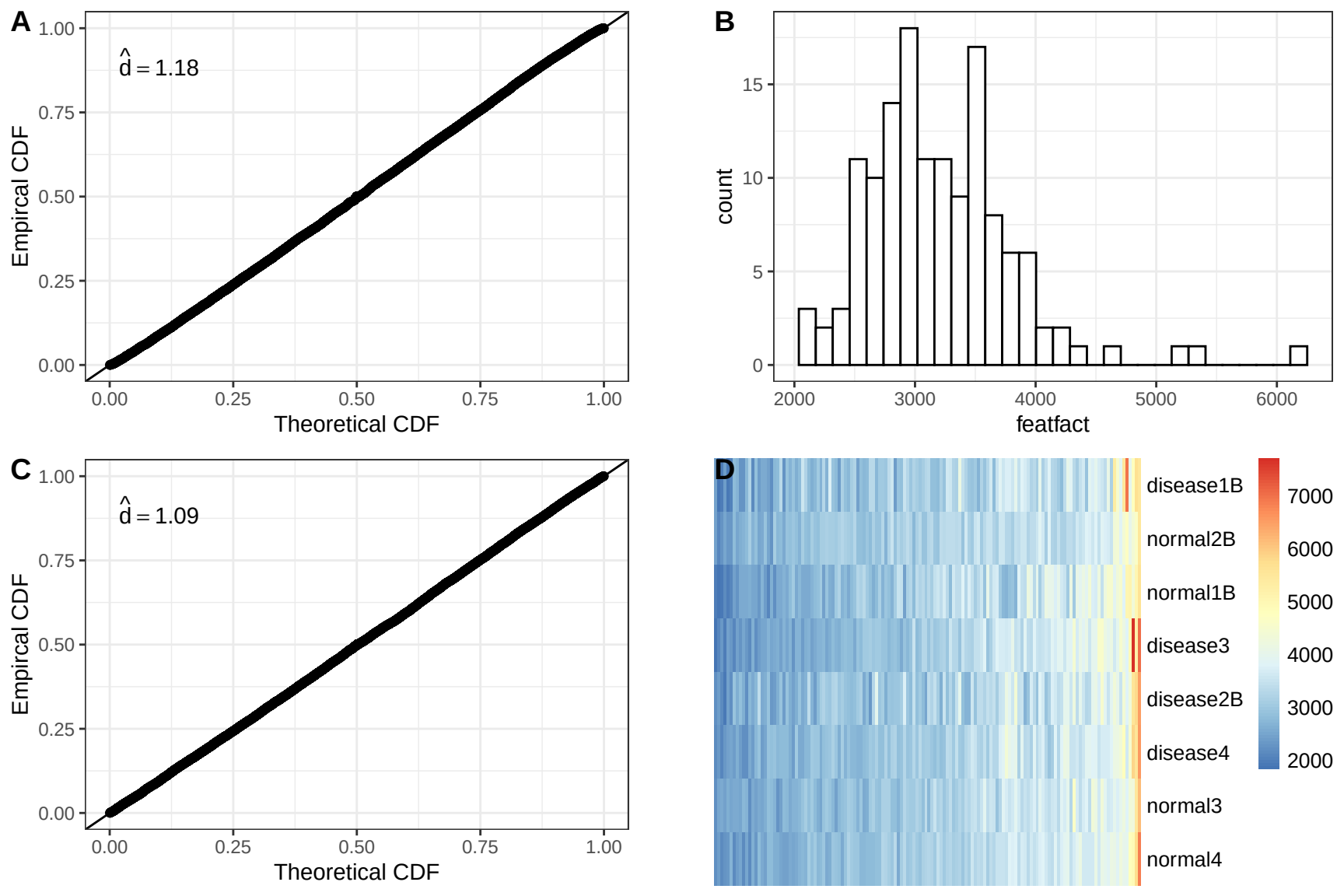
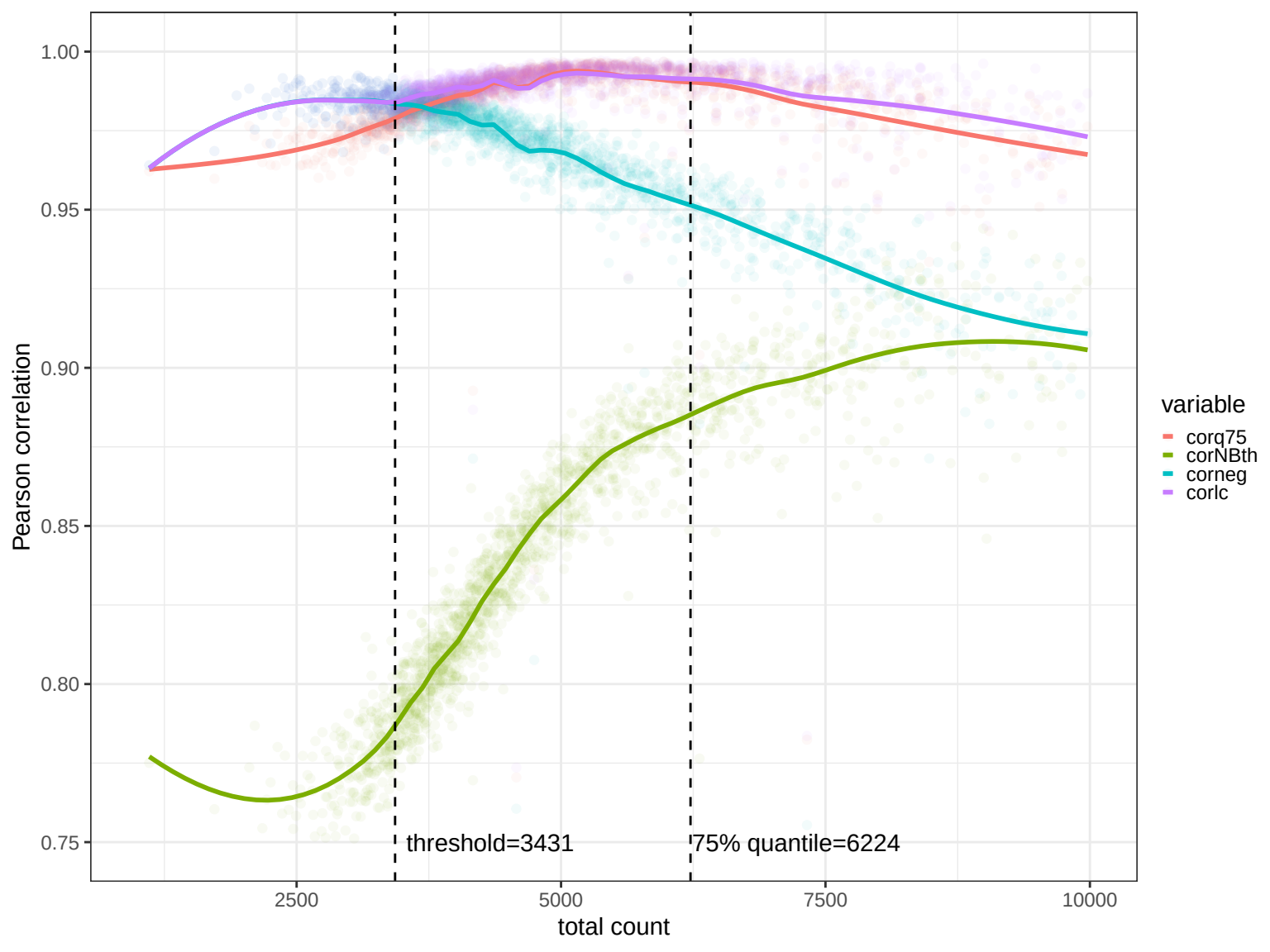
