Statistical Modeling for MicroRNA Sequencing Data

Seong-Hwan Jun[1] and Matthew McCall[1]

[1]Department of Biostatistics and Computational Biology, University of Rochester Medical Center

Abstract

MicroRNAs play a central role in regulating gene expression and modulating diseases. Despite the importance of micro RNAs, statistical methods for analyzing them have received far less attention compared to messenger RNAs. In fact, it is common practice to apply the methods developed for messenger RNA-seq data to analyze micro RNA-seq data. This study critically examines and challenges the assumptions of messenger RNA-based methods when applied to micro RNAs, highlighting the competitive nature of micro RNA expression. We propose a statistical method and modeling strategies better suited for the micro RNA sequencing data and demonstrates their efficacy on simulated and real datasets.

MicroRNA-seq (miRNA-seq) data analysis often utilize methods developed for bulk messenger RNA-seq (mRNA-seq), such as DESeq2 and edgeR [1, 2]. A fundamental aspect of these methods is the approximate independence assumption, allowing each feature (miRNA) to be analyzed independently. This assumption holds particularly well when there is a large number of unique RNA molecules, and the sequencing reads are fairly evenly distributed across these molecules. Under these conditions, the complex multivariate distribution of RNA sequences can be effectively approximated by a collection of univariate distributions [3]. Pertaining to microRNAs, we postulate that there is non-negligible competition for expression among microRNA species, where abundance of one miRNA necessarily diminishes the abundance of another miRNA.

To contextualize the compositional nature of miRNA-seq relative to mRNA-seq, first note the total number of miRNA species discovered in human is ~ 500 as compiled in the MirGeneDB [4] and $\sim 2,000$ by miRBase [5], which is 10-40 times less compared to over $20,000$ genes in human [6]. Furthermore, the distribution of reads is often highly skewed towards a small number of highly expressed miRNAs. In examining the microRNAome dataset [7], we observed that < 50 miRNAs often take up 90% of the reads (Figure 1 **A**). Examining bulk mRNA-seq samples in the recount dataset [6], we found that $> 5,000$ mRNAs account for 90% of the total reads (Figure 1 **B**). The coverage for expressed miRNAs further deviates from uniform and is skewed towards an even smaller number of miRNAs as measured by Shannon diversity index (Figure 1 **C**), while coverage is fairly evenly spread across mRNAs (Figure 1 **D**). We also noted a disproportionate amount of reads are attributed to the most highly expressed miRNAs, accounting for $\approx 20\%$ of the total reads (Figure 1 **E**). In contrast, the most highly expressed mRNAs typically only account for 1-2% of the total reads in bulk mRNA-seq data (Figure 1 **F**). Finally, the Spearman's correlation between the two most highly expressed miRNAs was negative across a range of datasets while no such pattern was observed for mRNA (Figure 1 **G**-**H**).

To address competition for expression, we propose a negative binomial softmax regression (NBSR) model, extending independent Negative Binomial regression by relating the underlying parameters using the softmax function [8]. Specifically, the softmax regression models the normalized abundance or fraction of all miRNA sequences in sample i originating from feature j, denoted π_{ij} (Methods). The observed counts, Y_{ij} for feature j in sample i is modeled using a Negative Binomial distribution, with the mean given by the product of its scale factor s_i and the normalized abundance level π_{ij}.

We evaluated the NBSR on a simulated dataset where the approximate independence assumption is violated, comparing it with DESeq2 and EdgeR across two conditions (control vs. treatment) with $n = 10$ samples for each condition (Methods). We computed the root mean squared error, confidence interval (CI) coverage, and precision. We compared the statistical power of the methods using false positive and negative rates; the Benjamini-Hochberg procedure was applied to the p-values obtained from each method at FDR thresholds of $\alpha \in \{0.01, 0.05, 0.1\}$. NBSR exhibited RMSE < 1 across different simulation settings and

average precision > 0.9 across all FDR threholds considered (Figure 2 **A**-**B**). The desired nominal CI of 95% was attained (Supplemental Figure 5). NBSR is both stringent and sensitive as it achieved false positive rates of 0.00201, 0.00879, 0.0187 and false negative rates of $0.0229, 0.0154, 0.0125$ for each of the thresholds considered, demonstrating statistical power gained by modeling the relative abundances (Figure 2 **C**).

Additionally, we found that modeling the dispersion parameters of the NBSR in relation to the normalized abundances further improved performance. The interpretation for the dispersion parameter is provided in terms of the biological coefficient of variation (CV), which measures the variability in π_{ij} in relation to its expectation between biological replicates [9, 10]. The mRNA-seq methods typically assume one dispersion parameter for each feature shared across all experimental conditions; this assumption may be violated in situations where there is a significant change in the miRNA expression across experimental conditions, altering the underlying distribution of the normalized abundances. While the mRNA-seq methods model dispersion in relation to the mean expression level – effective for features with high expression – this can obscure true biological variability for miRNAs expressed at lower levels. Specifically, the mean expression level may be confounding the canonical abundance level of a feature and the sample-specific scale factor. By modeling the dispersion in terms of the normalized abundance level, our aim is to decouple the canonical notion of abundance represented by π_{ij}, from the scale factor. We derived the exact biological CV for the simulated data and compared it to the dispersion estimates obtained from the NBSR model, DESeq2, and EdgeR (Supplementary Figure 7 **B**). As DESeq2 and EdgeR assume feature-specific dispersion, these methods produced estimates that tend to fall between the true biological CVs for conditions A and B. On the other hand, NBSR aligns closely with the true biological CV.

We evaluated NBSR on Hepatocellular carcinoma cell lines dk01, dks8, dld1 ($n = 3$ for each cell line) compiled in the microRNAome dataset [7]. The log2 fold changes show substantial differences between NBSR and mRNA-seq methods (Figure 2 **D**): a handful of miRNAs with estimated absolute $\log_2$ fold change > 1 by mRNA-seq methods but < 0.1 by NBSR. In-depth examination of these miRNAs revealed that these are sparsely and hence lowly expressed miRNAs (Figure 2 **E**). This use case illustrates that NBSR is robust to sparsity, eliminating the need for a pre-processing step of selecting miRNAs to include in the analysis. This is advantageous since filtering for miRNAs is not as well understood as in the case of the mRNAs. All of the miRNAs exclusively selected as significant by NBSR were expressed in all samples and with the differences in expression clearly observable across the cell lines (Figure 2 **E** and Supplementary Figure 8). The standard practice for selecting differential features considers two criteria, (1) p-values (e.g., < 0.05) and $\log_2$ fold change exceeding a certain threshold. Using $\log_2$ fold change can be problematic when highly expressed features take up a large proportion of the read as a large fold increase is highly unlikely for these features. In the case of Hepatocellular carcinoma cell lines, the top feature accounts for 40% of the reads (Supplementary Figure 9 **A**). To alleviate this concern, we propose to consider the absolute difference (e.g., absolute change of probabilities exceeding > 0.0001). This provides an additional context on top of p-values and $\log_2$ fold change to identify meaningful changes in the miRNA species composition (Figure 2 **F**-**G**).

Overall, NBSR is shown to offer several desirable properties: (1) differential expression across experimental conditions has an intuitive interpretation as the change in probability of expression using either the ratio (log relative risk) or difference of probabilities; (2) it does not require filtering for miRNAs as sparsely expressed miRNAs are considered in relation to other miRNAs; (3) the relationship between the biological coefficient of variation and the normalized abundance level of the feature can be directly modelled, increasing statistical power while achieving a tighter confidence interval compared to mRNA-seq methods. There are two extensions to this model that can improve the performance: (i) considering positive correlation between miRNAs sharing the same precursor transcript and (ii) considering the isoformsof miRNAs called isomiRs. The NBSR and its associated methodologies introduced in this paper provide a new perspective for viewing and analyzing miRNAs. This method, implemented in Python, is available as open-source software for public use.

Methods

Negative Binomial Softmax Regression

Following the assumptions underlying the counts data, the standard models (e.g., DESeq2, edgeR) for observed counts for feature j for sample i is modelled using Negative Binomial distirbution,

$$Y_{ij} \sim \text{NegBinom}(\mu_{ij}, \phi_j). \tag{1}$$

The mean is given by $\mu_{ij} = s_i q_{ij}$ with $q_{ij} = \exp \boldsymbol{\beta}_j^T \boldsymbol{x}_i$ for $\boldsymbol{x}_i$ is a P-dimensional vector denoting the covariates for sample i, $\boldsymbol{\beta}_j$ the coefficient effects for j, and s_i denoting the estimated scale factor of sample i. The dispersion parameters ϕ_j accounts for overdispersion in the counts and it is also related to biological coeffcient of variation [9]. To account for competition for expression, we propose to model the mean by $\mu_{ij} = s_i \pi_{ij}$ where $\pi_{ij} \in [0, 1]$ such that $\sum_j \pi_{ij} = 1$. We model π_{ij} using softmax regression,

$$\pi_{ij} = \frac{\exp(\boldsymbol{\beta}_j^T \boldsymbol{x}_i)}{\sum_{j'} \exp(\boldsymbol{\beta}_{j'}^T \boldsymbol{x}_i)}. \tag{2}$$

While the scale factor estimation can be used, we simply use s_i to denote the total number of counts $s_i = \sum_j Y_{ij}$. We place Normal prior on β_j centered around 0 with standard error σ_p for each covariate $p = 1, ..., P$. The Normal prior works to regularize parameter estimation towards zero. The complete model is specified as follows,

$$Y_{ij} \sim \text{NegativeBinomial}(\mu_{ij}, \phi_j) \tag{3}$$

$$\mu_{ij} = s_i \times \pi_{ij} \tag{4}$$

$$\pi_{ij} = \frac{\exp(\beta_j^T x_i)}{\sum_{j'} \exp(\beta_{j'}^T x_i)} \tag{5}$$

$$\beta_{j,p} | \sigma_p^2 \sim \text{Normal}(0, \sigma_p^2) \tag{6}$$

$$\sigma_p^2 \sim \text{InvGamma}(a_p, b_p). \tag{7}$$

The parameters of inverse gamma are chosen by the user; we use the default values of $a_p = 3, b_p = 2$ in the experiments.

