miRglmm: a generalized linear mixed model of isomiR-level counts improves estimation of miRNA-level differential expression and uncovers variable differential expression between isomiRs

Andrea M. Baran[1], Arun H. Patil[2], Ernesto Aparicio-Puerta[1], Marc K. Halushka[3], Matthew N. McCall[1,4,5,*]

1. Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY, USA
2. Lieber Institute for Brain Development, Baltimore MD, USA
3. Cleveland Clinic, Pathology and Laboratory Medicine Institute, Cleveland, OH
4. Department of Biomedical Genetics, University of Rochester Medical Center, Rochester, NY, USA
5. Wilmot Cancer Institute, University of Rochester Medical Center, Rochester, NY, USA

* Correspondence
Matthew N. McCall, Ph.D.
University of Rochester
Dept of Biostatistics and Computational Biology
265 Crittenden Boulevard, CU 420630
Rochester, New York 14642-0630
mccallm@gmail.com

Abstract

Background: MicroRNA (miRNA) expression can be characterized by small RNA sequencing. Typically, the resulting sequencing reads are aligned to known miRNAs, and read counts of sequences that align to the same miRNA are summed to produce miRNA-level read counts. The aggregated miRNA counts are then used for differential expression (DE) analyses, often using tools developed for mRNA-seq data analysis. There are two key drawbacks to this approach: (1) aggregation to miRNA-level counts discards information about different miRNA transcript isoforms, called isomiRs, and (2) miRNA-seq data violate key assumptions of the DE methods developed for mRNA-seq data. Together these necessitate the development of DE methods designed specifically for miRNA-seq data that can model isomiR-level data.

Methods: We establish miRglmm, a DE method, implemented in an R package of the same name, that uses a generalized linear mixed model (GLMM) of isomiR-level counts to account for dependencies due to reads coming from the same sample and/or from the same isomiR. The isomiR random effect variances can be used to quantify variability in differential expression between isomiRs that align to the same miRNA, thereby facilitating detection of miRNAs with differential expression or differential isomiR usage.

Results: Using biological data to simulate miRNA with variable isomiR DE, we demonstrate that miRglmm provides markedly lower mean squared error (MSE) and much better confidence interval coverage than other DE tools. Using an experimental benchmark data set of synthetic miRNAs, we show that even in the absence of isomiR variability, miRglmm provides the lowest MSE among all DE tools while maintaining confidence interval coverage at the nominal level. In real biological data, miRglmm finds significant isomiR-level DE variability for most miRNAs in our analyses. While fold change estimates are similar to those from commonly used DE tools (intraclass correlation coefficients ≥ 0.8), miRglmm differs the most from existing methods and is able to estimate both isomiR- and miRNA-level DE.

Conclusions: The use of isomiR-level counts, rather than summed miRNA-level counts, in a GLMM framework, represents a notable improvement in miRNA differential expression analysis. Our method, miRglmm, outperforms current DE methods in estimating DE for miRNA, whether or not there is significant isomiR variability, and simultaneously estimates isomiR-level DE.

Introduction

MicroRNA (miRNA) are a class of small, noncoding RNAs that perform a role in transcriptional regulation. They are typically 18-24 nucleotides long single-stranded RNA molecules that bind to mRNA causing translational suppression or mRNA degradation (Ambros 2004, Bartel 2004). Through this mechanism, miRNA can regulate entire pathways and drive disease pathogenesis (Zhang and Farwell 2008, Mendell and Olson 2012). The specificity of each miRNA:mRNA interaction leads to discrete downstream consequences (Valencia-Sanchez, Liu et al. 2006). miRNA expression patterns can be determined by small RNA sequencing (RNA-seq). This process involves enriching samples for small RNA species, sequencing the RNA contained in the sample, and aligning the sequences to known miRNAs, tRNA fragments, or other RNA species (Baras, Mitchell et al. 2015).

The specific mature sequence listed in miRNA databases is considered the canonical miRNA sequence (Tomasello, Distefano et al. 2021). Different isoforms of individual miRNAs, called isomiRs, can arise from both variation in the nucleotide sequence or from variation in the transcript length (Ameres and Zamore 2013). Sequence length variants have more or fewer nucleotides at the 5' and/or the 3' end of the canonical sequence, whereas polymorphic (internal) isomiRs include different nucleotides within the mature sequence (Tomasello, Distefano et al. 2021). A third form of isomiR is the addition of an adenosine or uracil tail by a terminal uridylyl transferase (TUT) or similar enzyme (Yang, Bofill-De Ros et al. 2022). We will use the general term isomiR to describe the various sequence isoforms mapped to a miRNA, without distinction between biologically or technically-derived isomiRs, with the assumption that most sequences observed at sufficient levels in biological samples to model are biologically relevant isomiRs.

The collection of isomiRs that align to a given miRNA are often referred to as the miRgroup (Schmauch, Attia et al. 2024). As an example of isomiR diversity within miRgroup, Supplemental Tables 1 and 2 display isomiR-level read counts produced by small RNA sequencing for two miRgroups: hsa-let-7a-5p and hsa-miR-26a-5p. Typically, read counts of each isomiR in a miRgroup are aggregated by summation and summarized as a single read count for each miRNA (as shown in the last row of Supplemental Tables 1 and 2). There are numerous combinations of isomiR counts that produce the same miRNA counts; therefore, this aggregation leads to the loss of any information contained in individual isomiR expression.

IsomiRs are thought to possess unique biological roles (Bofill-De Ros, Yang et al. 2020). Like canonical sequences, isomiRs are conserved throughout evolution, and biogenesis of isomiRs is tightly regulated. The region of heterogeneity between isomiRs may have unique implications for miRNA-mediated translation regulation (Ameres and Zamore 2013). 5' isomiRs have modifications at the 5' end of the miRNA resulting from differential processing of paralogous pre-miRNA and can result in regulation of distinct target mRNA (Tan, Chan et al. 2014). 3' isomiRs are more common and result from post-transcriptional trimming or tailing sequence modifications (Yang, Bofill-De Ros et al. 2019). These modifications can alter miRNA function by impacting target recognition or extent of target repression. Naturally occurring isomiRs have been shown to play distinct roles in a variety of biological processes including cytokine expression, virus proliferation, apoptosis, and tumor progression (Bofill-De Ros, Yang et al. 2020). Differential expression at the isomiR-level has identified isomiRs with cancer-specific expression (Telonis, Magee et al. 2017). The biological importance of isomiRs highlights the need for a miRNA-seq analysis method that can account for distinct isomiR expression patterns in estimating miRNA-level differential expression, while also producing more granular estimates of isomiR-level differences.

Similar to differential expression analyses typically performed using bulk mRNA-seq data, the goal of most small RNA sequencing studies is to study if, and which, miRNA differ in their expression between groups of samples. Differential expression tools commonly used for bulk mRNA-seq data have the goal of determining which mRNA exhibit differential expression between groups of interest and provide an estimate of fold change to quantify the magnitude of the difference. Common tools include DESeq2 (Love, Huber et al. 2014), edgeR (Robinson, McCarthy et al. 2010) and limma-voom (Law, Chen et al. 2014). All of these methods have been frequently used to analyze miRNA-seq data; however, there are some key differences between miRNA-seq data and mRNA-seq data that may make the assumptions of these methods invalid when applied to miRNA-seq data. Due to the compositional nature of RNA sequencing data, the reads can be viewed as a random sample of a fixed size from the pool of all RNA in the library, which can be modeled by a multinomial distribution. In bulk mRNA-seq data, the number of unique mRNA expressed is large and the reads are distributed relatively evenly across the mRNA. In this setting, if one assumes independence between features, the true underlying multinomial distribution can be approximated by marginal binomial distributions. This approximation is a key assumption of the tools mentioned above but is violated in miRNA-seq data in two important ways. First, there are generally fewer than 500 miRNA compared to over 10,000 mRNA expressed in a sample, which makes it more likely that random fluctuation in the expression of one miRNA substantially affects the expression

of other miRNAs. Second, the distribution of reads in miRNA-seq data is often skewed toward a small number of highly expressed miRNA compared to the more uniform distribution of reads seen in mRNA-seq data (Figure 1). This can induce negative correlation between highly expressed miRNA (due to competition of being counted) regardless of their underlying biological correlation. This also results in sparse data with a small number of non-zero features making traditional normalization approaches, such as size factor normalization or counts-per-million (CPM) normalization, unstable or highly skewed.

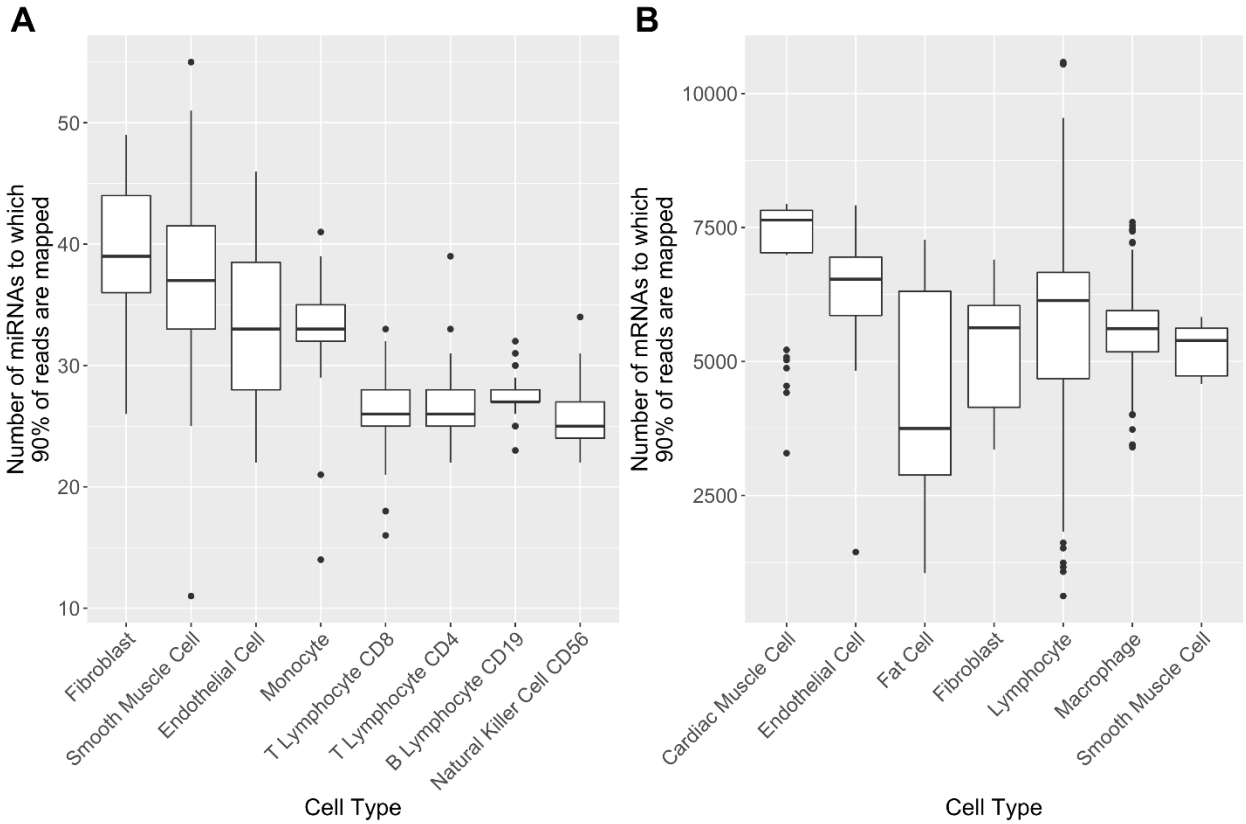


Figure 1: Comparison of the number of features that account for 90% of the reads in a sample between miRNA-seq data (panel A) and mRNA-seq data (panel B) across several cell types.

Differences between miRNA and mRNA make the analysis of miRNA counts at the isomiR level feasible. First, miRNAs and all isomiRs are fully sequenced (18-24 nt), whereas typically only fragments of mRNAs (100s of nt out of ~5,000-50,000 nt species) are sequenced. Therefore, quantifying the expression of miRNAs based on the entire sequence is possible (Baras, Mitchell et al. 2015). Second, as previously stated, there are far fewer miRNA expressed compared to mRNAs. Maintaining and analyzing sequence level data at the scale needed for mRNA analysis would be computationally burdensome. Recent work

has highlighted the importance of analyzing miRNA-seq data at the isomiR level (Schmauch, Attia et al. 2024). Through the analysis of 28 public miRNA-seq datasets and a newly generated human endothelial cell hypoxia data set, the authors show substantial differences between isomiR expression and their corresponding canonical miRNAs. They conclude by calling for "a comprehensive re-evaluation of the miRNA-seq analysis practices." However, this study simply compares DE results from applying DESeq2 to the aggregate miRNA level data with DE results from applying DESeq2 to the isomiR level data. As noted above, aggregate miRNA level data violates the assumption of independence between features due to the small number of unique miRNAs expressed in a sample and the highly skewed distribution of miRNA expression. While analyzing isomiR level data greatly increases the feature space and reduces the overall skew of the expression distribution, the isomiR level data introduce a new source of dependence due to high correlation between isomiRs from the same miRNA, which also violates a core assumption of the DESeq2 model.

Negative binomial mixed models (NBMM) can be used to model overdispersed count data when there is a correlation structure among the counts (Booth, Casella et al. 2003). This is the case when working with miRNA-seq data at the isomiR level, as counts that are observed for the same isomiR or for the same sample are correlated. Modelling the raw (non-normalized) counts directly has advantages over the typical reads per million (RPM) normalization. RPM is a dependent normalization strategy, so a change in any one miRNA read count will lead to changes in all other miRNA values even in the absence of a change in absolute expression (Witwer and Halushka 2016). To account for the technical artifact of difference in the number of total reads between samples, one can incorporate an offset term into the NBMM to adjust for total overall read count within each sample. NBMMs have been proposed for other next-generation sequencing analyses, including RNA-seq, metagenomic sequencing, and single cell RNA-seq analyses (Zhang, Mallick et al. 2017, Zhang, Pei et al. 2018, He, Davila-Velderrain et al. 2021, Tsonaka and Spitali 2021). However, the NBMMs used in these methods all model aggregate feature level data rather than isoform level data and none have been applied to miRNA-seq data.