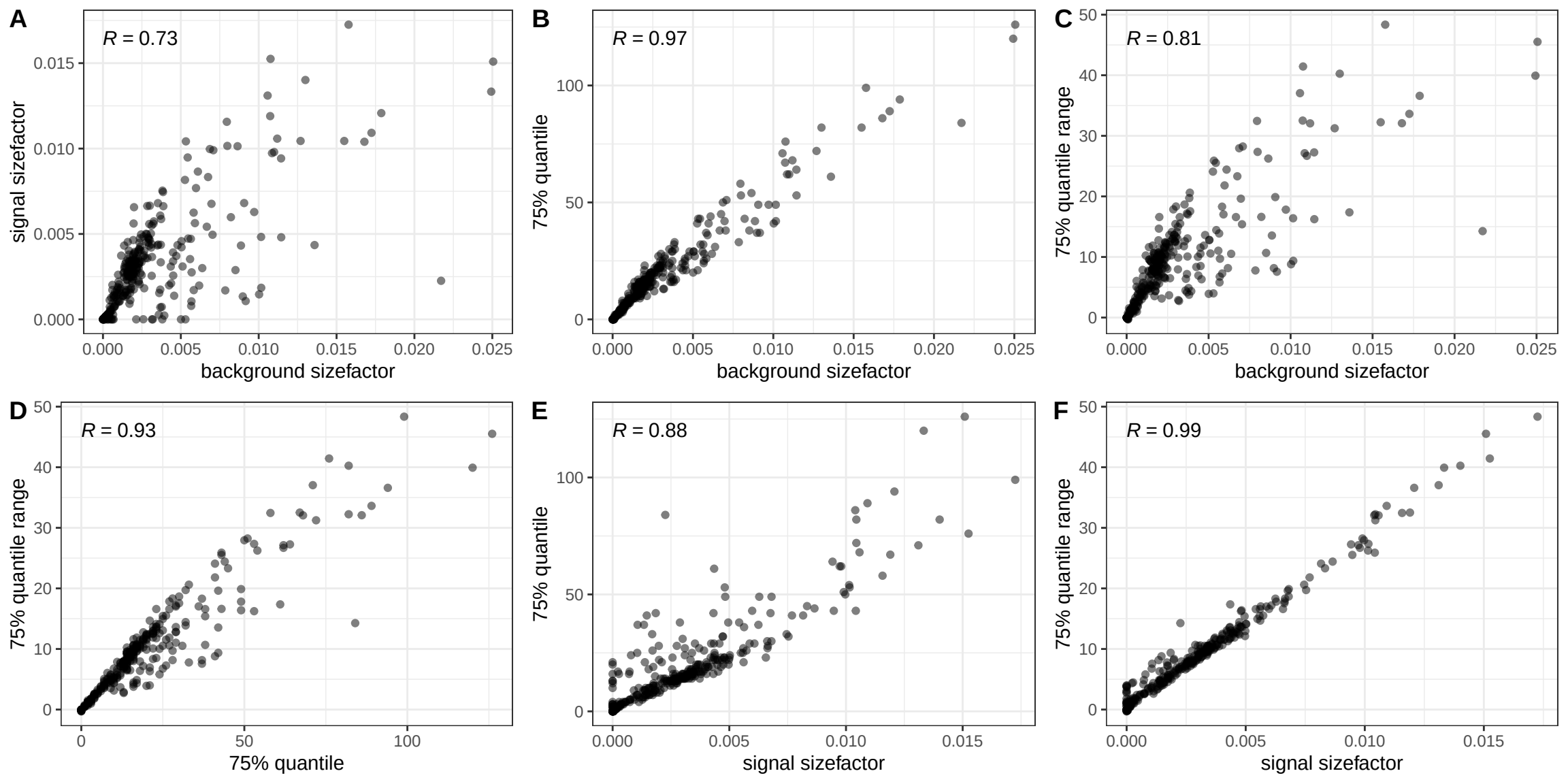