Inference on fold change

We are typically interested in comparing two experimental conditions, A and B. Without loss of generality, assume that A is the baseline level. We would like to test null hypothesis that the expression is the same across the two conditions against two-sided hypothesis that the expression is differential. One approach is to carry out hypothesis testing directly on β_j to determine the significance of feature j. An important feature of NBSR is that it is aimed at detecting signficant change in expression on feature j that arise indirectly as a result of changes to other features $j' \neq j$. To measure this indirect effect, we compare the conditional probability for condition B in relation to condition A, denoted $\pi_{j|B}$ and $\pi_{j|A}$ respectively. One possibility is to consider the log ratio,

$$\log \text{RR}_j = \log \frac{\pi_{j|B}}{\pi_{j|A}}. \tag{8}$$

This quantity has nice interpretation in statistics as log of *relative risk* as $\pi_{j|B}$ expresses the probability of an outcome for group B and similarly for $\pi_{j|A}$. We developed an inference methodology around the log relative risk in relation to the parameters of the negative binomial softmax regression. We derive the closed form for Hessian matrix of the NBSR model with respect to the model parameters β and use the

pseudo-inverse to obtain the Fisher information matrix. We found the pseudo-inverse to be fast and accurate even when the exact inverse may not be available. Statistical inference on the differential expression of each miRNA utilizes the log of the relative risk and full derivation of the standard error estimates of log2 fold change is provided in the Supplementary Information.

Derivation of biological CV for Dirichlet distribution

Assuming $\boldsymbol{\pi}_i \sim \text{Dirichlet}(\lambda_1, ..., \lambda_J)$ independent and identically distributed, the biological coefficient of variation is given by

$$\phi_j = \frac{\text{Var}(\pi_{ij})}{\mathbb{E}[\pi_{ij}]^2} = \frac{\bar{\lambda}_j(1-\bar{\lambda}_j)}{(\lambda_0+1)}\left(\frac{1}{\bar{\lambda}_j^2}\right), \tag{9}$$

where $\lambda_0 = \sum_{j'} \lambda_{j'}$ and $\bar{\lambda}_j = \lambda_j/\lambda_0$. We typically have experimental conditions that we want to compare. In such a case, the identical distribution assumption is no longer true and $\pi_{i|A}, \pi_{i|B}$ arise from Dirichlet distributions with parameters λ_j^A and λ_j^B. There is a unique biological coefficient of variation for condition A and B, ϕ_j^A, ϕ_j^B. We compute the biological coefficient of variation given in Equation 9 for conditions A and B in evaluating the dispersion model (Supplementary Figure 7 **B**).

Modeling dispersion

The mRNA-seq methods place a feature-specific dispersion, which is modelled as a function of mean expression level [2, 10, 11, 12]. In the original DESeq2 paper, the relationship between the dispersion is modelled using log-normal distribution centered around a trend line specified as a function of the mean of normalized read counts. EdgeR deploys a similar strategy by decomposing dispersion into three components, global level dispersion, trended dispersion, and tagwise (feature-specific) dispersion. These strategies have been adopted in the scRNA-seq domain with modifications, for example, using t-distribution and more flexible semi-parametric model (e.g., [11]). In all of these approaches, the dispersion is modelled in relation to the feature-specific mean expression levels μ_j or suitably normalized mean of the counts $\bar{\mu}_j$, both quantities are affected by scale factor of the samples. There are two potential limitations of modeling the dispersion in relation to mean expression levels. First, the mean expression level of feature j is a quantity that is proportional to the underlying fraction of feature j in sample i and the scale factor of sample i. In particular, as miRNA-seq is usually performed using small RNA sequencing (sRNA-seq) platform, there is an additional source of biological variability stemming from the total abundance of miRNAs in relation to otherl RNA molecules targetted by sRNA-seq in a given sample, as measured by percent miRNAs associated with each sample [13] (Supplementary Figure 7 **A**). This can be problematic for features that are lowly abundant as variation in scale factor across samples can affect the dispersion estimation. Second, by assuming feature-specific dispersion and taking average over the samples, we have lost potential difference in the distribution of the dispersion caused by the experimental condition and other covariates. Specifically, the coefficient of variation over Y_{ij} is decomposed into technical and biological [9]:

$$\text{CV}^2(Y_{ij}) = \frac{\text{Var}(Y_{ij})}{\mu_{ij}^2} \tag{10}$$

$$= \frac{1}{\mu_{ij}} + \phi_j, \tag{11}$$

where $1/\mu_{ij}$ is the coefficient of variation of Poisson distribution with mean μ_{ij} and ϕ_j denotes the biological variability over π_{ij} across different samples. This decomposition reveals an insight that when the mean expression is high, much of the variation is of biological nature while when the mean is low, the technical variability can dominate. Therefore, estimation of biological variability for highly expressed features is important and has received much attention; indeed, the current approaches implemented in popular packages perform very well for highly expressed features. On the other hand, estimating variability involving lowly expressed features in either of the experimental conditions is challenging. It is often the case that a large

coefficient of variation is estimated partially due to technical variability associated with sequencing depth and percent miRNAs captured; for example, sparsity is commonly observed among lowly expressed features resulting in larger than expected coefficient of variation. For these lowly expressed features, the normalized abundance π_{ij} may be more informative than the mean expression μ_{ij} because the true relationship between the dispersion and mean may be obscured by the size factor estimation. To address these problems, we consider modeling the dispersion in relation to π_{ij} and $R_i = \sum_j Y_{ij}$:

$$\mathbb{E}[\log \phi_{ij}] = b_{0j} + b_1 \log \pi_{ij} + b_2 \log R_i. \tag{12}$$

As π_{ij} represents canonical measure of the abundance, independent of the scale and R_i captures sample-specific variability due to randomness in the percent of miRNAs in the sample, we are able to decouple the effect of scale factor and the normalized measure of abundance.

One limitation of the above approach is that there are larger number of parameters to be estimated and therefore, we need sufficiently large number of samples. In our simulated study, with $n \in \{10, 20\}$, the above dispersion model successfully captured the biological CV (Supplementary Figures 7 **B**, 5, and 6). However, when the number of samples $n \in \{3, 5\}$, we observed convergence issues with the optimizer, likely due to insufficient number of data points leading to unidetifiable posterior distribution. For studies with small number of samples, we revert back to feature-specific dispersion assumption and to estimate dispersion using a multi-stage procedure as adopted in DESeq2 and EdgeR. Namely, we first estimate and fix the mean expression $\hat{\mu}_{ij}$ to obtain MLE $\hat{\phi}_j$. We then estimate the dispersion for each feature by regressing $\hat{\phi}_j$ on $\bar{\mu}_j = n^{-1} \sum_i \hat{\mu}_{ij}$ using local regression with Gaussian kernel [11]. While this approach does not resolve aforementioned problems, it alleviates convergence problems and permits estimation of the NBSR parameters and inference over the log relative risk, allowing to perform differential expression analysis on a feature in relation to other features. We observed that NBSR still reports the best result compared to DESeq2 and EdgeR (Supplementary Figures 3 and 4).

Simulated data generation

We view the counts data to arise from Dirichlet-Multinomial distribution or generalized Polya's urn experiment. Such model has been utilized to account for the randomness in the PCR amplification [14]. The Dirichlet model necessarily induces negative correlation between the features, which makes it well-suited for simulating competition for expression. The generative process for the data proceeds by first sampling the probability vectors $\boldsymbol{\pi}_i$ from a Dirichlet distribution with parameters $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J)$ followed by sampling of the counts given $\boldsymbol{\pi}_i$.

Estimation of Dirichlet parameters for microRNAome dataset

To simulate counts data resembling real datasets, we considered the cell type data originating from a single study (study id: 89) with at least 20 samples in the microRNAome datasets [7] (Supplementary Table 1). We filtered out samples with less than $100,000$ reads. We begin by obtaining maximum likelihood estimates of the parameters $\boldsymbol{\lambda}$ for each cell type data using the observed sample proportions $p_{ij} = Y_{ij} / \sum_{j'} Y_{ij'}$. The mean of the Dirichlet distribution given by $\bar{\lambda}_j = \lambda_j / \lambda_0$, where $\lambda_0 = \sum_j \lambda_k$ is referred to as the precision parameter. We estimate λ_j, λ_0 using fixed-point iteration described in [15]. The specific version of the estimation procedure that we implemented is presented in the Supplementary materials.

Simulating the counts

For each dataset, we selected miRNAs in the MirGeneDB [4] (596 miRNAs). We set the parameter values for the Dirichlet distribution representing the control as $\lambda_j^A = \lambda_j$. We classify the miRNAs for each datasets into two classes: $\lambda_j \leq 1$ (LOW) and $\lambda_j > 1$ (HIGH). We sample 40 miRNAs at random from each class and swap their parameters to form the ground truth parameters for the treatment condition B. The ground truth $\log_2$ fold change for miRNA j is given by $\log_2(\bar{\lambda}_j^B) - \log_2(\bar{\lambda}_j^A)$. To simulate the variability stemming from percent miRNAs, we sample $z_i \sim \text{Uniform}(0, 0.5)$ and sequencing depth $N_i \sim \text{Uniform}(10^6, 10^8)$ and set the reads mapping to miRNAs as $n_i = z_i N_i$. The read counts are then simulated from Multinomial distribution: $Y_i^{cond} \sim \text{Multinomial}(\boldsymbol{\lambda}^{cond}, n_i)$ for $cond \in \{A, B\}$.

Real data analysis

Filtering microRNAome dataset

We took a cell type subset of the microRNAome dataset with > 40 samples and eliminated samples with less than 100,000 reads. The resulting dataset was used in creating the Figure 1. The table of cell types and the number of samples are shown in Supplementary Table 2.

Filtering recount2 dataset

We removed samples from project DRP001797 and with average read length > 200. The cell type dataset with at least 10 samples were selected in the analysis (Figure 1). The table of cell types and the number of samples are shown in Supplementary Table 3.

Data availability

The miRNA-seq data used in the analysis is available from `microRNAome` package through Bioconductor `https://www.bioconductor.org/packages/3.16/data/experiment/html/microRNAome.html` v1.20.0. The mRNA-seq data used in the analysis is available from `recount` package through Bioconductor `https://bioconductor.org/packages/release/bioc/html/recount.html` accessed on 01/06/2020.