In this manuscript, we propose miRglmm, a generalized linear mixed model of isomiR-level counts that can be used to estimate miRNA-level DE while also obtaining estimates of isomiR DE and variability. We compare miRglmm to several commonly used DE tools on simulated data, an experimental benchmark data set, and real biological data sets. miRglmm has a lower MSE than other DE tools and better confidence interval coverage probabilities. Additionally, we find significant isomiR-level variability exists within most miRgroups, further motivating the use of miRglmm to analyze miRNA-seq data.

Methods

miRglmm: A Negative Binomial GLMM-based DE tool for miRNA-seq data

The statistical model implemented in miRglmm is as follows. Let C_{ij} denote the count for sample $i = 1, \ldots, n$ and isomiR $j = 1, \ldots, J$. Let $T_i = \sum_{j=1}^{J} C_{ij}$ be the total counts for sample i. This is used as an offset term in the model to adjust for variable sequencing depth across samples. For each sample i, let x_i be a p-vector of covariates. For each miRNA, miRglmm fits the following negative binomial mixed model (NBMM):

$$C_{ij} \sim \mathrm{NB}\left(C_{ij} \mid \mu_{ij}, \Theta\right)$$

$$log\left(\mu_{ij}\right) = log(T_i) + \left(\beta + \tau_{1j}\right)X_i + \tau_i + \tau_{0j}$$

$$\begin{bmatrix} \tau_{0j} \\ \tau_{1j} \end{bmatrix} \sim N(0, \Omega_i), \Omega_i = \begin{bmatrix} \sigma_{0j}^2 & \sigma_{01j} \\ \sigma_{01j} & \sigma_{1j}^2 \end{bmatrix}$$

$$\tau_i \sim N(0, \sigma_1^2)$$

β is the fixed effect of primary interest. τ_i is the random intercept term for sample, τ_{0j} is the random intercept for isomiR, and τ_{1j} is the random slope term for isomiR. The random effects for sample and isomiR are independent, while the intercept and slope random effects for isomiR have a dependence structure specified by Ω. σ_{1j}^2 is of primary interest in assessing the variability of the main effect of interest between isomiRs. We built miRglmm using the function glmer.nb in the lme4 package, which fits an NBMM via a Laplace approximation to the maximum likelihood with a variety of optimizer choices (Bates, Mächler et al. 2015). The analyses presented in this paper are reproducible using code in the GitHub repository found at https://github.com/mccall-group/miRglmm.

Software, data structures, and outputs

miRglmm is an R library that consists of one core function that can be easily integrated into DE analysis pipelines by replacing DE methods designed for mRNA-seq (such as DESeq2, edgeR, limma/voom) with miRglmm. IsomiR-level count matrices in the form of core Bioconductor structures, SummarizedExperiment objects, are taken as input, and a list of model fit summaries for each miRgroup analyzed is returned. The function is able to be run in parallel across miRgroups and can handle flexible design matrices. Additional functions are included in the package to extract and summarize values of interest from the model fit summaries, including miRNA-level estimates of DE with confidence intervals,

isomiR-level estimates of expression within a miRgroup, estimates of variability in DE across isomiRs within a miRgroup, and a likelihood ratio test for if the isomiR variability in the miRgroup is significant.

The microRNAome data resource

The miRNAome dataset was assembled to more fully understand the miRNA expression patterns across primary cell types (McCall, Kim et al. 2017, Patil, Baran et al. 2022). The dataset was built upon 2,077 samples from 175 public datasets across 196 primary cell types. miRNA annotation and quantification was performed using the miRge3.0 pipeline (Patil and Halushka 2021). Briefly, miRge3.0 is a multi-step miRNA alignment program. From a FASTQ file, miRge3.0 collapses identical sequences and processes them through repeated Bowtie alignment steps to identify canonical miRNAs, isomiR miRNAs, and other RNA species. In addition to a final read count per miRNA, it outputs an alignment file containing counts of all aligned isomiRs, which was used here.

Filtering of lowly expressed miRNAs and isomiRs

Counts per million (CPM) normalization of aggregated counts was used to assess overall miRNA expression, with the goal of retaining miRNA with sufficient expression to model. A threshold of log(median CPM)>5 was used to retain miRNA for modelling. Even after filtering at the miRNA-level, the resulting count matrix contains very sparse isomiR-level counts so miRglmm also filters lowly expressed isomiRs that contribute low/no amount of information to the model. Specifically, CPM normalization of isomiR-level counts was used to assess isomiR expression, and an isomiR filter based on log(median CPM) is implemented as an input argument of miRglmm (default = -1).

Inducing known effects in real biological data to simulate differential expression

We created ground truth data using real biological data inducing a known artificial effect into real data, allowing us to assess the performance of miRglmm under conditions seen in real data. We searched the miRNAome data (McCall, Kim et al. 2017) for one cell type with large sample counts coming from one study to have a relatively homogenous starting dataset. We used 39 monocyte samples from one study (Juzenas, Venkatesh et al. 2017) and retained 122 miRNA with sufficient expression to model using a log(median CPM) cutoff of 5.

To induce the artificial “group” effect, we randomly split the samples into 2 groups, with 19 samples labelled as Group A and 20 samples labelled as Group B (Supplemental Figure 1). Of 122 total miRNA, we used stratified sampling to select 20 to be overexpressed in Group A, and another 20 to be

underexpressed in Group A. The sampling was stratified by total miRNA expression to manipulate miRNA across the full range of expression values. We used a log(median CPM) cutoff of -1 to retain isomiRs for analysis. For the 20 miRNA overexpressed in Group A, we multiply the counts of all retained isomiRs in the miRgroup by a random truncated normal variable with mean value of 2, variance of 1, lower bound of 1 and upper bound of 3 in Group A samples only. These 20 miRNA now have a known miRNA-level fold change (B vs A) = 0.5, with random variability between isomiRs. For the 20 miRNA underexpressed in Group A, we multiply the counts of all retained isomiRs in the miRgroup by a random truncated normal variable with the same parameters in Group B samples only. These 20 miRNA now have a known miRNA-level fold change (B vs A) = 2, with random variability between isomiRs. Importantly, we recalculate the total counts used in the offset term after the artificial signal is induced. The stratification used in sampling miRNA helps ensure the effect of the signal being added is consistent across groups, even though total counts increase for all samples. The entire procedure, including randomly splitting samples into two groups, is repeated 100 times.

For each miRNA, we estimated a miRNA-level differential group effect, measured via log fold-change (logFC) estimates, using miRglmm. We compared these estimates to differential group effects produced by other commonly used differential expression tools used on aggregated data: DESeq2 (Love, Huber et al. 2014), edgeR (Robinson, McCarthy et al. 2010), and limma-voom (Law, Chen et al. 2014). We also included a Negative Binomial Generalized Linear Models (NB GLM), which is similar to the miRglmm model but without random effects that can model the isomiR variability. Data was aggregated to the miRNA-level prior to running DESeq2, edgeR and limma-voom, and hence we collectively call these methods “aggregated methods”. Aggregated count values were produced by summing counts from all isomiRs that align to a given miRNA.

We used Mean Squared Error (MSE) and 95% coverage probabilities to assess the performance. The MSE compared the estimated logFC to the induced effect for each miRNA, and the coverage probability assessed the probability that the true logFC fell in the 95% confidence interval estimated by the model. We cannot estimate a coverage probability for edgeR as that algorithm does not provide standard error (SE) estimates. We can also separately look at ability to estimate/cover miRNA where the effect was induced up or down, and those where no effect was induced.

To identify miRNA with significant variability in the group effect between isomiRs, we can test whether the random slope parameter τ_{1j} in the miRglmm model is necessary with a 1-degree of freedom (1-df) Likelihood Ratio Test (LRT) comparing likelihoods from the model specified above and a model removing

only the τ_{1j} parameter. With this test, we can identify the miRNA with significant random slope effects and summarize by induced effect groups.

Analysis of a benchmark experiment using synthetic miRNA pools

We sought to use data with known ground truth estimates to demonstrate that miRglmm can recover the true fixed effect estimates of primary interest. The Extracellular RNA Communication Consortium (ERCC) was established to facilitate expansion of the field of extracellular RNA (exRNA) biology, and consisted of collaborative projects to develop robust methods for isolation and analysis of exRNA data (Das, Ansel et al. 2019). One of these projects was a multi-protocol, multi-institution assessment of the bias of four small RNAseq library preparation methods, using ratiometric pools of synthesized small RNAs (Giraldez, Spengler et al. 2018). Chemically synthesized RNA oligonucleotides were added in varying ratios to pool A and pool B, from 10 to 1 and 1 to 10, leading to 15 levels of known logFC (from -2.3 to 2.3), providing a known true fold change value for each miRNA that can be used to evaluate the accuracy of miRglmm and other methods (Supplemental Figure 2). The ERCC data was processed by 6 different laboratories (Lab), and we used dimensionality reduction using non-metric multidimensional scaling (NMDS) to assess if there was sufficient laboratory variability to require adjustment for Lab in the analysis. The samples separated along the first NDMS dimension by Lab, and the second NMDS dimension appeared to capture the Pool effect (Supplemental Figure 3). We included a fixed effect for Lab in the models for all methods to adjust for the technical effect of the differences in lab-specific sample handling, processing and sequencer on the counts. The ERCC dataset included synthetic RNAs that were much longer at 50-90 nucleotides than the typical 18-24 nucleotide miRNAs. We excluded any RNA specified to have a length > 45, resulting in a set of 303 small RNAs.

This synthetic dataset also provided justification for establishing filters at both the miRNA and isomiR level. Since the oligonucleotides were synthetically added, we expect all included miRNA to be expressed. The distribution of miRNA expression via log(median CPM) of aggregated counts in this data can be used to determine a suitable threshold for miRNA expression. We use log(median CPM)>5 to retain miRNA for analysis in all analyses. To filter sequences, we aimed to separate biologically relevant isomiR counts from random background expression. Since the ERCC contains sequence isoforms that do not map to known miRNA, we can consider these background counts and compare the distribution of expression levels to isomiR counts that map to known miRNA to identify an appropriate range for an isomiR-level filter. When we compared the distribution of background sequence expression (reads not mapping to any known miRNA) to expression of sequences that map to miRNA, we saw that the

distributions separated around log(median CPM) of -1 (Supplemental Figure 3). We ran miRglmm without any sequence filtering, and also compared the performance of miRglmm under a range of reasonable filters (log(median CPM)>-1 (default) to 2).

For each miRNA, we estimated a differential Pool effect, adjusted for Lab and measured via logFC estimates, using miRglmm under a variety of isomiR filters. We also aggregated the data to the miRNA count level and ran existing (aggregated) methods for comparison. We used MSE and 95% coverage probabilities to assess the performance as described above. We utilized LRT as previously described to assess if there was significant variability in the Pool effect between isomiRs.

We assessed if there was any effect of filtering prior to aggregation for the aggregated methods. For the same set of filtering levels used for miRglmm, we filtered prior to aggregating to the count level and used MSE to compare the estimated effects from the aggregated methods to the truth.

Differential expression analyses of real biological data of tissues and cell types

Our goal in selecting samples for these analyses was to minimize possible technical variation, and therefore capture true biologically relevant isomiR-level differences. We chose bladder and testes tissues to compare because they represented a large set of tissues available from a single study (Li, Chen et al. 2011), where sample processing and sequencer would be consistent across samples. We chose a set of immune cell types that were also derived from a single study as our cell type differential expression analysis set (Juzenas, Venkatesh et al. 2017). We aimed to produce miRNA-level differential expression (DE) estimates using miRglmm and compare these results to other commonly used differential expression tools that run on aggregated miRNA-level count data

For each miRNA, we estimated a miRNA-level differential group (tissues or cell type) effect, measured via log fold-change (logFC) estimates, using miRglmm. We compared these estimates to differential group effects produced by existing methods, and compared the estimates using an intraclass correlation coefficient (ICC). Nominal significance at an alpha=0.05 level was compared using Cohen's Kappa. We also used an upset plot to simultaneously compare miRNA considered differentially expressed between all 5 methods. We can produce isomiR-level estimates of expression within group by summing fixed and random effects for each isomiR. We utilized LRT as previously described to identify miRNA within which there is significant variability in the group effect between isomiRs.

Results

The methodology developed in this manuscript was motivated by two initial observations. First, different isomiRs of the same miRNA can behave very differently between groups of samples. To illustrate this, we selected a study of bladder and testes samples (Li, Chen et al. 2011), with the goal of minimizing possible technical variation, and therefore capture true biologically relevant isomiR-level differences. We observed isomiRs with uniformly zero counts within bladder and large non-zero counts in testes, and other isomiRs with uniformly zero counts within testes and large non-zero counts in bladder (Supplemental Table 1). Aggregation to miRNA counts masks these isomiR differences and results in a loss of information. Additionally, we observe evidence of differential isomiR usage between bladder and testes (Figure 2). We consider the canonical sequence (which is typically the most highly expressed sequence) and the next two highest expressing isomiRs, as these would contribute most to the aggregate miRNA count. Even in this small subset of isomiRs, we observe differences in DE between tissues across sequences, indicating the need for a model that can capture this variability in isomiR DE.

Second, while miRNA-level data exhibit an artificially induced negative correlation between highly expressed miRNAs, isomiR-level data do not exhibit this bias. Data generated by high-throughput RNA sequencing is fundamentally compositional in nature. In other words, the resulting reads can be viewed as a random sample of a fixed size from the pool of RNA generated during library preparation. This produces a competition-to-be-counted (Silverman, Durand et al. 2018) in which randomly measuring more of one feature decreases the chance of measuring other features. Consider the extreme example of only two features; with a fixed number of total reads, it is clear that a higher count for feature 1 implies a lower count for feature 2, resulting in perfect negative correlation between the two features regardless of their underlying biological correlation. Frequently, this induces a negative correlation between the two most highly expressed miRNAs in each cell type; however, this induced correlation is not observed for the most highly expressed isomiRs. To illustrate this, we selected a study of immune cell types (Juzenas, Venkatesh et al. 2017), with the goal of eliminating compositional differences seen in tissue-level data. We compared the correlation between the two highest expressed miRNAs with the correlation between the two highest expressed isomiRs for each cell type. While there was generally a negative correlation between the highest expressed miRNAs, the same was not true at the isomiR level (Figure 3). This indicates that utilizing the isomiR-level count data could overcome technical biases seen with aggregated miRNA-level count data.