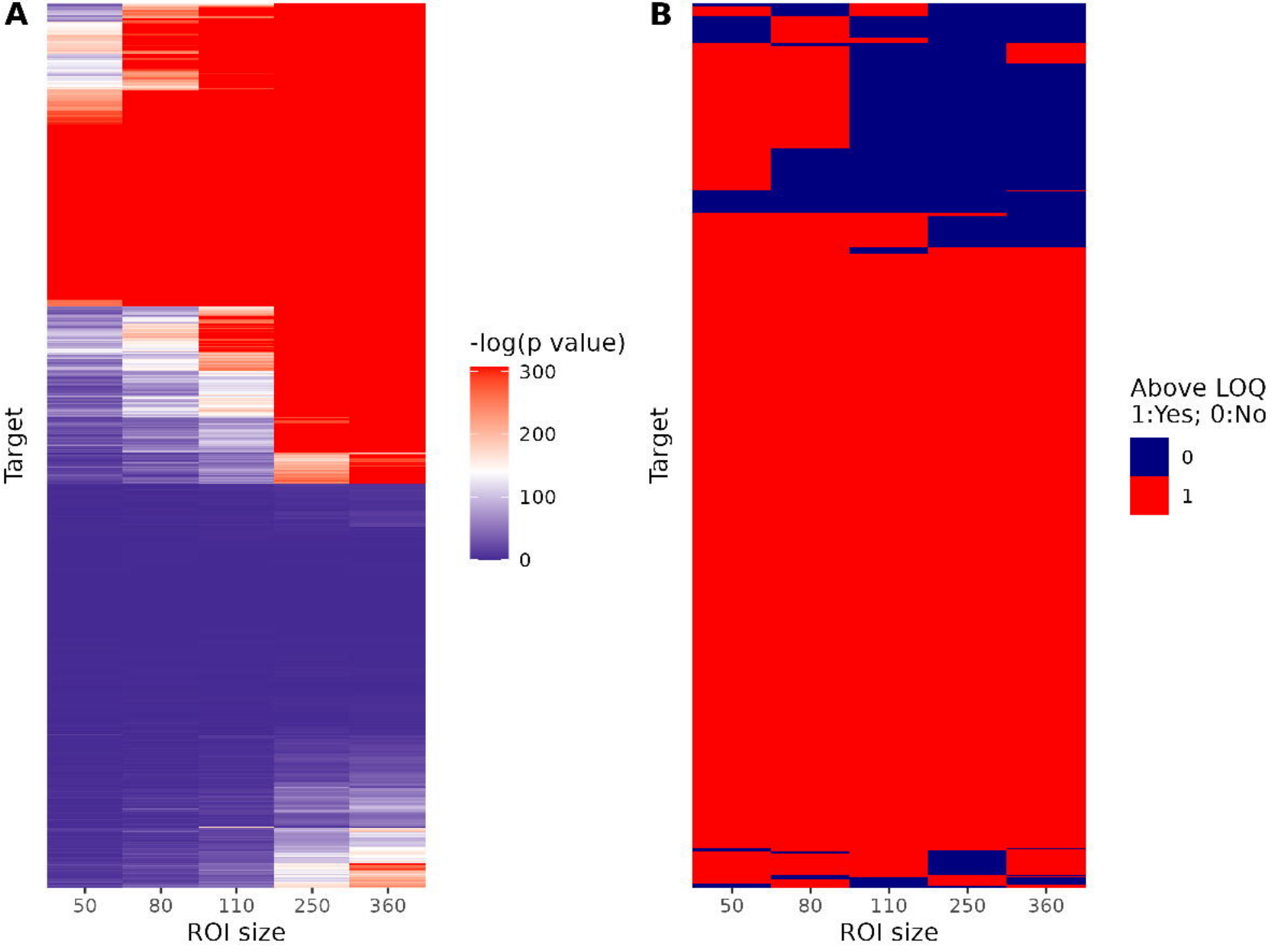
Figure 10: Correlation of expected expression vs the raw counts