Code availability

We implemented NBSR model using PyTorch and utilize automatic gradient computation to optimize the model parameters [16]. The code is available on Github `https://github.com/junseonghwan/nbsr/`. The analysis code is available on `https://github.com/junseonghwan/nbsr/`

References

[1] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010.

[2] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550, 2014.

[3] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.*, 20(1):295, December 2019.

[4] Bastian Fromm, Eirik Høye, Diana Domanska, Xiangfu Zhong, Ernesto Aparicio-Puerta, Vladimir Ovchinnikov, Sinan U Umu, Peter J Chabot, Wenjing Kang, Morteza Aslanzadeh, Marcel Tarbier, Emilio Mármol-Sánchez, Gianvito Urgese, Morten Johansen, Eivind Hovig, Michael Hackenberg, Marc R Friedländer, and Kevin J Peterson. MirGeneDB 2.1: toward a complete sampling of all major animal phyla. *Nucleic Acids Res.*, 50(D1):D204–D210, January 2022.

[5] Sam Griffiths-Jones, Russell J Grocock, Stijn van Dongen, Alex Bateman, and Anton J Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, 34(Database issue):D140–4, January 2006.

[6] Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.*, 35(4):319–321, April 2017.

[7] Matthew N McCall, Min-Sik Kim, Mohammed Adil, Arun H Patil, Yin Lu, Christopher J Mitchell, Pamela Leal-Rojas, Jinchong Xu, Manoj Kumar, Valina L Dawson, Ted M Dawson, Alexander S Baras, Avi Z Rosenberg, Dan E Arking, Kathleen H Burns, Akhilesh Pandey, and Marc K Halushka. Toward the human cellular microRNAome. *Genome Res.*, 27(10):1769–1781, October 2017.

[8] Kieran R Campbell, Adi Steif, Emma Laks, Hans Zahn, Daniel Lai, Andrew McPherson, Hossein Farahani, Farhia Kabeer, Ciara O'Flanagan, Justina Biele, Jazmine Brimhall, Beixi Wang, Pascale Walters, IMAXT Consortium, Alexandre Bouchard-Côté, Samuel Aparicio, and Sohrab P Shah. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biol.*, 20(1):54, March 2019.

[9] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, 40(10):4288–4297, May 2012.

[10] Yunshun Chen, Aaron T L Lun, and Gordon K Smyth. Differential expression analysis of complex RNA-seq experiments using edger. In Somnath Datta and Dan Nettleton, editors, *Statistical Analysis of Next Generation Sequencing Data*, pages 51–74. Springer International Publishing, Cham, 2014.

[11] Nils Eling, Arianne C Richard, Sylvia Richardson, John C Marioni, and Catalina A Vallejos. Correcting the Mean-Variance dependency for differential variability testing using Single-Cell RNA sequencing data. *Cell Syst*, 7(3):284–294.e12, September 2018.

[12] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.*, 9(1):284, January 2018.

[13] Yin Lu, Alexander S Baras, and Marc K Halushka. mirge 2.0 for comprehensive analysis of microRNA sequencing data. *BMC Bioinformatics*, 19(1):275, July 2018.

[14] Hazal Koptagel, Seong-Hwan Jun, Joanna Hård, and Jens Lagergren. Scuphr: A probabilistic framework for cell lineage tree reconstruction. December 2022.

[15] Thomas Minka. Estimating a Dirichlet distribution, 2000.

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Supplementary materials

Laplace's approximation

We derive standard error estimates via Laplace's appxorimation of the posterior distribution over the covariate coefficients $\beta_{j,d}$'s, evaluated at the MAP estimate given by, Normal($\boldsymbol{\beta}_{MAP}, \Sigma$). Typically, the covariance matrix is estimated by the inverse of the observed information, which is given by the negative of the Hessian of the log posterior distribution evaluated at the MAP estiamte.

Gradients for NBSR

We begin by deriving the gradient of the log posterior distribution, which is given by

$$\log p(\beta|Y) \propto \log p(Y|\beta) + \log p(\beta).$$

We derive the first order gradient of the log likelihood of the NBSR model with respect to $\beta_{k,d}$, which we denote $\nabla_{k,d} \log p(Y|\beta)$.

$$\nabla_{k,d} \log p(Y|\beta) = \sum_i \sum_j \nabla_{k,d} \log p(y_{i,j}|\beta). \tag{13}$$

Note that $p(y_{i,j}|\beta)$ is given by Negative Binomial distribution with parameters $p_{i,j} = \mu_{i,j}/\sigma^2_{i,j}$ and $r_j = 1/\phi_j$:

$$\log p(y_{i,j}|\beta) = \log C_{i,j} + r_j \log p_{i,j} + y_{i,j} \log(1 - p_{i,j}), \tag{14}$$

where $C_{i,j}$ is a constant with respect to β. We will further simplify the expression by plugging-in $p_{i,j} = \mu_{i,j}/\sigma^2_{i,j}$:

$$\log p(y_{i,j}|\beta) = r_j(\log \mu_{i,j} - \log \sigma^2_{i,j}) + y_{i,j}(\log(\mu^2_{i,j}) - \log \sigma^2_{i,j}). \tag{15}$$

Since $\sigma^2_{i,j} = \mu_{i,j} + \phi_j \mu^2_{i,j}$, computing the gradient amounts to computing the gradient of $\mu_{i,j}$ w.r.t $\beta_{k,p}$:

$$\begin{aligned} \nabla_{k,d}\mu_{i,j} &= s_i \nabla_{k,d}\pi_{i,j} && (16)\\ &= x_{i,d}\mu_{i,j}(1[j = k] - \pi_{i,k}), && (17) \end{aligned}$$

where the gradient for $\pi_{i,j}$ is given by,

$$\begin{aligned} \nabla_{k,d}\pi_{i,j} &= \nabla_{k,d} \frac{\exp(\beta_j^T x_i)}{\sum_{j'} \exp(\beta_{j'}^T x_i)} && (18)\\ &= \begin{cases} x_{i,d}\pi_{i,j}(1 - \pi_{i,k}) & \text{if } j = k \\ -x_{i,d}\pi_{i,j}\pi_{i,k} & \text{if } j \neq k \end{cases} && (19)\\ &= x_{i,d}\pi_{i,j}(1[j = k] - \pi_{i,k}). && (20) \end{aligned}$$

The gradient of $\mu^2_{i,j}$ can be derived by applying the chain rule:

$$\begin{aligned} \nabla_{k,d}\mu^2_{i,j} &= 2\mu_{i,j}\nabla_{k,d}\mu_{i,j} && (21)\\ &= 2x_{i,d}\mu^2_{i,j}(1[j = k] - \pi_{i,k}). && (22) \end{aligned}$$

Finally, the gradient of $\sigma^2_{i,j}$:

$$\begin{aligned} \nabla_{k,d}\sigma^2_{i,j} &= \nabla_{k,d}\mu_{i,j} + \phi_j \nabla_{k,d}\mu^2_{i,j} && (23)\\ &= x_{i,d}\mu_{i,j}(1[j = k] - \pi_{i,k})(1 + 2\phi_j\mu_{i,j}). && (24) \end{aligned}$$

Putting all of this together, we have the following expression for the gradient of log likelihood:

$$\begin{aligned} \nabla_{k,d} \log p(y_{i,j}|\beta) =& r_j x_{i,d}(1[j = k] - \pi_{i,k}) \left(1 - \frac{\mu_{i,j}(1 + 2\phi_j\mu_{i,j})}{\sigma^2_{i,j}}\right) && (25)\\ &+ y_{i,j}x_{i,d}(1[j = k] - \pi_{i,k}) \left(2 - \frac{\mu_{i,j}(1 + 2\phi_j\mu_{i,j})}{\sigma^2_{i,j}}\right) && (26) \end{aligned}$$

The gradient of $\log p(\beta)$ is,

$$\begin{aligned}\nabla_{k,d} \log p(\beta) &= \nabla_{k,d} \log p(\beta_{k,d}) \\ &= -\frac{\beta_{k,d}}{\sigma_d^2}.\end{aligned}$$

Note that

$$\mu_{i,j}(1 + 2\phi_j \mu_{i,j}) = \sigma_{i,j}^2 + \phi_j \mu_{i,j}^2,$$

allowing us to simplify Eq 25 further:

$$\nabla_{k,d} \log p(y_{i,j}|\beta) = r_j x_{i,d}(1[j = k] - \pi_{i,k}) \left(-\frac{\phi_j \mu_{i,j}^2}{\sigma_{i,j}^2} \right) \tag{27}$$

$$+ y_{i,j} x_{i,d}(1[j = k] - \pi_{i,k}) \left(1 - \frac{\phi_j \mu_{i,j}^2}{\sigma_{i,j}^2} \right) \tag{28}$$

Hence, the gradient of the log posterior distribution is given by,

$$\nabla_{k,d} \log p(\beta|Y) = \sum_i \sum_j -x_{i,d} r_j (1[j = k] - \pi_{i,k}) D_{i,j} + y_{i,j} x_{i,d}(1[j = k] - \pi_{i,k})(1 - D_{i,j}) - \frac{\beta_{k,d}}{\sigma_d^2}, \tag{29}$$

where $D_{i,j} = \frac{\phi_j \mu_{i,j}^2}{\sigma_{i,j}^2}$.