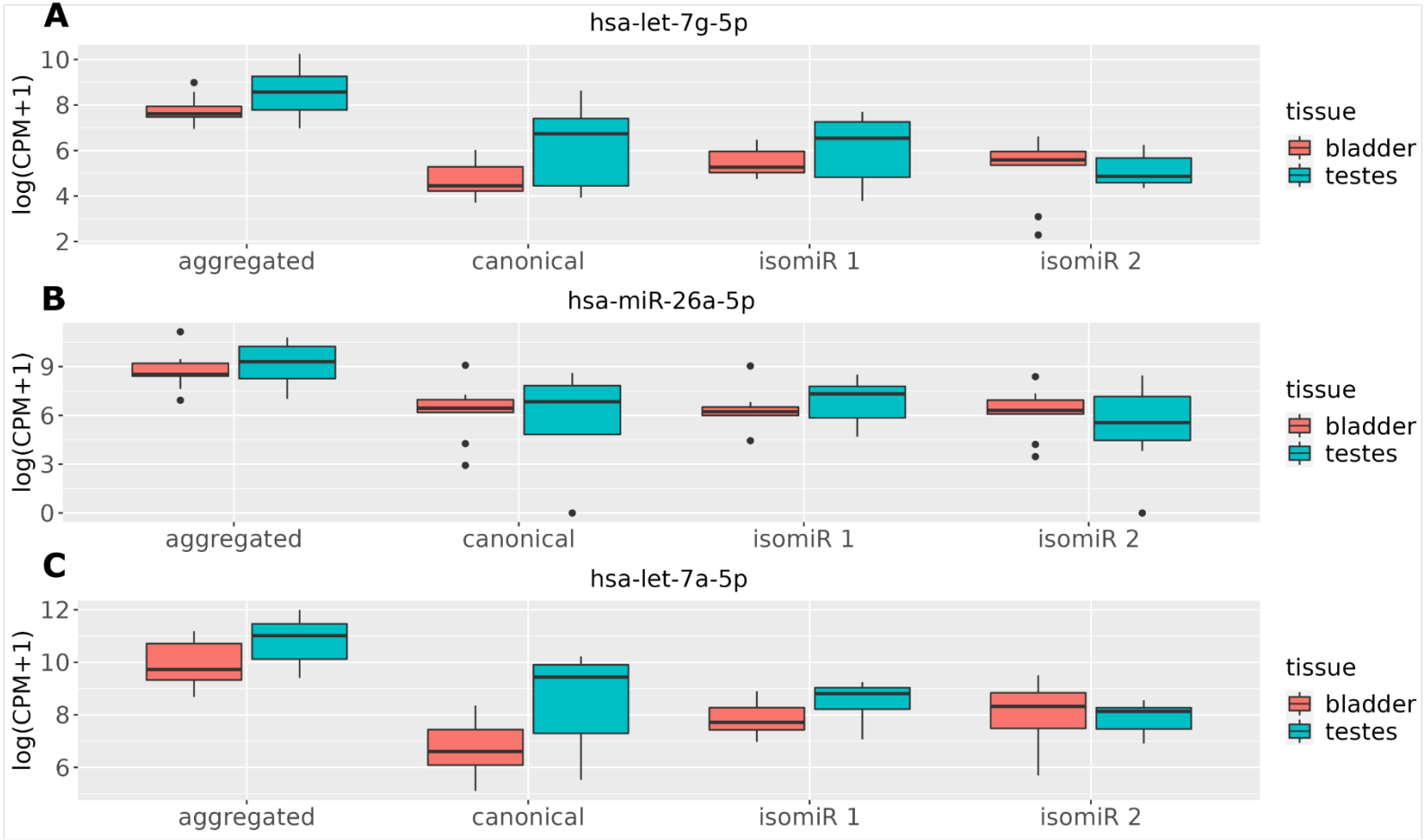


Figure 2: Tissue-specific counts are summarized by boxplots for three miRNAs (panel A: hsa-let-7g-5p, panel B: hsa-miR-26a-5p, panel C: hsa-let-7a-5p). Aggregated miRNA-level counts from summing counts across isomiRs within a miRNA are compared to the canonical/representative sequence and the two highest expressed non-canonical isomiRs (isomiR 1 and isomiR 2). The canonical sequence counts, and isomiR 1 are all higher in testes than bladder, but isomiR 2 has the opposite trend, which is masked when counts are aggregated to miRNA-level. CPM: counts per million.

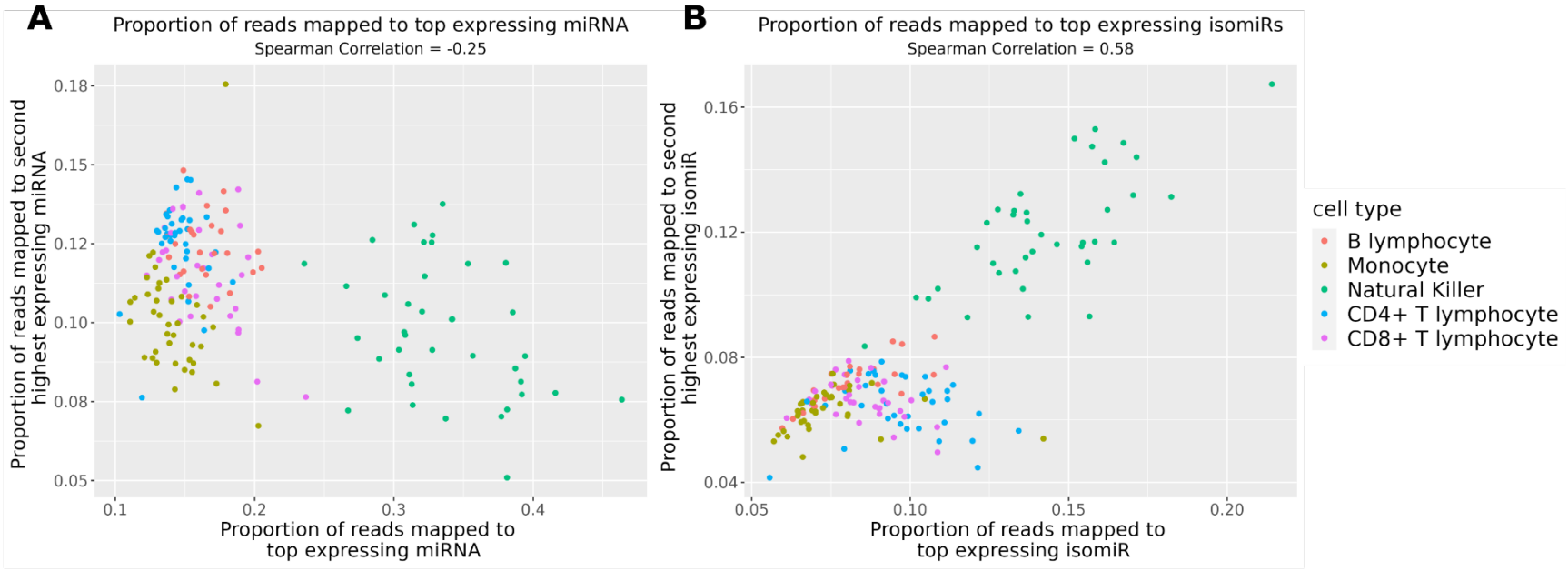


Figure 3: Negative correlation bias is observed when considering the top two highest expressing miRNA after aggregation to miRNA-level counts (panel B), whereas negative correlation is not observed when considering the two highest expressing isomiRs (panel A).

miRglmm: a generalized linear mixed effects model for miRNA-seq data analysis

We developed a method and corresponding software to analyze isomiR-level counts from miRNA-seq data. Our method, miRglmm, accounts for dependencies introduced by reads coming from the same sample and from the same isomiR, by using a negative bionomial generalized linear mixed model with random effects for sequence and sample. We utilized a negative binomial model to directly model the counts without the need for transformation and include an offset term to normalize for sequencing depth (see Methods for details). Unlike existing methods, miRglmm provides estimates of differential expression at both the miRNA-level and the isomiR-level. miRglmm is implemented in a free and open-source R package available at: https://github.com/mccall-group/miRglmm.

miRglmm outperforms other methods in detecting differentially expressed miRNAs in the presence of isomiR variability

To assess the performance of miRglmm in comparison to existing methods, we used a collection of monocyte samples to simulate differential expression at both the miRNA- and isomiR-level (see Methods). Over 100 simulations, miRglmm provided the lowest mean MSE (mean = 0.015) and median MSE (median = 0.014) among all the methods (Table 1, Supplemental Table 3). miRglmm was also the method with the lowest standard deviation across simulations (SD = 0.005) and the lowest maximum MSE (maximum= 0.033). miRglmm minimized MSE in 97% of simulations (97/100) when looking within a given simulation. DESeq2 minimized the MSE in the remaining 3 simulations. On average, for these 3 simulations, the MSE for DESeq2 was only smaller than the MSE for miRglmm by 0.001. miRglmm provided the highest mean coverage probability (mean = 91.4%) and highest median coverage probability (median = 92.6%) among the 4 methods that produce SE estimates which can be used to calculate coverage (Table 1, Supplemental Table 4).

Table 1: MSE and coverage probability summary statistics across 100 simulations

method	mean MSE (10^{-3})	minimum MSE (10^{-3})	maximum MSE (10^{-3})	# of times method minimizes MSE	mean coverage probability	minimum coverage probability	maximum coverage probability
miRglmm	14.85	7.84	32.92	97	0.91	0.71	0.98
DESeq2	21.92	12.29	36.08	3	0.80	0.66	0.89
limmavoom	23.10	13.81	45.37	0	0.79	0.55	0.89
NB GLM	23.15	13.26	48.83	0	0.77	0.53	0.89
edgeR	23.15	13.25	48.83	0	NA	NA	NA

MSE: Mean Squared Error, NA: not applicable

We can also assess performance within each “truth” group of miRNA: induced positive effect (N=20), induced negative effect (N=20), or no effect induced (N=82). When we calculate MSE within each group and summarize across simulations, we see that miRglmm provides markedly lower MSE in the groups with the induced effect compared to all aggregated methods (Figure 4A). In the group of miRNA with no change induced, DESeq2 tends to provide a lower MSE in most simulations. Similarly, the coverage probability in the two groups of miRNA with induced effects tends to be much closer to the 95% nominal level than for the aggregated methods (Figure 4B). The aggregated methods have very poor coverage probabilities (30%-40%). For the miRNA with no induced effect, DESeq2 tends to have the highest coverage probability, though miRglmm still provides coverage near the 95% nominal level.

An additional advantage of miRglmm is that it provides an estimate of the variability in the tissue effect between isomiRs. The other methods cannot estimate or account for this variability. We observe that the proportion of significant miRNA is very high in the groups with the induced effect (Supplemental Figure 4), indicating that our simulation procedure correctly implemented random isomiR DE variability in the group effect. Taken together, these results support the conclusion that the aggregation and resulting loss of information is detrimental to performance in cases where significant isomiR DE variability exists. miRglmm, which can account for this variability, provides consistently high performance whether or not there is significant isomiR DE variability.

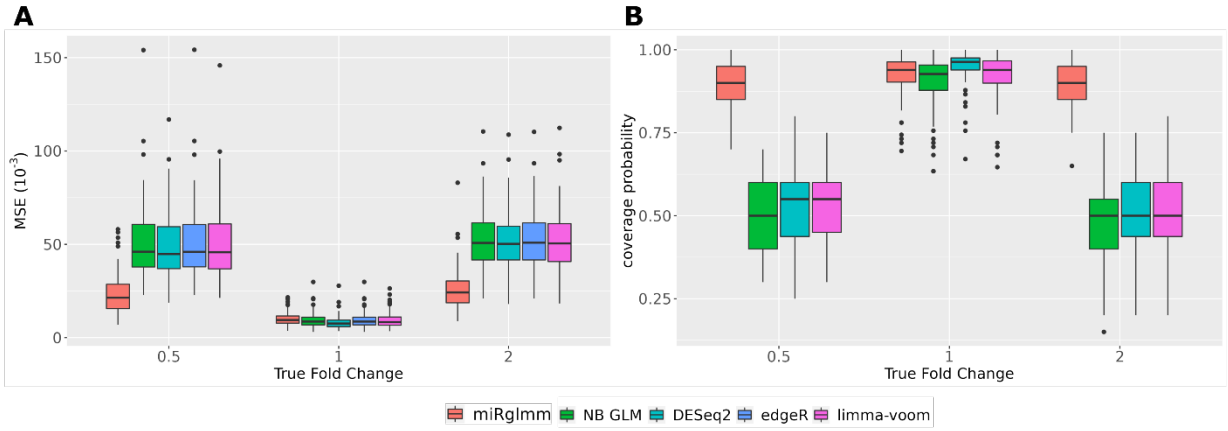


Figure 4: miRglmm outperforms the aggregated methods in terms of Mean Squared Error (MSE) for miRNA in which an artificial effect is induced (panel A). MSE across 100 simulations is summarized. DESeq2 provides the lowest MSE when there is no effect induced. The coverage probability of miRglmm is much higher than the coverage probability of the aggregated methods when an artificial effect is induced (panel B). Coverage probability across 100 simulations is summarized.

miRglmm provides accurate estimates of differential expression for isomiRs

Along with providing miRNA-level estimates of differential expression, miRglmm also provides isomiR-level estimates of differential expression. While rarely done, DESeq2 can also provide isomiR-level estimates if run on the isomiR-level count matrix. Summarizing over 100 simulations, miRglmm (MSE=0.022) provides better estimates of isomiR-level differential expression than DESeq2 (MSE=0.041). When we assess performance within each "truth" group, we see that miRglmm outperforms DESeq2 when there is a negative effect (0.035 vs 0.044), when there is a positive effect (0.035 vs 0.044) and also when there is no differential expression (0.015 vs 0.040). This is in contrast to what we saw at the miRNA-level, where DESeq2 slightly outperformed miRglmm with respect to MSE when there was no differential expression.

miRglmm outperforms other methods in detecting differentially expressed miRNAs even when there is no isomiR variability

While variable isomiR usage is common in real biological data, we wanted to assess how miRglmm would perform in the absence of isomiR variability. To assess the performance of miRglmm in this context, we used a multi-protocol, multi-institution experimental benchmark data originally designed to compare the performance of four different small RNAseq library preparation methods (Giraldez, Spengler et al. 2018). This experiment used ratiometric pools of synthesized small RNAs with known variable amounts of differential expression (see Methods), providing a known true fold change value for each miRNA that can be used to evaluate the accuracy of methods. Because the data came from mixtures of synthesized miRNA pools, there should be no biological variability in isomiR usage.