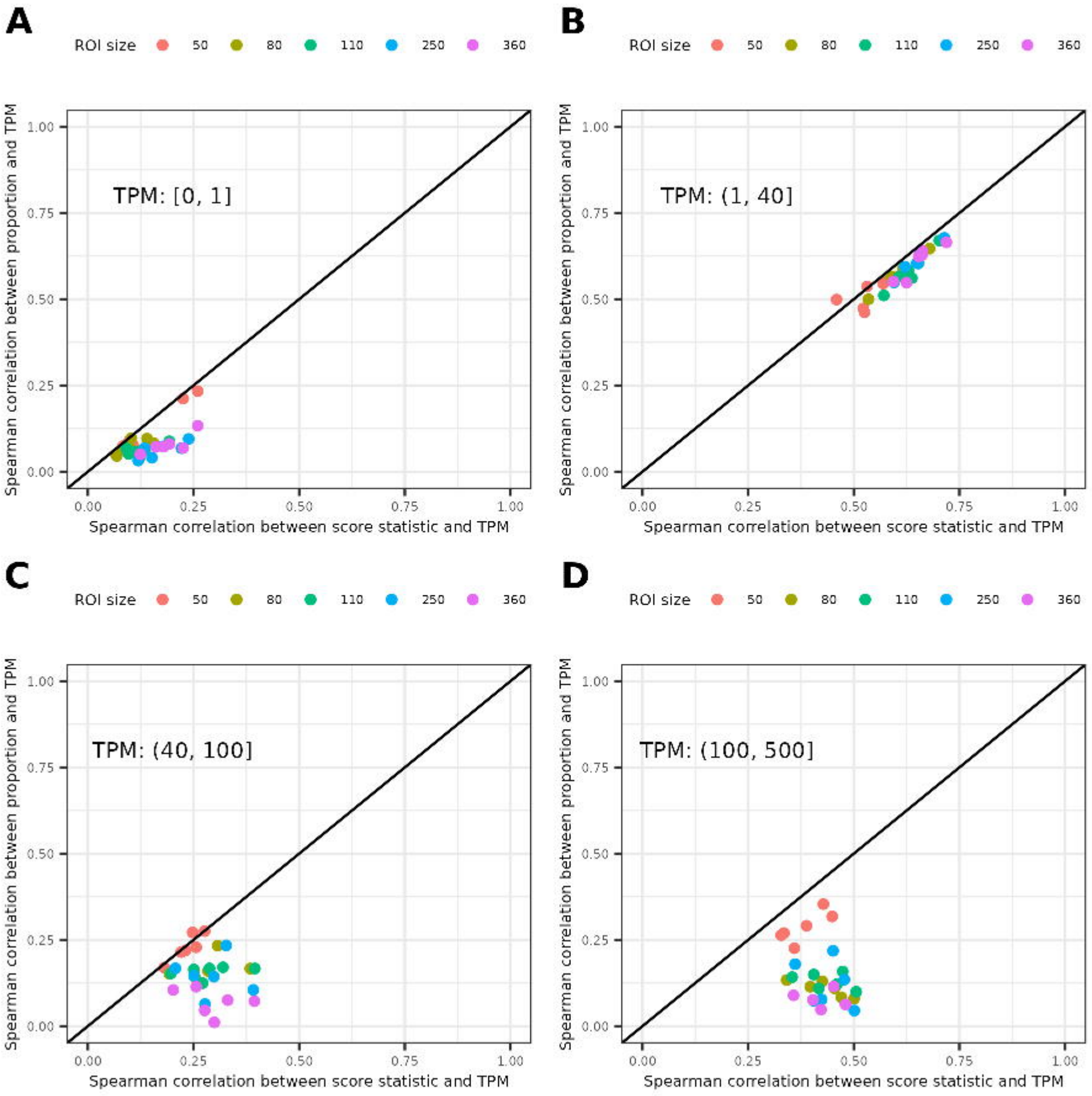
A: the correlation of expected raw count from Poisson threshold normalization vs raw count, the X axis is with EB prior and the Y axis is without EB prior. B, the correlation of expected raw count from Poisson threshold normalization with EB prior(X axis) vs raw count and 75% quantile(Y axis) vs raw count.