We now derive the second derivative of the log posterior.

$$\begin{aligned}\nabla_{k',d'} \nabla_{k,d} \log p(y_{i,j}|\beta) = &- x_{i,d} r_j \left[-\nabla_{k',d'} \pi_{i,k} D_{i,j} + (1[j = k] - \pi_{i,k}) \nabla_{k',d'} D_{i,j} \right] \\ &+ x_{i,d} y_{i,j} \left[-\nabla_{k',d'} \pi_{i,k}(1 - D_{i,j}) + (1[j = k] - \pi_{i,k})(-\nabla_{k',d'} D_{i,j}) \right].\end{aligned}$$

This amounts to computing the gradient of $D_{i,j}$ as we have derived the gradient for $\pi_{i,j}$ in Eq (18). The gradient of $D_{i,j}$ can be obtained by using a trick. For any function f

$$f(x) \nabla_x \log f(x) = \nabla_x f(x), \tag{30}$$

combined with the gradient of $\log \mu_{i,j}^2 / \sigma_{i,j}^2$ we have in Eq (15), the gradient can be obtained as,

$$\nabla_{k,d} D_{i,j} = x_{i,d}(1[j = k] - \pi_{i,k}) D_{i,j}(1 - D_{i,j}). \tag{31}$$

The Hessian for the log likelihood is given by,

$$H_{loglik} = \nabla_{k',d'} \nabla_{k,d} \log p(y_{i,j}|\beta) = r_j D_{i,j} x_{i,d} \pi_{i,k} x_{i,d'} (1[k = k'] - \pi_{i,k'}) \tag{32}$$

$$- y_{i,j}(1 - D_{i,j}) x_{i,d} \pi_{i,k} x_{i,d'} (1[k = k'] - \pi_{i,k'}) \tag{33}$$

$$- r_j D_{i,j}(1 - D_{i,j}) x_{i,d}(1[j = k] - \pi_{i,k}) x_{i,d'} (1[j = k'] - \pi_{i,k'}) \tag{34}$$

$$- y_{i,j} D_{i,j}(1 - D_{i,j}) x_{i,d}(1[j = k] - \pi_{i,k}) x_{i,d'} (1[j = k'] - \pi_{i,k'}). \tag{35}$$

And the Hessian for the log posterior is given by,

$$H = H_{loglik} - \frac{1}{\sigma_d^2} I_{JP \times JP}. \tag{36}$$

Finally, we set the covariance matrix as,

$$\Sigma = (-H)^{-1}. \tag{37}$$

Gradients for NBSR with dispersion

In an alternative modeling approach, we propose to model the dispersion parameters in relation to the underlying abundance level $\pi_{i,j}$. The gradients flows back to beta from the loss function (log posterior) through the dispersion parameters ϕ and mean parameters μ via abundance levels π. The gradient for $\log p(\beta)$ is unchanged so we show here the gradient for the log likelihood. Note that under this model, we have dispersion for each sample-feature, ϕ_{ij} and hence, Negative Binomial distribution parameter $r_{ij} = 1/\phi_{ij}$:

$$\log p(y_{i,j}|\beta) = r_{i,j} \log p_{i,j} + y_{ij} \log(1 - p_{ij}) + \tag{38}$$

$$\log \Gamma(y_{ij} + r_{ij}) - \log \Gamma(r_{ij}) - \log \Gamma(y_{ij}), \tag{39}$$

The gradient is given by:

$$\nabla_{k,d} \log p(y_{i,j}|\beta) = r_{i,j} \nabla_{k,d} \log p_{i,j} + y_{ij} \nabla_{k,d} \log(1 - p_{ij}) + \tag{40}$$

$$\nabla_{k,d} r_{i,j} (\psi(y_{ij} + r_{ij}) - \psi(r_{ij}) + \log p_{i,j}), \tag{41}$$

where ψ denotes the digamma function. First, we derive $\nabla_{k,d} r_{ij}$.

$$\nabla_{k,d} r_{ij} = -\frac{\nabla_{k,d} \phi_{ij}}{\phi_{ij}^2}, \tag{42}$$

where $\phi_{ij} = \exp(b_0 + b_1 \log \pi_{ij} + w' z_i)$.

$$\nabla_{k,d} \phi_{ij} = \phi_{ij} \left(b_1 \frac{\nabla_{k,d} \pi_{ij}}{\pi_{ij}} \right) \tag{43}$$

$$= \phi_{ij} b_1 x_{i,d} (1[j = k] - \pi_{ik}). \tag{44}$$

Plug-in above, we get,

$$\nabla_{k,d} r_{ij} = -\frac{b_1 x_{i,d} (1[j = k] - \pi_{ik})}{\phi_{ij}}. \tag{45}$$

Therefore, Eq 41 is given by,

$$\frac{-b_1 x_{i,d} (1[j = k] - \pi_{ik})}{\phi_{ij}} (\psi(y_{ij} + r_{ij}) - \psi(r_{ij}) + \log p_{i,j}). \tag{46}$$

To derive Eq 40, we first express Eq 38 (note: $1 - p_{ij} = \phi_{ij} \mu_{ij}^2 / \sigma_{ij}^2$):

$$r_{i,j} \log p_{i,j} + y_{ij} \log(1 - p_{ij}) = r_{i,j} (\log \mu_{ij} - \log \sigma_{ij}^2) + y_{ij} (\log \phi_{ij} + \log \mu_{ij}^2 - \log \sigma_{ij}^2). \tag{47}$$

Now the gradient of σ_{ij}^2,

$$\nabla_{k,d} \sigma_{i,j}^2 = \nabla_{k,d} \mu_{i,j} + \phi_{ij} \nabla_{k,d} \mu_{i,j}^2 + \mu_{i,j}^2 \nabla_{k,d} \phi_{ij} \tag{48}$$

$$= x_{i,d} (1[j = k] - \pi_{i,k}) (\mu_{i,j} + 2\phi_{ij} \mu_{i,j}^2 + b_1 \phi_{ij} \mu_{ij}^2) \tag{49}$$

$$= x_{i,d} (1[j = k] - \pi_{i,k}) (\sigma_{i,j}^2 + \phi_{ij} \mu_{i,j}^2 (1 + b_1)) \tag{50}$$

$$= x_{i,d} (1[j = k] - \pi_{i,k}) (\sigma_{i,j}^2 + (\sigma_{ij}^2 - \mu_{i,j})(1 + b_1)). \tag{51}$$

Hence, the gradient of $\log \sigma_{ij}^2$ is given by,

$$\nabla_{k,d} \log \sigma_{i,j}^2 = x_{i,d} (1[j = k] - \pi_{i,k}) (1 + (1 - p_{ij})(1 + b_1)). \tag{52}$$

The gradient of Eq 40 is given by,

$$r_{ij} x_{id} (1[j = k] - \pi_{ik}) (1 - p_{ij})(1 + b_1) + y_{ij} x_{id} (1[j = k] - \pi_{ik}) (p_{ij}(1 + b_1)). \tag{53}$$

$$
\begin{aligned}
r_{ij}\nabla_{k,d}\log p_{i,j} &= r_{ij}(\nabla_{k,d}\log\mu_{ij} - \nabla_{k,d}\log\sigma^2_{ij}) && (54)\\
&= r_{ij}\left(\frac{\nabla_{k,d}\mu_{ij}}{\mu_{ij}} - \frac{\nabla_{k,d}\sigma^2_{ij}}{\sigma^2_{ij}}\right) && (55)\\
&= r_{ij}\left(\frac{x_{id}\mu_{ij}(1[j=k]-\pi_{ik})}{\mu_{ij}} - \frac{\nabla_{k,d}\sigma^2_{ij}}{\sigma^2_{ij}}\right) && (56)\\
&= r_{ij}\left(x_{id}(1[j=k]-\pi_{ik}) - \frac{x_{id}(1[j=k]-\pi_{ik})(\sigma^2_{ij}+\phi_{ij}\mu^2_{ij}(1+b_1))}{\sigma^2_{ij}}\right) && (57)\\
&= r_{ij}\left(x_{id}(1[j=k]-\pi_{ik}) - x_{id}(1[j=k]-\pi_{ik})(1+(1-p_{ij})(1+b_1))\right) && (58)\\
&= r_{ij}x_{id}(1[j=k]-\pi_{ik})(1-(1+(1-p_{ij})(1+b_1))) && (59)\\
&= -r_{ij}x_{id}(1[j=k]-\pi_{ik})(1-p_{ij})(1+b_1) && (60)
\end{aligned}
$$

$$
\begin{aligned}
y_{ij}\nabla_{k,d}\log(1-p_{i,j}) &= y_{ij}(\nabla_{k,d}\log\phi_{ij} + \nabla_{k,d}\log\mu^2_{ij} - \nabla_{k,d}\log\sigma^2_{ij}) && (61)\\
&= y_{ij}\left(\frac{\nabla_{k,d}\phi_{ij}}{\phi_{ij}} + \frac{\nabla_{k,d}\mu_{ij}}{\mu^2_{ij}} - \frac{\nabla_{k,d}\sigma^2_{ij}}{\sigma^2_{ij}}\right) && (62)\\
&= y_{ij}\left(\frac{\phi_{ij}b_1x_{id}(1[j=k]-\pi_{ik})}{\phi_{ij}} + \frac{2x_{id}\mu^2_{ij}(1[j=k]-\pi_{ik})}{\mu^2_{ij}} - \frac{x_{id}(1[j=k]-\pi_{ik})(\sigma^2_{ij}+\phi_{ij}\mu^2_{ij}(1+b_1))}{\sigma^2_{ij}}\right) && (63)\\
&= y_{ij}\left(b_1x_{id}(1[j=k]-\pi_{ik}) + 2x_{id}(1[j=k]-\pi_{ik}) - x_{id}(1[j=k]-\pi_{ik})(1+(1-p_{ij})(1+b_1))\right) && (64)\\
&= y_{ij}x_{id}(1[j=k]-\pi_{ik})(b_1+2-(1+(1-p_{ij})(1+b_1))) && (65)\\
&= y_{ij}x_{id}(1[j=k]-\pi_{ik})(1+b_1)p_{ij}. && (66)
\end{aligned}
$$

Inference over log relative risk

Let $w \in \Omega$ be a categorical variable (nominal) indicating experimental condition. We are interessted in the relative change in the feature j across two or more experimental conditions. Let us denote by $\mathbf{x}_i = (\mathbf{w}_i, \mathbf{z}_i)$ where $\mathbf{w}_i$ is a one-hot encoding denoting the experimental condition for sample i. And furthermore, further decompose the model parameters by $\beta_j = (\theta_j, \gamma_j)$ where θ_j denotes the parameters corresponding to $\mathbf{w}_i$ and γ_j for $\mathbf{z}_i$. Consider a multivariate function $g : \boldsymbol{\beta} \to \mathbb{R}^J$ with $g_j(\boldsymbol{\beta}) = \log RR_{j|\mathbf{z}}(\boldsymbol{\beta})$, where

$$
\log RR_{j|\mathbf{z}} = \log\frac{\pi_{j|w_1,\mathbf{z}}}{\pi_{j|w_0,\mathbf{z}}}, \tag{67}
$$

with $w_0, w_1 \in \Omega$ denoting the two experimental conditions. We can derive standard errors for g_j by relying on the delta method and asymptotic normality of the MAP estimate $\boldsymbol{\beta}$:

$$
\nabla g(\boldsymbol{\beta})^T\Sigma\nabla g(\boldsymbol{\beta}).
$$

Similarly to above, we provide the derivative of g_j w.r.t. $\beta_{k,d}$.

$$
\log\frac{\pi_{j|w_1,\mathbf{z}}}{\pi_{j|w_0,\mathbf{z}}} = \log\pi_{j|w_1,\mathbf{z}} - \log\pi_{j|w_0,\mathbf{z}}. \tag{68}
$$

Using the gradient for π_j derived above:

$$
\log\frac{\pi_{j|w_1,\mathbf{z}}}{\pi_{j|w_0,\mathbf{z}}} = z_{w_1,d}(1[j=k]-\pi_{j|w_1,\mathbf{z}}) - z_{w_0,d}(1[j=k]-\pi_{j|w_0,\mathbf{z}}). \tag{69}
$$

Dirichlet parmeter estimation

We use mean-precision representation of Dirichlet. The log probability density for dataset $D = (\boldsymbol{p}_n)$ is given by,

$$\log p(D|\boldsymbol{\alpha}) = N \log \Gamma(s) - N \sum_k \log \Gamma(s m_k) + N \sum_k (s m_k - 1) \log \bar{p}_k, \tag{70}$$

where $\log \bar{p}_k = \frac{1}{N} \sum_n \log p_{n,j}$ denotes the sufficient statistics.