We used MSE and coverage probabilities to assess the performance of the methods (Table 2) and filtering strategies (Supplemental Table 5). miRglmm accurately estimated the known Pool effect while maintaining greater than nominal coverage probabilities. In terms of MSE, miRglmm at any filter level provided smaller MSE than all aggregated methods. The unfiltered NBMM had the highest MSE, and we observed that the estimates were systemically biased towards the null due to the inclusion of the lowly expressed isomiRs (Supplemental Figure 5). The four aggregated methods provide similar estimates but are underestimating the true effect at all levels. We chose log(median CPM) > -1 as the default filter for miRglmm because all filtering strategies provide similar MSE and there does not appear to be a more stringent filtering level that clearly outperforms it (Supplemental Figure 6). As filtering gets more restrictive, we lose the ability to model some miRNA if fewer than two isomiRs are retained.

Table 2: Comparing performance of miRglmm to aggregated methods

method	MSE (x 10^{-3})	coverage probability
miRglmm	7.81	0.98
DESeq2	10.19	0.99
NB GLM	12.11	0.97
edgeR	12.11	NA
limma-voom	13.78	0.99

MSE: Mean Squared Error, NA: not applicable

To determine whether other methods would improve if we filtered lowly expressed isomiRs, we compared MSEs based on estimates from running the other methods on aggregated counts from all isomiRs versus filtering prior to aggregation of the isomiR counts. We saw that the gain in performance seen due to filtering in miRglmm is not seen for the aggregated models (Supplemental Figure 5). Since aggregating is done via summation, the low counts do not influence the sum to the same degree that they influence being averaged into the estimate from the NBMM.

Due to the synthetic nature of the data, we do not expect any biological isomiRs and presume that the sequence isoforms we are modeling arise solely from technical variation. Hence, the random slope parameter was not found to be significant in the synthetic data (Supplemental Figure 5). If we remove this extraneous random slope parameter for isomiR from the model, the performance is similar in terms of the MSE, coverage probabilities and bias (Supplemental Table 6 and Supplemental Figure 7). This indicates that retaining the random slope parameter, even if not needed, does not diminish the ability of the model to estimate the effect of interest.

Differential miRNA expression and differential isomiR usage between immune cell types

We performed differential expression analyses comparing miRNA expression between five immune cell types: monocytes (N=39), Natural Killer Cells (N=38), CD4+ T lymphocytes (N=35), CD8+ T lymphocytes (N=32) and CD19+ B lymphocytes (N=26). The samples derive from one single experiment in the miRNAome data(Juzenas, Venkatesh et al. 2017), limiting the possibility of technical/experimental variation due to library preparation or sequencer differences. Based on the isomiR filtering strategies assessed using the ERCC data, we chose to implement miRglmm using a threshold of log(median CPM) < -1 to remove isomiRs with insufficient expression. After isomiR filtering, we retained 7,389 isomiRs from 104 miRNA, and the number of isomiRs mapping to any one miRNA ranged from 6 to 633 with a median of 35 isomiRs.

We estimated log fold-change (logFC) estimates for each miRNA using miRglmm and compared the estimates to existing methods (see Methods). B lymphocytes were used as the baseline in each logFC estimate. We used intra-class correlation coefficients to assess the agreement in logFC estimates between methods (Figure 5A). While all methods provided very good agreement in logFC estimates (ICC≥0.95), we observed only fair to moderate agreement in assessment of significant (at alpha=0.05 level) differential expression between methods (Cohen's Kappa 0.38-0.55). There does not appear to be

a substantial difference in ICC based on contrast, but some contrasts have very poor agreement in significance (minimum Kappa=0.11 when comparing NB GLM and miRglmm for monocyte vs B lymphocyte contrast). An upset plot of miRNA-contrast pairs deemed significantly different across the methods provides a visual comparison of all 5 methods simultaneously (Figure 5B). We used likelihood ratios tests (LRTs) to assess if there is variability in the cell type effect between isomiRs within a given miRNA. All miRNA have significant LRT p-values for the random tissue parameter for isomiRs (Figure 5C). The other methods cannot estimate or account for this variability.

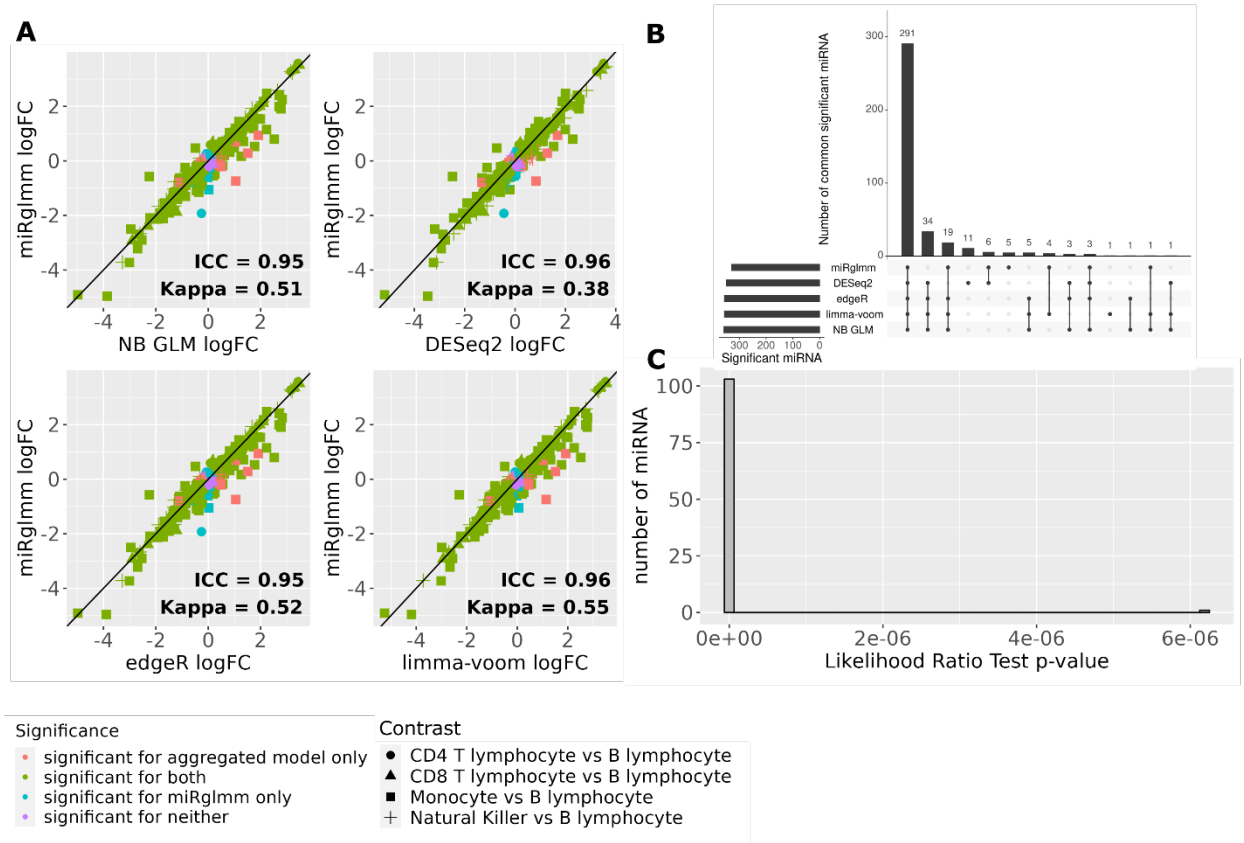


Figure 5: Log fold change (logFC) estimates of miRNA-level expression in cell types compared to B lymphocytes from miRglmm are plotted versus logFC estimates from each aggregated count-based model (panel A). Each point is a unique contrast for a unique miRNA. Significant DE vs B lymphocytes is indicated by the color (nominal alpha=0.05 level). Intra-class correlation coefficients (ICC) are used to assess agreement in the values of the estimated logFC between methods. Cohen's kappa is used to assess agreement in the significance of the estimate between methods. An upset plot shows the common number of miRNAs differentially expressed across the 5 methods (panel B). A histogram of the p-values of an LRT testing the significance of the random isomiR cell-type effect indicates that most of the miRNA have significant variability in the cell-type effect across isomiRs (panel C).

Differential miRNA expression and differential isomiR usage between tissues

We performed differential expression analyses comparing miRNA expression between bladder (N=9) and testes (N=7). These tissues types represented the largest set of two distinct tissues in one single experiment in the miRNAome data. Since the data came from one experiment, there was no known technical source of variation to adjust for, and tissue-type was the only fixed effect in the model. We want to analyze miRNA that are sufficiently expressed in these tissues, and ignore miRNA not expressed in these tissue types. We used a threshold of log(median CPM)>5 to threshold miRNA expression and remove the lowly or not expressed miRNA from the analysis. Based on the isomiR filtering strategies assessed using the ERCC data, we chose to implement a threshold of log(median CPM) < -1. After isomiR filtering, we retained 3,036 isomiRs from 90 miRNA, and the number of isomiRs mapping to any one miRNA ranged from 5 to 436 with a median of 18 isomiRs.

We estimated log fold-change (logFC) tissue estimates for each miRNA using miRglmm and compared the estimates to existing methods (see Methods). We used intra-class correlation coefficients to assess the agreement in logFC estimates between methods (Figure 6A).All methods provided very good agreement in logFC estimates (ICC≥0.8). These Agreements in miRNA considered significantly differentially expressed, as measured via Cohen's Kappa, ranged from moderate with limma-voom (κ=0.71) to weak with NB GLM and DESeq2 (κ=0.44 and κ=0.47, respectively). The estimates from the aggregated methods tend to be slightly larger than the NBMM estimates for the majority of the miRNA, and for limma-voom this difference is most pronounced. We similarly compared all aggregated methods in a pairwise fashion (Supplemental Figure 8), and observed high ICC (≥0.94) and moderate kappas (≥0.58) between methods. An upset plot of miRNA deemed significantly different between bladder and testes across the methods provides a visual comparison of all 5 methods simultaneously (Figure 6B). 67 of 90 miRNA have significant LRT p-values for the random tissue parameter for isomiRs (Figure 6C).

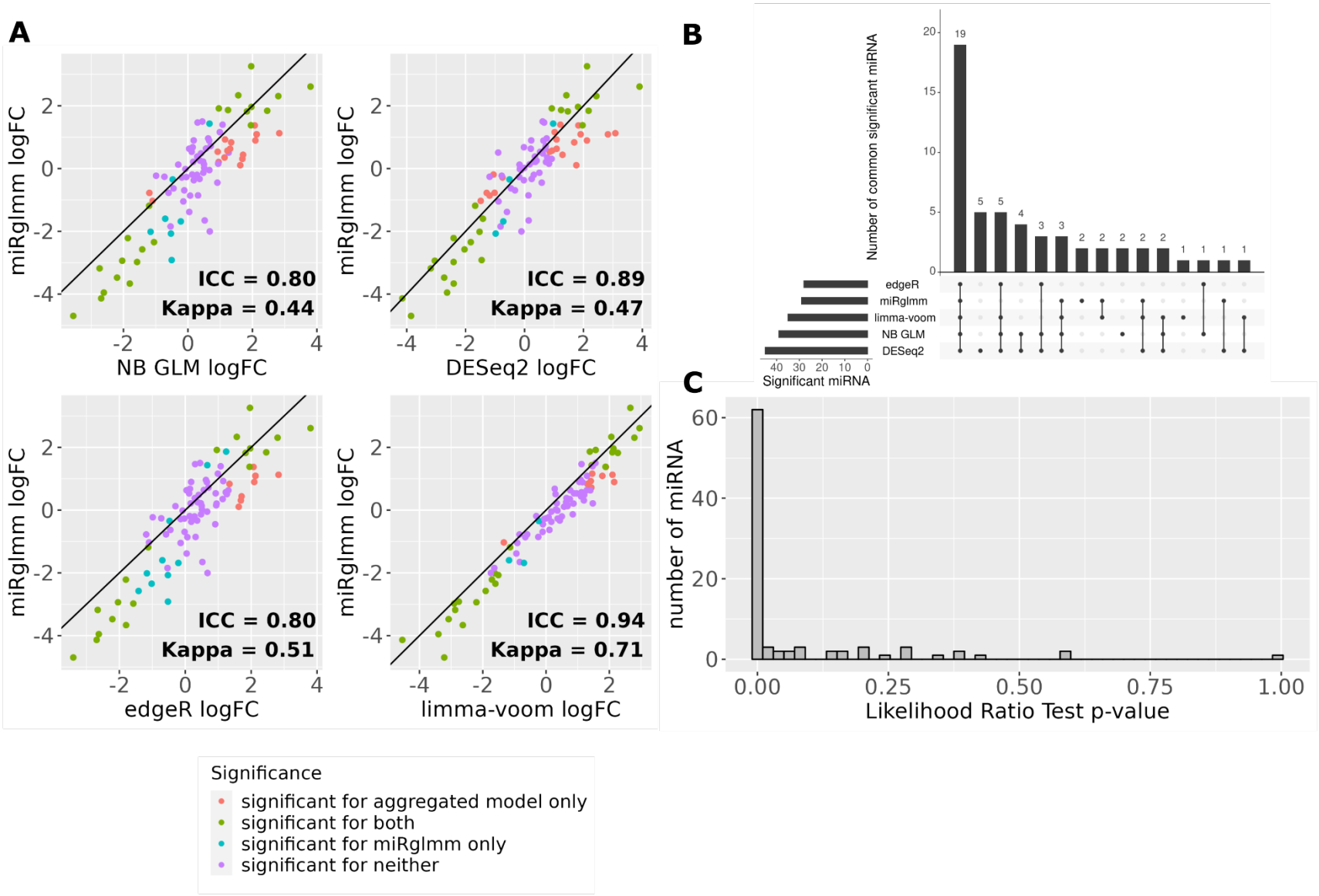

A
miRglmm logFC
NB GLM logFC
ICC = 0.80
Kappa = 0.44
DESeq2 logFC
ICC = 0.89
Kappa = 0.47
edgeR logFC
ICC = 0.80
Kappa = 0.51
limma-voom logFC
ICC = 0.94
Kappa = 0.71
B
Number of common significant miRNA
Significant miRNA
edgeR
miRglmm
limma-voom
NB GLM
DESeq2
C
number of miRNA
Likelihood Ratio Test p-value
Significance
significant for aggregated model only
significant for both
significant for miRglmm only
significant for neither