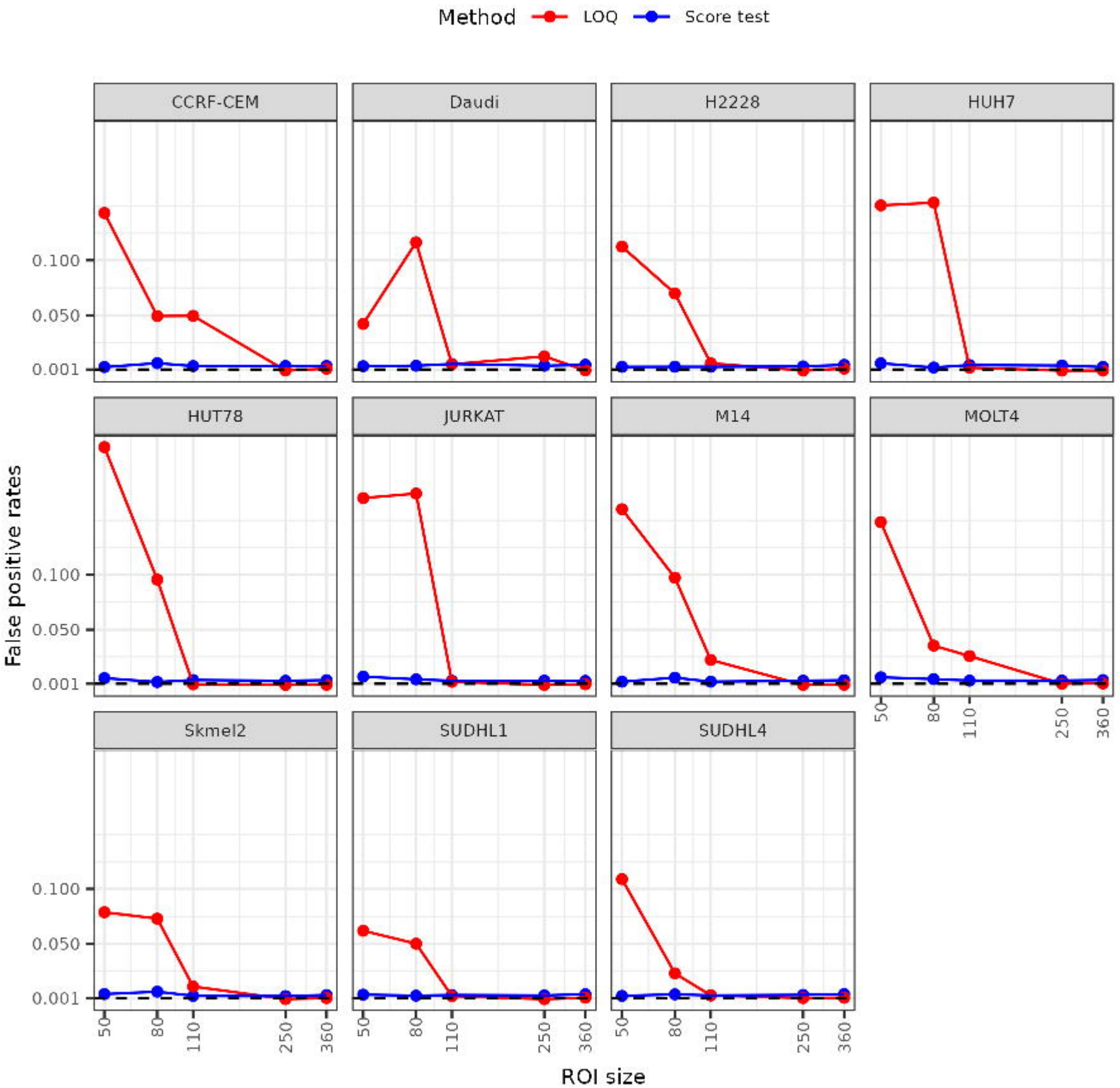

A
$\hat{d} = 1.18$
Empircal CDF
Theoretical CDF
0.00
0.25
0.50
0.75
1.00
B
count
featfact
0
5
10
15
2000
3000
4000
5000
6000
C
$\hat{d} = 1.09$
Empircal CDF
Theoretical CDF
D
disease1B
normal2B
normal1B
disease3
disease2B
disease4
normal3
normal4
7000
6000
5000
4000
3000
2000