The fixed-point iteration for s given by,

$$\frac{1}{s_{new}} = \frac{1}{s} + \frac{1}{s} \left(\frac{d^2}{ds^2} \log p(D|s) \right)^{-1} \left(\frac{d}{ds} \log p(D|s) \right). \tag{71}$$

The fixed-point iteration for α_k is given by,

$$\Psi(\alpha_{k,new}) = \log \bar{p}_k - \sum_j m_j (\log \bar{p}_j - \Psi(s m_j)), \tag{72}$$

followed by setting

$$m_{k,new} = \frac{\alpha_{k,new}}{\sum_j \alpha_{j,new}}. \tag{73}$$

The fixed-point iteration for α_k requires inverting Ψ, which can be posed as finding the root of:

$$F(x) = \Psi(x) - \log \bar{p}_k + \sum_j m_j (\log \bar{p}_j - \Psi(s m_j)) = 0. \tag{74}$$

This can be accomplished using Newton's method:

$$x_{new} = x_{old} - \frac{F(x_{old})}{F'(x_{old})} = x_{old} - \frac{F(x_{old})}{\Psi'(x_{old})}. \tag{75}$$

We initialize $\mathbf{m}_0 = \frac{1}{N} \sum_n \boldsymbol{p}_n$ and use Stirling's approximation for s:

$$s_0 = \frac{(K-1)/2}{-\sum_k m_k \log \bar{p}_k / m_k}. \tag{76}$$

Table 1: Sample counts for each cell type used in the microRNAome dataset analysis.

Cell type	N
B lymphocyte CD19	39
Monocyte	43
Natural killer cells CD56	41
Neutrophil	20
Red blood cell	25
T lymphocyte CD4	42
T lymphocyte CD8	41

Table 2: The number of samples for each cell type used in the microRNAome cell type data analysis.

Cell type	N
Fibroblast	45
Smooth muscle	43
Endothelial	43
Monocyte	51
T lymphocyte CD8	48
T lymphocyte CD4	48
B lymphocyte CD19	42
Natural Killer Cells CD56	44

Table 3: The number of samples for each cell type used in the recount2 messenger RNA-seq data analysis.

Cell type	N
Cardiac Muscle Cell	33
Endothelial Cell	81
Lymphocytes	685
Fibroblast	13
Macrophage	318
Fat Cell	74
Smooth Muscle Cell	15

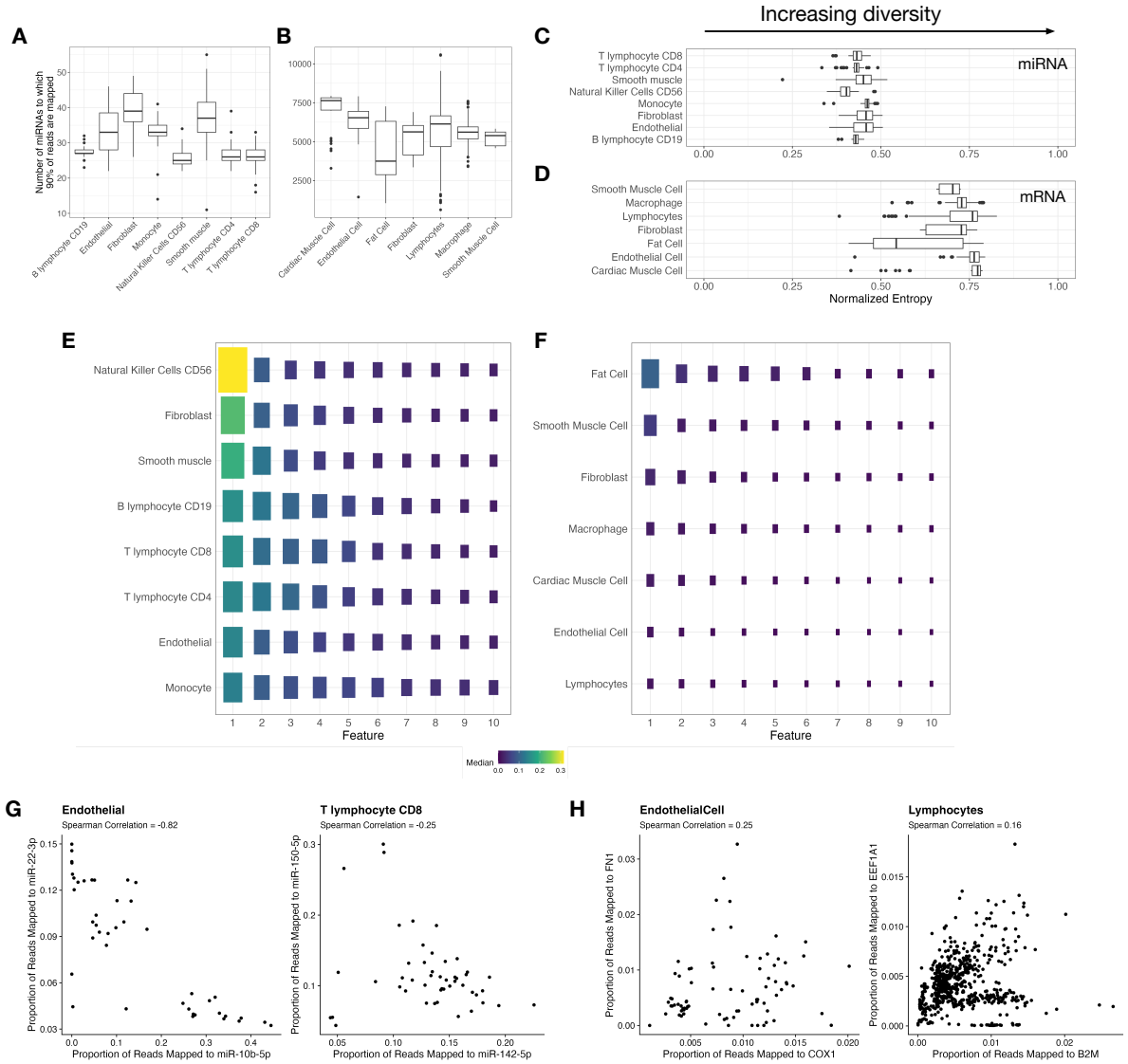

Figure 1: Number of features to account for 90% of the reads. **A** The number of features needed is below 50 for all miRNA samples considered. **B** With an exception of fat cell, the median number of features needed to express 90% of the reads exceeds 5,000 for mRNA. The normalized Shannon index for **C** miRNA is smaller than **D** mRNA, indicating less diverse number of species are expressed. The boxplots show IQR1-3. Median of the proportion of reads allocated to the most highly expressed 10 features across samples for **E** miRNA and **F** mRNA. The top features in the miRNA samples account for the majority of the expression. Spearman's correlation for **G** Endothelial and CD8 T lymphocyte miRNA samples and **H** corresponding Endothelial and Lymphocyte mRNA samples. The miRNA data commonly exhibits negative correlation.