Figure 6: Log fold change (logFC) estimates of miRNA-level expression in testes compared to bladder from miRglmm are plotted versus logFC estimates from each aggregated count-based model (panel A). Each point is a unique miRNA. The significance based on each model for each miRNA is indicated by the color. Intra-class correlation coefficients (ICC) are used to assess agreement in the values of the estimated logFC between methods. Cohen's kappa is used to assess agreement in the significance of the estimate between methods. An upset plot shows the common number of miRNAs significantly different between bladder and testes across the 5 methods (panel B). There are 5 miRNA found to be significant for all aggregated methods but not miRglmm. There are 3 miRNA found to be significant for miRglmm but none of the aggregated methods. A histogram of the p-values of an LRT testing the significance of the random isomiR tissue effect indicates that most of the miRNA have significant variability in the bladder-testes effect across isomiRs (panel C).

miRglmm can also provide isomiR-level estimates of differential expression. We visualized isomiR-level expression and variability within several miRNA by plotting the estimated effects for each isomiR mapping to the miRNA and overlaying for all isomiRs (Figure 7). We overlaid estimated fixed effects from miRglmm to compare the miRNA-level differential expression estimates (fixed effect estimates from miRglmm) with underlying individual isomiR effects (random isomiR effect estimates from miRglmm). We also overlaid NB GLM estimated effect as an example of aggregated method results to compare aggregated effect estimates versus estimates that account for underlying isomiR variability (other aggregated method results are similarly shown in Supplemental Figure 9). Figure 7A and 7B show 2 miRNAs in which only miRglmm identifies as differentially expressed between bladder and testes. Figure 7C and 7D show 2 miRNAs that are significantly differentially expressed by all aggregated methodsbut not by miRglmm. Figure 7E and Figure 7F show 2 miRNAs with large magnitude of differences in miRNA-level differential expression estimates between NB GLM (the highlighted aggregated method) and miRglmm. Hsa-miR-100-5p was significantly different using miRglmm (and limma-voom), but not NB GLM (or DESeq2). Hsa-miR-195-5p was not significantly different using miRglmm (or DESeq2 or limma-voom) but was significantly different using NB GLM. All miRNA represented in Figure 5 with the exception of hsa-miR-143-5p (Figure 7B) were found to have significant random slope variability.

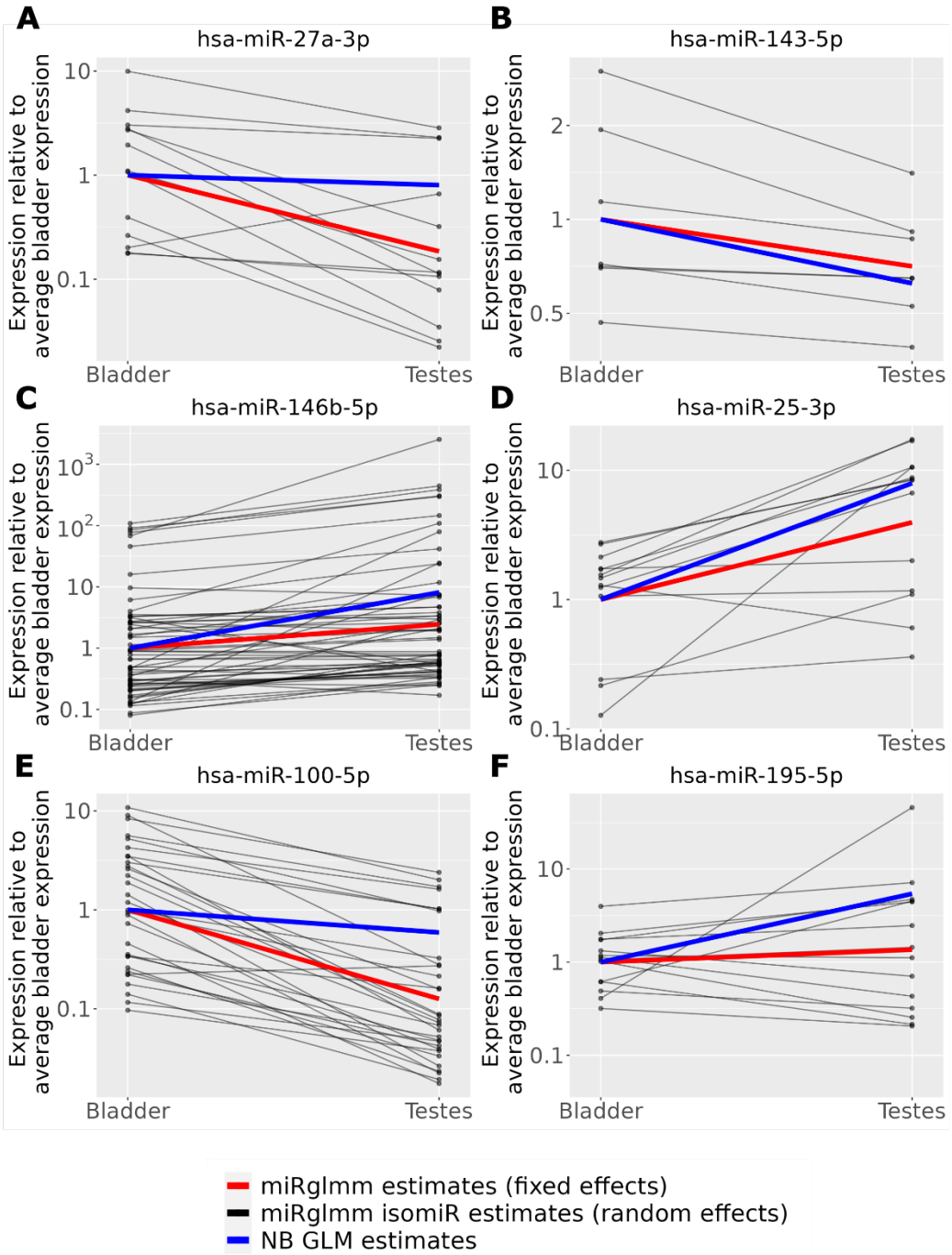


Figure 7: miRglmm estimates of relative expression are compared to estimates of relative expression from aggregated-count NB GLM models. We obtain estimates of isomiR variability within miRNA from the random effects components of miRglmm. Panel A and B represent two miRNA found to be

differentially expressed using miRglmm but not found by any of the aggregated models. Panel C and D represent two miRNA found to be differentially expressed using all aggregated models but not found by miRglmm. Panel E and F represent two miRNA with large differences in estimates from miRglmm and aggregated models. With the exception of hsa-miR-143-5p (panel B), all miRNA presented have significant isomiR variability in the bladder vs testes effect.

Discussion

With miRglmm, we address the need for a DE tool specifically suited for miRNA-seq data that can also utilize information from isomiRs in estimating miRNA-level DE estimates. Commonly used DE tools, such as DESeq2 (Love, Huber et al. 2014), edgeR (Robinson, McCarthy et al. 2010) and limma-voom (Law, Chen et al. 2014), were built for analysis of mRNA-seq data, and some key features of miRNA-seq cause violation of the independence between features assumption of these existing methods. Additionally, these tools run on aggregated miRNA-level counts, resulting in the loss of information contained in individual isomiRs. There is evidence of the biological importance of isomiRs (Bofill-De Ros, Yang et al. 2020), so miRglmm allows for the discovery of heterogeneous DE among isomiRs of the same miRNA via estimation of the random group effect for isomiR (and associated likelihood ratio test).

We show that miRglmm can identify miRNA with significant isomiR variability. When such variability exists, miRglmm far outperforms other DE tools in terms of MSE and coverage probability, indicating that miRglmm provides a better estimate of DE at the miRNA-level than those produced by other DE tools. Even in settings where isomiR variability does not exist, performance of miRglmm remains superior at a lesser extent to other DE tools. Consistent isomiR expression reduces variability and leads to a precise estimate of miRNA-level differential expression. miRglmm also produces isomiR-level estimates that can be used for exploration of individual isomiR effects.

A limitation of miRglmm is that it does not aim to distinguish a true biological isomiR from a technically arising isomiR and treats all isomiRs within a miRgroup equally. It has been shown that differences in library preparation can induce bias in quantification of isomiRs (Wright, Rajpurohit et al. 2019, Gómez-Martín, Aparicio-Puerta et al. 2023). By filtering lowly expressed isomiRs, we eliminate some technical variability due to background count alignment. Our study design in the DE analysis of bladder vs testes was implemented to minimize technical variation by selecting tissue samples processed by a single laboratory using consistent library preparation techniques and sequencing equipment. Importantly, we observed the majority of miRNA did have significant isomiR variability between tissues in this setting where we assume we are capturing true biological variability. If we were to analyze larger datasets that

encompass multiple studies, we would expect even higher rates of miRNA with significant isomiR variability, though some of this would be purely technical. In this case, miRglmm is flexible and can adjust for technical design variables, such as laboratory, library preparation method, sequencer, etc, as we did in analysis the ERCC benchmarking data (Giraldez, Spengler et al. 2018).

In conclusion, we provide a new method and analysis tool, miRglmm, that uses isomiR variability to improve differential expression analysis of miRNAs from miRNA-seq datasets. miRglmm provides superior performance to alternative DE tools, whether or not significant isomiR variability exists.

References

Ambros, V. (2004). "The functions of animal microRNAs." Nature **431**(7006): 350-355.

Ameres, S. L. and P. D. Zamore (2013). "Diversifying microRNA sequence and function." Nat Rev Mol Cell Biol **14**(8): 475-488.

Baras, A. S., C. J. Mitchell, J. R. Myers, S. Gupta, L. C. Weng, J. M. Ashton, T. C. Cornish, A. Pandey and M. K. Halushka (2015). "miRge - A Multiplexed Method of Processing Small RNA-Seq Data to Determine MicroRNA Entropy." PLoS One **10**(11): e0143066.

Bartel, D. P. (2004). "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function." Cell **116**(2): 281-297.

Bates, D., M. Mächler, B. Bolker and S. Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4." Journal of Statistical Software **67**(1): 1 - 48.

Bofill-De Ros, X., A. Yang and S. Gu (2020). "IsomiRs: Expanding the miRNA repression toolbox beyond the seed." Biochim Biophys Acta Gene Regul Mech **1863**(4): 194373.

Booth, J. G., G. Casella, H. Friedl and J. P. Hobert (2003). "Negative binomial loglinear mixed models." Statistical Modelling **3**(3): 179-191.

Das, S., K. M. Ansel, M. Bitzer, X. O. Breakefield, A. Charest, D. J. Galas, M. B. Gerstein, M. Gupta, A. Milosavljevic, M. T. McManus, T. Patel, R. L. Raffai, J. Rozowsky, M. E. Roth, J. A. Saugstad, K. Van Keuren-Jensen, A. M. Weaver and L. C. Laurent (2019). "The Extracellular RNA Communication Consortium: Establishing Foundational Knowledge and Technologies for Extracellular RNA Research." Cell **177**(2): 231-242.

Giraldez, M. D., R. M. Spengler, A. Etheridge, P. M. Godoy, A. J. Barczak, S. Srinivasan, P. L. De Hoff, K. Tanriverdi, A. Courtright, S. Lu, J. Khoory, R. Rubio, D. Baxter, T. A. P. Driedonks, H. P. J. Buermans, E. N. M. Nolte-'t Hoen, H. Jiang, K. Wang, I. Ghiran, Y. E. Wang, K. Van Keuren-Jensen, J. E. Freedman, P. G. Woodruff, L. C. Laurent, D. J. Erle, D. J. Galas and M. Tewari (2018). "Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling." Nature Biotechnology **36**(8): 746-757.

Gómez-Martín, C., E. Aparicio-Puerta, M. A. J. van Eijndhoven, J. M. Medina, M. Hackenberg and D. M. Pegtel (2023). "Reassessment of miRNA variant (isomiRs) composition by small RNA sequencing." Cell Rep Methods **3**(5): 100480.

He, L., J. Davila-Velderrain, T. S. Sumida, D. A. Hafler, M. Kellis and A. M. Kulminski (2021). "NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data." Communications Biology **4**(1): 629.