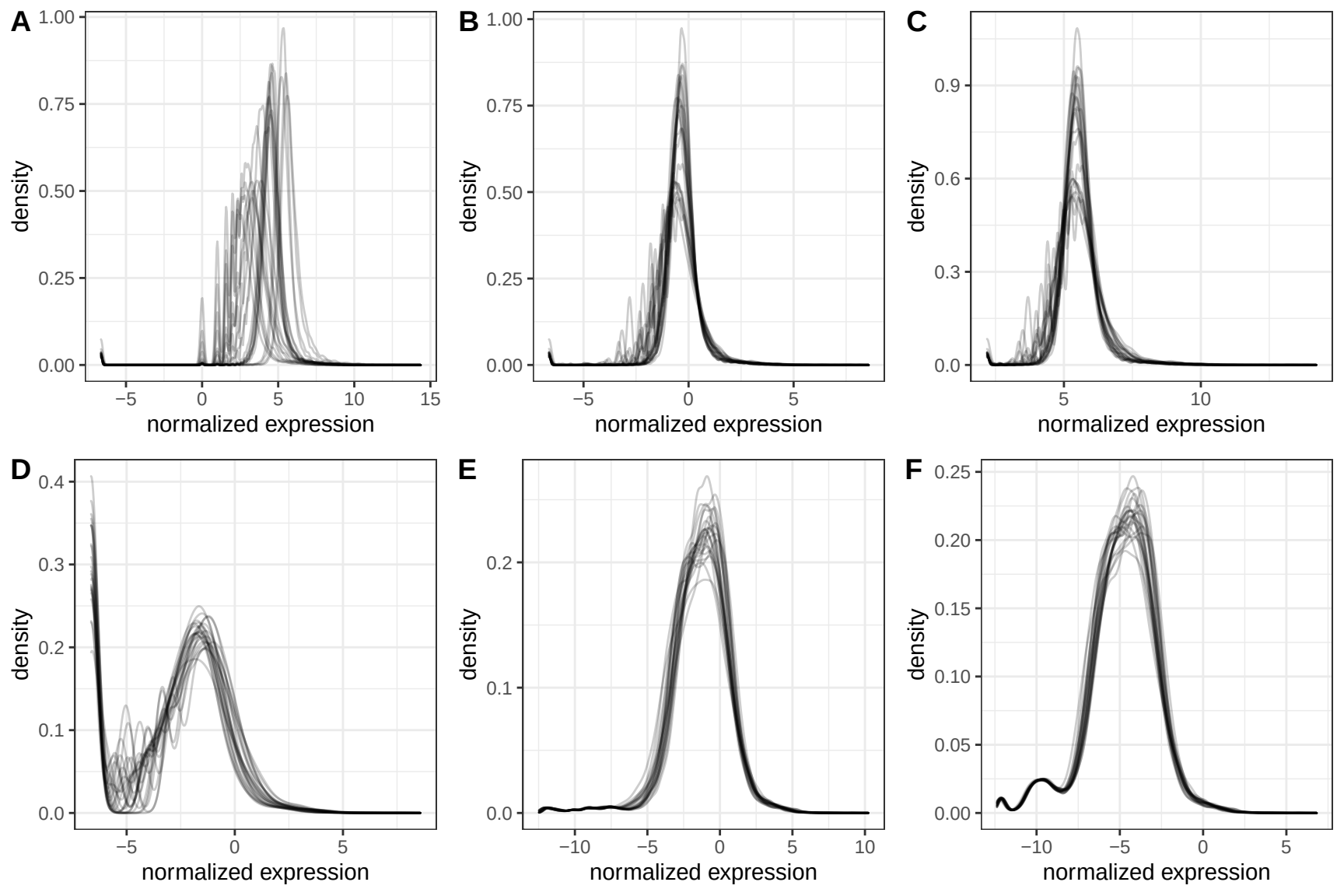

1.00
0.95
0.90
0.85
0.80
0.75
Pearson correlation
2500
5000
7500
10000
total count
threshold=3431
75% quantile=6224
variable
corq75
corNBth
corneg
corlc