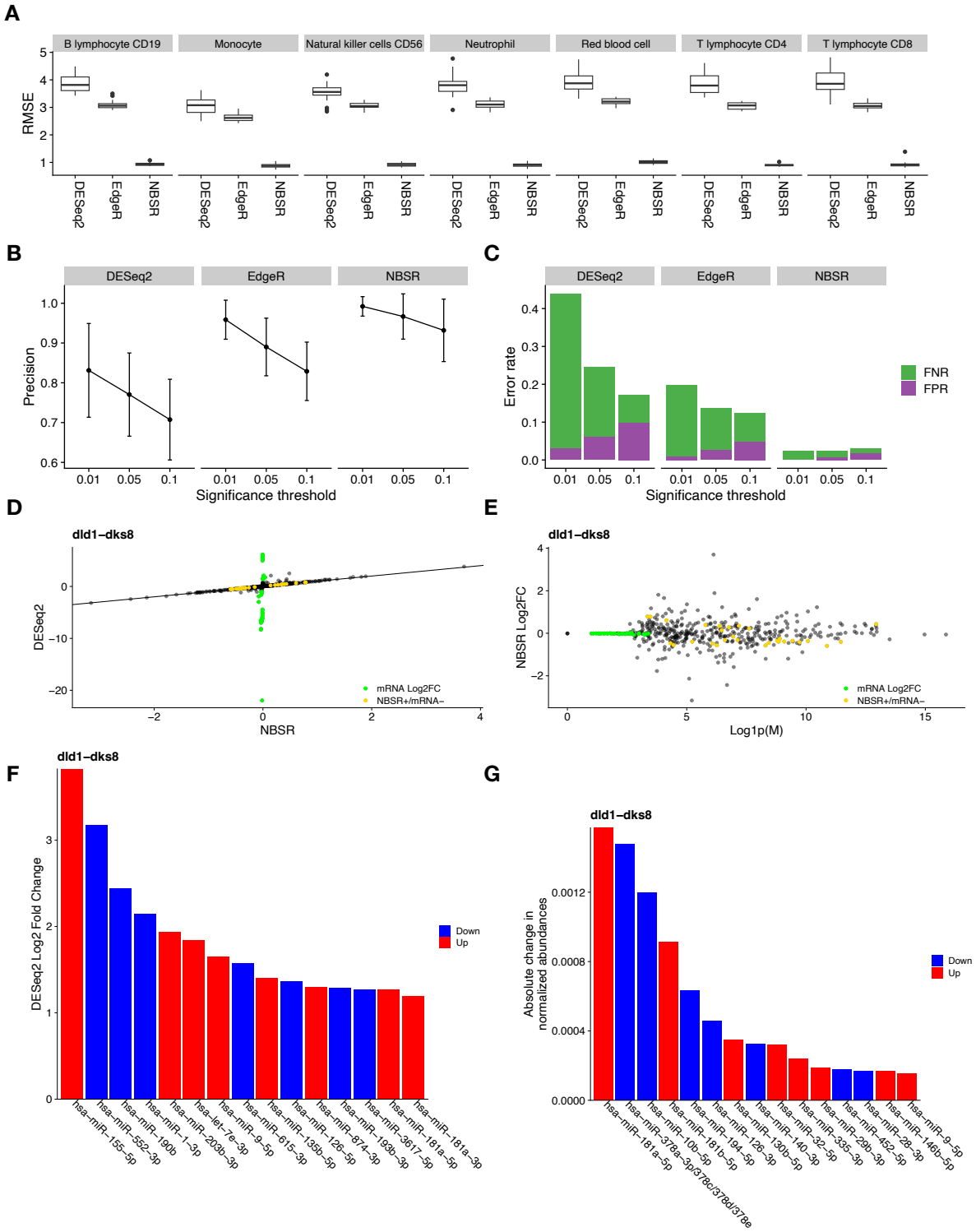

Figure 2: **A** Root mean squared error for three methods on simulated data. Boxplot shows IQR1-3 across 20 replicates. **B** Comparison of precision across the three methods. The confidence interval shows ± 2 standard errors drawn from 20 replicates. **C** False positive and negative rates for the three methods averaged across 20 replicates. **D** Comparison of the estimated log2 fold change values for DESeq2 and NBSR. The two methods mostly agree except for the points shown in blue (mRNA+). **E** The mRNA+ features are sparse and lowly expressed as shown in the MA-plot. **F** Significant features at $\alpha = 0.05$ selected by DESeq2, sorted by log2 fold change to show the top 15 features. **G** Top 15 significant features at $\alpha = 0.05$ sorted by absolute change in normalized abundance as selected by NBSR.

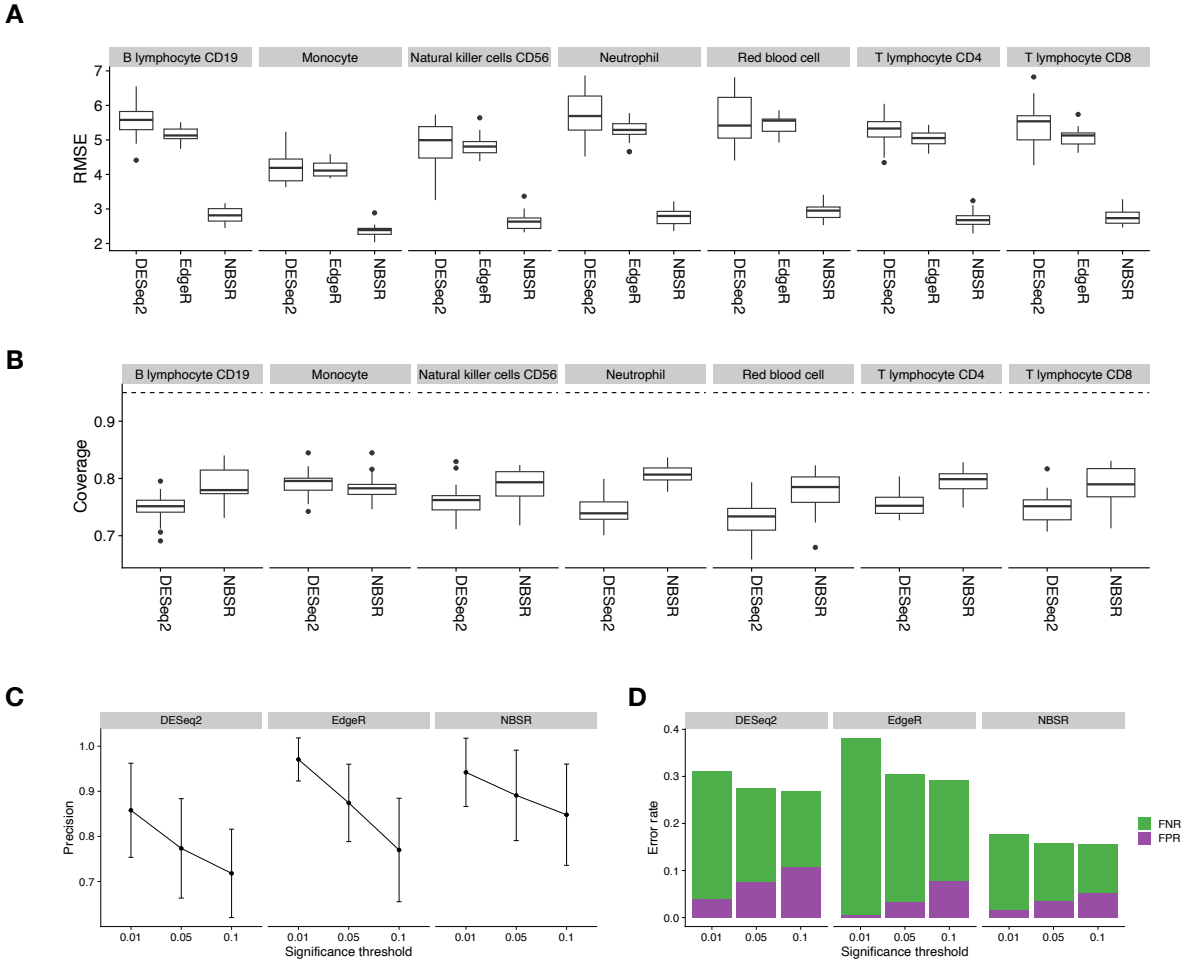

Figure 3: Boxplots with IQR1-3 depicting spread of performance for NBSR vs DESeq2 with $n = 3$: A. root mean squared error and B. coverage of the ground truth log2 fold change obtained from 20 replicates C. Precision with confidence bounds given by ± 2 standard errors obtained from 20 replicates. D. False positive and negative rates at various thresholds averaged over 20 replicates.

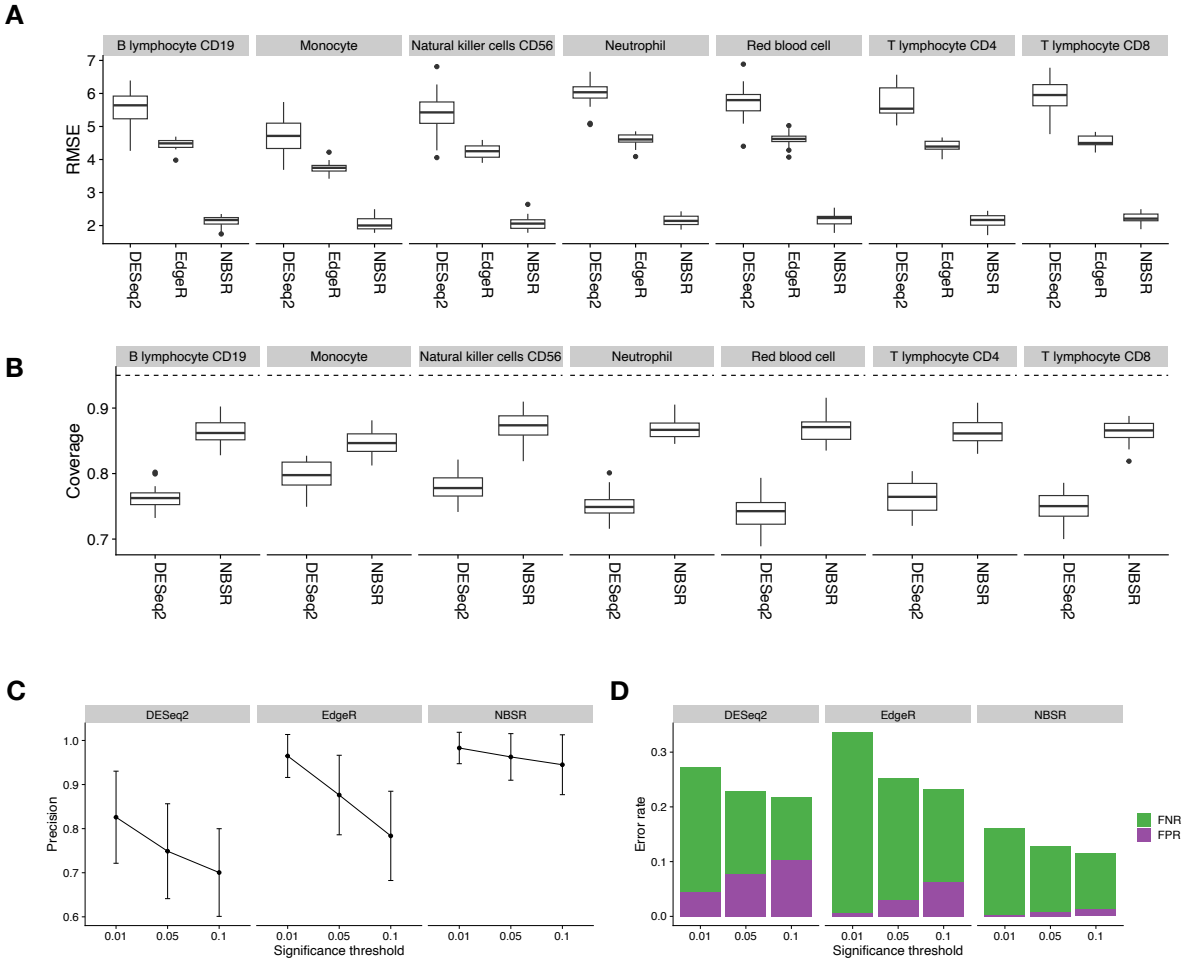

Figure 4: Boxplots with IQR1-3 depicting spread of performance for NBSR vs DESeq2 with $n = 5$: A. root mean squared error and B. coverage of the ground truth log2 fold change obtained from 20 replicates C. Precision with confidence bounds given by ± 2 standard errors obtained from 20 replicates. D. False positive and negative rates at various thresholds averaged over 20 replicates.