Juzenas, S., G. Venkatesh, M. Hübenthal, M. P. Hoeppner, Z. G. Du, M. Paulsen, P. Rosenstiel, P. Senger, M. Hofmann-Apitius, A. Keller, L. Kupcinskas, A. Franke and G. Hemmrich-Stanisak (2017). "A

comprehensive, cell specific microRNA catalogue of human peripheral blood." Nucleic Acids Res **45**(16): 9290-9301.
Law, C. W., Y. Chen, W. Shi and G. K. Smyth (2014). "voom: precision weights unlock linear model analysis tools for RNA-seq read counts." Genome Biology **15**(2): R29.
Li, X., J. Chen, X. Hu, Y. Huang, Z. Li, L. Zhou, Z. Tian, H. Ma, Z. Wu, M. Chen, Z. Han, Z. Peng, X. Zhao, C. Liang, Y. Wang, L. Sun, J. Chen, J. Zhao, B. Jiang, H. Yang, Y. Gui, Z. Cai and X. Zhang (2011). "Comparative mRNA and microRNA expression profiling of three genitourinary cancers reveals common hallmarks and cancer-specific molecular events." PLoS One **6**(7): e22570.
Love, M. I., W. Huber and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biol **15**(12): 550.
McCall, M. N., M. S. Kim, M. Adil, A. H. Patil, Y. Lu, C. J. Mitchell, P. Leal-Rojas, J. Xu, M. Kumar, V. L. Dawson, T. M. Dawson, A. S. Baras, A. Z. Rosenberg, D. E. Arking, K. H. Burns, A. Pandey and M. K. Halushka (2017). "Toward the human cellular microRNAome." Genome Res **27**(10): 1769-1781.
Mendell, J. T. and E. N. Olson (2012). "MicroRNAs in stress signaling and human disease." Cell **148**(6): 1172-1187.
Patil, A. H., A. Baran, Z. P. Brehm, M. N. McCall and M. K. Halushka (2022). "A curated human cellular microRNAome based on 196 primary cell types." Gigascience **11**.
Patil, A. H. and M. K. Halushka (2021). "miRge3.0: a comprehensive microRNA and tRF sequencing analysis pipeline." NAR Genom Bioinform **3**(3): lqab068.
Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.
Schmauch, E., Y. Attia, P. Laitinen, T. A. Turunen, P. Bartos, M.-A. Vaananen, T. Malm, P. Tavi, M. Kellis, M. U. Kaikkonen and S. Linna-Kuosmanen (2024). "The critical role of isomiRs in accurate differential expression analysis of miRNA-seq data." bioRxiv: 2024.2003.2028.587190.
Silverman, J. D., H. K. Durand, R. J. Bloom, S. Mukherjee and L. A. David (2018). "Dynamic linear models guide design and analysis of microbiota studies within artificial human guts." Microbiome **6**(1): 202.
Tan, G. C., E. Chan, A. Molnar, R. Sarkar, D. Alexieva, I. M. Isa, S. Robinson, S. Zhang, P. Ellis, C. F. Langford, P. V. Guillot, A. Chandrashekran, N. M. Fisk, L. Castellano, G. Meister, R. M. Winston, W. Cui, D. Baulcombe and N. J. Dibb (2014). "5' isomiR variation is of functional and evolutionary importance." Nucleic Acids Res **42**(14): 9424-9435.
Telonis, A. G., R. Magee, P. Loher, I. Chervoneva, E. Londin and I. Rigoutsos (2017). "Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types." Nucleic Acids Res **45**(6): 2973-2985.
Tomasello, L., R. Distefano, G. Nigita and C. M. Croce (2021). "The MicroRNA Family Gets Wider: The IsomiRs Classification and Role." Frontiers in Cell and Developmental Biology **9**.
Tsonaka, R. and P. Spitali (2021). "Negative Binomial mixed models estimated with the maximum likelihood method can be used for longitudinal RNAseq data." Brief Bioinform **22**(4).
Valencia-Sanchez, M. A., J. Liu, G. J. Hannon and R. Parker (2006). "Control of translation and mRNA degradation by miRNAs and siRNAs." Genes Dev **20**(5): 515-524.
Witwer, K. W. and M. K. Halushka (2016). "Toward the promise of microRNAs - Enhancing reproducibility and rigor in microRNA research." RNA Biol **13**(11): 1103-1116.
Wright, C., A. Rajpurohit, E. E. Burke, C. Williams, L. Collado-Torres, M. Kimos, N. J. Brandon, A. J. Cross, A. E. Jaffe, D. R. Weinberger and J. H. Shin (2019). "Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods." BMC Genomics **20**(1): 513.
Yang, A., X. Bofill-De Ros, T. J. Shao, M. Jiang, K. Li, P. Villanueva, L. Dai and S. Gu (2019). "3' Uridylation Confers miRNAs with Non-canonical Target Repertoires." Mol Cell **75**(3): 511-522.e514.

Yang, A., X. Bofill-De Ros, R. Stanton, T.-J. Shao, P. Villanueva and S. Gu (2022). "TENT2, TUT4, and TUT7 selectively regulate miRNA sequence and abundance." Nature Communications **13**(1): 5260.
Zhang, B. and M. A. Farwell (2008). "microRNAs: a new emerging class of players for disease diagnostics and gene therapy." Journal of Cellular and Molecular Medicine **12**(1): 3-21.
Zhang, X., H. Mallick, Z. Tang, L. Zhang, X. Cui, A. K. Benson and N. Yi (2017). "Negative binomial mixed models for analyzing microbiome count data." BMC Bioinformatics **18**(1): 4.
Zhang, X., Y.-F. Pei, L. Zhang, B. Guo, A. H. Pendegraft, W. Zhuang and N. Yi (2018). "Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data." Frontiers in Microbiology **9**.

Supplementary Materials

Supplemental Table 1: Example of isomiR data for miRNA hsa-let-7a-5p

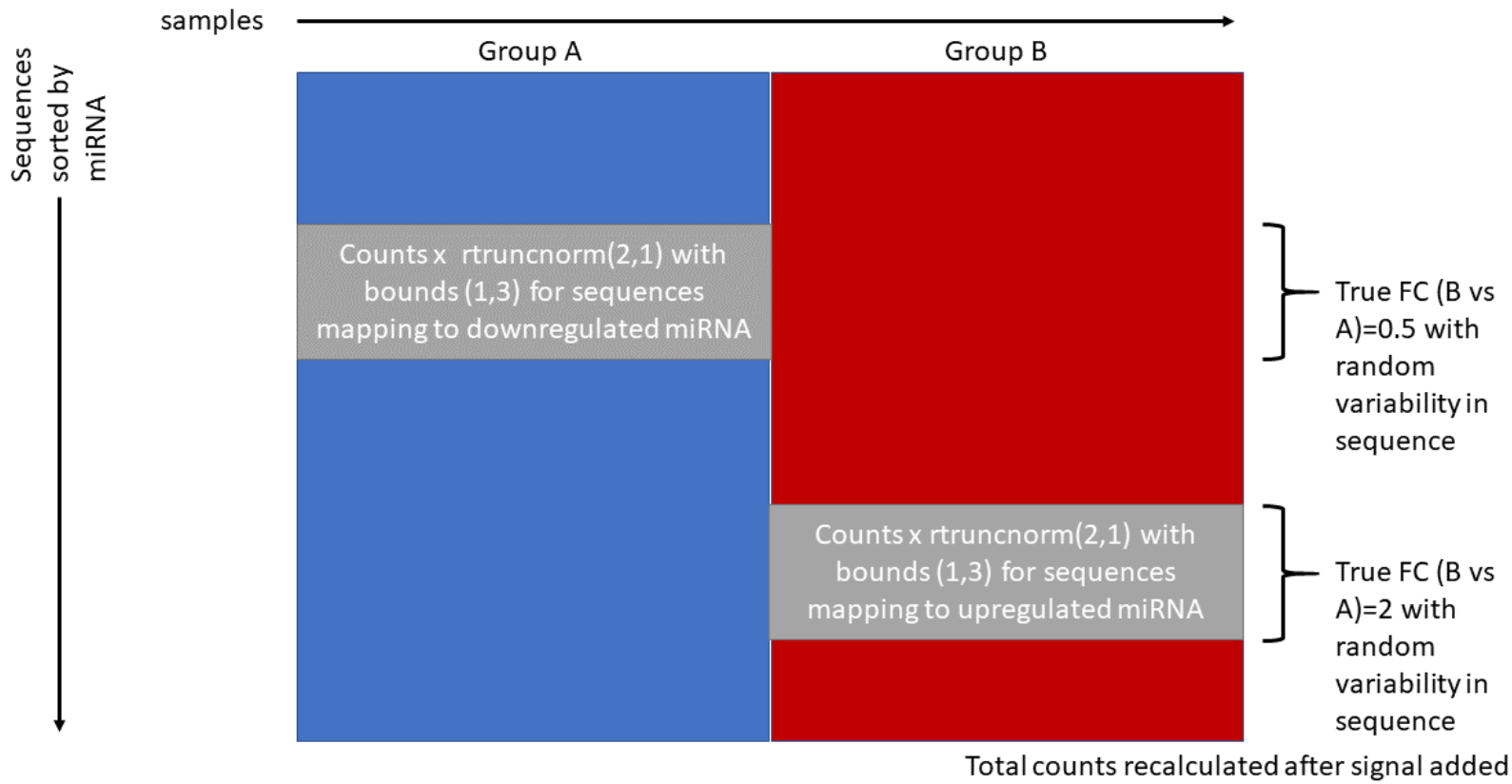
	SRR333658	SRR333672	SRR333674	SRR333678	SRR333680	SRR333682
	bladder	bladder	bladder	testes	testes	testes
TGAGGTAGTAGGTTGTATAGTT*	**768**	**321**	**461**	**1988**	**3463**	**1534**
TGAGGTAGTAGGTTGT	2216	2269	1869	322	186	244
TGAGGTAGTAGGTTGTATAG	1327	812	1060	587	1169	605
TGAGGTAGTAGGTTGTATAGTTT	86	35	43	2173	2312	890
TGAGGTAGTAGGTTGTATAGT	1169	481	764	602	1000	310
TGAGGTAGTAGGTTGTATA	990	722	762	175	334	157
TGAGGTAGTAGGTTGTATAGTTTCG	0	0	0	789	973	492
TGAGGTAGTAGGTTGTATAGTTTC	25	11	9	535	616	839
TGAGGTAGTAG**A**TTGTATAGTTTCG	0	0	1	51	1451	493
TGAGGTAGTAGGTTGTATAGT**A**TCTC	447	431	947	0	0	0
TGAGGTAGTAGGTTGTATAGTTAT	400	303	336	229	283	145
TGAGGTAGTAGGTTGTATAGT**A**TCT	414	370	688	0	0	1
TGAGGTAGTAGGTTGTATAGTTATCT	668	258	477	0	0	5
TGAGGTAGTAGGTTGTATAGT**A**TC	248	239	521	32	54	50
TGAGGTAGTAGGTTGTATAGT**A**T	223	229	453	59	75	32
TGAGGTAGTAGGTTGTATAGTTCGT	1	0	0	140	420	174
TGAGGTAGTAGGTTGTATAGTTTT	1	0	1	192	225	92
TGAGGTAGTAGGTTGTATAGT**C**GTAT	3	0	0	29	175	14
TGAGGTAGTAGG**C**TGT	17	87	69	0	0	0
TGAGGTAGTAGGTTGTATAGT**C**GT	0	1	0	23	97	15
TGAGGTAGTAGGTTGTATAG**A**TCTC	35	42	56	0	0	0
TGAGGTAGTAGGTTGT**G**TAGT	46	21	11	32	16	6
TGAGGTAG**G**AGGTTGTATAGT**A**TCT	37	29	38	0	0	0
TGAGGTAGTAGGTTGTATAG**A**TCT	20	38	42	0	0	0
hsa-let-7a-5p (aggregated)	**13141**	**9185**	**12254**	**9626**	**15328**	**9461**

*denotes canonical/representative miRNA sequence, red text denote internal modifications leading to isomiRs

Supplemental Table 2: Example of isomiR data for miRNA hsa-miR-26a-5p

Sequence Read Archive Number	SRR333662	SRR333664	SRR333674	SRR333680	SRR333684	SRR333688
Tissue Type	bladder	bladder	bladder	testes	testes	testes
TTCAAGTAATCCAGGATAGGCT*	**303**	**660**	**388**	**163**	**1171**	**383**
TTCAAGTAATCCAGGATAGGCTA	330	544	412	7	650	231
TTCAAGTAATCCAGGATAGGCTATCG	0	0	0	4	257	102
TTCAAGTAATCCAGGATAGGCTATCT	297	227	300	0	0	0
TTCAAGTAATCCAGGATAGGCTATCTC	179	226	305	0	0	0
TTCAAGTAATCCAGGATAGGCTC	7	7	7	1360	356	101
TTCAAGTAATCCAGGATAGGCTCG	0	0	2	39	177	48
TTCAAGTAATCCAGGATAGGCTCGT	0	0	0	190	63	15
TTCAAGTAATCCAGGATAGGCTCGTA	1	0	0	8	98	22
TTCAAGTAATCCAGGATAGGCTT	0	7	5	139	93	53
TTCAAGTAATCCAGGATAGGCTTCG	0	0	0	80	45	33
TTCAAGTAATCCAGGATAGGCTTCGT	0	0	0	17	35	21
TTCAAGTAATCCAGGAT	76	287	115	16	417	123
TTCAAGTAATCCAGGATA	163	497	164	80	673	233
TTCAAGTAATCCAGGATAG	161	559	154	115	976	358
TTCAAGTAATCCAGGATAGG	236	823	248	189	762	346
TTCAAGTAATCCAGGATAGGAT	5	20	10	1	1	0
TTCAAGTAATCCAGGATAGGATCT	12	23	17	0	0	0
TTCAAGTAATCCAGGATAGGATCTC	14	21	15	0	0	0
TTCAAGTAATCCAGGATAGGC	273	854	421	69	722	328
TTCAAGTAATCCAGGATAGGCA	10	17	7	6	17	24
TTCAAGTAATCCAGGATAGGCAG	1	4	2	1	15	10
TTCAAGTAATCCAGGATAGGCATCT	15	13	6	0	0	0
TTCAAGTAATCCAGTATAGGC	12	1	1	0	1	2
hsa-miR-26a-5p (aggregated)	**2827**	**6118**	**3452**	**2814**	**8277**	**3386**

*denotes canonical/representative miRNA sequence, red text denote internal modifications leading to isomiRs

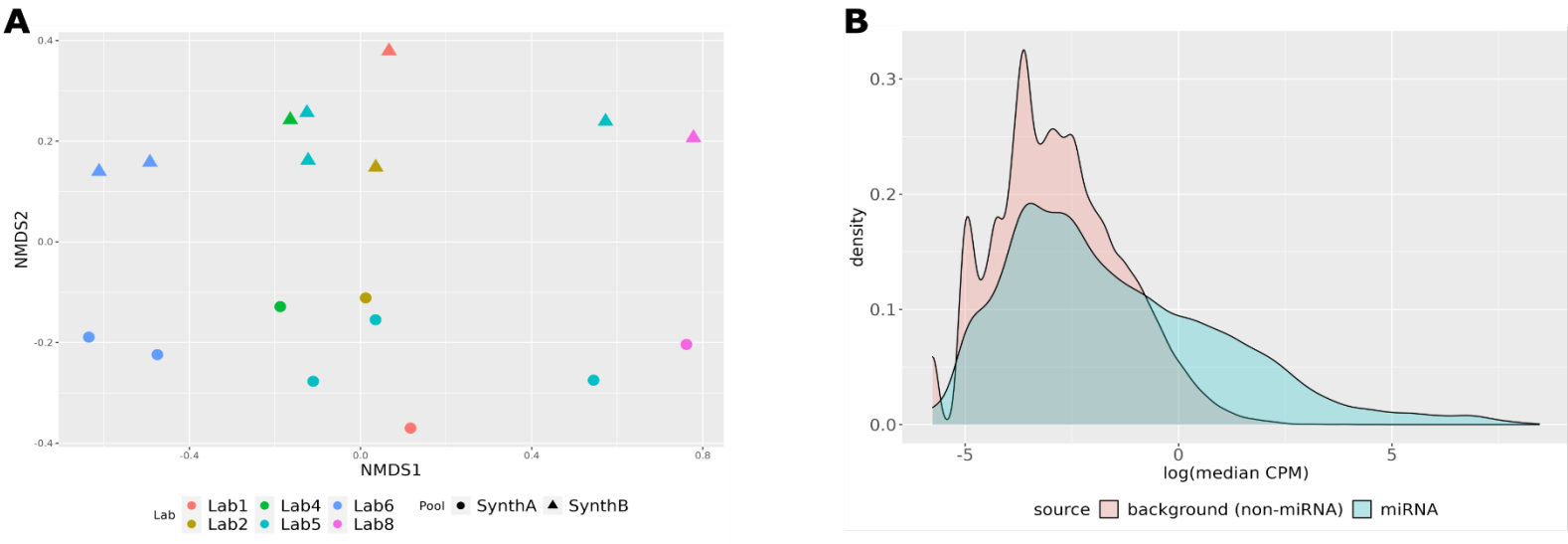


Supplemental Figure 1: Inducing an artificial group effect with random sequence variability into monocyte miRNA-seq data. The entire procedure from random splitting of samples into groups is repeated 100 times.