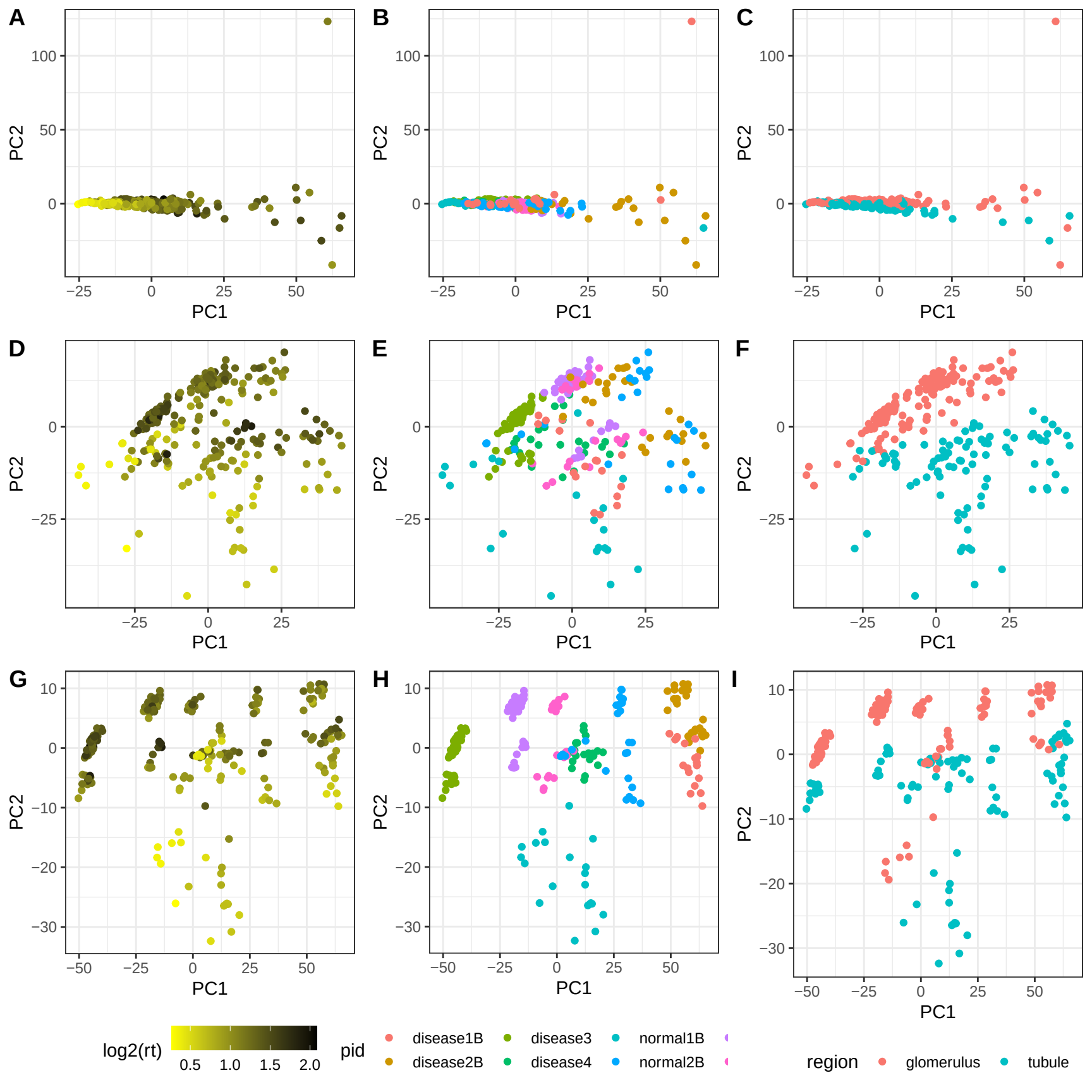

A
R = 0.73
signal sizefactor
background sizefactor
B
R = 0.97
75% quantile
background sizefactor
C
R = 0.81
75% quantile range
background sizefactor
D
R = 0.93
75% quantile range
75% quantile
E
R = 0.88
75% quantile
signal sizefactor
F
R = 0.99
75% quantile range
signal sizefactor