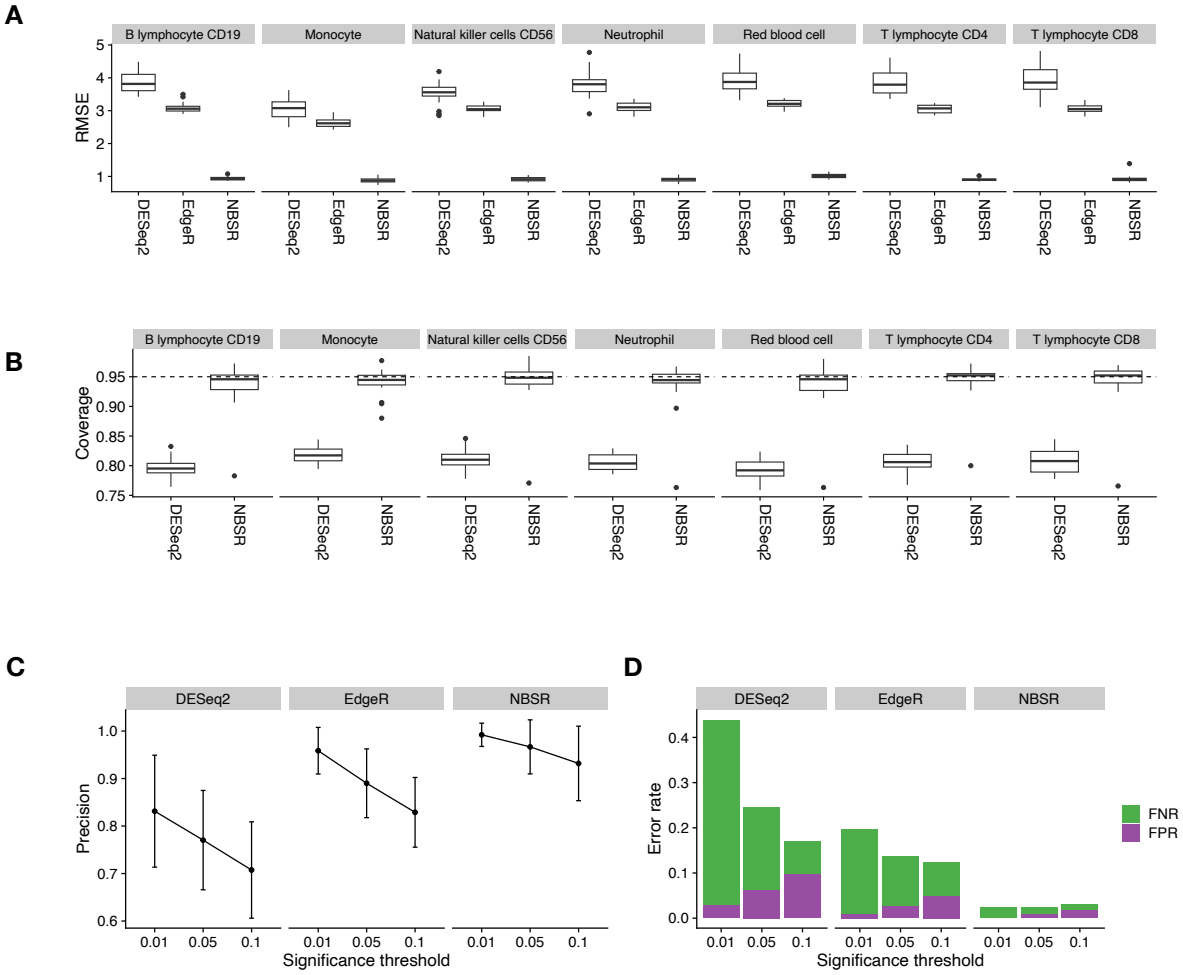

Figure 5: Boxplots with IQR1-3 depicting spread of performance for NBSR vs DESeq2 with $n = 10$: A. root mean squared error and B. coverage of the ground truth log2 fold change obtained from 20 replicates C. Precision with confidence bounds given by ± 2 standard errors obtained from 20 replicates. D. False positive and negative rates at various thresholds averaged over 20 replicates.

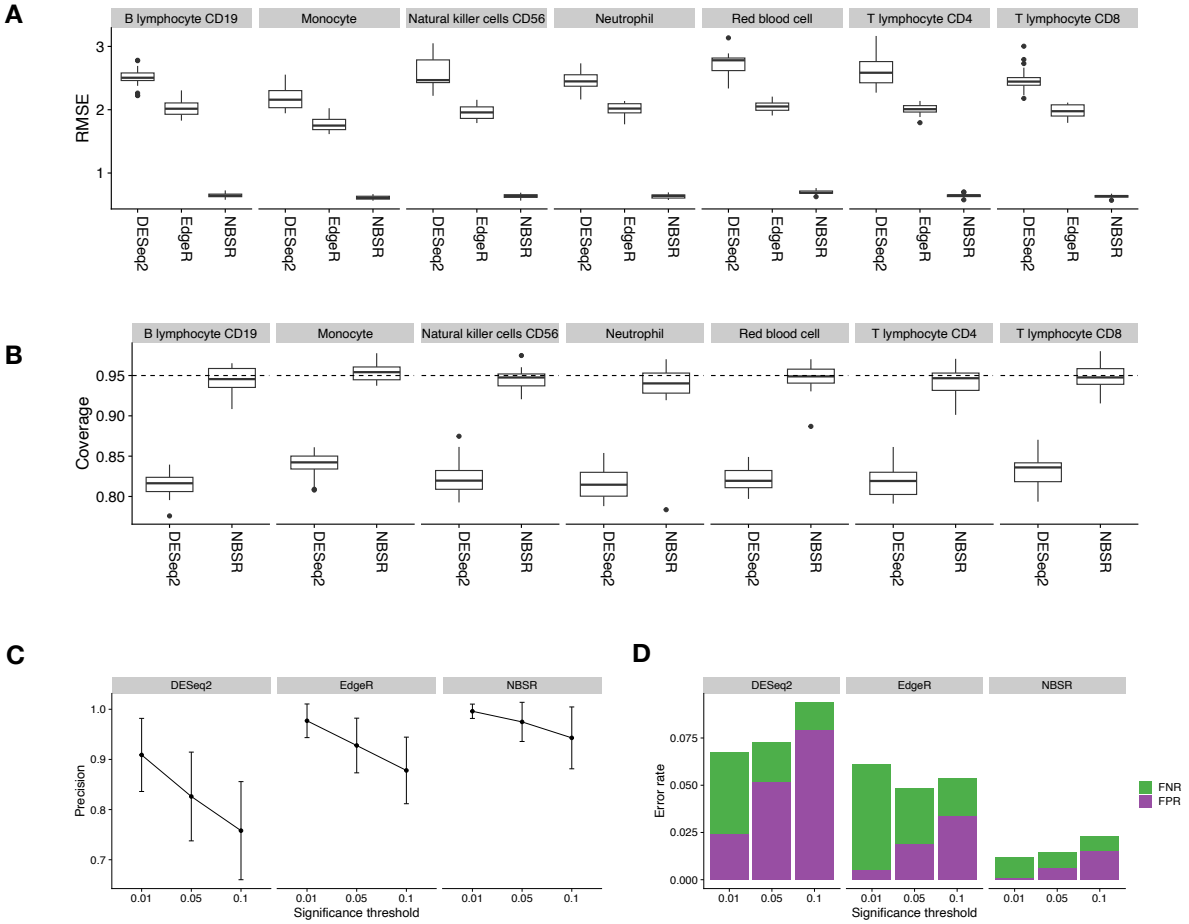

Figure 6: Boxplots with IQR1-3 depicting spread of performance for NBSR vs DESeq2 with $n = 20$: A. root mean squared error and B. coverage of the ground truth log2 fold change obtained from 20 replicates C. Precision with confidence bounds given by ± 2 standard errors obtained from 20 replicates. D. False positive and negative rates at various thresholds averaged over 20 replicates.