Ratiometric Pool Summary

Pool A	Pool B	LogFC (B vs A)	N miRNA
10x	1x	-2.3	18
8x	1x	-2.08	18
5x	1x	-1.61	18
4x	1x	-1.39	20
3x	1x	-1.1	18
2x	1x	-0.69	20
1.5x	1x	-0.41	20
1x	1x	0	35
1x	1.5x	0.41	20
1x	2x	0.69	19
1x	3x	1.1	18
1x	4x	1.39	20
1x	5x	1.61	19
1x	8x	2.08	19
1x	10x	2.3	21

Supplemental Figure 2: ERCC experimental design. Samples from two pools were generated by six laboratories. Each synthetic miRNA was added to Pool A and Pool B at one of fifteen possible ratios.



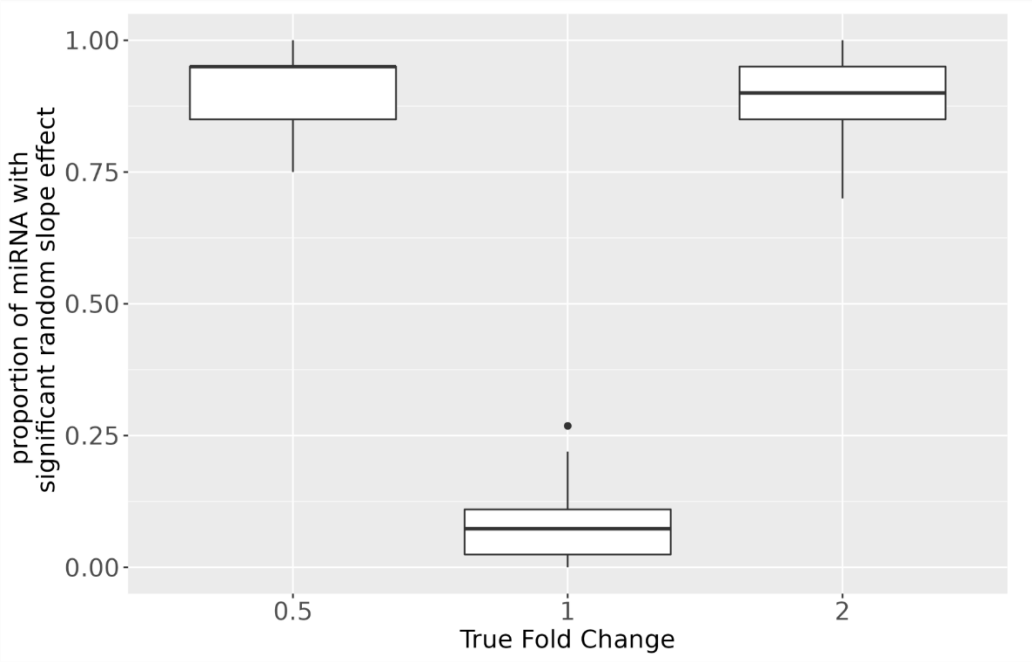
Supplemental Figure 3: Non-metric multidimensional rescaling (NMDS) of raw counts reveals separation of samples based on laboratory along the first dimension and separation based on Pool along the second dimension (panel A). Density plots comparing the distributions of expression between sequences mapped to miRNA and sequences not mapped to miRNA can be used to define a threshold for retaining sequences with expression above background levels (panel B).

Supplemental Table 3: MSE summary statistics across 100 simulations

method	mean	SD	median	minimum	maximum	number of times minimizing MSE
miRglmm	0.01485	0.00463	0.01360	0.00784	0.03292	97
NB GLM	0.02315	0.00612	0.02239	0.01326	0.04883	0
DESeq2	0.02192	0.00511	0.02153	0.01229	0.03608	3
edgeR	0.02315	0.00612	0.02239	0.01325	0.04883	0
limmavoom	0.02310	0.00622	0.02234	0.01381	0.04537	0

Supplemental Table 4: Coverage probability summary statistics across 100 simulations

method	mean	SD	median	minimum	maximum
miRglmm	0.9143	0.0559	0.9262	0.7131	0.9836
NB GLM	0.7727	0.0641	0.7869	0.5328	0.8852
DESeq2	0.8048	0.0449	0.8074	0.6557	0.8852
limmavoom	0.7869	0.0669	0.7992	0.5492	0.8934

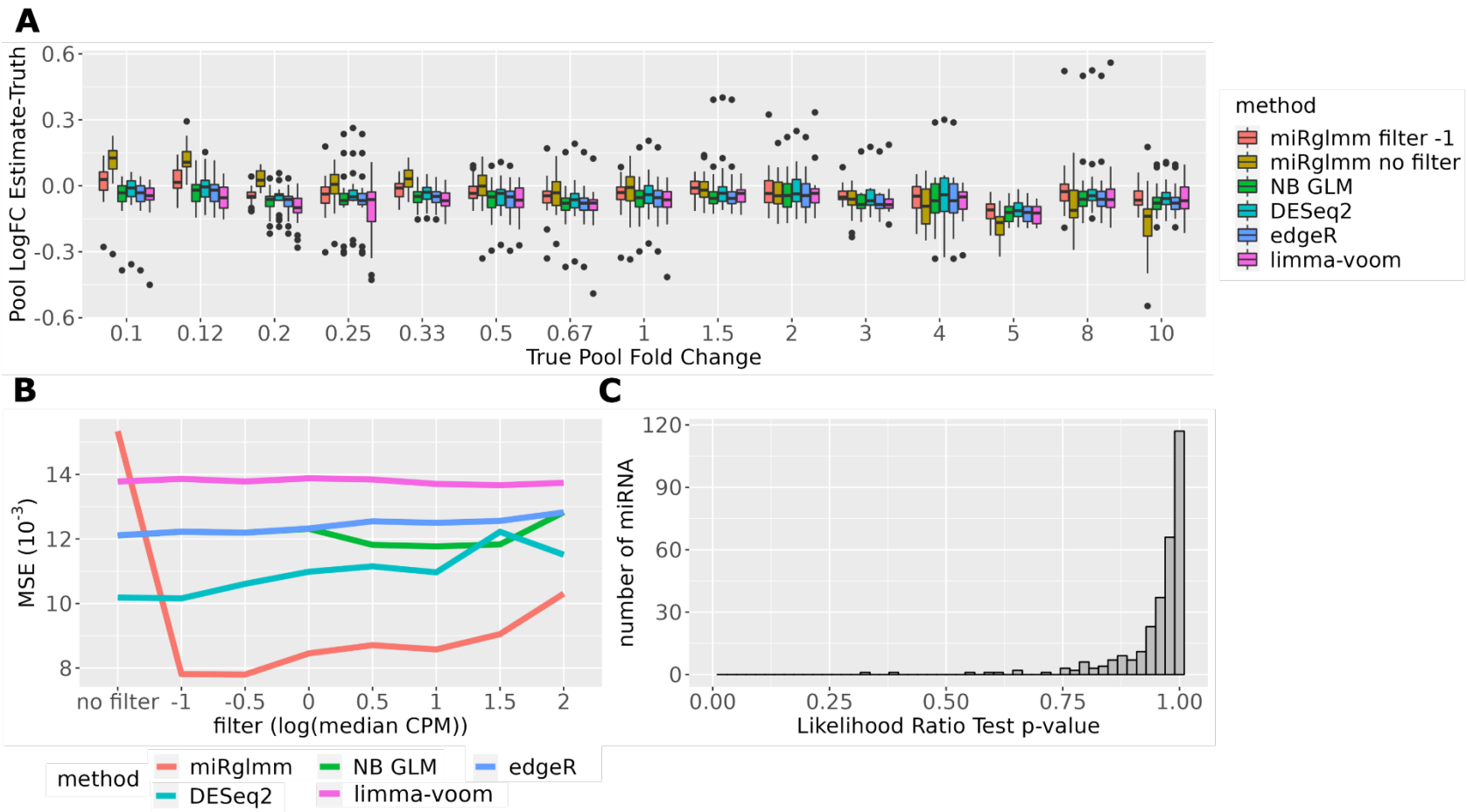


Supplemental Figure 4: The simulation procedure creates significant variability in the group effect across isomiRs.

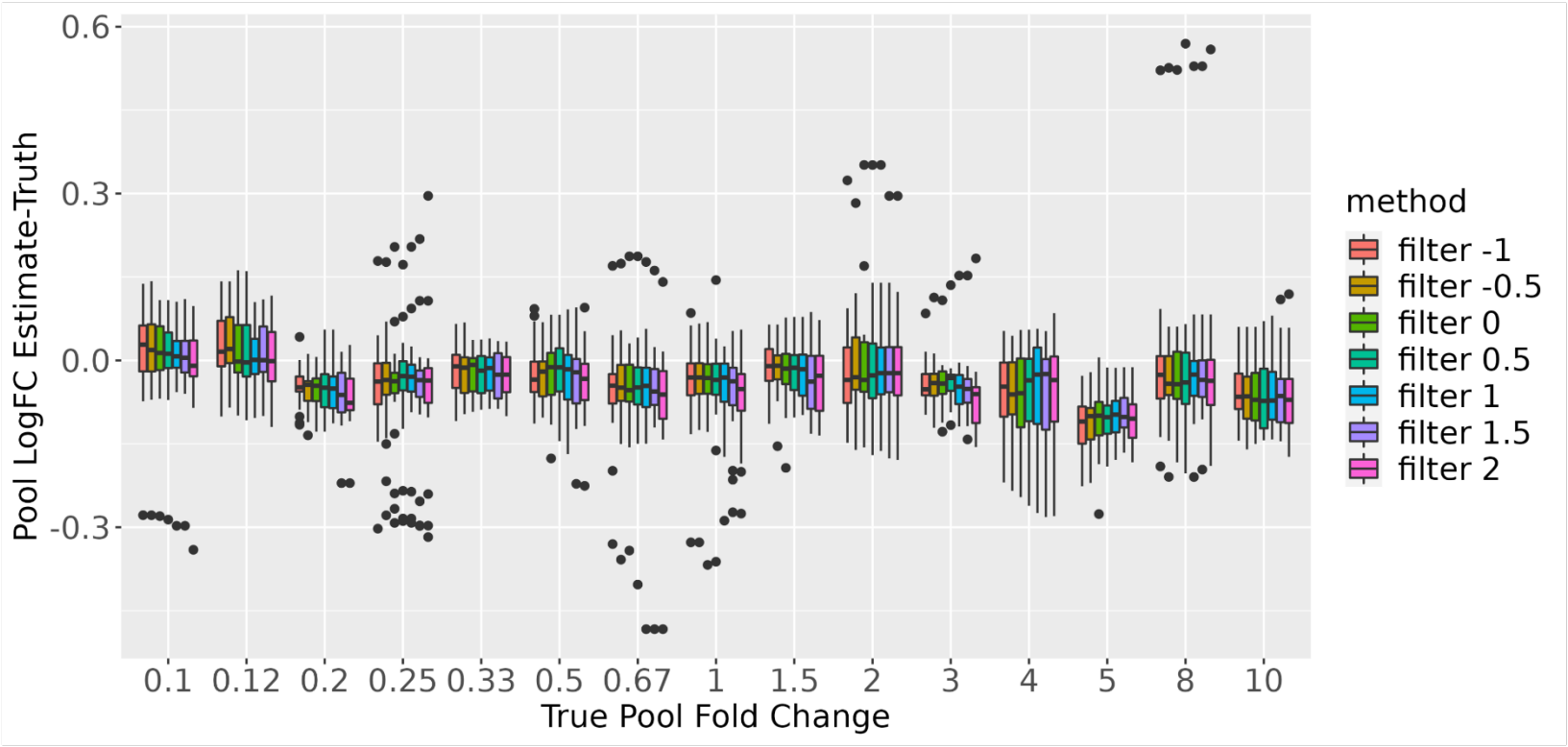
Supplemental Table 5: Comparing performance by miRglmm filters

method	MSE (x 10^{-3})	number of miRNA estimated	minimum number of isomiRs	maximum number of isomiRs	coverage probability
miRglmm filter -1*	7.81	303	7	40	0.98
miRglmm filter -0.5	7.80	303	6	33	0.97
miRglmm filter 0	8.45	303	4	30	0.97
miRglmm filter 0.5	8.71	302	2	24	0.98
miRglmm filter 1	8.58	301	2	22	0.98
miRglmm filter 1.5	9.05	301	2	18	0.99
miRglmm filter 2	10.31	300	2	15	0.99
miRglmm no filter	15.34	303	12	263	0.97