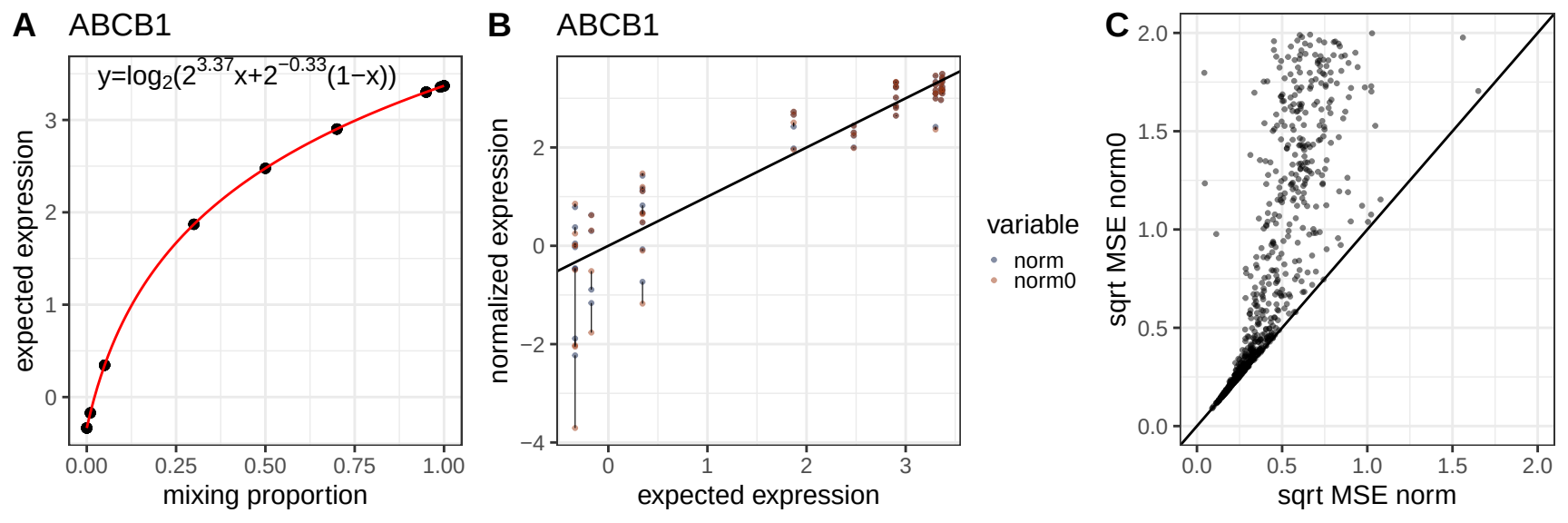

A
Target
-log(p value)
300
200
100
0
50
80
110
250
360
ROI size
B
Target
Above LOQ
1:Yes; 0:No
0
1
50
80
110
250
360
ROI size

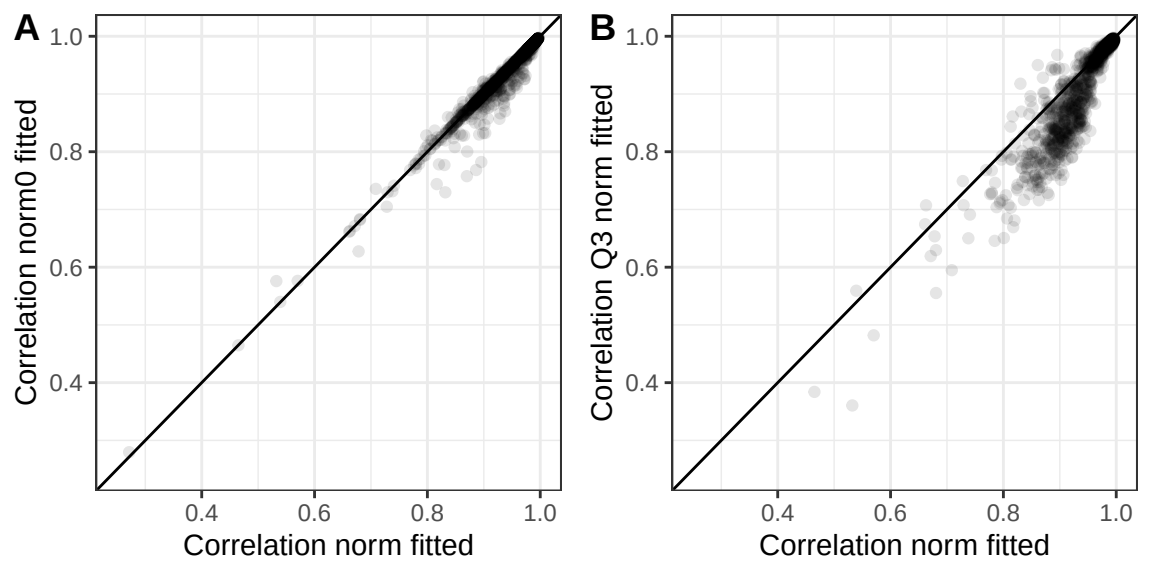

A
B
C
D
ROI size
50
80
110
250
360
TPM: [0, 1]
TPM: (1, 40]
TPM: (40, 100]
TPM: (100, 500]
Spearman correlation between proportion and TPM
Spearman correlation between score statistic and TPM
0.00
0.25
0.50
0.75
1.00


Method
LOQ
Score test
CCRF-CEM
Daudi
H2228
HUH7
HUT78
JURKAT
M14
MOLT4
Skmel2
SUDHL1
SUDHL4
False positive rates
0.100
0.050
0.001
50
80
110
250
360
ROI size


A
B
C
D
E
F
density
normalized expression


A
B
C
D
E
F
G
H
I
PC1
PC2
log2(rt)
0.5
1.0
1.5
2.0
pid
disease1B
disease2B
disease3
disease4
normal1B
normal2B
region
glomerulus
tubule


A
ABCB1
y=log2(2^3.37x+2^−0.33(1−x))
expected expression
mixing proportion
0.00
0.25
0.50
0.75
1.00
B
ABCB1
normalized expression
expected expression
variable
norm
norm0
C
sqrt MSE norm0
sqrt MSE norm


A
B
Correlation norm0 fitted
Correlation Q3 norm fitted
Correlation norm fitted
Correlation norm fitted
1.0
0.8
0.6
0.4
0.4
0.6
0.8
1.0