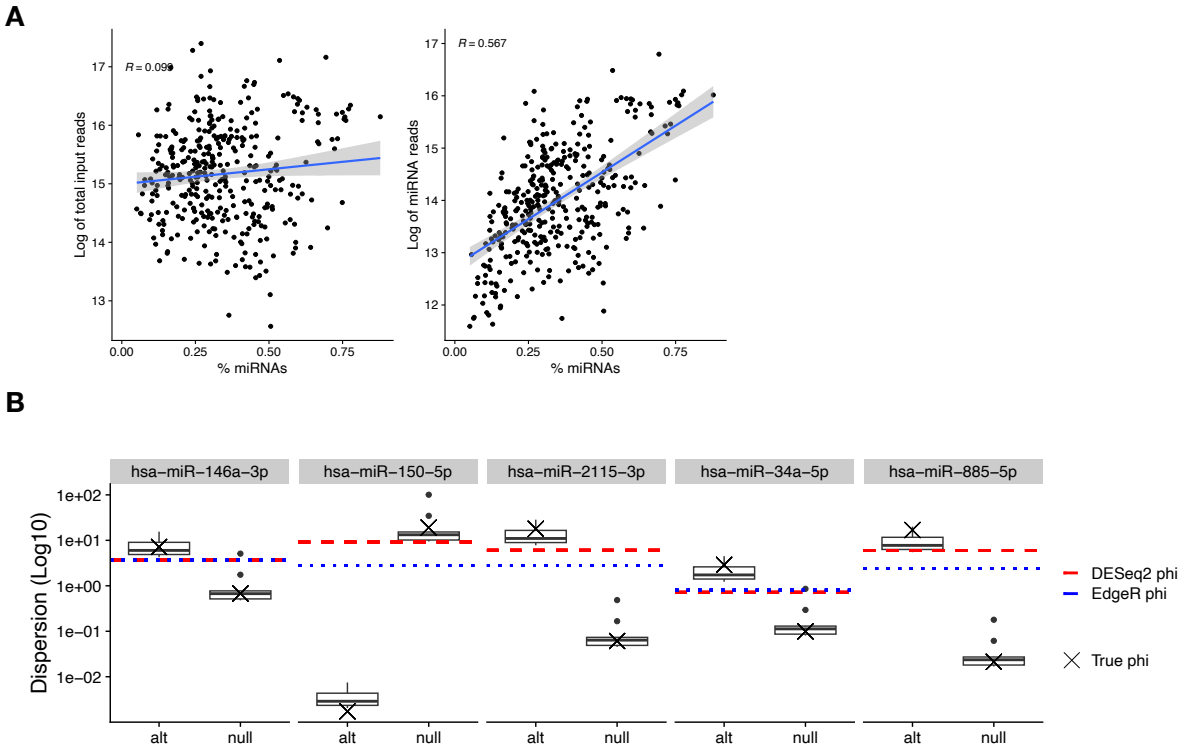

Figure 7: A. (Left) Relationship between the number of miRNA reads and total sequencing depth. The amount of miRNAs captured does not seem to be related to the sequencing depth but rather a sample-specific variability. (Right) Percent miRNA and log of total miRNA reads. The two figures depict that small RNA-sequencing has a layer of variability in capturing the miRNAs, which is proportional to the total sequencing depth and amount of miRNAs present in the sample. B. Comparison of the dispersion estimation for three methods. EdgeR and DESeq2 are not able to account for variability across experimental conditions due to feature-specific assumption of dispersion. Our proposed dispersion model can be utilized in settings where there is significant variability across experimental conditions to account for the effects of biological coefficient of variation on the counts data.

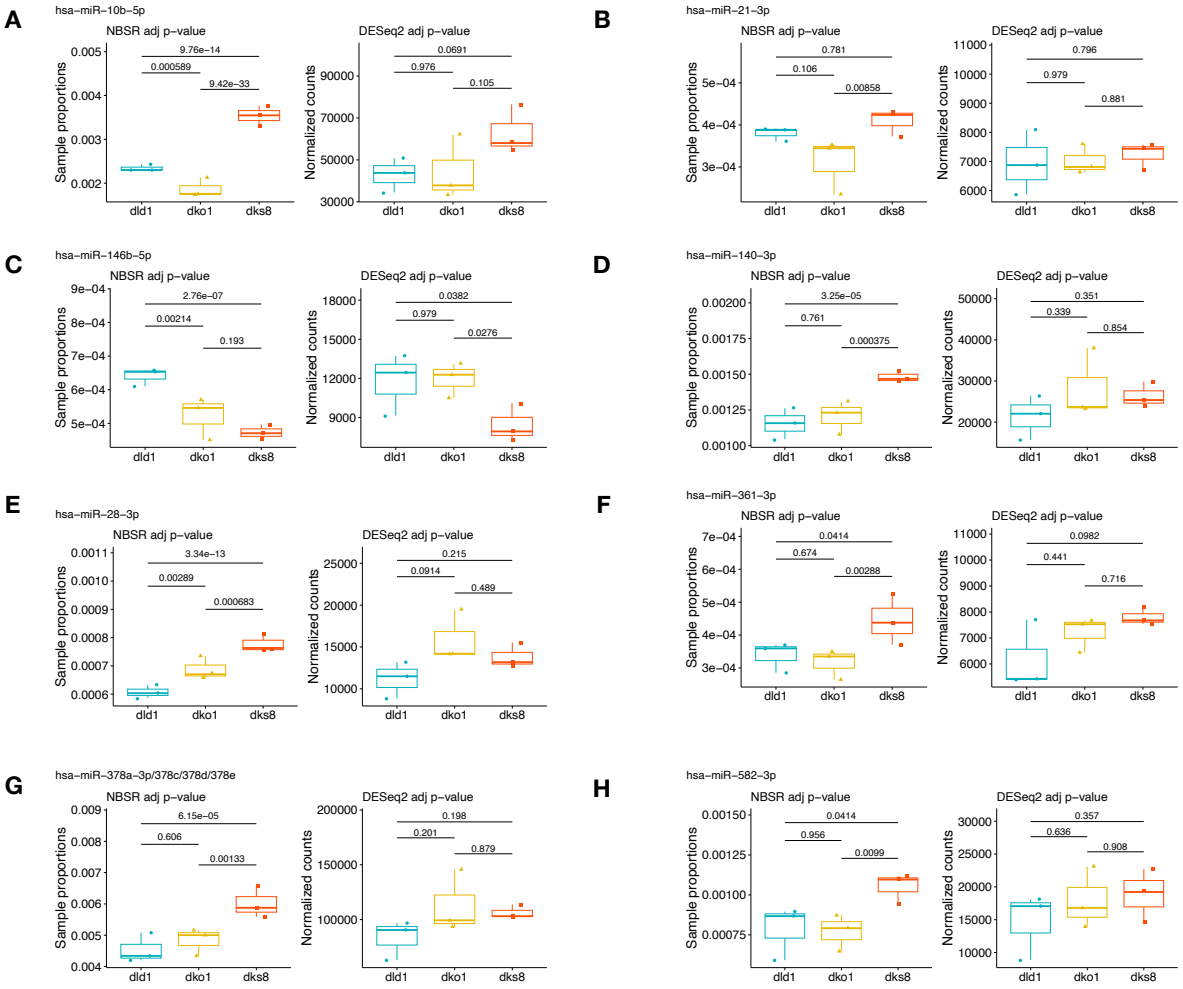

Figure 8: Boxplots of miRNAs identified as significant by NBSR at threshold $\alpha = 0.01$ but not significant by DESeq2 at threshold $\alpha = 0.1$. The miRNAs whose estimated absolute change in probabilities exceeding 0.0001. This panel of figures illustrates that estimating the normalized abundance can reveal differences undetectable by (normalized) counts and mean expression levels.

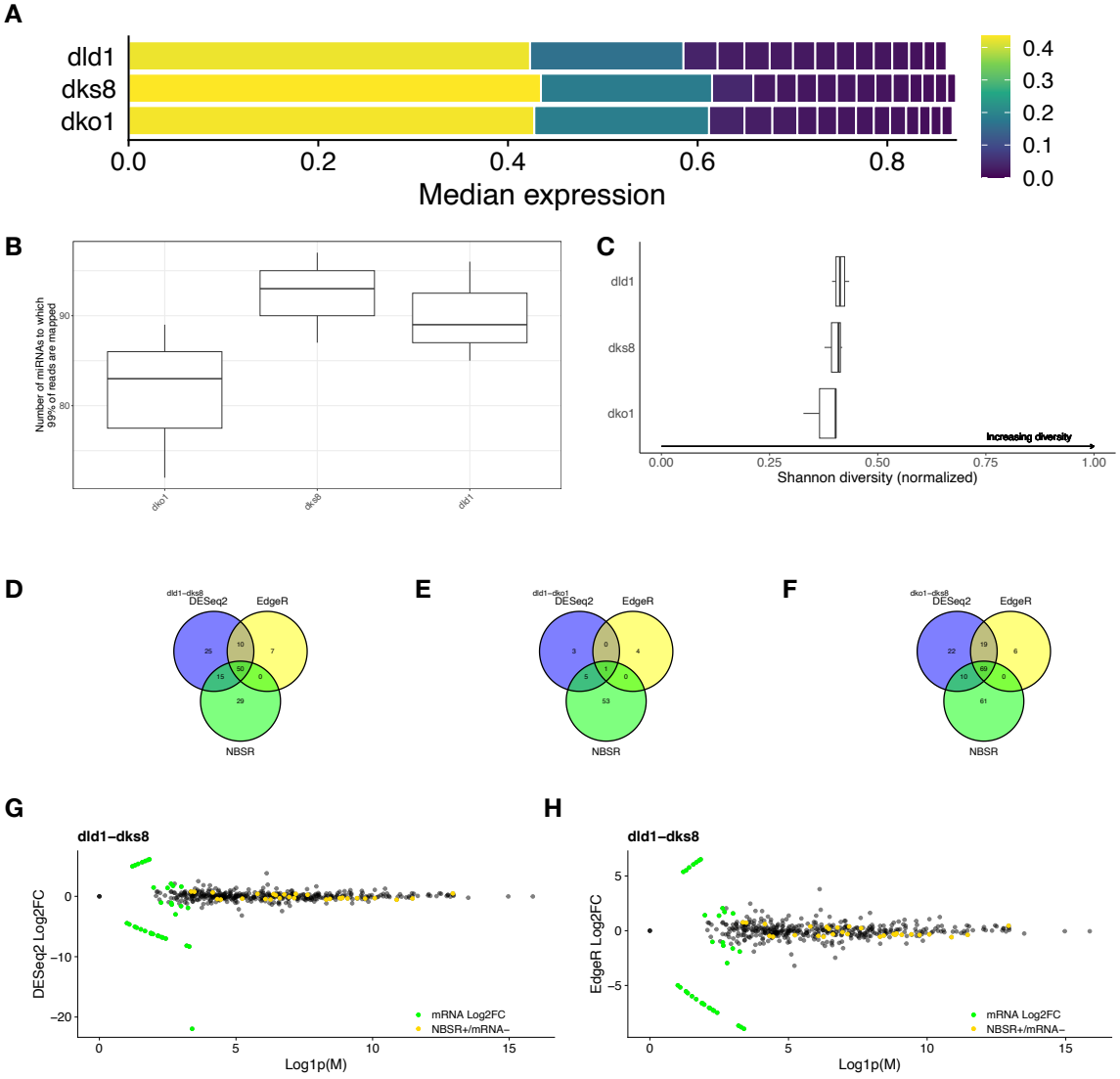

Figure 9: Summary of the Hepatocullular carcinoma cell line data. A. Plot of the top 15 most highly expressed miRNAs. The first two miRNAs account for about 60% of the read counts. B. The number of miRNAs needed to account for 99% of the read counts. C. Shannon diversity index depicts that the reads are not uniformly distributed across the miRNAs. D-F. Venn diagram depicting the commonalities and differences in the three methods in calling significantly differential miRNAs at threshold $\alpha = 0.05$. The MA-plots for dld1 vs dks8 for G. DESeq2 and H. EdgeR. Features identified as significant by mRNA-seq methods tend to be lowly expressed with sparsity (zeros) while NBSR significant features are highly expressed and dense (non-zeros across all samples).