MSE: Mean Squared Error, NA: not applicable, *denotes default filter implemented in miRglmm



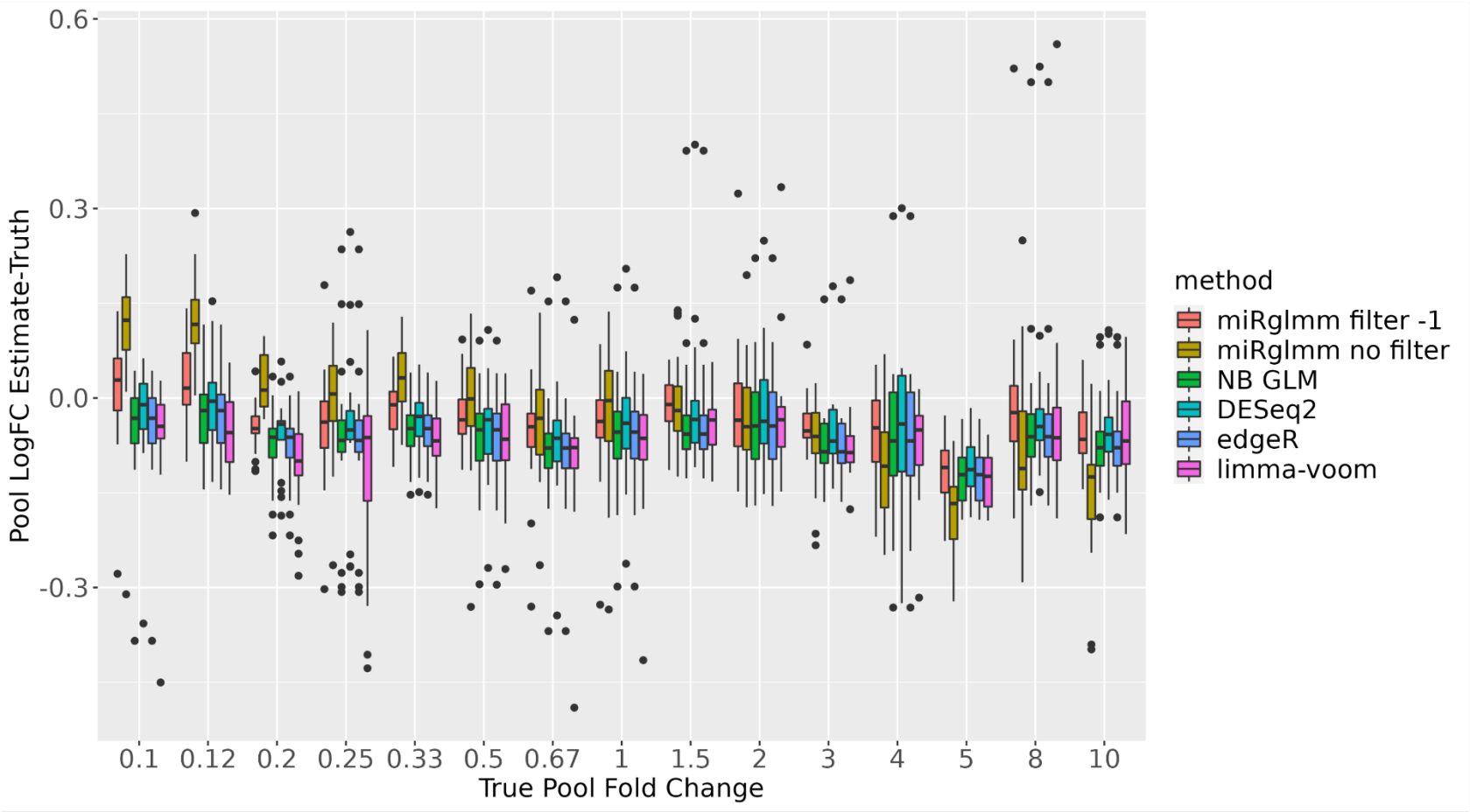
Supplemental Figure 5: Log Fold Change (logFC) estimates of Pool B expression compared to Pool A expression are compared to known true values for each miRNA and are summarized within each True Pool Fold Change group by method (Panel B). miRglmm after filtering out isomiRs with log(median CPM < -1), miRglmm with no filtering and 4 aggregated methods (with no filtering) were used. Filtering results in a dramatic decrease in Mean Squared Error (MSE) when using miRglmm, but there is little change in performance after filtering for the aggregated methods (panel B). There are no miRNA with significant variability in the Pool effect across isomiRs as measured via a Likelihood Ratio Test p-value (panel C).



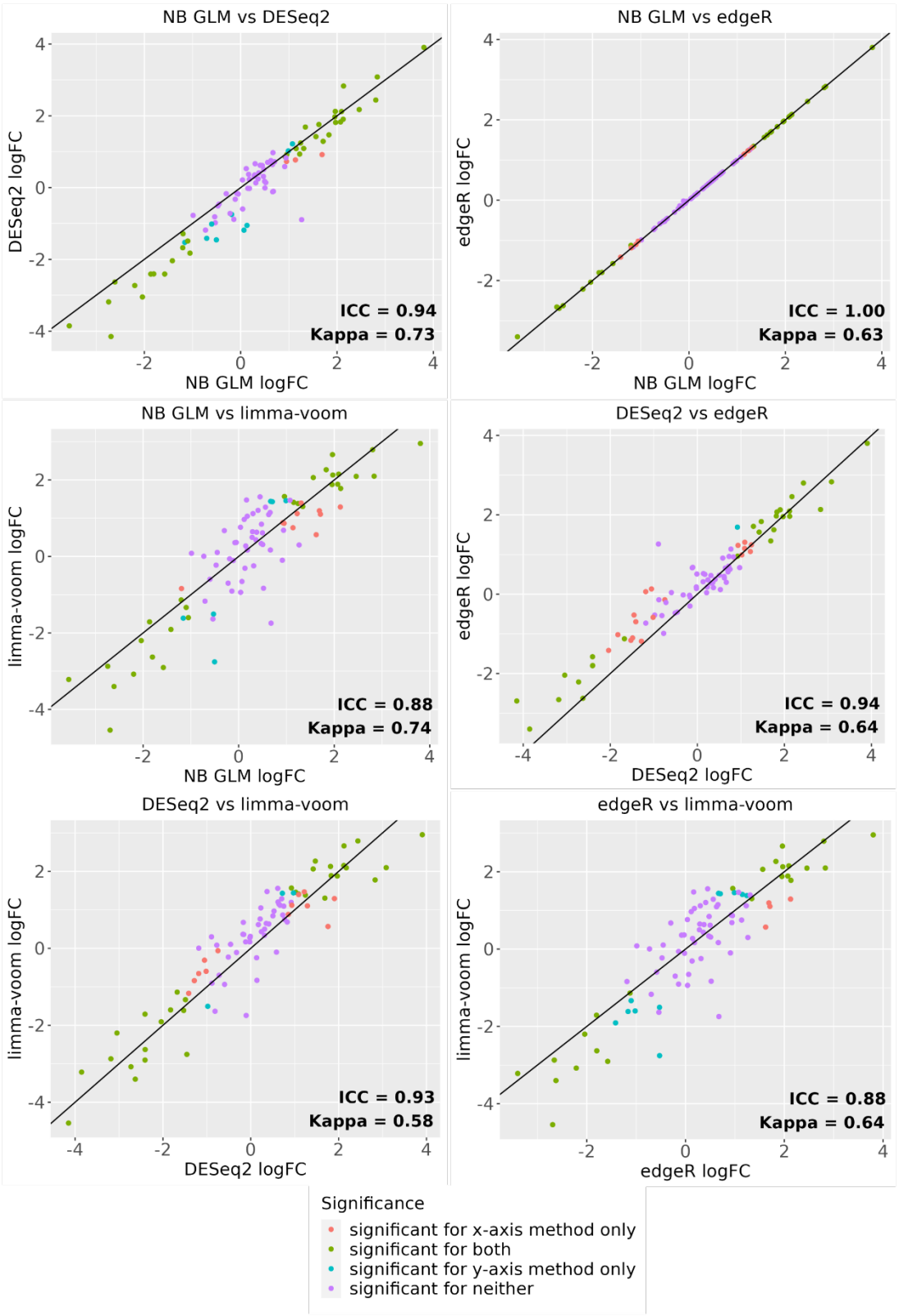
Supplemental Figure 6: The performance of miRglmm after a variety of isomiR filters based on log(median CPM) expression appears fairly uniform. Log fold change (logFC) estimates of Pool B expression compared to Pool A expression are compared to known true values for each miRNA and summarized within True Pool Fold Change group.

Supplemental Table 6: Comparing performance by miRglmm filters and comparing to aggregated methods when random slope effect removed

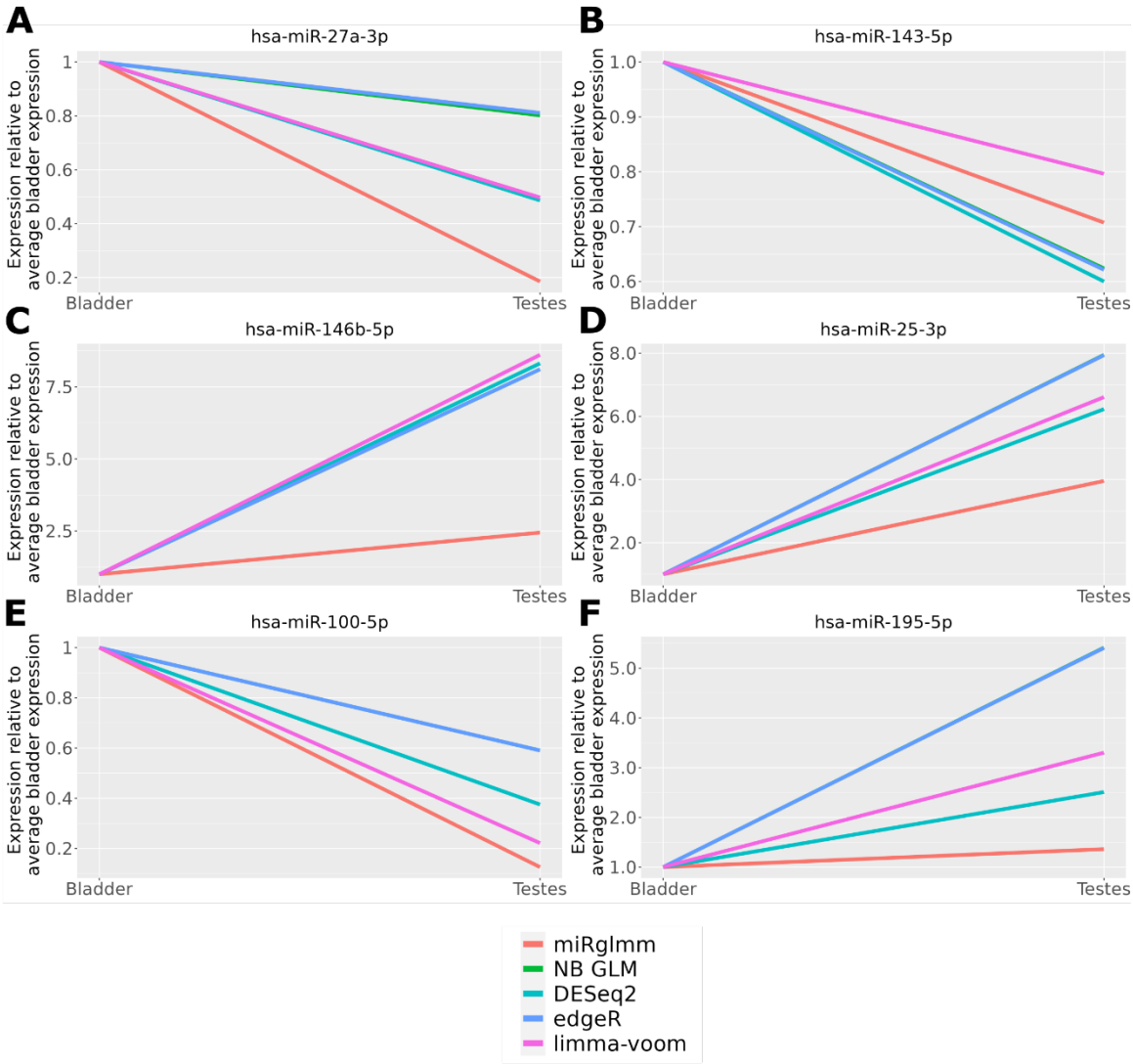
method	MSE (10^{-3})	number of miRNA estimated	minimum number of sequences	maximum number of sequences	coverage probability
miRglmm filter -1	7.80	303	7	40	0.98
miRglmm filter -0.5	7.79	303	6	33	0.98
miRglmm filter 0	8.44	303	4	30	0.97
miRglmm filter 0.5	8.67	302	2	24	0.98
miRglmm filter 1	8.56	301	2	22	0.98
miRglmm filter 1.5	9.05	301	2	18	0.99
miRglmm filter 2	10.29	300	2	15	0.99
miRglmm no filter	14.80	303	12	263	0.97
miRglmnb	12.11	303	NA	NA	0.97
DESeq2	10.19	303	NA	NA	0.99
edgeR	12.11	303	NA	NA	NA
limma-voom	13.78	303	NA	NA	0.99



Supplemental Figure 7: The performance of miRglmm without the random Pool effect for isomiR is compared to the performance of aggregated methods. Log fold change (logFC) estimates of Pool B expression compared to Pool A expression are compared to known true values for each miRNA and summarized within each True Pool Fold Change group.



Supplemental Figure 8: Log fold change (logFC) estimates of expression in testes compared to bladder are compared between aggregated methods with differences in significance indicated by color



Supplemental Figure 9: Estimates of testes expression relative to bladder are compared between miRglmm and all aggregated methods.